UNIVERSIDAD
NACIONAL
DE COLOMBIA

Multimodal representation learning with neural networks

John Edilson Arevalo Ovalle

National University of Colombia
Engineering School, Systems and Industrial Engineering Departament
Bogotá, Colombia
2018

Multimodal representation learning with neural networks

John Edilson Arevalo Ovalle

Thesis submitted as requirement to obtain the title of:
PhD. in Systems and Computer Engineering

Advisor:
Fabio A. González, Ph.D.
Co-Advisor:
Thamar Solorio, Ph.D.

Research Field:
Machine learning
Research group:
MindLab

National University of Colombia
Engineering School, Systems and Industrial Engineering Departament
Bogotá, Colombia
2018

Multimodal representation learning with neural networks

John Arevalo

Abstract

Representation learning methods have received a lot of attention by researchers and practitioners because of their successful application to complex problems in areas such as computer vision, speech recognition and text processing [1]. Many of these promising results are due to the development of methods to automatically learn the representation of complex objects directly from large amounts of sample data [2]. These efforts have concentrated on data involving one type of information (images, text, speech, etc.), despite data being naturally multimodal. Multimodality refers to the fact that the same real-world concept can be described by different views or data types. Addressing multimodal automatic analysis faces three main challenges: feature learning and extraction, modeling of relationships between data modalities and scalability to large multimodal collections [3, 4].

This research considers the problem of leveraging multiple sources of information or data modalities in neural networks. It defines a novel model called **gated multimodal unit (GMU)**, designed as an internal unit in a neural network architecture whose purpose is to find an intermediate representation based on a combination of data from different modalities. The GMU learns to decide how modalities influence the activation of the unit using multiplicative gates. The GMU can be used as a building block for different kinds of neural networks and can be seen as a form of intermediate fusion. The model was evaluated on four supervised learning tasks in conjunction with fully-connected and convolutional neural networks. We compare the GMU with other early and late fusion methods, outperforming classification scores in the evaluated datasets. Strategies to understand how the model gives importance to each input were also explored. By measuring correlation between gate activations and predictions, we were able to associate modalities with classes. It was found that some classes were more correlated with some particular modality. Interesting findings in genre prediction show, for instance, that the model associates the visual information with animation movies while textual information is more associated with drama or romance movies. During the development of this project, three new benchmark datasets were built and publicly released. The BCDR-F03 dataset which contains 736 mammography images and serves as benchmark for mass lesion classification. The MM-IMDb dataset containing around 27000 movie plots, poster along with 50 metadata annotations and that motivates new research in multimodal analysis. And the Goodreads dataset, a collection of 1000 books that encourages the research on success prediction based on the book content. This research also facilitates reproducibility of the present work by releasing source code implementation of the proposed methods.

Contents

Abstract **IV**

List of Figures **VII**

List of Tables **VIII**

1 Introduction **1**
1.1 Problem statement . . . 1
1.2 Main contributions . . . 3
1.3 Outline . . . 5

2 Background and related work **6**
2.1 Representation learning . . . 6
2.2 Learning with multimodal representations . . . 7
2.2.1 Representation . . . 7
2.2.2 Multimodal fusion . . . 9
2.2.3 Applications . . . 10
2.3 Scalable representation learning . . . 11

3 Gated multimodal unit **13**
3.1 Introduction . . . 13
3.2 Gated multimodal unit . . . 15
3.2.1 Noisy channel model . . . 16
3.3 Conclusions . . . 18

4 GMU for genre classification **19**
4.1 Multimodal IMDb dataset . . . 19
4.2 GMU for genre classification . . . 21
4.3 Data representation . . . 21
4.3.1 Text representation . . . 22
4.3.2 Visual representation . . . 22
4.3.3 Multimodal fusion baselines . . . 23
4.4 Experimental setup . . . 23
4.4.1 Neural network training . . . 24
4.5 Results . . . 25
4.6 Conclusions . . . 27

5 GMU for image segmentation **30**
5.1 Introduction . . . 30
5.2 Deep scene dataset . . . 31
5.3 Convolutional GMU for segmentation . . . 31
5.4 Experimental setup . . . 32
5.5 Results . . . 33
5.6 Conclusions . . . 33

6 GMU for medical imaging **36**
6.1 Introduction . . . 36
6.2 Breast cancer digital repository . . . 37
6.2.1 Benchmarking Dataset . . . 38
6.3 Gated multimodal networks for feature fusion . . . 39
6.4 Data representation . . . 40
6.4.1 Baseline descriptors . . . 40
6.4.2 Supervised feature learning . . . 41
6.5 Experimental setup . . . 45
6.6 Results . . . 46
6.6.1 Learned features . . . 46
6.6.2 Classification results . . . 46
6.7 Conclusions . . . 50

7 GMU for book analysis **52**
7.1 Introduction . . . 52
7.2 Goodreads dataset . . . 53
7.3 GMU for feature fusion . . . 54
7.4 Data representation . . . 55
7.4.1 Hand-crafted text features . . . 55
7.4.2 Neural network learned representations . . . 56
7.5 Experimental setup . . . 58
7.6 Results . . . 60
7.6.1 GMU fusion . . . 61
7.7 Conclusions . . . 61

8 Summary and conclusions **62**

Bibliography **64**

List of Figures

2.1 Standard workflow reported in the literature 8

3.1 Illustration of gated units 16
3.2 Noisy channel model 16
3.3 Generative model for the synthetic task 17
3.4 Activations for GMU internals 18

4.1 Co-ocurrence matrix of genre tags 20
4.2 Size distribution of movie posters 20
4.3 Length distribution of movie plots 20
4.4 Integration of the GMU in a multilayer perceptron 21
4.5 p-values distribution 26
4.6 Percentage of gates activations 27

5.1 Deep scene sample images 31
5.2 GMU in convolutional architecture 32
5.3 Visualization of segmentation 34
5.4 Percentage of gates activations 35

6.1 Samples of lesions presented in the dataset 38
6.2 GMU for feature fusion 39
6.3 Mammography images after the preprocessing step 43
6.4 Best convolutional neural network evaluated on mass classification 44
6.5 Filters learned in the first layer of the CNN model 46
6.6 ROC curve for evaluated representations 47
6.7 Boxplots of different runs for each representation method 48
6.8 GMU results for BCDR dataset 49

7.1 GMU for book success prediction 55
7.2 Multitask method for book success prediction 57

List of Tables

2-1 Summary of multimodal applications 11

4-1 Summary of classification task on the MM-IMDb dataset 25
4-2 Macro F-Score for single and multimodal 26
4-3 Examples of predictions in test set 28

5-1 Summary of segmentation performance 33
5-2 Intersection-over-union per class 33

6-1 Set of hand-crafted features 40
6-2 Summary of results in terms of AUC 49
6-3 p-values for statistical test 50

7-1 Goodreads Data Distribution 54
7-2 Confusion matrix between two different definitions of success 54
7-3 Results for classification 59
7-4 Feature combination results for goodreads dataset 60
7-5 Feature fusion results 61

1 Introduction

Recently, machine learning methods have received a lot of attention by researchers and practitioners because of its successful application to the solution of complex problems in areas such as computer vision, speech recognition and text processing. Many of these promising results are due to the development of methods to automatically learn the representation of complex objects directly from large amounts of sample data [2]. These methods are an evolution of neural networks and are known as deep learning. Deep learning leads the state-of-the-art in different areas with success cases in object recognition, scene image labeling, autonomous car driving and speech recognition among others [2]. The success of deep learning methods has to do with two main reasons: the development on hardware and software technology that allows to train large models with millions, and even billions, of parameters; and the availability of huge amounts of data.

Many databases relate different information sources to describe the same real-world concept. Collaborative encyclopedias (such as Wikipedia) describe a famous person through a mixture of text, images and, in some cases, audio. Users from social networks comment events like concerts or sport games with small phrases and multimedia attachments (images/videos/audios). Medical records are represented by a collection of images, sound, text and signals, among others. The increasing availability of multimodal databases from different sources has motivated the development of automatic analysis techniques to exploit the potential of these data as a source of knowledge in the form of patterns and structures that reveal complex relationships [5, 6]. Such automatic analysis faces three main challenges: feature learning and extraction, modeling of relationships between data modalities and scalability to large multimodal collections [3, 4].

1.1 Problem statement

The main aim of this research is to devise methods to effectively learn representations of multimodal data. Recent surveys have shown the feasibility and advantages of learning the representation automatically from the data [2, 7–10]; however the majority of those works are focused on particular types of data: images, audio, text and video. The main research question that orientates the proposed research is: How to automatically learn effective representations from multimodal data that allow exploiting them better for automatic analysis tasks? The combination of different information sources to discover relevant patterns and latent concepts leads to a better understanding of data collections [11]. Such abstract or latent concepts can be better detected by modeling relationships and correlations from different data sources. Since data comes from different input channels, modalities, in general, have different statistical properties. This makes harder to design a fusion strategy that works in all the cases [12]. Even though the number of publications related to representation learning in multimodal scenarios has grown in the last few years, the problem is still an open challenge for the research community and has been addressed in a quite standard way

as detailed in Chapter 2. These approaches do not exploit hierarchical representations, such as the ones learned by deep learning methods [9]. In addition, the growth of databases demands efficient algorithms to perform automatic analysis in a feasible way using representation learning strategies [2]. Thus, in order to address the main research problem of linking heterogeneous representation modalities to make automatic analysis easier, it is important to answer the following research questions:

- How to simultaneously learn the representation of data with multiple modalities?
- How to automatically extract multimodal representation correlations and interactions?
- In which learning tasks does multimodal representation learning shows advantages over monomodal representation learning?
- How to use the learned representation to improve interpretability and support automatic analysis of learned models?
- How to scale up the proposed algorithms to deal with large multimodal database collections?

Specifically, this research proposes the following goals:

Main goal To develop a scalable model for automatic representation learning in multimodal data collections.

Specific goals

- To propose a conceptual framework to combine multimodal information using representation learning strategies.
- To design an algorithm for combining representations from multimodal data.
- To build scalable implementations of the proposed method that can be extended to large volumes of data.
- To systematically evaluate the proposed strategies in terms of effectiveness for automatic analysis tasks.

Research impact domains The impact of this research is twofold: on the one hand, this project develops new alternative methods to deal with multimodal data, learn useful representations and improve the performance in automatic data analysis tasks. On the other hand, this research explores different applications of the developed methods in fields where multimodality is present. Some examples of potential application fields include:

Medical Analysis: Medical information involves several source modalities that can be exploited in different scenarios. The training of specialists could be supported by exploration and understanding of clinical cases documented with images and writing reports. Computer aided diagnosis systems could support medical decisions based on multimodal clinical data.

Information Retrieval: Indexing information using features from different modalities to improve accuracy of results would allow users to explore large collections in a more efficient way.

Recommendation Systems: The use of multimodal sources during training recommendation systems would yield to suggest more accurate content for users/clients.

Computational Biology: Applying data analysis to understand biological phenomena can be improved by discovering mid-level features from different sources like images and text.

1.2 Main contributions

This research presents novel strategies and systematic evaluations to perform representation and automatic analysis of multimodal information. The following is the outline of the main contributions of this work.

- Systematic evaluation of supervised and non-supervised strategies for learning image representations. Such representations were evaluated in the medical context for two classification problems: basal cell carcinoma detection [13] and breast mass lesion classification [14].
- Exploration of different representation learning strategies for text classification.
- Construction of the first multimodal dataset for movie genre classification. The dataset comprises 27,000 movies along with their plots, posters and more than 50 additional features. The dataset is publicly available at `http://lisi1.unal.edu.co/mmimdb`.
- Formulation and evaluation of a novel neural network unit that automatically learns to combine different sources of information. The strategy surpasses standard early and late fusion models. It was evaluated on four different tasks obtaining the state-of-the-art results.

We contributed the data and code for reproducibility and benchmarking of this research. These are new resources to facilitate and encourage new research in this direction. As result, we released the three datasets used in this dissertation:

- The Breast cancer digital repository - BCDR-F03 (`http://bcdr.inegi.up.pt/`)
- The Multimodal IMDb (MM-IMDb) dataset (`http://lisi1.unal.edu.co/mmimdb/`)
- The book success prediction dataset (`http://ritual.uh.edu/resources/`)

The following is a list of papers that have been published during the development of this research:

1. **Arevalo, John** and Cruz-Roa, Angel and others, "Histopathology image representation for automatic analysis: A state-of-the-art review", Revista Med 22, 2 (2014), pp. 79–91. [15]
2. Vanegas, Jorge A and **Arevalo, John** and Otálora, Sebastian and Páez, Fabián and Pérez-Rubiano, Santiago A. , "MindLab at ImageCLEF 2014: Scalable Concept Image Annotation", in CLEF 2014 Evaluation Labs and Workshop, Online Working Notes. Sheffield, UK (September 15-18 2014) (2014). [16]

3. Vanegas, Jorge A and **Arevalo, John** and Gonzalez, Fabio A, "Unsupervised feature learning for content-based histopathology image retrieval", in Content-Based Multimedia Indexing (CBMI), 2014 12th International Workshop on (2014), pp. 1–6.[17]

4. **Arevalo, John** and González, Fabio A and Ramos-Pollán, Raúl and Oliveira, Jose L and Lopez, Miguel Angel Guevara, "Convolutional neural networks for mammography mass lesion classification", in Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE (2015), pp. 797–800. [18]

5. **Arevalo, John** and Cruz-Roa, Angel and Arias, Viviana and Romero, Eduardo and González, Fabio A, "An unsupervised feature learning framework for basal cell carcinoma image analysis", Artificial intelligence in medicine 64, 2 (2015), pp. 131–145. [13]

6. Cruz-Roa, Angel and **Arevalo, John** and Basavanhally, Ajay and Madabhushi, Anant and González, Fabio, "A comparative evaluation of supervised and unsupervised representation learning approaches for anaplastic medulloblastoma differentiation", in Proc. SPIE vol. 9287, (, 2015). [19]

7. Cruz-Roa, Angel and **Arévalo, John** and Judkins, Alexander and Madabhushi, Anant and González, Fabio, "A method for medulloblastoma tumor differentiation based on convolutional neural networks and transfer learning", in 11th International Symposium on Medical Information Processing and Analysis (SIPAIM 2015) (2015). [20]

8. Otálora, Sebastian and Cruz-Roa, Angel and **Arevalo, John** and Atzori, Manfredo and Madabhushi, Anant and Judkins, Alexander, "Combining unsupervised feature learning and riesz wavelets for histopathology image representation", in International Conference on Medical Image Computing and Computer-Assisted Intervention (2015), pp. 581–588. [21]

9. Pellegrin, Luis and Vanegas, Jorge A and **Arevalo, John** and Beltrán, Viviana and Escalante, Hugo Jair, "INAOE-UNAL at ImageCLEF 2015: Scalable Concept Image Annotation.", in CLEF (Working Notes) (2015). [22]

10. Pellegrin, Luis and Vanegas, Jorge A and **Arevalo, John** and Beltrán, Viviana and Escalante, Hugo Jair and Montes-y-Gómez, Manuel, "A Two-Step Retrieval Method for Image Captioning", in International Conference of the Cross-Language Evaluation Forum for European Languages (2016), pp. 150–161. [23]

11. **Arevalo, John** and González, Fabio A and Ramos-Pollán, Raúl and Oliveira, Jose L and Lopez, Miguel Angel Guevara, "Representation learning for mammography mass lesion classification with convolutional neural networks", Computer methods and programs in biomedicine 127 (2016), pp. 248–257. [14]

12. **Arevalo, John** and Solorio, Thamar and Montes-y-Gómez, Manuel and González, Fabio A, "Gated Multimodal Units for Information Fusion", in 5th International conference on learning representations 2017 workshop (2017). [24]

13. Maharjan, Suraj and **Arevalo, John** and Montes, Manuel and González, Fabio A and Solorio, Thamar, "A Multi-task Approach to Predict Likability of Books", in Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Pap... vol. 1, (2017), pp. 1217–1227. [25]

Another collaborations with representation learning methods and scalable implementations were also published:

1. **Arevalo, John** and Ramos-Pollan, Raúl and González, Fabio A, "Distributed Cache Strategies for Machine Learning Classification Tasks over Cluster Computing Resources", in High Performance Computing (2014), pp. 43–53. [26]

2. Perdomo, Oscar and Otalora, Sebastian and Rodríguez, Francisco and **Arevalo, John** and González, Fabio A, "A Novel Machine Learning Model Based on Exudate Localization to Detect Diabetic Macular Edema", (2016). [27]

3. Perdomo, Oscar and **Arevalo, John** and González, Fabio A, "Convolutional network to detect exudates in eye fundus images of diabetic subjects", in 12th International Symposium on Medical Information Processing and Analysis (2017). [28]

1.3 Outline

The remainder of this document is organized as follows. Next chapter gives an overall background of multimodal learning. In Chapter 3 the gated multimodal unit (GMU) is presented and its behavior is empirically evaluated in synthetic experiments. In Chapter 4, the GMU is evaluated in the automatic genre classification task. The model improved the traditional early and late fusion strategies. In Chapter 5, the GMU is integrated with convolutional architectures to address image segmentation using multimodal information. Then, the model is evaluated for combining handcrafted and learned features for mass lesion classification in mammography images (Chapter 6). In Chapter 7, the GMU is used to predict the success of books based on their content. This model learns to combine features obtained with representation learning techniques and a set of hand-engineering features to outperform previous state-of-the-art results. Finally, Chapter 8 summarizes the main aspects of this research, discuses the conclusions and provides future directions in multimodal representation learning.

2 Background and related work

The proposed approach in this document is based on three main areas: representation learning, multimodality and large scale machine learning. This chapter present a review of previous works reported in such areas and their open challenges.

2.1 Representation learning

When applying machine learning strategies for automatic analysis tasks, the representation of the data is a fundamental stage. The main goal of the representation is to transform the original data to extract features that facilitate the automatic analysis task of the learning algorithm (SVM, K-means, GMM, etc.). Traditional approaches involve the guidance of experts in the task to design specific feature extractors; e.g. texture and intensity features are frequently used for image analysis. An alternative for the design of such representation strategies at hand, is to include the representation stage in the learning procedure. Representation learning seeks to automatically learn transformations of the data that make easier automatic analysis tasks. These strategies include supervised dictionary learning [29], matrix factorization [30], different forms of clustering and deep learning among others. In particular, deep learning has shown to be one of the most effective approaches to learn useful transformations for automatic analysis tasks. The dominance of deep learning, in comparison with the other methods is mainly alluded to three factors: massive amount of data to train the models, increase of the complexity of the models and scalability to deal with large datasets and large models. These factors also limit the scenarios where deep learning can be applied. On the one hand, The computational cost is higher with respect to the other representation models. On the other hand, deep learning models can to easily overfit small datasets and low-dimensional input data.

Inspired by how the brain works, deep learning has been a very successful strategy to learn representations from data. Bengio, Courville, and Vincent [2] defines deep learning methods as those that are formed by the composition of multiple linear and non-linear transformations of the data, with the goal of yielding more abstract and ultimately more useful representations. A recent state-of-the-art review in representation learning and deep learning [10] describes how these methods have increased rapidly in popularity and research activity due a remarkable string of empirical successes both in academy and industry, beating traditional approaches in each application domain with breakthrough results. Microsoft released in 2012 a new version of their MAVIS (Microsoft Audio Video Indexing Service) speech recognition system based on deep learning [31], where they reduced the word error rate on four major benchmarks by about 30% compared to state-of-the-art models based on Gaussian mixtures for the acoustic modeling. In object recognition, first deep learning approaches were addressed over MNIST digit image classification task [32, 33], breaking the predominance of Support Vector Machines (1.4% error), whereas in natural images the lat-

est breakthrough has been achieved on the Imagenet dataset for object classification claiming the state-of-the-art with 4.94% error rate [34] surpassing the human level performance (5.1%), also using deep learning models to classify a dataset of 1.2 million of images and 1000 classes.

In summary, representation learning strategies have proved to be efficient finding functions that map from one modality to relatively small set of categorical variables. For some of these particular tasks, the machine learning performance have surpassed the human performance [34]. These advances do not mark the frontier in this area, but the beginning of a set of new challenges and applications that can help to both improving quality of life and boost human knowledge. In computer vision area, for instance, image captioning task has recently received a lot of attention [35–39]. The goal is to generate a syntactically and semantically correct phrase that explains the content of the image. This kind of tasks, by definition, requires the interaction of two modalities: images and text.

2.2 Learning with multimodal representations

In recent years, Internet has given rise to complex end-user interactions by describing a single concept in different ways. As an example consider a youtube video; It is composed by the video itself, which in turn contains not only its audio streaming, but also user comments, rating, user profile, among other information that explain, in some way or another, the content of the video. A similar phenomenon is presented in Wikipedia or News websites, where the articles content is supported and linked with multimedia resources. Like these, there are plenty of scenarios where people and other system interact with complex information. These interactions as well as other technological advances has increased the amount of these kind of multimodal information [5]. Such growth is resulting in widespread attention to find automatic analysis techniques that allow to exploit those multimodal databases.

Different reviews [5–8] have summarized strategies that addressed multimodal analysis. Most of the collected works claimed the superiority of multimodal over unimodal approaches for automatic analysis tasks. Figure 2.1 shows the standard workflow of a model and their components reported in the literature to perform automatic analysis tasks in multimodal scenarios using representation learning strategies. A conventional multimodal analysis system receives as input two or more modalities that describe a particular concept. The most common multimodal sources are video, audio, images and text. The first step (Section 2.2.1) is to transform the raw data into a set of useful features such that they can explain the content of the data in a compact way. The second step (Section 2.2.2) fusions extracted features to find correlations and patterns that links both modalities into a single concept. Loss functions formulations, neural networks and probabilistic models are the standard approaches used to fuse the information. Finally, in the third step (Section 2.2.3) a supervised model is trained to perform the automatic analysis task.

2.2.1 Representation

Representation is a fundamental process for machine learning. Its goal is to extract useful features from training data which are later fed to a learning algorithm. In computer vision, representation corresponds to calculate values from input images. In audio signal processing, representation

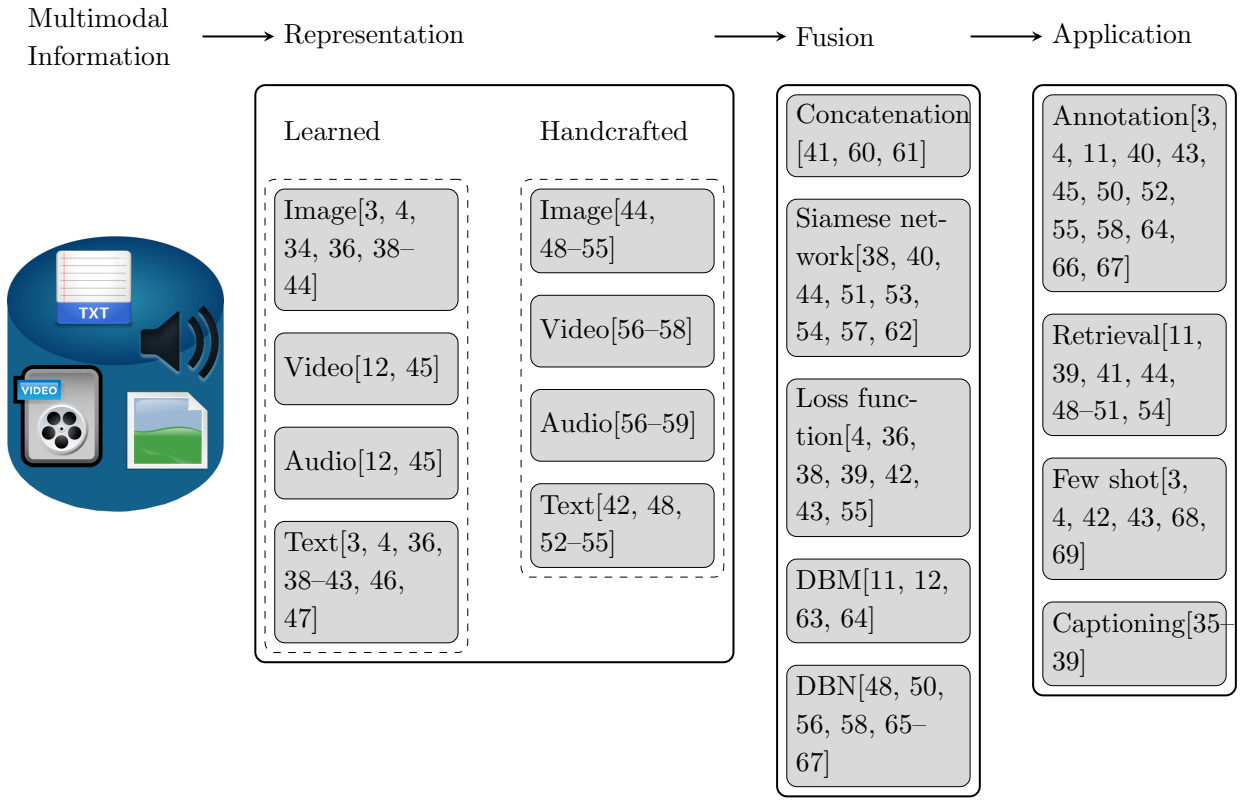


Figure 2.1: Standard workflow reported in the literature. Feature extraction stage aims to represent in an efficient way each modality separately. Fusion stage combines representations trying to find correlations between them. Finally, the combined representation is used to solve the automatic analysis task.

corresponds to compute numerical values that characterizes the complex nature of audio signals. These values (features) represent particular characteristics of the original data and are calculated from the raw values. The functions used to compute such features are called feature detectors.

Traditional approaches are based on standard or hand-crafted feature detectors which are manually selected to fit the problem at hand using expert knowledge in the domain. SIFT variants [11, 44, 52, 53] and MPEG-based [11, 53, 58] for images; Bag of words [52, 53] for textual; and Mel frequency cepstral coefficients (MFCCs) [56, 59, 70] for audio are the traditional feature detectors in multimodal environments. A main drawback in hand-crafted features is the high cost of such expert intervention. Experts usually have to design a different set of features for each problem.

Many efforts have focused on improving the performance of automatic analysis tasks or enhancing the representation using domain knowledge. Representation learning tackle this problem from a different perspective. Instead of designing custom feature detectors, representation learning learns them from data. There are two seminal works that applied representation learning models to fusion different modalities. Firstly, Ngiam et al. [12] tried several strategies to combine audio and visual data to perform speech recognition discovering useful multimodal features, however they couldn't outperform state-of-the-art based on hand engineering features. Secondly, Srivastava and Salakhutdinov [11] applied a generative model to learn features from the joint space of images and text to perform image classification and retrieval. Here, images were represented by 3857 hand-engineering features. Despite other works has been developed extending both works [11, 12], two main challenges remain open: modeling of mid-level relationships between modalities and generalization of feature extraction. These challenges, among others, has been also discussed in the Bengio's foresight [2].

In multimodal scenarios, representations learning strategies have been reasonably standard for each modality. Probabilistic models such as Deep Boltzmann Machines (DBMs) [11, 12, 45, 65] have been successfully applied to learn representations from video and audio. Conventional text representations based on count vectors, such as bag of words, has been gradually replaced by neural language models: Models learned by huge amount of documents that exploit context information to get a word embedding space using neural networks [3, 4, 36, 38–43, 46, 47]. Analogously, image representation has been replaced by Convolutional Neural Networks (CNN), in particular, the architecture proposed by Krizhevsky, Sutskever, and Hinton [71] which achieved a remarkable milestone on the ImageNet Challenge. Due to its popularity, CNNs have become the de facto standard to represent natural images. Recent works have used pretrained network [3, 4, 36, 38, 41, 42] as black-box representation methods. Other multimodal models use precomputed representations as an initialization strategy, or set as fixed. To the best of our knowledge there are not attempts to model relationships in the very low level feature representation, but only in higher ones.

2.2.2 Multimodal fusion

Previous section details strategies to find useful representations for each modality. Fusion stage seeks one single representation such that makes easier automatic analysis tasks when building classifiers or other predictors. A naive approach is to concatenate features to get a final representation [41, 60, 61]. Although it is a straightforward strategy, it ignores inherent correlations between different modalities.

More complex fusion strategies includes Restricted Boltzmann Machines and autoencoders. Ngiam et al. [12] concatenated higher level representations and train two RBMs to reconstruct the original audio and video representations respectively. Additionally, they trained a model to reconstruct both modalities given only one of them as input. In an interesting result, Ngiam et al. [12] were able to mimic a perceptual phenomenon that demonstrates an interaction between hearing and vision in speech perception known as McGurk effect. A similar approach was proposed by Srivastava and Salakhutdinov [11] modifying feature learning and reconstruction phases by Deep Boltzmann Machines. Authors claimed that such strategy is able to exploit large amounts of unlabeled data by improving the performance in retrieval and annotation tasks. Other similar strategies propose to fusion modalities using neural network architectures [38, 40, 44, 51, 53, 54, 57, 62] with two input layers separately and including a final supervised layer such as softmax regression classifier.

An alternative approach involves an objective or loss function suited for the target task [4, 36, 38, 39, 42, 43, 55]. These strategies usually assume that there exists a common latent space where modalities can express the same semantic concept through a set of transformations of the raw data. The semantic embedding representations are such that two concepts are similar if and only if their semantic embeddings are close [3]. In [43] a multimodal strategy to perform zero-shot classification was proposed. They trained a word-based neural network model [46] to represent textual information, whilst use unsupervised feature learning models proposed in [72] to get image representation. The fusion was done by learning an image linear mapping to project images into the semantic word space learned in the neural network model. Additionally a Bayesian framework was included to decide whether an image is of a seen or unseen class. Frome et al. [4] learn the image representation using a CNN trained with the Imagenet dataset and a word-based neural language model [47] to represent the textual modality. To perform the fusion they re-train CNN using text representation as targets. This work outperform scalability with respect to [43] from 2 to 20,000 unknown classes in the zero-shot learning task. A modified strategy of [4] was presented by [3]. Instead of re-train the CNN network, they built a convex combination with probabilities estimated by the classifier and semantic embedding vector of the unseen label. This simple strategy outperforms state-of-the-art results. However it should notice that the success of such approach relies totally in the power of the pre-trained models to disentangle latent concepts in large collections.

2.2.3 Applications

A set of comprehensive reviews regarding to representation learning applications has been recently published [6, 8, 10, 73] highlighting works related with multimodality. Particularly, they have encompassed such works in 3 broad group: Transfer learning, multi-task learning and zero-shot learning. Table **2-1** shows a summary of published papers related to multimodal representation learning, detailing used modalities as well as their application tasks.

Bengio, Courville, and Vincent [2] defines transfer learning as "the ability of a learning algorithm to exploit commonalities between different learning tasks to share statistical strength and transfer knowledge across tasks". This ability can be exploited in a multimodal scenario if the algorithm can discover those statistical relationships that exists in different sources and that are explaining the same abstracted concept. Under this setup, a competition was organized [77] to motivate how unsupervised strategies can be exploited in the transfer learning scenario. Such competition was

	Modality 1	Modality 2	Application
[12, 56, 70, 74]	audio	video	speech-recognition
[45, 58]	audio	video	annotation
[59]	audio	images	cross-modal
[3, 11, 40, 43, 50, 52, 53, 55, 67]	images	text	annotation
[35–39]	images	text	image captioning
[11, 39, 41, 44, 48–51, 54]	images	text	retrieval
[75]	images	text	clustering
[60]	images	depth	segmentation
[66]	text	stats-info	annotation
[76]	images	depth	object detection

Table **2-1**: Summary of modalities and applications reported in multimodal learning representation strategies

won by [78] where a mix of different autoencoders architectures were assembled. Under the zero-shot learning scenario the goal is to build a strategy to learn from a labeled dataset, and in test stage, the target labels are not in training set. A natural way to address this task is to find a set of transformations for input data and target labels such that in a semantic space both, inputs and targets, are close together. Interesting works has been reported using representation learning strategies. Frome et al. [4] used text information to improve an image recognition system. An independent system to classify images using only visual information was pre-trained using a CNN architecture, then a semantic embedding model was trained to find transformations of input data such that labels and visual information fall close together, under a particular similarity metric, in the new space. A modified version of the previous strategy was proposed in [3]. They do not modify model learned by the CNN, instead, they defined a deterministic transformation from the outputs of the classifier to the semantic space. This approach avoids the re-training of the CNN, which is a computationally expensive process. A last recent work was published in [43]. Herein, standard image classification and zero-shot learning tasks were addressed jointly by merging visual and textual information in a semantic space.

2.3 Scalable representation learning

The power of representation learning models comes from their capabilities to scale up in terms of both, large architectures and large amounts of data. On the one hand, a large enough architecture allows to model complex patterns and relationships inherent in the data. On the other hand, when the model learns from large datasets, it is seeing many different scenarios that help to deal with variability and to prevent overfitting.

However, these two characteristics include the scalability challenge. In order to train such models, one would be able to both, store the model in memory and do computations in short time. With large amounts of data also comes the challenge to find efficient strategies that make possible to exploit them.

The first explorations that addressed the scalability were supported on cloud computing. One of the first works that scales representation learning models was done by Dean et al. [79]. They obtained specialized neurons to detect pedestrians, faces and cats by learning a 9-layered sparse autoencoder with 1 billion connections from a set of 10 million of 200×200 unlabeled images on a cluster of 1000 machines (16.000 cores). Despite the outstanding obtained results, such models are easily reproducible, because of both, complexity and hardware architecture cost.

Nowadays, high performance computing strategies such as GPUs has made possible to train such large models in a reasonable way. Xu et al. [35] combined neural networks with visual attention model to automatically caption images. The dataset used contains around 87.000 images with 5 sentences per image. The training took 3 days with a GPU. Socher et al. [39] trained a neural network using 14M images with 22000 categories; and a neural language model to address image captioning problem. The whole training took 8 days on a large cluster of machines (not specified). Karpathy et al. [80] trained a neural network for video classification with 1 million of youtube videos for 487 classes. They start from raw pixels in frames, ignoring audio signal. Its training took 1 month.

Despite their interesting results, the amount of time required to train these models, usually days, makes unfeasible to explore different ideas. Other works have tried to perform parallel computing on GPUs [81], obtaining promising results. However these strategies are tightly coupled to the target problem, and thus, new implementations to adapt them to other domains are required.

3 Gated multimodal unit

This chapter presents a neural-network-based strategy for addressing supervised tasks with multimodal data. The key component is a novel type of hidden unit, the Gated Multimodal Unit (GMU) that learns to weight how the modalities influence to the output activation using multiplicative gates. The first part of this chapter defines the model. The second part presents a formal definition of the GMU, their components along with its bi-modal and multi-modal variants. The final part analyzes the GMU behavior in a synthetic task for denoising a multimodal input. Part of this work was published in the *International conference on learning representations* [24] and submitted to the *Transactions in neural newtworks and learning systems* journal [82].

3.1 Introduction

Representation learning methods have received a lot of attention by researchers and practitioners because of their successful application to complex problems in areas such as computer vision, speech recognition and text processing [1]. Most of these efforts have concentrated on data involving one type of information (images, text, speech, etc.), despite data being naturally multimodal. Multimodality refers to the fact that the same real-world concept can be described by different views or data types. Collaborative encyclopedias (such as Wikipedia) describe a famous person through a mixture of text, images and, in some cases, audio. Users from social networks comment about events like concerts or sport games with small phrases and multimedia attachments (images/videos/audios). Patient's medical records are represented by a collection of images, text, sound and other signals. The increasing availability of multimodal databases from different sources has motivated the development of automatic analysis techniques to exploit the potential of these data as a source of knowledge in the form of patterns and structures that reveal complex relationships [5, 6]. In recent years, multimodal tasks have received attention by the representation learning community. Strategies for visual question answering [83], or image captioning [35, 37, 84] have developed interesting ways of combining different representation learning architectures.

Different reviews [5–8] have summarized strategies that addressed multimodal analysis. Most of the reviewed works claimed the superiority of multimodal over unimodal approaches for automatic analysis tasks. A conventional multimodal analysis system receives as input two or more modalities that describe a particular object. The most common multimodal sources are video, audio, images and text. In recent years there has been a consensus with respect to the use of representation learning models to characterize the information of this kind of sources [1]. However, the way that such extracted features are combined is still in exploration.

Multimodal combination seeks to generate a single representation that eases automatic analysis tasks when building classifiers or other predictors. A basic approach is to concatenate features to get a final representation [41, 60, 61]. Although it is a straightforward strategy, given that the

nature of data for each modality is different, their statistical properties usually are not shared across modalities [11], and thus the predictor needs to model complex interactions between them. Instead, more elaborated combination strategies has been proposed, in which prior knowledge is exploited, additional information is included or multimodal interactions are explicitly modeled. Some of those strategies include Restricted Boltzmann Machines (RBMs) and autoencoders. Ngiam et al. [12] concatenated higher level representations and trained two RBMs to reconstruct the original audio and video representations respectively. A similar approach was proposed by Srivastava and Salakhutdinov [11]. They modified feature learning and reconstruction phases with Deep Boltzmann Machines. An alternative approach involves an objective or loss function suited for the target task [4, 36, 38, 39, 42, 43, 55]. Because the cost function involves both multimodal combination and supervision, these family of models are tied to the task of interest. Thus, if the domain or task conditions change, an adaptation of the model is required.

Also, most of these models are focused on mapping from one modality to another or solving an auxiliary task to create a common representation with the information of all modalities. In this work, we design a novel module that combines multiple sources of information, which is optimized with respect to the end goal objective function. Our proposed module is based on the idea of using gates for combining input modalities giving a higher importance to the ones that are more likely to contribute for correctly generating the desired output. We use multiplicative gates that assign importance to various features simultaneously, creating a rich multimodal representation that does not require manual tuning, but instead it learns directly from the training data. We show in the experimental evaluation that our gated model can be reused in different network architectures for solving different tasks, and can be optimized end-to-end with other modules in the architecture using standard gradient-based optimization algorithms. Such behavior was evidenced in the experimental analysis that suggested that the gain is based on giving more weight to specific modalities for specific problems.

We initially explored two application use cases: genre movie prediction, and image segmentation. On the one hand, genre prediction has several application areas like document categorization [85], recommendation systems [86], and information retrieval systems, among others. On the other hand, image segmentation is heavily used in autonomous drive systems [87], medical imaging [88] and other computer vision tasks. The main hypothesis of this work is that a model using GMU, in contrast to conventional multimodal late and early fusion architectures, will be able to learn an input-dependent gate-activation pattern that determines how each modality contributes to the output of hidden units. The motivations to chose the above tasks are twofold: 1) to evaluate the model in different and unrelated scenarios in order to support that the model is suitable for different multimodal learning tasks, and 2) to integrate the proposed unit in the most popular network architectures: convolutional and fully connected.

The proposed model is closely related to the mixture of experts (MoE) approach [89]. However, the common usage of MoE is focused on performing decision fusion, i.e. combining predictors to address a supervised learning problem [90]. Similar late-fusion models have been extended to deep architectures with bagging methods [91]. Our model is devised as a new component in the representation learning scheme, making it independent from the final task (e.g. classification, regression, unsupervised learning, etc) provided that the defined cost function be differentiable. On the other hand, It is noteworthy that extending current models to deal with more than two

modalities is a complex challenge [92]. Our proposed method addressed this multimodal challenge by generalizing the gate approach with independent parameters per modality.

3.2 Gated multimodal unit

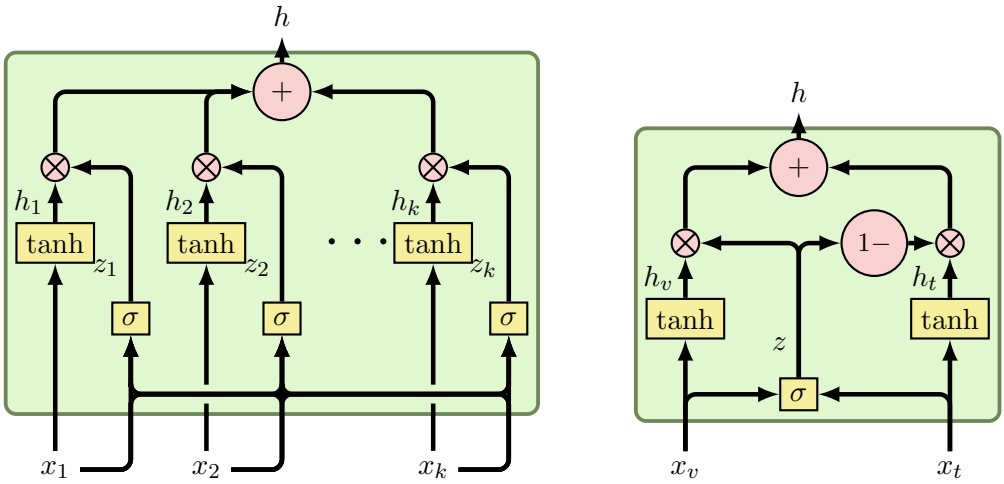
Multimodal learning is closely related to data fusion. Data fusion looks for optimal ways of combining different information sources into an integrated representation that provides more information than the individual sources [5]. This fusion can be performed at different levels and can be categorized into two broad categories: feature fusion and decision fusion. Feature fusion, also called early fusion, looks for a subset of features from different modalities, or combinations of them, that better represent the information needed to solve a particular problem. On the other hand, decision fusion, or late fusion, combines decisions from different systems, e.g. classifiers, to produce consensus. This consensus may be reached by a simple average, a voting system or a more complex Bayesian framework.

In this work we present a model, based on gated neural networks, for data fusion that combines ideas from both feature and decision fusion. The model, called Gated Multimodal Unit (GMU), is inspired by the control flow in recurrent architectures like gated recurrent units [93] or the long short-term memory unit Hochreiter and Schmidhuber [94]. A GMU is intended to be used as an internal unit in a neural network architecture whose purpose is to find an intermediate representation based on a combination of data from different modalities. Figure 3.1.a depicts the structure of a GMU. Each x_i corresponds to a feature vector associated with modality i. Each feature vector feeds a neuron with a tanh activation function, which is intended to encode an internal representation feature based on the particular modality. For each input modality, x_i, there is a gate neuron (represented by σ nodes in the diagram), which controls the contribution of the feature calculated from x_i to the overall output of the unit. When a new sample is fed to the network, a gate neuron associated to modality i receives as input the feature vectors from all the modalities and uses them to decide whether the modality i may contribute, or not, to the internal encoding of the particular input sample.

Figure 3.1.b shows a simplified version of the GMU for two input modalities, x_v (visual modality) and x_t (textual modality). It should be noted that models from Figure 3.1.a and 3.1.b are not completely equivalent, since in the bimodal case the gates are tied. Such weight tying constraints the model, so that the units control the trade off between both modalities while they use less parameters than the multimodal case. The equations governing this GMU are as follows:

$$
\begin{aligned}
h_v &= \tanh\left(W_v \cdot x_v\right) \\
h_t &= \tanh\left(W_t \cdot x_t\right) \\
z &= \sigma\left(W_z \cdot [x_v, x_t]\right) \\
h &= z * h_v + (1 - z) * h_t \\
\Theta &= \{W_v, W_t, W_z\}
\end{aligned}
$$

with Θ the parameters to be learned and $[\cdot, \cdot]$ the concatenation operator. Since all are differentiable operations, this model can be easily coupled with other neural network architectures and trained with stochastic gradient descent.



(a) The model to use with more than two modalities. (b) A simplification for the bimodal approach.

Figure 3.1: Illustration of gated units

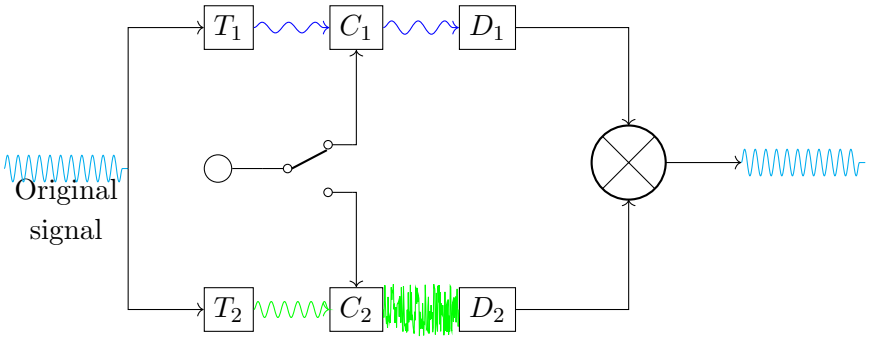


Figure 3.2: Noisy channel model. The switch determines which signal will carry the information.

3.2.1 Noisy channel model

In order to analyze the behavior of the GMU, we built a synthetic scenario to determine which modality carries the most relevant information. Consider the channel model illustrated in Figure 3.2. There is an original source signal that is transformed by two independent components T_1 and T_2. The signals from T_1 and T_2 are transmitted by two channels, C_1 and C_2 respectively, that have two operation modes. In mode one, the channel transmits the original signal, in mode two, it transmits noise. A switch controls which channel will carry the signal. In one position, C_1 carries the signal and C_2 carries noise, in the other position, the situation is inverted. The switch may change its position at any time. The goal is to get the information of the original signal from the combination of the signals C_1 and C_2 without knowing which one is carrying the information and which one is carrying noise at a given time.

We implemented the noisy channel scenario through the generative model depicted in Figure 3.3. In this model we define the random binary variable C as the target and $x_v, x_t \in \mathbb{R}$ as the input features. M is a random binary variable that decides which modality will contain the relevant information that determines the class. The input features of each modality can be generated by a

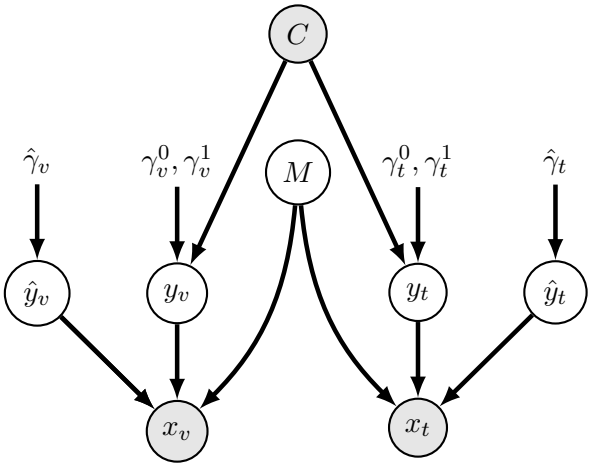


Figure 3.3: Generative model for the synthetic task. Grayed nodes represent visible variables, the other nodes represent hidden variables.

random source, $\hat{y}_v$ and $\hat{y}_t$, or by an informed source, y_v and y_t. The generative model is specified as follows:

$$
\begin{aligned}
C &\sim \text{Bernoulli}(p_C) & x_v &= My_v + (1-M)\hat{y}_v \\
M &\sim \text{Bernoulli}(p_M) & y_t &\sim \mathcal{N}(\gamma_t^C) \\
y_v &\sim \mathcal{N}(\gamma_v^C) & \hat{y}_t &\sim \mathcal{N}(\hat{\gamma}_t) \\
\hat{y}_v &\sim \mathcal{N}(\hat{\gamma}_v) & x_t &= M\hat{y}_t + (1-M)y_t
\end{aligned}
$$

We trained a model with a single GMU and applied a sigmoid function over h, then the binary cross entropy was used as loss function. Using the generative model, 200 samples per class were generated for each experiment. 1000 synthetic experiments with different random seeds were run and the GMU outperformed a logistic regression classifier in 370 of them, while obtaining equal results in the remainder ones. Our goal in these simulations was to show that the model was able to learn a latent variable that determines which modality carries the useful information for the classification. An interesting result is that between M and the activations of the gate z there is a correlation of 1. This means the model was capable of learning such latent variable by only observing the x_v and x_t input features.

We also wanted to project back the z activations to the feature space in order to visualize regions depending on the modality. Figure 3.4 shows the activations in a synthetic experiment generated by the setup of Figure 3.3 for $x_v, x_t \in \mathbb{R}$. Each axis represents a modality, red and blue dots are the samples generated for the two classes and black Gaussian curves represent the $\hat{\gamma}_v$ and $\hat{\gamma}_t$ noises. The gray (white) regions of the left figure represent the activation of z. Notice that in the white region ($z = 1$), the model gives more importance to the x_v modality while in gray regions ($z = 0$) the x_t modality is more relevant; i.e. the z gate is isolating the noise. The contour of the right figure (blue-red) represents the model prediction. It is noteworthy that the boundary defined by the gates still holds when the model solves the task. This also encourages the inclusion of non-linearities to the z gate so that it is able to discriminate more complex interactions between modalities.

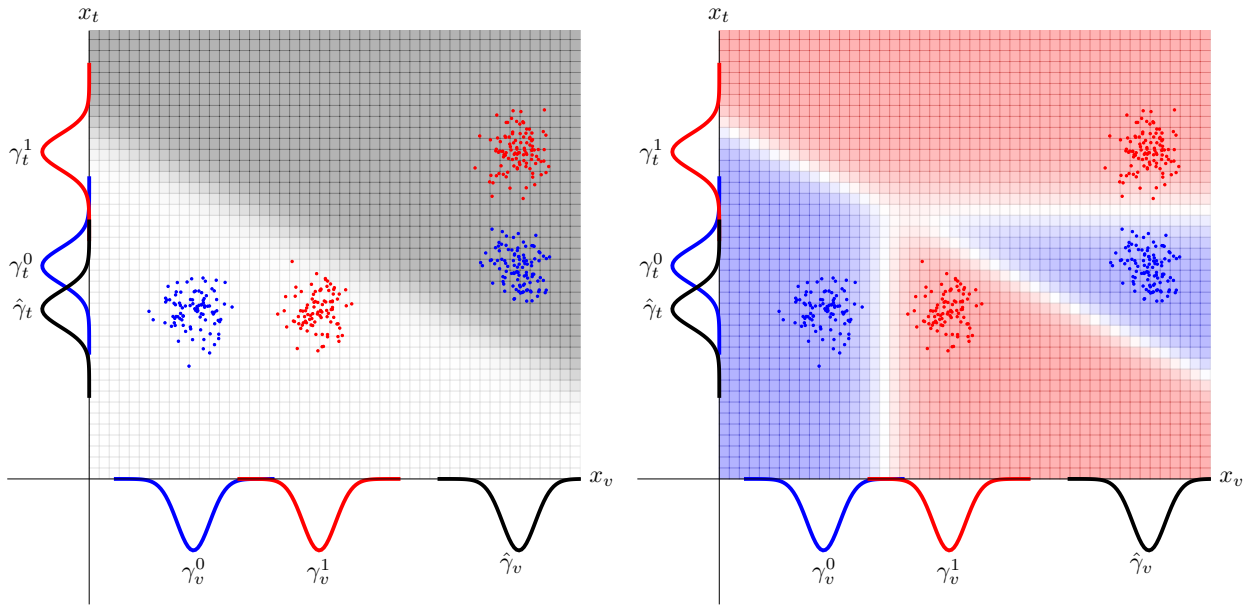


Figure 3.4: Activations of z (left) and prediction (right) for a synthetic experiment with $x_v, x_t \in \mathbb{R}$. Each axis represents a modality.

3.3 Conclusions

This chapter presented a strategy to learn fusion transformations from multimodal sources. Similarly to the way recurrent models control the information flow, the proposed model is based on multiplicative gates. The Gated Multimodal Unit (GMU) receives two or more input sources and learns to determine how much each input modality affects the unit activation. This contrasts the traditional fusion methods that adjust weights for each modality and are fixed for all instances, while the GMU weights are determined by the input. In synthetic experiments the GMU was able to learn hidden latent variables. A key property of the GMU is that, being a differentiable operation, it is easily coupled in different neural network architectures and trained with standard gradient-based optimization algorithms.

4 GMU for genre classification

This chapter explores the movie genre prediction task in a multimodal scenario. This is a multilabel task since most of the movies belong to more than one genre, (e.g. Matrix (2000) is a Sci-fi/Action movie). In this setup, Anand [95] explores the efficiency of using keywords and users' tags to perform multilabeling using the movies from MovieLens 1M dataset which contains $1,700$ movies. Also Ivasic-Kos, Pobar, and Mikec [96] and Ivasic-Kos, Pobar, and Ipsic [97] performed multilabel classification using handcrafted features from posters, with $1,500$ samples for 6 genres. Makita and Lenskiy [86, 98] use movie ratings matrix and genre correlation matrix to predict the genre. It used a smaller version of the Movielens dataset with 18 movie genres. Most of the above works have used the publicly available MovieLens datasets. However, there is not a single experimental setup defined so that all methods can be systematically compared. Also, to the best of our knowledge, none of the previous works contain more than $10,000$ samples. With this work we released a dataset created with the movies of the MovieLens 20M dataset. We included not only genre, poster and plot information used in this work, but also the poster of the movie as well as more than 50 characteristics taken from the IMDb website. Part of this work was published in the *International conference on learning representations* [24] and submitted to the *Transactions in neural newtworks and learning systems* journal [82].

4.1 Multimodal IMDb dataset

With this work we make publicly available the Multimodal IMDb (**MM-IMDb**)[1] dataset. MM-IMDb dataset is built with the IMDb id's provided by the Movielens 20M dataset [2] that contains ratings of $27,000$ movies. Using the IMDbPY [3] library, movies which do not contain their poster image were filtered out. As the final result, the MM-IMDb dataset comprises $25,959$ movies along with their plot, poster, genres and other 50 additional metadata fields such as year, language, writer, director, aspect ratio, etc.

Notice that one movie may belong to more than one genre. Figure 4.1 shows the co-occurrence matrix, where the color bar indicates the representative co-occurrence per row, while Figure 4.2 and Figure 4.3 depict the distribution of the movie poster sizes and length of movie plots respectively. Each plot contains on average 92.5 words, while the longest one contains $1,431$ words and the average of genres per movie is 2.48. In this work, we defined the task of movie genre prediction based on its plot and image poster. Nevertheless, the additional metadata information encourages other interesting tasks such as rating prediction and content-based retrieval, among others.

[1] `http://lisi1.unal.edu.co/mmimdb/`
[2] `http://grouplens.org/datasets/movielens/`
[3] `http://imdbpy.sourceforge.net/`

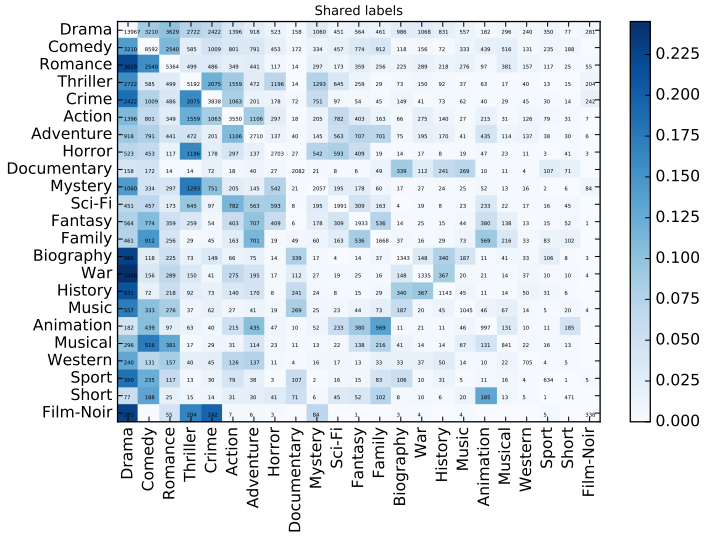


Figure 4.1: Co-ocurrence matrix of genre tags

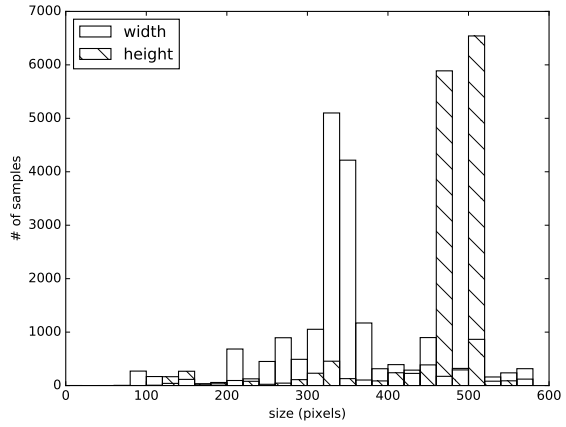


Figure 4.2: Size distribution of movie posters

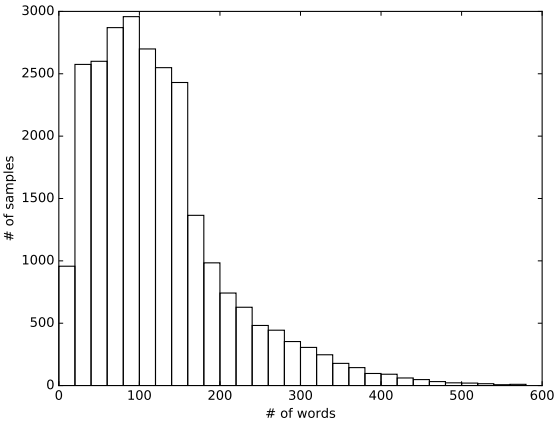


Figure 4.3: Length distribution of movie plots

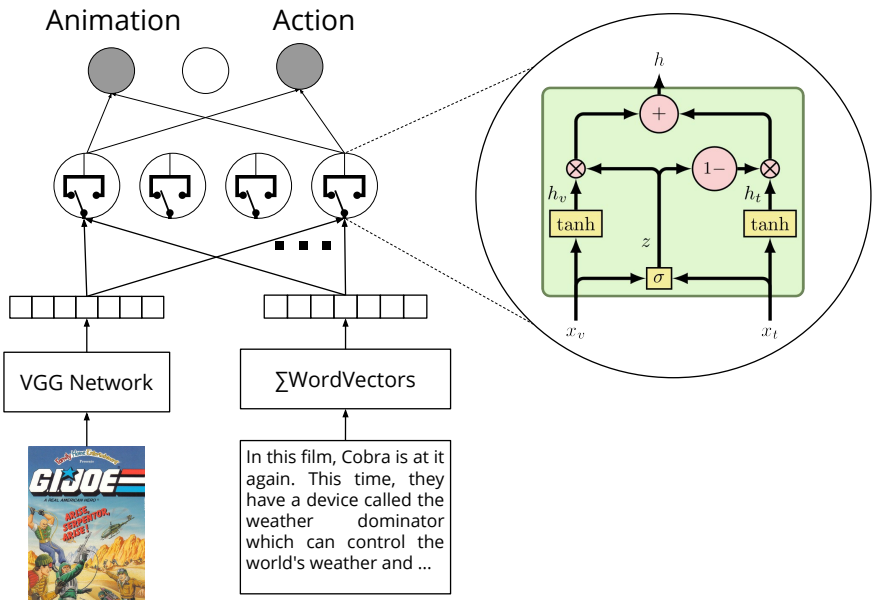


Figure 4.4: Integration of the GMU in a multilayer perceptron for genre classification

4.2 GMU for genre classification

The proposed model for genre classification is presented in Figure 4.4. Both modalities are represented with pretrained models. Then the feature vectors are fused using the GMU. Finally a multilayer perceptron (MLP) with maxout units is stacked on top. The maxout activation function $h_i : \mathbb{R}^n \rightarrow \mathbb{R}$ is a defined as:

$$h_i(\mathbf{s}) = \max_{j \in [1,k]} z_{i,j} \tag{4-1}$$

where $\mathbf{s} \in \mathbb{R}^n$ is the input vector, $z_{i,j} = \mathbf{s}^T \mathbf{W}_{\dots ij} + \mathbf{b}_{ij}$ is the output of the j-th linear transformation of the i-th hidden unit, and $\mathbf{W} \in \mathbb{R}^{d \times m \times k}$ and $\mathbf{b} \in \mathbb{R}^{m \times k}$ are learned parameters. It has been shown that maxout models with 2 hidden units behave as universal approximators, while are less prone to saturate units [99]. Since our intention is to measure how the network's depth affects the model performance, we evaluate the architecture with one and two fully connected layers.

4.3 Data representation

Given that the nature of data for each modality is different, their statistical properties usually are not shared across modalities [11]. Thus, an evaluation of different strategies for representing visual and textual content are required. For text information we evaluated word2vec models, n-grams models and RNN models. For processing visual data we evaluated two different convolutional neural networks. The details of each representation are discussed below.

4.3.1 Text representation

Text representation is a critical step when classification tasks are addressed using machine learning methods. Traditional approaches are based on counting frequencies of n-gram occurrences such as words or sequences of characters (e.g. bag-of-words models). The main drawback of such approaches is the difficulty to model relationships between words and their context. An alternative approach was initially proposed by Bengio et al. [100], by building a neural network language model (NNLM). The NNLM was able to learn distributed representations of words that capture contextual information. Later, this model was simplified to deal with large corpora by removing hidden layers in the neural network architecture (word2vec) [101]. This is a fully unsupervised model that takes advantage of large sets of unlabeled documents. Herein, three text representations were evaluated:

n-gram Following the strategy proposed by Kanaris and Stamatatos [85], we used the character n-gram strategy for representing text. Despite their simplicity, n-gram models have shown to be a competitive baseline.

Word2Vec Word2vec is an unsupervised learning algorithm that finds a vector representation for each word based on its context [101]. It has been shown that this model is able to find semantic and syntactic relationships using arithmetic operations between the vectors. Based on this property, we represent a movie as the average of the vectors of words in the plot outline. The main motivation to aggregate word2vec vectors is the property of additive compositionality that this representation has exposed over different sets of tasks such as word analogies. The usual way to aggregate the word vectors to represent a document is to perform arithmetic operations over the vectors. We take the average to avoid large input values to the neural network.

We used The pretrained Google Word2vec [4] embedding space. There were $41,612$ words from the MM-IMDb plots that are in the Google word2vec vocabulary. Other than lowercase, no text preprocessing was applied. This textual representation obtained comparable state-of-the-art results [85] in two publicly available datasets: *7genre* dataset that comprises $1,400$ web pages with 7 disjoint genres and *ki-04* dataset that comprises $1,239$ samples classified under 8 genres. Notice that the state-of-the-art model [85] used character n-grams with structured information from the HTML tags to predict the genre of web pages while ours only used the plain text.

Recurrent neural network This model takes as input a sequence of words to train a supervised recurrent neural network. Two variants were evaluated: 1) *RNN_w2v*, a transfer learning model that takes as input the word vectors of word2vec as representations; 2) *RNN_end2end*, which learns the word vectors from scratch.

4.3.2 Visual representation

In computer vision tasks, Convolutional neural networks have become the *de facto* standard. It has been shown that CNN models trained with a huge amount of data are able to learn common features shared across different domains. This characteristic is usually exploited by transfer learning

[4] `https://code.google.com/archive/p/word2vec/`

approaches. For visual representation we explored 2 strategies: transfer learning and end-to-end training.

VGG Transfer In this approach images are propagated through the VGG Network [102], a CNN trained with the ImageNet dataset, and the last hidden layer activations are used as the visual representation.

End2End CNN Here, a CNN with 5 convolutional layers and an MLP (see Section 4.2) on top was trained from scratch.

The first visual approach, *VGG_Transfer*, uses VGG network as feature extractor. The second approach takes as input the raw RGB images to a CNN. Since all the images do not have the same size, all images were scaled, and cropped when required, to 160×256 pixels keeping the aspect ratio. This CNN comprises 5 CNN layers of $5, 3, 3, 3, 3$ squared filters and 2×2 pool sizes. Each convolutional layer has 16 hidden units. The convolutional layers are connected with the *MaxoutMLP* classifier on top.

4.3.3 Multimodal fusion baselines

We evaluate 4 different ways to combine both modalities as baselines.

average probability This can be seen as a late-fusion strategy. The probabilities obtained by the best model of each modality are averaged and thresholded.

concatenation Different works have found that a simple concatenation of representations of different modalities are good for combining the information [41, 60, 61]. Herein, we concatenated both representations to train the *MaxoutMLP* architecture.

linear sum Following the way Vinyals et al. [37] combine text and images representation into a single space, this model adds a linear transformation for each modality so that both outputs have the same size to be summed up and then followed by the *MaxoutMLP* architecture.

MoE The mixture of experts (MoE) [89] model was adapted for multilabel classification. two gating strategies were explored: *tied*, where a single gate multiplies all the logistics outputs, and *untied* where every logistic output has its own gate. Logistic regression and *MaxoutMLP* were evaluated as experts.

4.4 Experimental setup

The MM-IMDb dataset has three subsets: Train, development and test subsets contain 15552, 2608 and 7799 respectively. The sample was stratified so that training, dev and test sets comprises 60%, 10%, 30% samples of each genre respectively.

In the multilabel classification the performance evaluation can be more complex than traditional *multi-class* classification and the differences can be significant among several measures [103]. Herein,

four averages of the f-score (f_1) are reported: *samples* computes the f-score per sample and then averages the results, *micro* computes the f-score using all predictions at once, *macro* computes the f-score per genre and then averages the results. *weighted* is the same as *macro* with a weighted average based on the number of positive samples per genre. F-scores are calculated as follows [103]:

$$\mathrm{p}^{micro} = \frac{\sum_{j=1}^{Q} tp_j}{\sum_{j=1}^{Q} tp_j + \sum_{j=1}^{Q} fp_j} \quad \mathrm{r}^{micro} = \frac{\sum_{j=1}^{Q} tp_j}{\sum_{j=1}^{Q} tp_j + \sum_{j=1}^{Q} fn_j}$$

$$f_1^{micro} = \frac{2 \times \mathrm{p}^{micro} \times \mathrm{r}^{micro}}{\mathrm{p}^{micro} + \mathrm{r}^{micro}}$$

$$f_1^{macro} = \frac{1}{Q} \sum_{j=1}^{Q} \frac{2 \times \mathrm{p}_j \times \mathrm{r}_j}{\mathrm{p}_j + \mathrm{r}_j}$$

$$f_1^{sample} = \frac{1}{N} \sum_{i=1}^{N} \frac{2 \times |\hat{y}_i \cap y_i|}{|\hat{y}_i| + |y_i|} \quad f_1^{weighted} = \frac{1}{Q^2} \sum_{j=1}^{Q} Q_j \frac{2 \times \mathrm{p}_j \times \mathrm{r}_j}{\mathrm{p}_j + \mathrm{r}_j}$$

With N the number of examples; Q the number of labels; Q_j the number of true instances for the j-th label; p the precision, r the recall; $\hat{y}_i, y_i \in (0,1)^Q$ the prediction and ground truth binary tuples respectively; $tp_j, fp_j \text{and} fn_j$ the number of true positives, false positives and false negatives for the j-th label respectively.

4.4.1 Neural network training

Neural network models were trained using batch normalization scheme [104]. This strategy applies a normalization step across samples that belong to the same batch, so that each hidden unit in the network receives a zero-mean and unit variance. Stochastic gradient descent with ADAM optimization [105] was used to learn the weights of the neural network. Dropout and max-norm regularization were used to control overfitting. Hidden size ($\{64, 128, 256, 512\}$), learning rate ($\left[10^{-3}, 10^{-1}\right]$), dropout ($[0.3, 0.7]$), max-norm ($[5, 20]$) and initialization ranges ($\left[10^{-3}, 10^{-1}\right]$) parameters were explored by training 25 models with random (uniform) hyperparameter initializations and the best was chosen according to validation performance. It has been reported that this strategy is preferable over grid search when training deep models [106]. All the implementation was carried on with the Blocks framework [107] [5].

During the training process, we noticed that batch normalization considerably helped in terms of training time and convergence, resulting in less sensitivity to hyperparameters such as initialization ranges or learning rate. Also, dropout and max-norm regularization strategies helped to increase the performance at test time.

For classification stage, two methods to map from feature vectors to genre classification were explored: 1) Logistic regression and 2) a multilayer perceptron (MLP) with fully connected layers and maxout activation function.

Experiments are supported by the McNemar statistical test to determine whether the differences have statistical evidence ($p < 0.01$).

[5] `https://github.com/johnarevalo/gmu-mmimdb`

Table **4-1**: Summary of classification task on the MM-IMDb dataset

Modality	Representation	F-Score			
		weighted	samples	micro	macro
Multimodal	GMU	**0.624**	**0.634**	**0.636**	**0.549**
	Linear_sum	0.606	0.617	0.617	0.520
	Concatenate	0.599	0.607	0.609	0.520
	AVG_probs	0.604	0.616	0.615	0.491
	MoE_MaxoutMLP	0.592	0.593	0.601	0.516
	MoE_MaxoutMLP (tied)	0.579	0.579	0.587	0.489
	MoE_Logistic	0.541	0.557	0.565	0.456
	MoE_Logistic (tied)	0.483	0.507	0.518	0.358
Text	MaxoutMLP_w2v	0.604	0.607	0.612	0.528
	RNN_transfer	0.570	0.580	0.580	0.480
	MaxoutMLP_w2v_1_hidden	0.540	0.540	0.550	0.440
	Logistic_w2v	0.530	0.540	0.550	0.420
	MaxoutMLP_3grams	0.510	0.510	0.520	0.420
	Logistic_3grams	0.510	0.520	0.530	0.400
	RNN_end2end	0.490	0.490	0.490	0.370
Visual	VGG_Transfer	0.416	0.436	0.449	0.284
	CNN_end2end	0.370	0.350	0.340	0.210

4.5 Results

Table **4-1** shows the results in the proposed dataset. For the textual modality, the best performance is obtained by the combination of word2vec representation with an MLP classifier. The behavior of all representation methods are consistent across the performance measures. Learning from scratch the RNN model performed the worst. We hypothesize this has to do with the lack of data to learn meaningful relations among words. It has been shown that millions of words are required to train a model such as word2vec that is able to exploit common regularities between word co-occurrences.

For the visual modality, the usage of pretrained models works better than training the model from scratch. It seems it is still a small dataset to learn all the complexities of the posters. Now, comparing the performance independently per genre, as in Table **4-2**, it is interesting to notice that in *Animation* the visual modality outperforms the textual one.

In the multimodal scenario, by adding the GMU as building block to learn the fusion we obtained the best performance, improving independent modalities in the averaged measures and in 16 of out 23 genres and outperforming all other evaluated fusion strategies. The concatenation or the linear combination approaches were not enough to model the correlation between the modalities and MoE models did not perform better than simpler approaches. This is an expected behavior for MoE in a relatively small dataset because the data is fractionated over different experts, and thus it doesn't

Table **4-2**: Macro F-Score reported per genre for single and multimodal approaches.

Genre	**Textual**	**Visual**	**GMU**	**Genre**	**Textual**	**Visual**	**GMU**
Drama	0.75	0.68	**0.77**	Family	0.51	0.47	**0.58**
Comedy	0.63	0.59	**0.67**	Biography	**0.40**	0.01	0.25
Romance	0.52	0.32	**0.52**	War	0.65	0.16	**0.66**
Thriller	0.58	0.41	**0.61**	History	**0.41**	0.06	0.37
Crime	0.63	0.27	**0.65**	Music	0.57	0.04	**0.57**
Action	0.58	0.38	**0.62**	Animation	0.43	0.61	**0.65**
Adventure	**0.53**	0.32	0.51	Musical	0.22	0.19	**0.27**
Horror	0.65	0.43	**0.70**	Western	0.64	0.33	**0.68**
Documentary	0.75	0.18	**0.76**	Sport	**0.69**	0.14	0.68
Mystery	0.39	0.12	**0.39**	Short	0.29	0.20	**0.30**
Sci-Fi	0.66	0.31	**0.67**	Film-Noir	0.20	0.09	**0.30**
Fantasy	**0.45**	0.22	0.44				

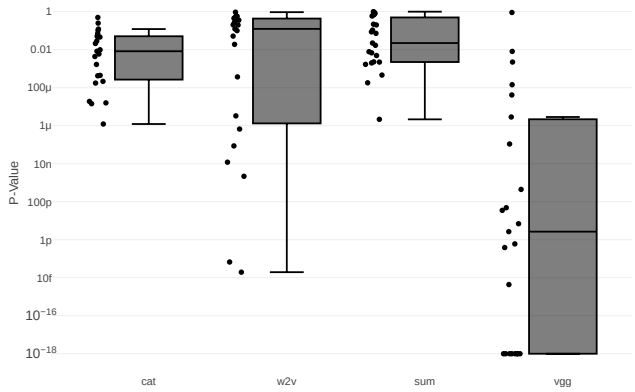


Figure 4.5: Distribution of the p-values for comparison per genre between the GMU and other models. Each point represents a genre.

make an efficient use of the training samples.

We applied the McNemar test for GMU model vs the rest in two scenarios. First, we built the contingency table using the 179377 values from the confusion matrix (7799 test samples for 23 genres). The p-values were less than 0.01 showing that the differences are significant. We also performed the comparison per genre. The distribution of p-values is shown in figure 4.5. In this scenario, the statistical evidence showed that there is a significant difference between the GMU and the second best model for *Action*, *Horror*, *Drama*, *Thriller* and *Crime* genres. The GMU shares the first place with other method in the remainder genres.

In order to evaluate which modality influences more the model when assigning a particular label, we averaged the activations of a subset of z gates of the test samples to which the model assigned them such label. We counted the number of samples that pays more attention to the textual modality ($z <= 0.5$) or to the visual modality ($z > 0.5$). The units were chosen taking into account the mutual information between the predictions and the z activations. The result of this analysis

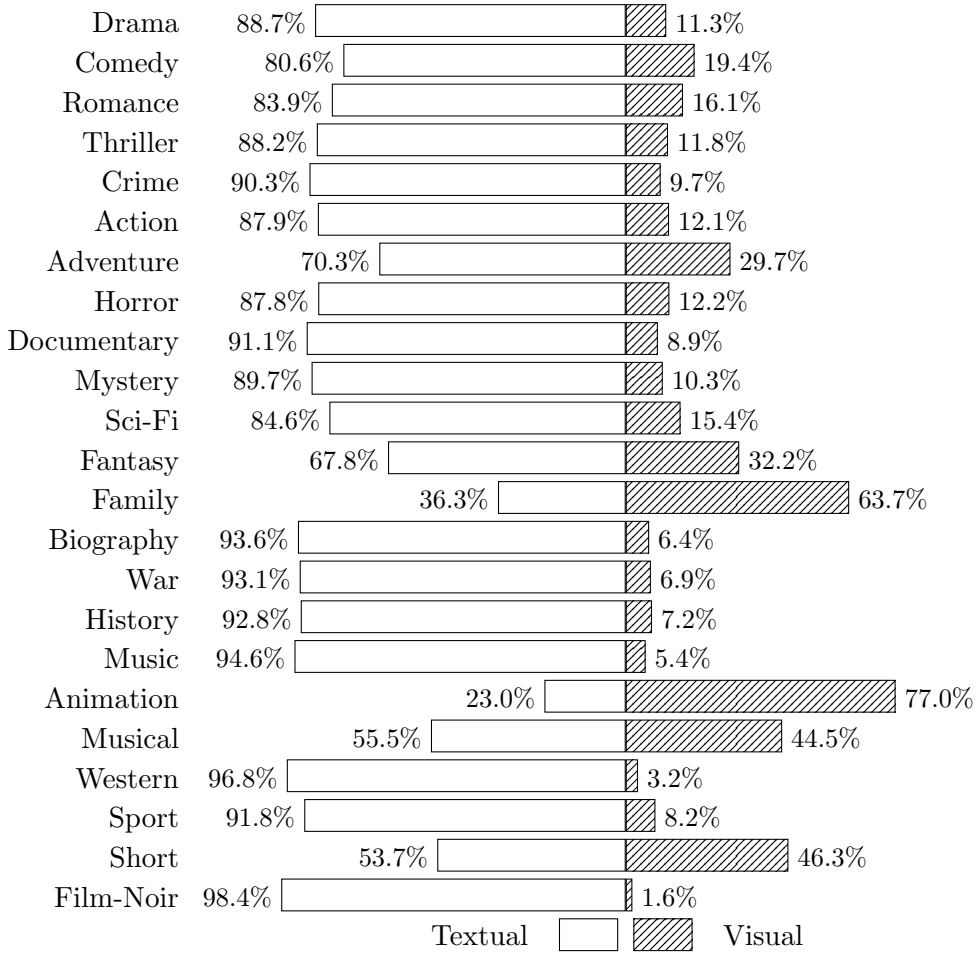


Figure 4.6: Percentage of gates activations ($z > 0.5$: Visual; $z <= 0.5$: textual) for each genre in the test set.

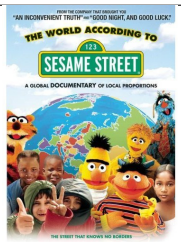
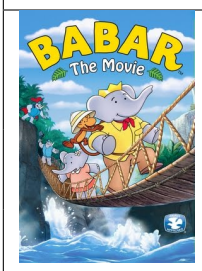
is depicted in Figure 4.6. As expected, the model is generally more influenced by the textual modality. But, in some specific genres such as *Animation* or *Family*, the visual modality affects more the model. This is also consistent with results of Table **4-2** which reports better performances for visual modality.

We wanted to qualitative explore test examples in which performance was improved by a relative large margin. Table **4-3** illustrates cases where the model takes advantage of the most accurate modality, and in some cases removes false positives. It is noteworthy that some of these examples can be confusing for a human if one modality is missing, or additional context information is not given.

4.6 Conclusions

The gated multimodal network involved a fully connected architecture taking as input the plot

Table **4-3**: Examples of predictions in test set. Red and blue genres are false positives and true positives respectively.

The World According to Sesame Street		
	a documentary which examines the creation and co - production of the popular children ' s television program in three developing countries: bangladesh , kosovo and south africa .	
	Ground Truth	Documentary
	Textual	Documentary, History
	Visual	Comedy, Adventure, Family, Animation
	GMU	Documentary
Babar: the movie		
	in his spectacular film debut , young babar , king of the elephants , must save his homeland from certain destruction by rataxes and his band of invading rhinos .	
	Ground Truth	Adventure, Fantasy, Family, Animation, Music
	Textual	Adventure, Documentary, War, Music
	Visual	Comedy, Adventure, Family, Animation
	GMU	Adventure, Family, Animation
Letters from Iwo Jima		
	the island of iwo jima stands between the american military force and the home islands of japan . (...) when the american invasion begins , both kuribayashi and saigo find strength , honor , courage , and horrors beyond imagination .	
	Ground Truth	Drama, War, History
	Textual	Drama, Action, War, History
	Visual	Thriller, Action, Adventure, Sci-Fi
	GMU	Drama, War, History

of the movie and the image poster to annotate (multi-label) 23 genres. Experimental evaluation showed the model learned to weight the modalities based on the input features, and outperformed early and late fusion approaches by 3% in terms of F-score

5 GMU for image segmentation

The proposed unit is easily adaptable to other architectures different from the traditional "Fully connected". Since the GMU is a differentiable operator, it can be applied to part of the input and still be optimized with gradient-based methods. This is the basic idea of convolutional architectures. This chapter adapts the GMU to convolutional neural networks for addressing image segmentation. The model learns to fuse RGB and depth information to outperform standard early and late fusion strategies. An analysis to the learned model highlights correlations between modalities and semantic concepts. Part of this work was submitted to the *Transactions in neural newtworks and learning systems* journal [82].

5.1 Introduction

Multimodal image segmentation has been addressed with representation learning techniques using RGB and depth images. Pei et al. [60] learned a dictionary from concatenated patches from RGB and depth images to extract features from small regions, then those features are used to trained a pixel-based classifier. In a similar setup, Valada, Dhall, and Burgard [108] integrated a mixture of experts model in a convolutional neural network to segment 6 concepts in outdoor images. They explored different modalities, obtaining their best results when RGB and depth images were combined. Our work is similar because it is also an end-to-end convolutional neural network, trained with gradient-based algorithms, but differs in the way the modalities are fused. While [108] used two predictors to combine the information, we instead used gates to combine intermediate representations. This allows our model to be applied also in unsupervised tasks such as image generation or feature learning, provided that the model can be trained with gradient-based approaches.

Convolutional architectures are widely used in image processing scenarios. Shortly after the Imagenet success [71], CNN became the de-facto standard architecture when using neural networks for image representation. In CNN, there are convolution and pooling transformations to the input image. Consider the input image $M \in \mathbb{R}^{p \times p}$, the first transformation applies a convolution with a filter $K \in \mathbb{R}^{k \times k}$ to obtain a feature map $S \in \mathbb{R}^{(p-k+1) \times (p-k+1)}$, followed by a non-linearity activation function $a : \mathbb{R} \Rightarrow \mathbb{R}$ applied in an element-wise fashion. The second transformation reduces the dimension of the feature map by applying a local subsampling function over the output feature map $a(S)$.

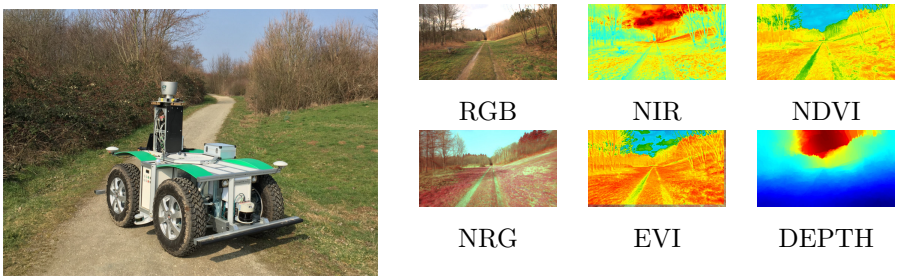


Figure 5.1: Left: Robot used to capture the images. Right: sample from the Deep scene dataset with the available modalities (taken from Valada et al. [109]).

5.2 Deep scene dataset

The convolutional architecture is evaluated in the DeepScene dataset [1] [109]. The dataset was collected using an autonomous mobile robot platform equipped with a stereo vision camera and a modified dashcam for acquiring RGB and Near-InfraRed (NIR) data respectively. Both cameras were time synchronized and frames were captured at 20Hz. Additional image post-processing was applied to match both images. Figure 5.1 shows the autonomous robot platform and one example with the available modalities.

The data was collected on three different days to have variability in lighting conditions as shadows and sun angles play a crucial role in the quality of acquired images. The DeepScene dataset comprises 366 images with pixel-level groundtruth annotations which were manually annotated with 1 out of the 6 concepts: {*grass*, *obstacle*, *tree*, *vegetation*, *road* and *sky*}. It also provides train and test sets with 230 and 135 scenes respectively [2].

Global-based vegetation indices such as Normalized Difference Vegetation Index (NDVI) and Enhanced Vegetation Index (EVI) to extract consistent spatial and global information were computed as shown by Huete, Justice, and Van Leeuwen [110]. Depth images were obtained using the approach from Liu, Shen, and Lin [111] that employs a deep convolutional neural field model for depth estimation by constructing unary and pairwise potentials of conditional random fields. the Multispectrum channel fusion NRG (Near-Infrared, Red, Green) image was also computed and included as another modality. We choose RGB and Depth images as input to the proposed multimodal approach because these are the most common and general modalities. The remainder ones are specific for environments with abundant presence of vegetation.

5.3 Convolutional GMU for segmentation

Some tasks involve multimodal sources that are suitable to be represented by a convolutional architecture. This is the case of image segmentation using RGB and depth images. Both of them represent the same scene, but using different information. Also, both of them can be naturally represented by a CNN. This is a convenient scenario to apply the GMU to let the model learn which parts of the image are more relevant to the classification. Concretely, this work integrated

[1] `http://deepscene.cs.uni-freiburg.de/`
[2] We discarded the image with ID *b275-311* from test set because it is incorrectly annotated.

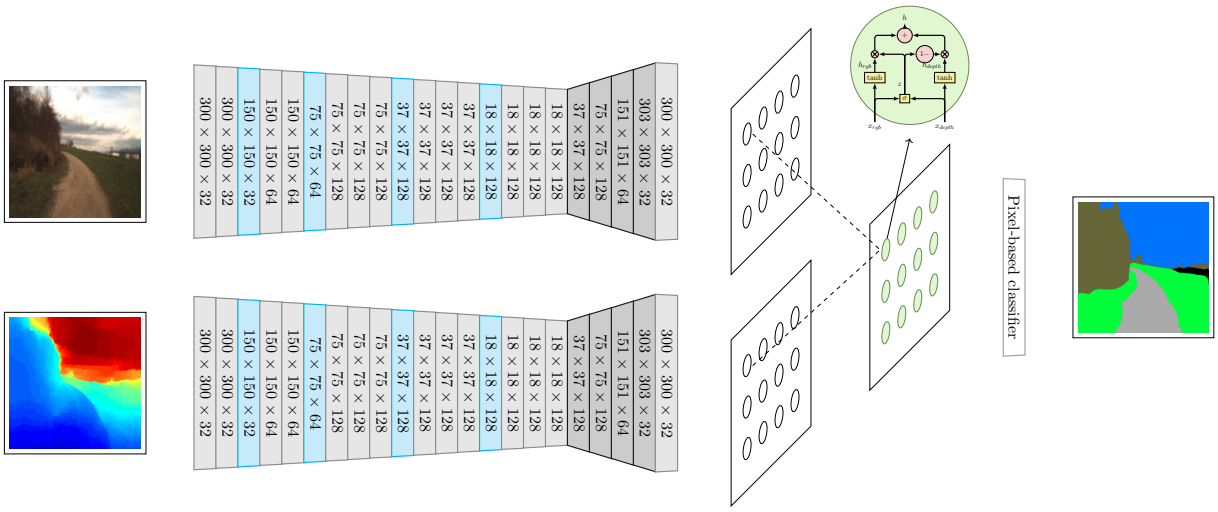


Figure 5.2: Integration of the proposed GMU in convolutional architectures. Light gray, dark gray and cyan represent convolutional, deconvolutional and pooling layers respectively. Output dimensions are denoted inside each layer. Convolutional kernels are 3×3 with padding of 1, except the last convolutional layer which has a kernel size 4×4 with zero-padding. Parameters of both convolutional networks are shared.

the GMU in a convolutional architecture as depicted in Figure 5.2, where the GMU layer takes S_{rgb} and S_{depth} feature maps as inputs and outputs a combined feature map.

5.4 Experimental setup

We took 46 scenes from train as our validation set to tune hyperparameters of the model. Hyperparameters were explored by training 25 models with random (uniform) hyperparameter initializations and the best was chosen according to validation performance.

Following the dataset authors' approach [109], images were preprocessed by resizing the original image to 300×300 pixels keeping the aspect ratio and cropping them when necessary. During training, images were oversampled by applying random rotations between $[-30, 30]$ degrees, random flipping and random cropping the images. Previous works [71] have reported this as a convenient way to artificially increase the number of training samples, which in turn helps to better generalization during the model training.

The convolutional architecture used in these experiments is detailed in Figure 5.2. The pixel-based classification layer after this deep model varies depending on the model used. For single-modality approaches, the last layer is a convolution with 6 kernels of 3×3 with border of 1 to keep the 300×300 size followed by a Softmax activation function. For the multimodal approach there is an additional layer with 32 kernels for each modality, then the ConvGMU layer that merges those 32 pairs of feature maps, followed by a Softmax activation layer.

Experiments are supported by the McNemar statistical test to determine whether the differences have statistical evidence ($p < 0.01$).

Table **5-1**: Summary of image segmentation results using single and multimodal approaches.

Method	**IoU**	**ACC**	**FPR**	**FNR**
RGB	0.840	0.964	0.029	0.083
Depth	0.630	0.914	0.064	0.239
AvgProb	0.818	0.964	0.028	0.097
Concatenate	0.851	0.969	0.025	0.084
LinearSum	0.855	0.970	0.025	0.082
ConvGMU	**0.861**	**0.971**	**0.022**	**0.077**

Table **5-2**: Intersection-over-union per class for unimodal and multimodal approaches.

Method	**Road**	**Grass**	**Vegetation**	**Sky**
RGB	0.784	0.822	0.891	0.863
Depth	0.392	0.574	0.774	0.780
ConvGMU	**0.828**	**0.842**	**0.893**	**0.880**

5.5 Results

Firstly, It is noteworthy that some inconsistencies in the original annotations were highlighted when visualizing the predictions. Figure 5.3 depicted that *obstacle* and *tree* concepts are correctly annotated in training set, but are missing in the test set. Due to such inconsistencies, in this experimentation those two concepts were discarded when methods were compared.

Following the original paper, Intersection over union (IoU), accuracy (ACC), false positive rate (FPR) and false negative rate (FNR) are used as performance measures. Table **5-1** summarizes the results for unimodal and multimodal approaches. Results showed RGB outperformed the depth modality for all classes. Also, the behavior of other multimodal approaches is consistent with the results for the MM-IMDb dataset. Here, again the GMU approach outperformed both unimodal and multimodal methods. We applied the McNemar statistical test for paired data in a pixel-wise manner. The statistical evidence showed that the differences between GMU and the remainder models are significant ($p < 0.01$) for all the classes.

As noted in Table **5-2**, IoU of *road* and *sky* concepts increased the most with the convolutional GMU model. This is consistent with the nature of the data, since closest and farthest concepts are closely related with the kind of information that depth images provide.

Likewise in the MM-IMDb task, an analysis of z activations with respect to the predictions is reported in Figure 5.4. For *road*, *grass* and *vegetation* the RGB modality is more dominant. In contrast, for *tree*, *sky* and *obstacle* the depth modality gives more information for the classifier. We believe this is consistent with the nature of the data, since concepts such as *sky* and *obstacle* would be easier to detect when additional information like distance to camera is provided.

5.6 Conclusions

The GMU has been integrated with convolutional and fully connected networks for two real su-

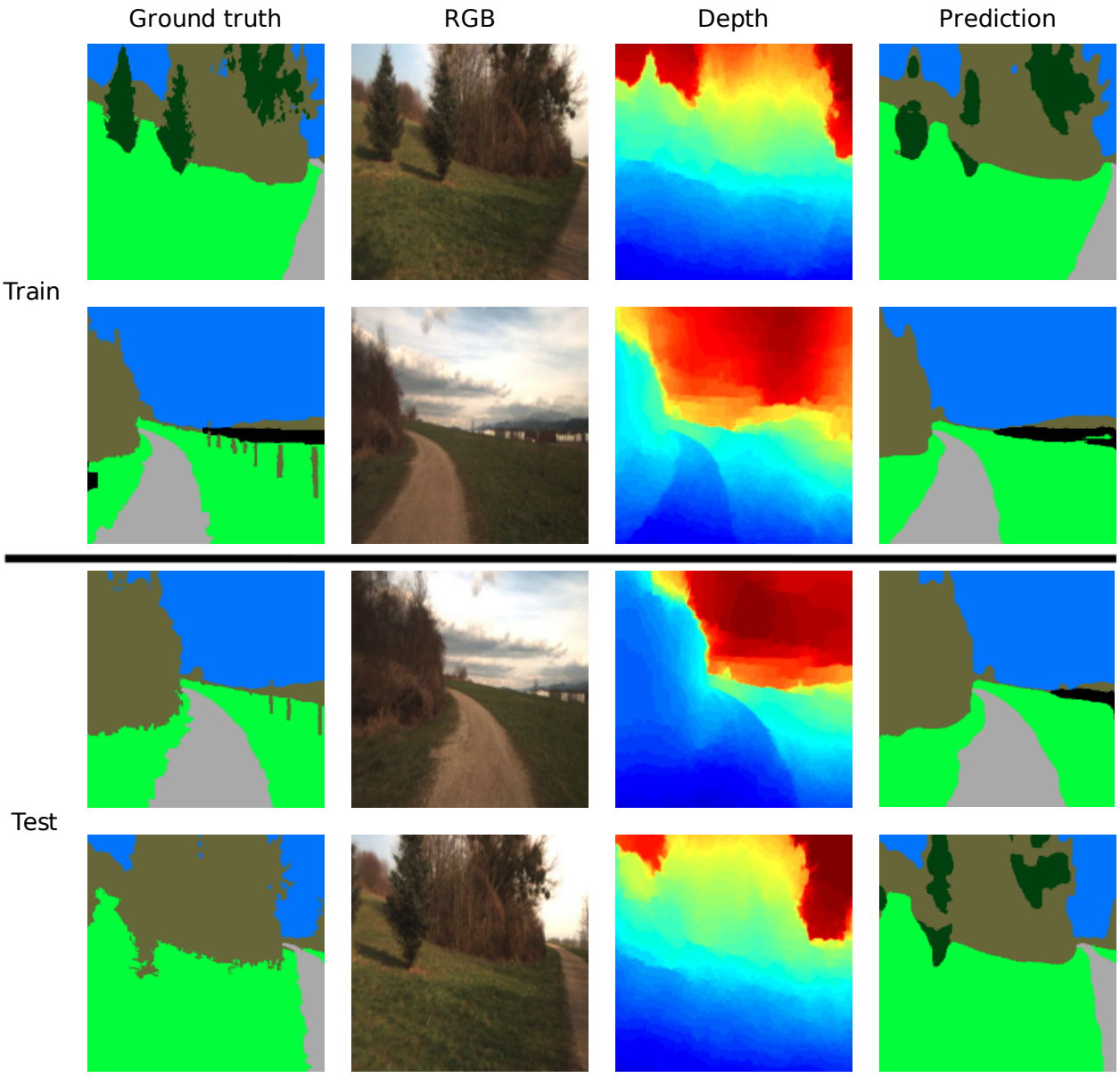


Figure 5.3: Segmentation results for the convolutional network with GMU. Concepts for the first(ground truth) and fourth (prediction) columns are colored as follows: *sky*: blue, *grass*: light green, *vegetation*: olive, *road*: light gray, *obstacle*: black, *tree*: dark green. Note that *obstacle* and *tree* concepts are correctly annotated in the training set, but at test set are absent.

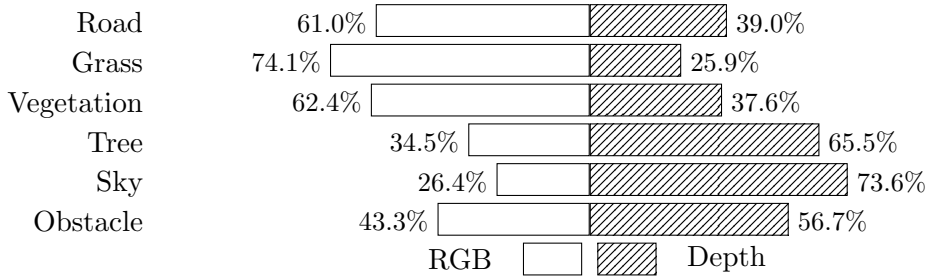


Figure 5.4: Percentage of gates activations in the image segmentation task ($z > 0.5$: RGB; $z <= 0.5$: Depth) for each concept in the test set.

pervised scenarios where it outperformed the single-modality, early and late fusion approaches. In the image segmentation task, the gated multimodal network involved an end-to-end convolutional architecture taking as input the RGB and depth images and output the segmented image with 6 semantic concepts. Likewise, the model outperformed other single and multimodal approaches measuring the Intersection-over-union score. The activations of the GMU layer were mapped to the output concepts finding correlations between input modalities and output concepts, e.g. depth information was more correlated with "sky" and "tree" while RGB is more correlated with "grass" and "vegetation". It should be noted that even though the model is capable of combining information, the content representation is critical to correctly take advantage of the different modalities. Including more complex transformations in the gate, an untied version of the bimodal case and dealing with missing modalities are interesting directions for future work.

6 GMU for medical imaging

This chapter explores the application of representation learning models for classification of mass lesions in mammography images. Different Convolutional neural network architectures were explored. Finally, the combination of learned and hand-crafted features obtained the best results. The GMU was used to weight the importance of each set of features for each sample. Part of this work was published in the ISBI conference [18] and the CMPB journal [14].

6.1 Introduction

Breast cancer is the most common cancer in women worldwide, with nearly 1.7 million new cases diagnosed in 2012 (second most common cancer overall); this represents about 12% of all new cancer cases and 25% of all cancers in women [1]. Breast cancer has a known asymptomatic phase that can be detected with mammography, and therefore, mammography is the primary imaging modality for screening. Double-reading (two radiologists independently read the same mammograms) has been advocated to reduce the proportion of missed cancers and it is currently included in most screening programs [112]. However, double-reading incurs in additional workload and costs. Alternatively, computer-aided diagnosis (CADx) systems can assist a single radiologist when reading mammograms providing support for their decisions. These systems can be used as second opinion criteria by radiologists, playing a key role in the early detection of breast cancer and helping to reduce the death rate among women with breast cancer in a cost-effective manner [113].

A successful approach to build CADx systems is to use machine learning classifiers (MLC). MLC are learned from a set of labeled data samples capturing complex relationships in the data [114–116]. In order to train a MLC for breast cancer diagnosis, a set of features describing the image is required. Ideally, features should have high discriminant power that allows inferring whether a given image is from a malignant finding or not. This is, however, a challenging topic that has gathered the focus of research in several sciences, from medicine to computer vision. Thus, several types of features may be used to infer the diagnosis. Many CADx systems use hand-crafted features based on prior knowledge and expert guidance. In particular, strategies based on feature selection [117] and hand-crafted features that characterize geometry and textures [118] has been proposed for mass classifications. As an alternative, the use of machine learning strategies to learn good features directly from the data is a new paradigm that has shown successful results in different computer vision tasks. One such paradigm is *deep learning*.

Deep learning methods have been widely applied in recent years to address several computer perception tasks [2]. Their main advantage lies in avoiding the design of specific feature detectors. In turn, deep learning models look for a set of transformations directly from the data. This

[1] World Cancer Research Fund International `http://www.wcrf.org/int/cancer-facts-figures/data-specific-cancers/breast-cancer-statistics`, Accessed May 20, 2015

approach has had remarkable results, particularly in computer vision problems such as natural scene classification and object detection [10]. Deep learning models have also been adapted to different medical tasks such as tissue classification in histology and histopathology images [119, 120], Alzheimer disease diagnosis [61, 64, 121, 122], and knee cartilage segmentation [123] among others.

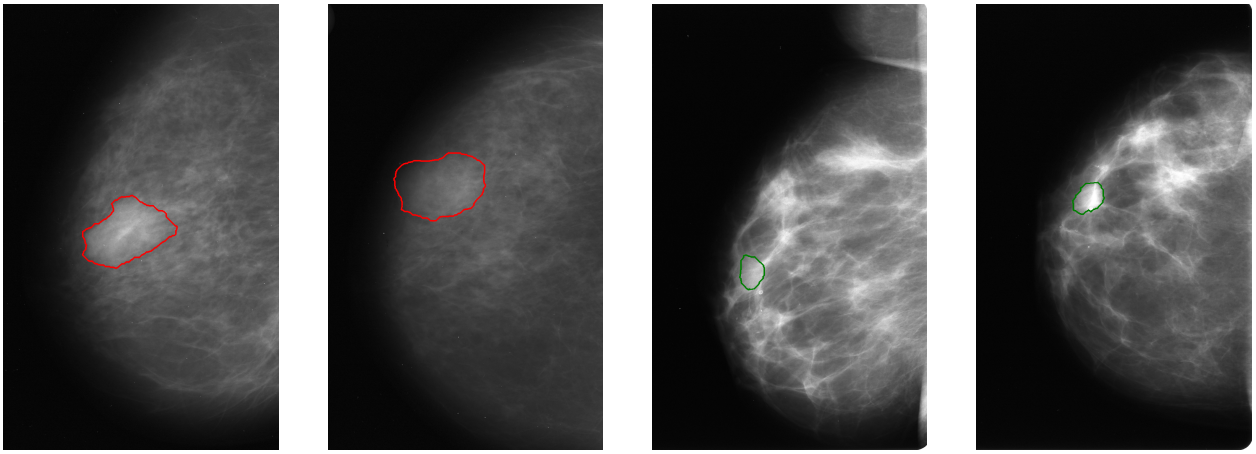
However, only few works have explored deep learning methods to address the automatic classification of identified lesions in mammography images [124]. In [125] stacked deep auto-encoders were used to estimate breast density score using multiscale features. Lately, this has been extended by including breast tissue segmentation and scoring of mammographic texture [126] with a convolutional neural network (CNN) model. CNN model is the most successful deep learning strategy applied to image understanding [10]. In [127, 128] CNNs are used as representation strategy to characterize microcalcifications. Finally, the most recent work developed in this area was done in [129] which uses Adaptive Deconvolutional Networks to learn the representation in order to classify malign/benign breast lesions. Such strategy was evaluated on 245 lesions in a bootstrap fashion, reporting the area under the ROC curve (AUC) $AUC = 0.71$. In this work, we also use convolutional architectures, however the features are learned in a supervised way during CNN training, taking advantage of expert knowledge represented by previously identified lesions in breast imaging, manually segmented by expert radiologists in both mammographic views (mediolateral oblique and craniocaudal).

The remainder of the chapter is organized as follows: Section 6.2 describes the dataset used in this exploration. Section 6.4 Details different representation strategies for visual content. Section 6.5 details the experimental setup used to evaluate the proposed approach. Finally, Sections 6.6 and 6.7 show results and present the main conclusions of this work.

6.2 Breast cancer digital repository

The benchmarking dataset used in this study is available on the Breast Cancer Digital Repository (BCDR) [2]. BCDR is a wide-ranging annotated public repository composed of Breast Cancer patient' cases in the northern region of Portugal. The BCDR is subdivided in two different repositories: (1) a Film Mammography-based Repository (BCDR-FM) and (2) a Full Field Digital Mammography-based Repository (BCDR-DM). Both repositories were created with anonymous cases from medical archives (complying with current privacy regulations as they are also used to teach regular and postgraduate medical students) supplied by the Faculty of Medicine – Centro Hospitalar São João, at University of Porto (FMUP–HSJ). BCDR provides normal and annotated patient cases of breast cancer including mammography lesions outlines, anomalies observed by radiologists, pre-computed image-based descriptors and related clinical data. The BCDR-FM is composed by 1010 patient cases (998 female and 12 male, with ages between 20 and 90 years old), including 1125 studies, 3703 mediolateral oblique (MLO) and craniocaudal (CC) mammography incidences and 1044 identified lesions clinically described (820 already identified in MLO and/or CC views). With this, 1517 segmentations were manually made and BI-RADS classified by specialized radiologists. MLO and CC images are grey-level digitized mammograms with a resolution of 720 (width) by 1168 (height)

[2] http://bcdr.inegi.up.pt



(a) (b) (c) (d)

Figure 6.1: Samples of lesions presented in the dataset. Malignant lesion in a) oblique view and b) craneo-caudal view. Benign lesion in c) oblique view and d) craneo-caudal view.

pixels and a bit depth of 8 bits per pixel, saved in the TIFF format. The BCDR-DM, still in construction, at the time of writing is composed by 724 Portuguese patient cases (723 female and 1 male, with ages between 27 and 92 years old), including 1042 studies, 3612 MLO and/or CC mammography incidences and 452 lesions clinically described (already identified in MLO and CC views). With this, 818 segmentations were manually made and BI-RADS classified by specialized radiologists. The MLO and CC images are grey-level mammograms with a resolution of 3328 (width) by 4084 (height) or 2560 (width) by 3328 (height) pixels, depending on the compression plate used in the acquisition (according to the breast size of the patient). The bit depth is 14 bits per pixel and the images are saved in TIFF format. As described below, this work is focused on the BCDR-FM Repository.

6.2.1 Benchmarking Dataset

A new dataset of the BCDR-FM repository has been made publicly available, at *http://bcdr.inegi.up.pt*, for comparison and research reproducibility purposes. The 8-bit resolution "Film Mammography Dataset Number 3" (BCDR-F03) was built as a subset of the BCDR-FM and it is composed of 344 patients with 736 film images containing 426 benign mass lesions and 310 malign mass lesions, including clinical data and image-based descriptors. Such lesions are associated with masses. The motivations to choose 8-bit resolution images over 12-bit or 14-bit are twofold: Firstly, in contrast to the BCDR-DM (currently under construction), almost all lesions in the BCDR-FM repository have a proven biopsy; and secondly, digital mammography (high resolution images) are not as widely available as film mammography images since the former are more expensive to acquire [130]. For all the experimentation clinical data were not included as features. Figure 6.1 shows examples of both classes with their respective segmentations. The dataset contains MLO and CC views.

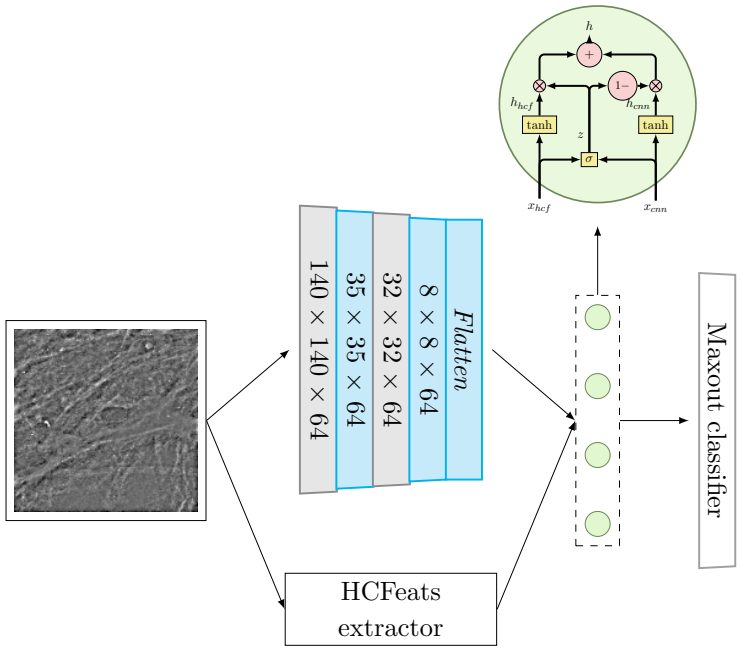


Figure 6.2: Integration of the proposed GMU for feature fusion. Light gray and cyan represent convolutional and pooling layers respectively. *HCFeats* is a set of handcrafted features and *Maxout classifier* is a MLP with maxout units and softmax layer on top.

6.3 Gated multimodal networks for feature fusion

For some particular problems, there is empirical evidence that combining hand-crafted features with learned features improves the performance of predictors. Otálora et al. [21] combined unsupervised feature learning and Riesz wavelets for histopathology image representation outperforming previous reported results in the differentiation between anaplastic and non-anaplastic medulloblastoma. This behavior was also visualized during the development of this research. In previous section some models took advantage of handcrafted features for mass lesion classification in mammography images. Notice, however that more complex models such as the deepest convolutional network didn't gain performance when other features were concatenated. Herein, it is hypothesized that the way this features should be combined should also be determined by the inputs, i.e. should be learned from the data.

Going further, we took out the proposed model from the multimodal scenario to the feature fusion scenario. In particular, the GMU was evaluated in the mass lesion classification task with multiple representations of the same modality. In this scenario the model receives two different representations of the same input, one learned with deep learning architectures and the second using highly specialized features.

The proposed model is depicted in Figure 6.2. The model extracts two set of features. The set of handcrafted features, detailed in Section 6.4.1, comprises different morphometric and statistical measures related to the lesion. The second is a set of features learned with a convolutional network, detailed in Section 6.4.2.

Type	Features
Intensities	Mean, median, maximum, minimum, standard deviation, skewness, kurtosis
Shape	Area, perimeter, circularity, elongation, y_center_mass, x_center_mass, form
Textures	Contrast, correlation, entropy

Table **6-1**: Set of hand-crafted features. For details see [114]

6.4 Data representation

6.4.1 Baseline descriptors

Based on the systematic evaluation presented by Moura and Guevara López [114], the histogram of oriented gradients (HOG) and the histogram of gradient divergence (HGD) were selected as descriptors for our baseline since they showed the best performance against other traditional descriptors. Additionally, a set of 17 hand-crafted features extracted from the segmented lesions (representative of shape, texture and intensities of the mammograms) are used for comparative purposes.

Hand-crafted features (HCfeats)

HCfeats is a set comprising 17 features selected from produced sets of high performance features proposed by Pérez et al. [131] that demonstrated a high impact in characterizing lesions corresponding to masses. Table **6-1** lists the features and their description. HCfeats is composed by intensity descriptors computed directly from the grey-levels of the pixels inside the lesion's contour identified by the radiologists; texture descriptors computed from the grey-level co-occurrence matrix related to the bounding box of the lesion's contour; and shape descriptors computed from the lesion's contour. Notice that computing this set of features requires not only the region of interest (ROI) detection, but also the manual segmentation provided by the expert.

Histogram of oriented gradients (HOG)

HOG describes images through the distribution of the gradients. Images are divided into a grid of blocks (e.g. 3×3), and each block is described by a histogram of the orientation of the gradient. Each histogram has a predefined number of bins dividing the range of possible orientations (from 0 to 2π radians, or from 0 to π radians), and the value of each bin is calculated by summing the magnitude of the gradient of the pixels which have gradient direction within the limits of the bin.

Histogram of gradient divergence (HGD)

Gradient divergence in a point i, j is measured as the angle between the vector of the intensity gradient on i, j and a vector pointing to the center of the image with origin in i, j. HGD describes images through the distribution of the gradient divergence. Images are divided into concentric regions, and each region is described by a histogram of the gradient divergence.

6.4.2 Supervised feature learning

Image representation is fundamental for automatic classification of lesions in mammography images. The goal is to describe the content of the image in a compact and discriminative way. Traditional CADx systems represent images with a carefully selected set of mathematical and heuristic features aiming to characterize the lesion. Recent studies have replaced this hand-crafted process with a learning-based approach where a model is trained in an unsupervised way using deep learning, to transform the raw pixels in a set of features that feeds a classifier algorithm [126, 129]. In contrast to previous work, we have herewith applied a hybrid approach in which CNNs are used to learn the representation in a supervised way. That is, we used lesions previously classified (labeled as benign or malignant) to guide the feature learning process.

The proposed method comprises two main stages: preprocessing and supervised training. The *preprocessing* stage aims to prepare the data in better conditions through a set of transformations so that the next stage takes advantage of relevant characteristics. *Supervised learning* is the second stage that involves two processes: feature learning and classification training. *Feature learning* is performed by training a CNN. It is noteworthy that feature learning is a supervised stage since the CNN training is guided by the labeled samples. The final stage is the SVM *classifier training* with the penultimate layer of the CNN as features.

Preprocessing

Preprocessing is a common stage in CADx systems. Its main goal is to enhance the characteristics of the image by applying a set of transformations that could help to improve performance in following stages. The first step in this work is to extract the ROI from the image. Secondly, an oversampling strategy is used to both get more samples artificially and help to prevent overfitting during training. Finally, a normalization process is carried out to prepare data for learning algorithms. It is widely known that feature learning and deep learning methods usually perform better when the input data has some properties such as decorrelation and normalization, mainly because such properties help gradient-based optimization techniques to converge [132].

Cropping CADx systems aim at classifying a previously identified ROI in the whole film image. This ROI can be obtained by a manual segmentation or automatically detected by a computer aided detection system. Because of lesions in BCDR-03 dataset are manually segmented, we fixed the input size to ROIs of $r \times r$ pixels. With this, ROIs can be easily extracted by taking the bounding box of the segmented region. Specifically, images were cropped to the bounding box of the lesions and rescaled to $r \times r$ pixels preserving the aspect ratio when either width or height of the bounding box are greater than r, otherwise the lesion is centered without scaling and preserving the surrounding region.

Data augmentation The expressiveness of neural network models, and particularly deep ones, comes mainly from the large number of parameters to learn. However, more complex models also increase the chance of overfitting the training data. Data augmentation is a good way to help to prevent this behavior [71]. Data augmentation is the process of artificially create new samples by applying transformations to the original data. In a lesion classification problem, data augmentation

makes sense because a lesion can be presented in any particular orientation. Thus, the model also should be able to learn from such transformations. In particular, For each training image, we have artificially generated 7 new label-preserving samples using a combination of flipping and $90, 180$ and 270 degrees rotation transformations.

Global contrast normalization Due to the digitalization process, the lighting conditions between different film images will be different, and all pixel values of the image are affected by that. A common way to overcome this effect, is to perform a global contrast normalization (GCN) by subtracting the mean of the intensities in the image to each pixel. Notice that the mean is not calculated per pixel, but per image, so it is perfectly fine to subtract it without worrying about whether the current image belongs to train, validation, or test set. Let $\mathbf{X} \in \mathbb{R}^{r \times r}$ be the image, the element-wise transformation is

$$\hat{\mathbf{X}}_{i,j} = \mathbf{X}_{i,j} - \bar{x} \tag{6-1}$$

with $\bar{x} \in \mathbb{R}; \bar{x} = \frac{1}{r^2} \sum_{i,j} \mathbf{X}_{i,j}$, the mean of the $\mathbf{X}$ image intensities, and $\mathbf{X}_{i,j} \in \mathbb{R}$ the intensity in the i, j pixel.

Local contrast normalization Local contrast normalization (LCN) is a transformation inspired by computational neuroscience models [133]. Its main idea is to mimic the behavior the V1 visual cortex. It is implemented by defining a $\mathbf{G} \in \mathbb{R}^{k \times k}$ normalized Gaussian window, i.e $\sum_{p,q} \mathbf{G}_{p,q} = 1$. Then, for each pixel in the global contrast normalized image $\hat{\mathbf{X}}$, the mean of its $k \times k$ neighborhood is removed:

$$\mathbf{V}_{i,j} = \hat{\mathbf{X}}_{i,j} - \sum_{p,q} \mathbf{G}_{p,q} \cdot \hat{\mathbf{X}}_{i+p,j+q} \tag{6-2}$$

with $\mathbf{V} \in \mathbb{R}^{k \times k}$ as the local normalized patch. Then the norm of each $k \times k$ neighborhood is scaled to 1 when it is greater than 1:

$$\tilde{\mathbf{X}}_{i,j} = \frac{\mathbf{V}_{i,j}}{\max\left(c, \sigma_{i,j}\right)} \tag{6-3}$$

where $\sigma_{i,j} \in \mathbb{R}; \sigma_{i,j} = \sqrt{\sum_{p,q} \mathbf{G}_{p,q} \cdot v^2_{i+p.j+q}}$ is the norm of the $k \times k$ neighborhood, and $c \in \mathbb{R}$ is a tolerance parameter to avoid floating point precision problems. It has been empirically shown that such divisive normalization reduces statistical dependencies [132, 134], which in turn accentuates differences between input features and accelerates gradient-based learning [135].

Improvement in both performance and training time when using such normalizations has been reported when the stochastic gradient descent algorithm is used to train deep networks [132]. This has been explained by the fact that, in the same way as whitening and other decorrelation methods, all variables end up with similar variances, making the model more likely to discover non-trivial relationships between spatially near inputs [136]. Also, it has been shown that similar strategies to locally normalize contrast in mammograms have enhanced performance of automatic analysis [137]. Figure 6.3 shows an original image and its corresponding output after applying the preprocessing stage. Again, this preprocessing is performed in an image-wise fashion, thus it is not necessary to store parameters in the training procedure.

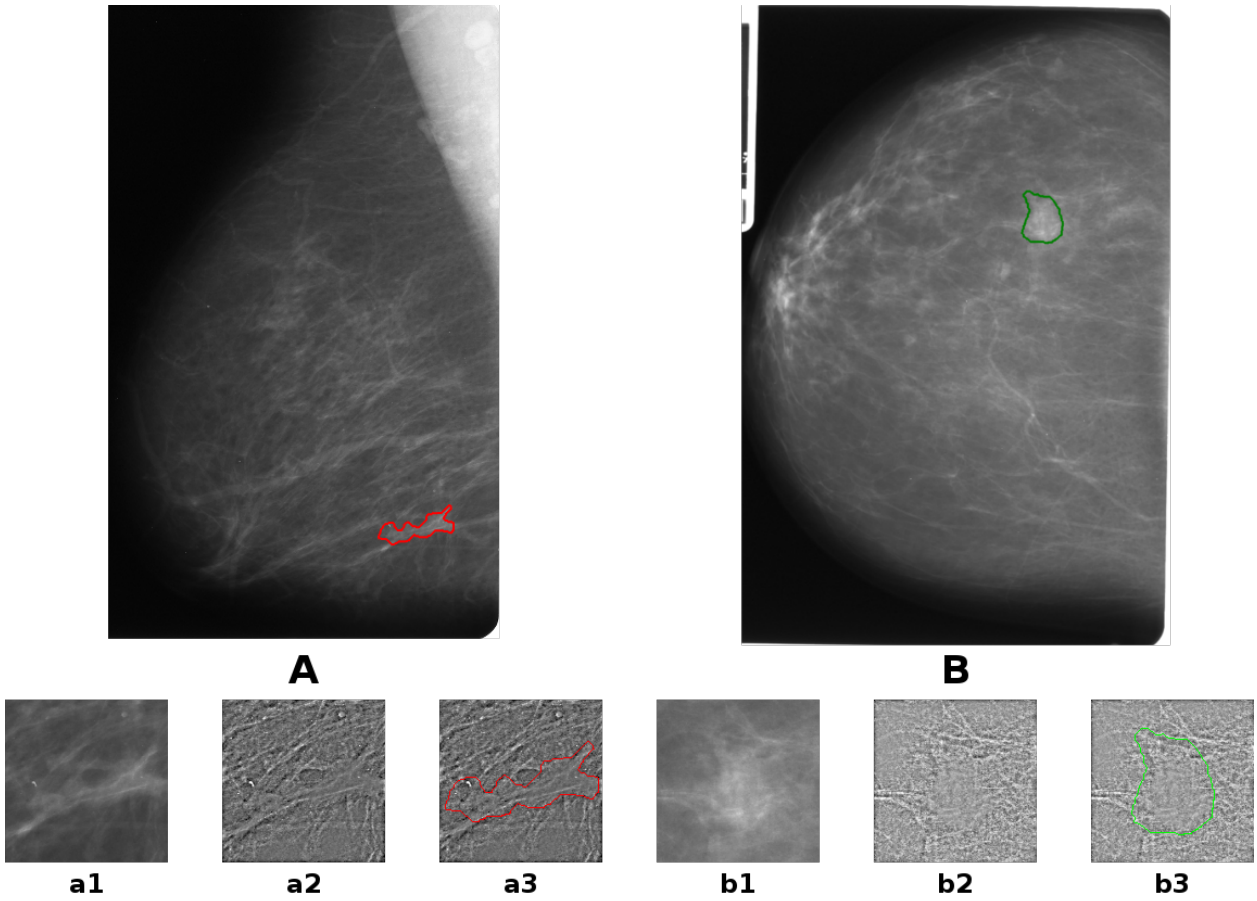


Figure 6.3: Mammography images after the preprocessing step. Images A and B represent malignant and benign lesions respectively. Images a1 and b1 are the bounding box of the lesions. Images a2 and b2 show the output of global and local contrast normalizations. Images a3 and b3 show outline of the lesions over the normalized images.

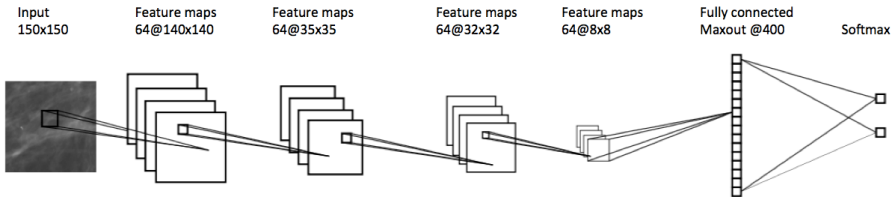


Figure 6.4: Best convolutional neural network evaluated on mass classification

Convolutional network for mass representation

A CNN is a neural network that shares connections between hidden units yielding low computational time and translational invariance properties. CNNs have been successfully applied in shape recognition problems [138] as well as medical diagnosis that involved texture as a discriminant feature [120]. Because mass characterization is highly correlated with shape and texture features [114, 124], a CNN model becomes a suitable strategy for mass lesion classification. The main components of the CNN and the applied strategies to train it are detailed below.

Architecture A CNN comprises 3 main components: a convolutional layer, an activation function and a pooling layer. To improve the capability of the model the three components are stacked iteratively so that the output of one component is the input for the next one, and the output of one set of components is the input for the next set, building a deep neural network with many layers. The convolutional layer is composed of several small matrices or "*kernels*" that are convolved throughout the whole input image working as filters. The output of this convolution is called "*feature map*". These feature maps are the input for the activation function which applies a non-linear transformation in an element-wise fashion. Finally, the pooling layer aggregates contiguous values to one scalar with functions like mean or max.

The proposed architecture, depicted in Figure 6.4, has 11×11 local kernels and the rectifier linear as activation function in the first convolutional layer followed by a 5×5 pooling layer with stride of 4×4 pixels. The second convolutional layer has 4×4 local kernels with the rectifier linear as activation function, with 4×4 pooling layer without overlapping. Then a fully connected layer with 400 units with maxout activation function is stacked to finally add a softmax classifier. In particular, the maxout activation function $h_i : \mathbb{R}^n \to \mathbb{R}$ is a defined as:

$$h_i(\mathbf{s}) = \max_{j \in [1,k]} z_{i,j} \tag{6-4}$$

where $\mathbf{s} \in \mathbb{R}^n$ is the input vector, $z_{i,j} = \mathbf{s}^T \mathbf{W}_{\cdots ij} + \mathbf{b}_{ij}$ is the output of the j-th linear transformation of the i-th hidden unit, and $\mathbf{W} \in \mathbb{R}^{d \times m \times k}$ and $\mathbf{b} \in \mathbb{R}^{m \times k}$ are learned parameters. It has been shown that maxout models with just 2 hidden units behave as universal approximators, while are less prone to saturate units [99].

Since it is our intention to measure how the network's depth affects the performance of the model, we first evaluate the architecture with a single convolutional layer with a fully connected layer and called it CNN2 in the experiments. Consequently, the whole architecture, i.e. two convolutional layers plus a fully connected layer, is referenced as CNN3.

Regularization The number of parameters in the model is directly related to capability to overfit the training data. Usually neural networks require different strategies to control this behavior. In this work dropout and max-norm regularization were used. Dropout randomly set to 0 the input of a unit, while max-norm regularization forces the norm of each vector of incoming weights in a unit to a maximum value. In [139] it was empirically shown that these two strategies help prevent co-adaptation between units, e.g., during error back propagation, a unit should not rely on other units to correct its mistakes since there is no certainty about their activations.

Optimization The proposed architecture has approximately 4.6 million of parameters. Training large models has to scale in both, memory requirements and computational time. The strategy used in this work to train the CNN is stochastic gradient descent with momentum. An early stopping strategy monitoring the area under the ROC curve (AUC) on the validation set was chosen as stop criterion. The implementation of the whole framework was carried out with the Pylearn2 framework [140]. This library uses the Theano framework [141], which in turn takes advantage of GPU technology obtaining up to 140x speedup with respect to CPU implementations, making feasible the training of architectures with millions of parameters.

Classification

Following the previous work, a linear SVM was selected as classification strategy. Train and validation sets were used to fine-tune the C parameter. To evaluate the CNN as a representation strategy, images are propagated through the network, then the penultimate layer activations are extracted and used as representation. This process is done to reduce processing time because, in terms of computational cost, training a single SVM is cheaper than training the whole CNN network. This stage can be seen as a fine-tuning process of the last layer, where a smaller model is adjusted.

6.5 Experimental setup

The dataset was split in training (50%) validation (10%) and test (40%) sets following a stratified sampling per patient, that is, we make sure all computed instances of a particular patient belongs to only one of the three subsets. This setup warranties that the model is not tested using patients seen during the training stage.

In the preprocessing stage, the size of the cropped region was fixed to $r = 150$ according to the distribution of the lesion size and computational capability; and the filter size for LCN is $k = 11$ pixels. Following previous results [114], 5×5 and 3×3 blocks sizes for HOG and 4 and 8 regions for HGD were explored. Histograms for both 8 and 16 bins were evaluated. The best configuration in train-validation setup was used to report test results.

The CNN parameter exploration was performed by training 25 models with random hyperparameter initializations and the best was chosen according to validation performance. It has been reported that this strategy is preferable over grid search when training deep models [106]. Exploration was conducted using the CETA-CIEMAT [3] Research Center infrastructure. Bigger models that requires more intensive computation were carried out using a NVidia Tesla K40 GPGPU card.

[3] http://www.ceta-ciemat.es/, accessed on February 17, 2015

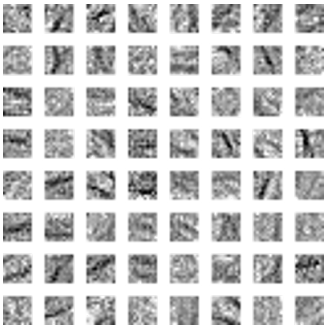


Figure 6.5: Filters learned in the first layer of the CNN model

Before training the SVM model, a zero-mean unit-variance normalization process is carried out. Train and validation sets were used to fine-tune the C parameter for the SVM classifier. Final performance is reported in terms of AUC in the test set.

Comparison of the methods was based on the average AUC of 5 runs using different random seeds for dataset splitting for each run. Experiments were supported by the Wilcoxon signed rank test to determine whether differences have statistical evidence ($\rho > 0.1$).

6.6 Results

6.6.1 Learned features

Recall that the CNN weights in the first layer are equivalent to local kernels that work as filters over the image. Thus, visualizing them would allow to describe the patterns that the model is looking for. Figure 6.5 shows the weights of the best learned model. This image exposes a set of edges in different orientations as well as some texture patterns. It seems the learned filters are affected by noise, probably because it is still few data for this kind of models. We experimentally found that normalization preprocessing was fundamental to obtain good-looking features and ultimately, good performances in the classification. Without normalization the models were not able to surpasses 0.7 of AUC.

6.6.2 Classification results

Figure 6.6 shows ROC curves for all the evaluated representations for the best run. The HCfeats set, which uses segmentation information, performs slightly better than HOG-based descriptors. This confirms the importance of shape information for mass characterization. Interestingly, CNN models, which use only the raw pixels, outperform the state-of-the-art features [114]. The training of CNN3 model took 1.4 hours on the Tesla K40 GPGPU card. It is also worthwhile noting that adding a second hidden layer to the CNN model improves the representation capability producing better results. Such behavior is consistent with theoretical foundations to choose deep architectures over shallow ones [142].

For comparative purposes, we included the evaluation of DeCAF [143], a pre-trained model with the Imagenet dataset [144]. DeCAF is a model with greater complexity than all the other

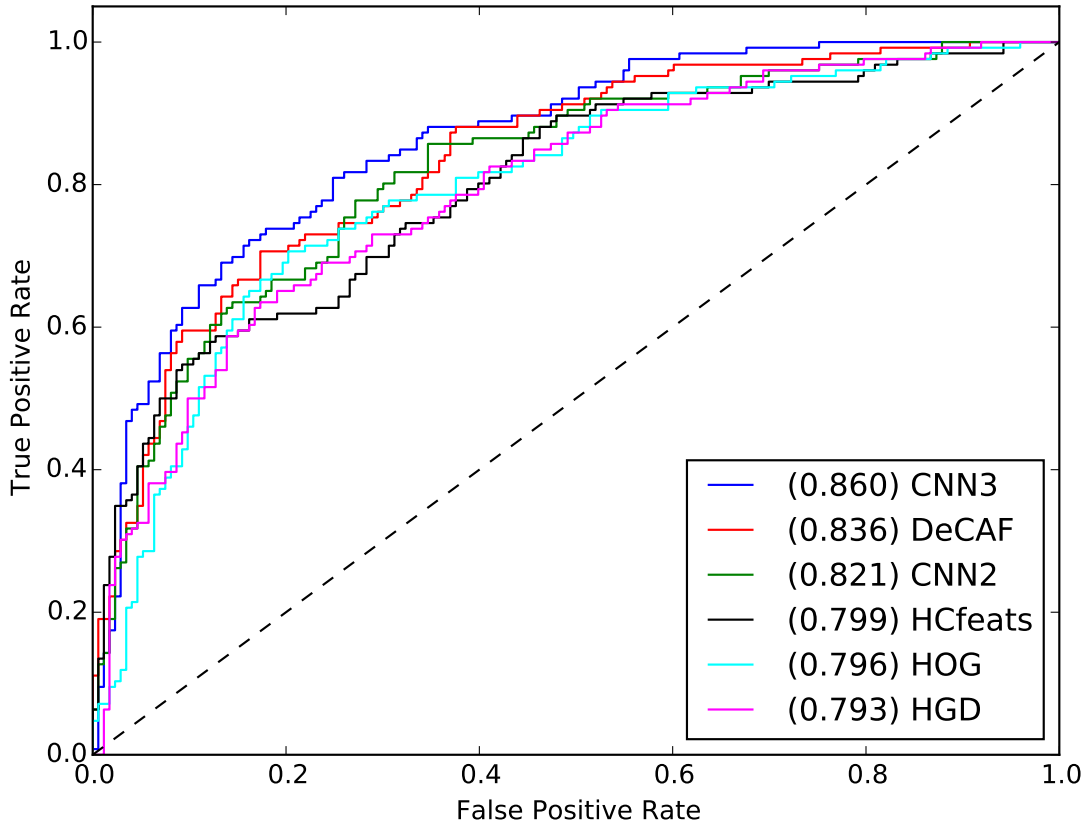


Figure 6.6: ROC curve for evaluated representations in test set for the best run.

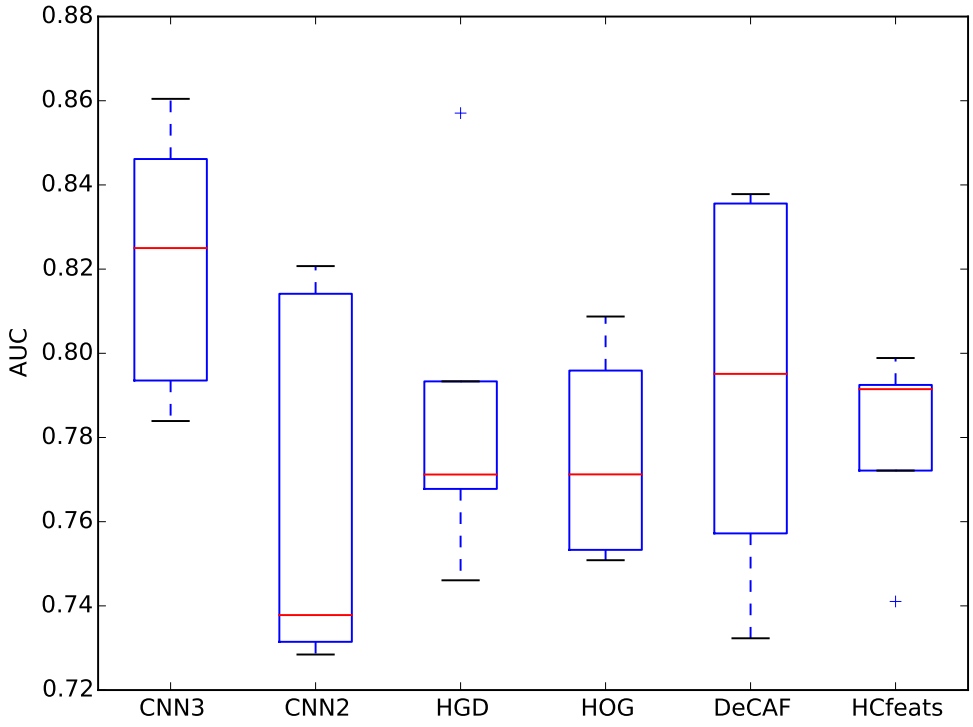


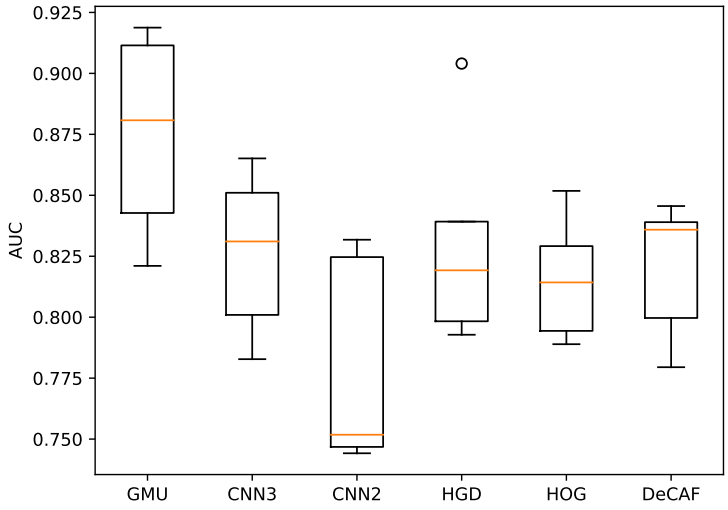
Figure 6.7: Boxplots of different runs for each representation method

representations evaluated on this work. Thus, it is expected to perform better than using hand-crafted features. However, a smaller CNN model trained with the images of the domain performs the best. This behavior, similarly reported when CNNs are trained with small datasets [145], leads to the two main conclusions of this work: On one hand, CNN models outperform state-of-the-art representations for automatic lesion classification in mammography image analysis. On the other hand, such automatic mammography image analysis is a problem with its own particularities, and thus it is not enough to learn the representation using a large CNN model. The learning process should also be guided by a training set with a wide visual variability to show the model texture and shape features presented in mass lesions. Figure 6.7 shows boxplots results in terms of AUC for each representation. According to the Wilcoxon test hypothesis, the CNN3 model performs best as compared to other evaluated representations ($\rho < 0.1$).

In order to combine the image-based features with additional information given in the segmentation, HCfeats, described in section 6.4.1, were concatenated to each CNN representation and baseline descriptors (HOG, HGD and DeCAF). The resultant vector feature of each image has 417 elements, 400 from the last fully connected layer in the CNN plus 17 features from the HCfeats set. Table **6-2** shows a summary of these experiments. In general, this combination improves the results. It specially helps to augment the performance of the hand-crafted representations, while CNN models are not very affected. This suggests that CNN models are already capable of

Representation	Standalone	Combined with HCfeats
CNN3	**0.82+/-0.03**	0.82+/-0.03 (*)
CNN2	0.76+/-0.05	0.78+/-0.04
HGD	0.78+/-0.04	**0.83+/-0.04**
HOG	0.77+/-0.03	0.81+/-0.03 (*)
DeCAF	0.79+/-0.05	0.82+/-0.03 (*)
HCfeats	0.77+/-0.02	–

Table **6-2**: Summary of results in terms of AUC in the test set. Best results are shown in bold typeface and (*) signals scores with no evidence of differences from the highest ($\rho < 0.1$).



Model	AUC
GMU	0.875 ± 0.038
CNN3	0.826 ± 0.030
CNN2	0.780 ± 0.040
HGD	0.831 ± 0.040
HOG	0.815 ± 0.023
DeCAF	0.820 ± 0.026

Figure 6.8: Results of the GMU model used to classify mass lesions in mammography images.

capturing shape information, which is consistent with the learned filters depicted in Figure 6.5, and thus giving such information explicitly could be redundant. Again, this experimentation was supported by the Wilcoxon test, which showed no significant statistical evidence in the differences between representations combined with HCfeats. However, comparing standalone vs combined with HCfeats, all representations except CNN3, obtained evidence for a statistically significant improvement($\rho < 0.05$).

Finally, we evaluated the GMU as alternative to fuse handcrafted features with features learned from the CNN. Results are shown in Figure 6.8. The GMU boost the performance of the CNN obtaining the state-of-the-art results for this dataset. Notice that all of them use the same input information: raw image and morphometric measures from the segmented lesion, but the combination is done differently. GMU learns to combine while other models concatenate both representations.

Kruskal and Wilcoxon statistical test between all the possible pairs were performed. Each classifier was trained 5 times and the results are presented in Table **6-3**. GMU and CNN3 show bigger significance with respect to the remainder models. The GMU presents p-values < 0.1 for all comparisons with the Wilcoxon test.

Method 1	Method 2	Kruskal p-value	Wilcoxon p-value
GMU	CNN3	0.117185	0.079616
GMU	CNN2	0.028280	0.043114
GMU	HGD	0.075800	0.043114
GMU	HOG	0.047202	0.043114
GMU	DeCAF	0.075800	0.043114
CNN3	CNN2	0.117185	0.043114
CNN3	HGD	0.916815	0.892738
CNN3	HOG	0.601508	0.685830
CNN3	DeCAF	0.601508	0.685830
CNN2	HGD	0.174525	0.043114
CNN2	HOG	0.250592	0.079616
CNN2	DeCAF	0.075800	0.043114
HGD	HOG	0.601508	0.079616
HGD	DeCAF	0.916815	0.685830
HOG	DeCAF	0.754023	0.500184

Table **6-3**: P-values for two statistical tests on 5 runs for each classifier.

An open question regarding these results is how this method would perform in high resolution images (12 or 14-bit images). Based on preliminary experimentation, we hypothesize that the model would obtain superior performance using higher resolution images, since the learning model will have more available information. However, we still do not have enough data to report statistically significant results. On the other hand, it is noteworthy that the neural network design would face new challenges such as higher dimensional input, fewer number of examples and different primitive patterns, among others. Thus, we believe new network architectures should be explored to address high resolution images.

6.7 Conclusions

This chapter presented a framework to address classification of mass lesions in mammography film images. Instead of designing particular descriptors to explain the content of mammography images, the proposed approach learns them directly from data in a supervised way. CNNs were used as the representation learning strategy. The proposed neural network architecture takes the raw pixels of the image as input, to learn in a hierarchical way a set of nonlinear transformations to represent the visual content of an image. The model is composed of a set of local filters with a rectified linear unit activation function, maxpooling layers, a fully-connected layer with maxout activation function

and a softmax layer. Our approach outperformed the state-of-the-art image features, HOG and HGD descriptors [114], increasing the performance from 0.787 to 0.822 in terms of AUC. The GMU was also used to combine handcrafted with learned features. Interestingly, this model also took advantage of the additional information of segmentation given by the radiologist. The combination of both representations, learned and hand-crafted, resulted in the best descriptor for mass lesion classification, obtaining 0.875 in the AUC score.

Our future work includes larger architectures as well as the inclusion of other image modalities to enhance the representation. It also would be worth to evaluate the proposed strategy on BCDR-DM images since this suppose a new challenge due to the high resolution images.

7 GMU for book analysis

This chapter explores the combination of feature engineering and neural network models for predicting successful writing. Similar to previous work, it is addressed as a binary classification task. New strategies to automatically learn representations from book contents were explored. Using the GMU as combination strategy for hand-crafted and learned representations we obtained the best performance of 76.10% weighted F1-score. Part of this work was published in the EACL conference [25].

7.1 Introduction

Every year millions of new books are published, but only a few of them turn into commercial successes, and even fewer achieve critical praise in the form of prestigious awards or meaningful sales. Editors have the difficult task of making the go/no-go decision for all manuscripts they receive, and the revenue for their publishing house depends on the accuracy of that judgment. The website `www.litrejections.com` documents some of the biggest mistakes in the history of the publishing industry, including Agatha Christie, J.K. Rowling, and Dr. Seuss, all of whom received many rejection letters before landing their first publishing deal.

Many factors contribute to the eventual success of a given book. Internal factors such as plot, story line, and character development all have a role in the likability of a book. External factors such as author reputation and marketing strategy are arguably equally relevant. Some factors might even be out of the control of an author or publishing house, such as the current trends, the competition from books released simultaneously, and the historical and contextual factors inherent to society.

Previous work by Ganjigunte Ashok, Feng, and Choi [146] demonstrated relevant results using stylistic features to predict the success of books. Their definition of success was a function of the number of downloads from Project Gutenberg. However downloading a book is not by itself an indicator of a highly liked or a commercially successful book. We instead propose to use the rating from reviewers collected from Goodreads as a measure of success. We also propose features and deep learning techniques that have not been used before on this problem, and validate their usefulness in two different tasks: success prediction and genre classification.

Predicting the success of books is a difficult task, even for an experienced editor. Researchers have studied related tasks, for example predicting the quality of text from lexical features, syntactic features and different measures of density. Pitler and Nenkova [147] found a strong correlation between user-perceived text quality and the likelihood measures of the vocabulary as computed by a language model, as well as the likelihood measures of discourse relations, as determined by a language model trained on discourse relations. Louis and Nenkova [148] proposed a combination of genre-specific and readability features with topic-interest metrics for the prediction of great writing

in science articles. While some of the features in this prior work were relevant to our task, our goal is different and more aligned to [146], since we aim to model success in books of different genres.

Ganjigunte Ashok, Feng, and Choi [146] investigated the correlation between writing style and number of downloads. The authors analyzed lexical features, production rules, constituents, and sentiment features of books downloaded from Project Gutenberg [1]. They obtained an average accuracy of 70.38% using only unigram features with Support Vector Machines (SVM) as the classifier.

Deep learning representations have seen their share of successes in Natural Language Processing (NLP) tasks [149–153]. In particular, RNN models have been successfully applied in several scenarios where temporal dependencies provide relevant information [1, 142]. Kiros et al. [154] used RNN models to learn language models from books using an unsupervised approach. Also, word embedding [101] and Paragraph Vector [155] have been shown to achieve state-of-the-art performance in several text classification and sentiment classification tasks. These techniques are able to learn distributed vector representations that capture semantic and syntactic relationships between words. Collobert and Weston [156] trained jointly a single Convolutional Neural Network (CNN) architecture on different NLP tasks and showed that multitask learning increases the generalization of the shared tasks. Other researchers [142, 157, 158] have also reached to similar conclusions.

We provided a new benchmark dataset for predicting successful books in a more realistic class distribution. This data set is available to the community from this link [2]. We provide the first results on using recurrent neural networks (RNN) to discover book content representations that are useful for classification tasks such as success prediction and genre detection. We show that the GMU model benefits the training to obtain better performance than the single success prediction task approach.

7.2 Goodreads dataset

The EMNLP13 collection Ganjigunte Ashok, Feng, and Choi [146] [3] contained Project Gutenberg books from eight different genres. We manually reviewed the dataset and found missing or irrelevant content in 58 books: a total of 53 books contained Project Gutenberg license information repeated verbatim, and five books contained only the audio recording certificate in place of the actual book content. We also identified some odd adjudications. For example, 'The Prince And The Pauper' is a popular book by Mark Twain that was adapted into various films and stage plays. Also, 'The Adventures of Captain Horn' was the third best selling book of 1895 [159]. Both these books are labeled as unsuccessful due to their low download counts. We suspect as well that some of the counts are inflated by college students doing English or Literature assignments that may not be directly related to the potential commercial success of a book.

To address these concerns, we propose a new approach to creating gold labels for successful books based on public reviews rather than download counts. We collected a new set of Project Gutenberg books for this benchmarking. This data also came from Project Gutenberg. We mapped the books

[1] `https://www.gutenberg.org/`

[2] The data can be downloaded from `http://ritual.uh.edu/resources/` page.

[3] The data can be downloaded from `http://www3.cs.stonybrook.edu/songfeng/success/`

Genre	Unsuccessful	Successful	Total
Detective Mystery	60	46	106
Drama	29	70	99
Fiction	30	81	111
Historical Fiction	16	65	81
Love Stories	20	60	80
Poetry	23	158	181
Science Fiction	48	39	87
Short Stories	123	135	258
Total	349	654	1,003

Table **7-1**: Goodreads Data Distribution

		EMNLP13 Success definition	
		Unsuccessful	Successful
Goodreads Success definition	Unsuccessful	73	32
	Successful	110	184

Table **7-2**: Confusion matrix between two different definitions of success

to their review pages on Goodreads[4], a website where book lovers can search, review, and rate books. We consider only those books that have been rated by at least 10 people. We use the average star rating and total number of reviews for labeling each book. We then set an average rating of 3.5 as the threshold for success, such that books with average rating < 3.5 are classified as *Unsuccessful*. Table **7-1** shows the data distribution of our books. To our knowledge, we have one of the largest collection of books, as researchers generally work with a low number of books [160–162].

Success Definitions Comparison: After compiling and labeling both the datasets, we drew a comparison between the two definitions of success. To do this, we downloaded the Project Gutenberg download counts for the books in Goodreads dataset and labeled them using the Ganjigunte Ashok, Feng, and Choi [146] definition of success. Since they only considered books in the extremes of download counts, we could only label 399 books in the Goodreads dataset using their definition. We found that 142 books had different labels according to the two definitions. 19.7% of these mismatched books were labeled as unsuccessful despite having ratings ≥ 3.5 and being reviewed by more than 100 reviewers. Table **7-2** details the discrepancies between the two definitions.

7.3 GMU for feature fusion

Previous results gave two main insights in the way the GMU model can be used to perform feature fusion. First, the usage of genre information improved results in every experiment[25]. Second, the combination of hand-crafted features and learned features benefits the success prediction [25]. Accordingly to these insights, we proposed the integration of learned and handcrafted features

[4] `https://www.goodreads.com/`

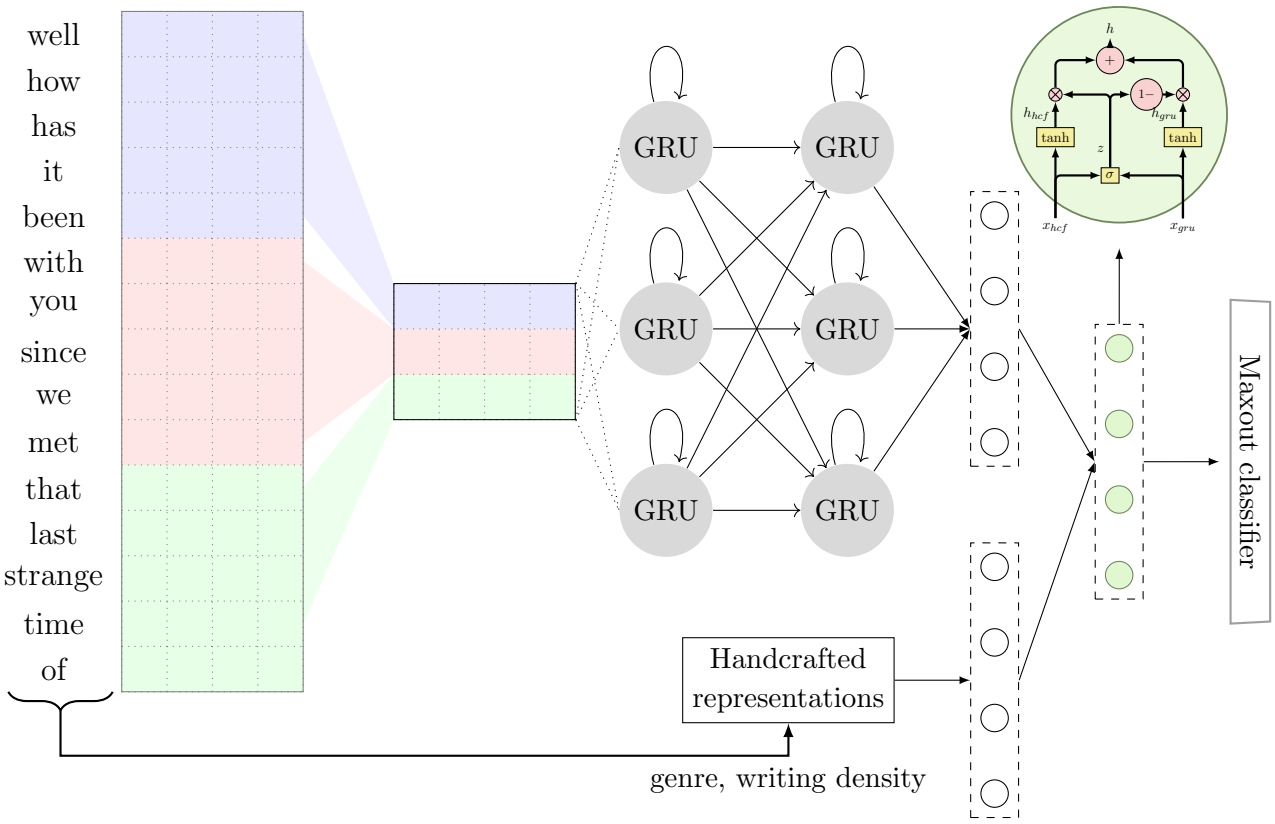


Figure 7.1: GMU integration with learned features and handcrafted features.

through a GMU layer. The proposed model is presented in Figure 7.1. The model builds two representations, one extracted with handcrafted features (see Section 7.4.1) and one learned with word embeddings and recurrent neural networks (see Section 7.4.2). Then, both features are fused through a GMU layer. A logistic classifier is stacked on top. Recurrent and word embeddings modules are pretrained and kept fixed during the training of the multimodal approach.

7.4 Data representation

We investigated a wide range of textual features in an attempt to capture the topic, sentiment, writing style, and readability for each book. This set included both new and previously used features. We also explored techniques for automatically learning representations from text using neural networks, which have been shown to be successful in various text classification tasks [1, 154]. These techniques include word embeddings, document embeddings, and recurrent neural networks.

7.4.1 Hand-crafted text features

Lexical: We used skip-grams, word and char n-grams, and typed char n-grams [163] with term frequency-inverse document frequency (TF-IDF) as the weighting scheme. Sapkota et al. [163]

showed that classical character n-grams lose some information in merging instances of n-grams like *the* which could be a prefix (***the****sis*), a suffix (*brea****the***), or a standalone word (***the***). They separated character n-grams into ten categories representing grammatical classes, like affixes, and stylistic classes, like beg-punct and mid-punct which reflect the position of punctuation marks in the n-gram. The purpose of these features is to correlate success with an author's word choice.

Constituents: We computed the normalized counts of *'SBAR', 'SQ', 'SBARQ', 'SINV',* and *'S'* syntactic tag sets from the parse tree of each sentence in each book, following the method of Ganjigunte Ashok, Feng, and Choi [146] to determine the syntactic style of the authors.

Sentiment: We computed sentence neutrality, positive and negative, using SentiWordNet [164] along with the counts of nouns, verbs, adverbs, and adjectives. We averaged these scores for every 50 consecutive sentences in order to evaluate change in sentiment throughout the course of each book, because we anticipate consecutive change in these scores by substantiating scores of two consecutive sections. These deltas are then feed to the classifier. emotions, like suspense, anger, and happiness to contribute to the success of the book.

SenticNet Concepts: We extracted sentiment concepts from the books using the Sentic Concept Parser [5]. The parser chunks a sentence into noun and verb clauses, and extracts concepts from them using Part Of Speech (POS) bigram rules. We modeled these as binary bag-of-concepts (BoC) features. We also extracted average polarity, sensitivity, attention, pleasantness, and aptitude scores for the concepts defined in the SenticNet-3.0 knowledgebase, which contains semantics and sentics associated with 30,000 common-sense concepts [165].

Writing density: We computed the number of words, characters, uppercase words, exclamations, question marks, as well as the average word length, sentence length, words per sentence, and lexical diversity of each book, with the expectation that successful and unsuccessful writings will have dissimilar distributions of these density metrics. them relevant features in our computational models.

Readability: We computed multiple readability measures including Gunning Fog Index [166], Flesch Reading Ease [167], Flesch Kincaid Grade Level [168], RIX, LIX [169], ARI [170], and Smog Index [171] and used their mean normalized values for training. Intuitively, the use of simple language will resonate with a larger audience and contribute to book success.

7.4.2 Neural network learned representations

Representation learning techniques are able to learn a set of features automatically from the raw data. Our hypothesis is that the learned representation can capture the complex factors that influence the success of a book. In the case of textual data, word embeddings learned by using neural networks have been found to be very useful in various natural language processing applications.

Word embeddings with Book2Vec: In contrast with Word2Vec, which learns a representation for individual words, Doc2Vec learns a representation for text fragments or even for full documents. We trained the Doc2Vec module of the Gensim [172] Python library, on all the books in the Goodreads dataset to obtain a 500 dimensional dense vector representation for each book. Using Doc2Vec, we first trained a distributional memory (DM) model with two approaches: concatenation of context vectors (DMC) and sum of context word vectors (DMM). Then we trained a distributional

[5] `https://github.com/pbhuss/Sentimental/blob/master/parser/SenticParser.py`

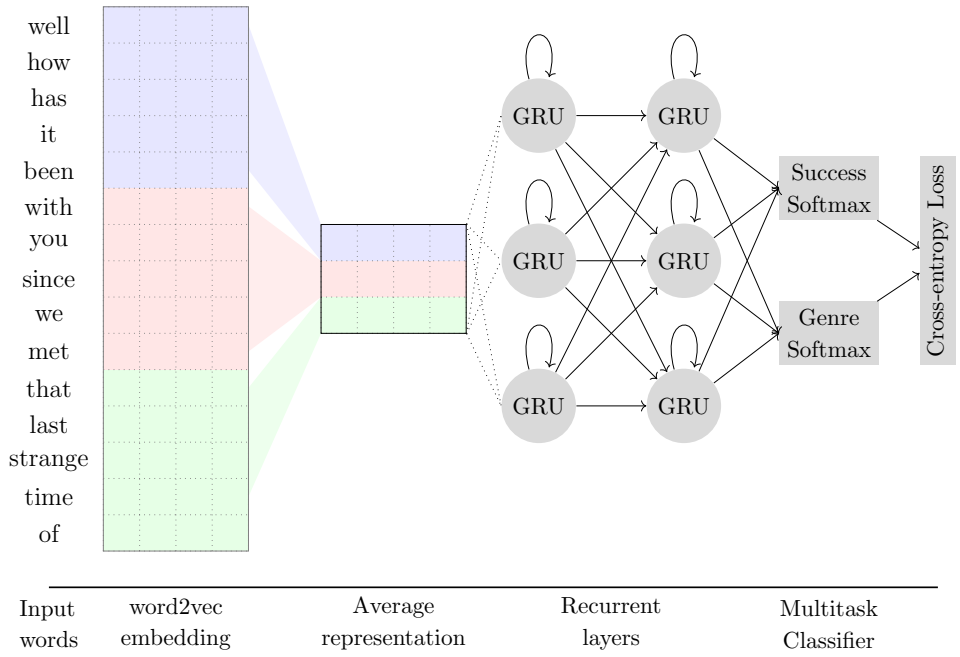


Figure 7.2: Multitask method. Words are represented in the Word2Vec space. Such representations are averaged per window. Sequences are feed to GRU network. Finally, the features are feed to two softmax components to predict genre and success simultaneously.

bag of words (DBoW) model and combined it with the DMC and the DMM for a total of five different models. We set the number of iterations to 50 epochs and shuffled the training data in each pass. We called these book vectors *Book2Vec*. Furthermore, we created two 300 dimensional vector representations for each book by averaging the vectors of each word in the book using pre-trained Word2Vec vectors from the Google News dataset[6] and our own Word2Vec trained with ∼350M words from 5,000 random books crawled from Project Gutenberg.

Multitask RNN method: When dealing with variable length data such as time series or plain text, traditional approaches like feed-forward neural networks are not easily adapted since they expect fixed-size input to model sequential data. One limitation of RNNs is that it has problems dealing with long sequences [173]. We propose a strategy to represent large documents, such as books, with an aggregated representation. Figure 7.2 depicts the proposed multitask method. The overall strategy uses a RNN to learn a model of sequences of sentences. Each sentence is represented by the average of the Word2Vec representation of its constituent words. The RNN is composed of 2 hidden layers with 32 hidden gated recurrent units (GRU) [174] each, and the output is a softmax layer. We train the RNN in a supervised fashion using the success categorization and the book genre as labels. The RNN serves a feature extractor and the last hidden states for each sequence acts as its representation. At training time, all sentences from one book are extracted and divided in chunks of 128 sentences. The book's success/genre labels are assigned to each sequence. A sentence is then represented as the average of its constituent word vectors. To make the book label assignment at testing time, we average the predictions of all sequences extracted from each book.

[6]The pre-trained Word2Vec was downloaded from `https://code.google.com/p/word2vec/`

Using 128 sentences has threefold a motivation: (1) mitigate vanishing gradient problem [173], (2) obtain more examples from one book, and (c) be a power of 2 to efficiently use the GPU.

An interesting property of neural networks is that the same learning approach, i.e stochastic gradient descent, still holds for more complex architectures as long as the objective cost function is differentiable. We take advantage of this property to build a unified neural network that addresses both genre and success prediction using a single model. These kinds of multitask architectures are also useful as regularizers [142]. In particular, our cost function $J(X, Y)$ is defined as follows:

$$h_i = rnn(x_i)$$

$$\hat{y}_i^{succ} = \frac{e^{z_i^{succ}}}{\sum_k e^{z_k^{succ}}}$$

$$\hat{y}_i^{gen} = \frac{e^{z_i^{gen}}}{\sum_l e^{z_l^{gen}}}$$

$$J(X, Y) = -\sum_i (y_i^{succ} \ln \hat{y}_i^{succ} + y_i^{gen} \ln \hat{y}_i^{gen})$$

where x_i represents the i-th sample and y^{succ} and y^{gen} are success and genre labels respectively. The $rnn(\cdot)$ function represents the forward propagation over the recurrent neural network and h represents the last hidden state. $\hat{y}^{succ}$ and $\hat{y}^{gen}$ represent predictions for the two labels. Notice that both of them are computed using the same unified representation h. z^{succ} and z^{gen} represent two different linear transformations over h that map to the number of classes.

7.5 Experimental setup

We merged books from different genres, and then randomly divided the data into a 70:30 training/test ratio, while maintaining the distribution of *Successful* and *Unsuccessful* classes per genre. As a preprocessing step we converted all words to lowercase and removed infrequent tokens having document frequency ≤ 2. For our tagging and parsing needs, we used the Stanford parser [175]. We then trained a LibLinear Support Vector Machine (SVM)[7] classifier with L2 regularization using the hand-crafted features described in Section 7.4. We tuned the C parameter in the training set with 3-fold grid search cross-validation over different values of 1e{-4,...,4}.

With the features used by Ganjigunte Ashok, Feng, and Choi [146], we obtained the highest weighted F1-score of 0.659 with word bigram features. We set this value as our baseline. In order to study the effect of the multitask approach, we devised analogous experiments to our proposed multitask RNN method and predicted both genre and success together for the features described in Section 7.4. Hence we have two settings for the classification experiments, Single task (ST) and Multitask (MT).

Since we had average rating information, we also modeled the problem as a regression problem and predicted the average rating using only the content of the books. Our work differs from other researchers in this aspect, as most of them [176–178] use review content instead of the actual book content to predict the average rating. We used the Elastic Net regression algorithm with $l1\_ratio$

[7] We use LibLinear SVM wrapper from `http://scikit-learn.org/stable/`

Features	ST (F1)	MT (F1)	MSE
Word Bigram	0.659	0.685	0.152
2 Skip 2 gram	0.645	0.688	0.156
2 Skip 3 gram	0.506	0.680	0.156
Char 3 gram	0.669	0.700	0.155
Char 4 gram	0.676	0.689	0.155
Char 5 gram	0.683	0.699	0.154
Typed beg_punct 3 gram	0.621	0.672	0.151
Typed mid_punct 3 gram	0.598	0.641	0.151
Typed end_punct 3 gram	0.626	0.677	0.151
Typed mid_word 3 gram	0.653	0.687	0.156
Typed whole_word 3 gram	0.658	0.666	0.154
Typed multi_word 3 gram	0.607	0.657	0.154
Typed prefix 3 gram	0.624	0.624	0.154
Typed space_prefix 3 gram	0.589	0.646	0.155
Typed suffix 3 gram	0.624	0.637	0.154
Typed space_suffix 3 gram	0.626	0.664	0.154
Clausal	0.506	0.558	0.156
Writing Density (WR)	0.605	0.640	0.156
Readability (R)	0.506	0.634	0.144
SentiWordNet Sentiments(SWN)	0.582	0.610	0.156
Sentic Concepts and Scores (SCS)	0.657	0.670	0.155
GoogleNews Word2Vec	0.669	0.692	0.156
Gutenberg Word2Vec	0.672	0.673	0.140
Book2Vec (DBoW)	0.643	0.654	0.130
Book2Vec (DMM)	0.686	**0.731**	0.142
Book2Vec (DMC)	0.640	0.674	0.131
Book2Vec (DBoW+DMC)	0.647	0.677	0.131
Book2Vec (DBoW+DMM)	**0.695**	0.729	0.142
RNN	0.529	0.686	**0.125**

Table **7-3**: Results for classification (ST = Single task setting, MT = Multi-task setting) and regression tasks on Goodreads dataset. MSE = Mean Square Error, F1 score is weighted F1 scores across *Successful* and *Unsuccessful* classes.

tuned over range {0.01, 0.05, 0.25, 0.5, 0.75, 0.95, 0.99} with 3-fold grid search cross-validation of the training data.

Parameter tuning for RNN: We trained 25 models with random hyper-parameter initialization for learning rate, weights initialization ranges and regularization parameters. We chose the best validation performance model. This is preferable over grid search when training deep models [106]. We used the ADAM algorithm [105] to update the gradients. Since these models are prone to overfitting because of the high number of parameters, we applied clip gradient, max-norm weights, early stopping and dropout regularization strategies. still with these strategies the model was able to achieve perfect classification in training data.

Features	ST (F1)	MT (F1)	MSE
Unigram+Bigram	0.660	0.691	0.15
Unigram+Bigram+Trigram	0.660	0.700	0.149
Char 3,4,5 gram	0.682	0.689	0.153
All Typed ngram	0.663	0.691	0.144
SCS+WR+Typed mid word	**0.720**	0.710	0.155
SCS+Book2Vec	0.695	0.731	0.139
R+Book2Vec	0.695	0.729	0.139
WR+Book2Vec	0.693	0.726	0.139
Word Ngram+ RNN	0.691	0.688	0.125
Skip gram + RNN	0.689	0.683	0.125
Typed char ngram+ RNN	0.689	0.702	0.125
Char 3 gram + RNN	0.689	0.688	0.125
Clausal+ RNN	0.689	0.688	0.125
SCS + RNN	0.691	0.688	0.125
WR+Book2Vec+ RNN	0.701	**0.735**	0.129
SCS+WR+RNN	0.675	0.696	**0.123**
All hand-crafted	0.670	0.689	0.148
All hand-crafted+neural	0.667	0.712	0.129

Table **7-4**: Feature combination results for goodreads dataset. (ST = Single Task, MT =Multi-task, SCS = Sentic concept+average scores of sensitivity, attention, pleasantness, aptitude, polarity, WR = Writing Density, R = Readability)

7.6 Results

Table **7-3** shows the results with the proposed feature sets for the classification and regression tasks. In the ST setting, except for the character n-gram features, all proposed hand-crafted features individually had a weighted F1-score less than the word bigram baseline. On the other hand, the neural network methods obtained better results than the baseline. We obtained the highest weighted F1-score of 0.695 and 0.731 with the *Book2Vec* method in the ST and MT settings, respectively. The results show that the MT approach is better than the ST approach. The genre prediction task must have acted as a regularizer for the success prediction task. Also, we found that modeling the entire book as a vector, rather than modeling it as the average of word vectors, gave better performance. Although the ST *Book2Vec* performs better than the MT RNN method, the difference is very small. We performed McNemar's test on these methods and found that the results were not statistically significant, with $p=0.5$. The MT RNN method had the lowest mean square error (MSE) for the regression task, at 0.125.

The character n-gram proved to be one of the most important hand-crafted features, whereas clausal feature was the least important one. Individually, writing density and readability features seemed to be weak features. We assumed that the sentiment changes in books would be an important characteristic for the task. However, the results in Table **7-3** show an unimpressive F1-score of 0.610 for sentiment features. On the other hand, the bag of sentic concepts model with average scores for sensitivity, attention, pleasantness, aptitude, and polarity gave a more impressive F1-score of 0.670, much higher than the baseline. This result points to the relevance of performing a more nuanced sentiment analysis beyond lexical statistics for this task.

Model	Accuracy	F1-weighted	ROC_AUC
GMU_ST	**0.772**	**0.761**	0.745
GMU_MT	0.766	0.751	**0.755**

Table **7-5**: Results for feature fusion strategy using GMU model. Single Task worked better than Multi-task approach.

Our next set of experiments included the combinations of hand-crafted and neural network representations. Some of the best combination results are shown in Table **7-4**. Out of the different possible feature combinations, we obtained the highest weighted F1 score of 0.735 by combining hand-crafted and learned representations in the MT setting. We observed that combining low performing hand-crafted features like readability, syntactic clauses, and skip grams with neural representation boosted their performance. Likewise for the regression task, the MT RNN representation proved to be a better choice, as its combination with other features generally lowered the MSE. The best combinations for the regression task lowered the MSE to 0.123. Deep learning and hand-crafted methods may capture complementary sources of information, which upon combination boost performance.

7.6.1 GMU fusion

We evaluated the model in two different setups: ST and MT. Table **7-5** shows the results of this approach. This result outperformed all other previous results. The model learned to effectively combine handcrafted features and learned features by dynamically weighting each sample. This behavior is consistent with results reported in Chapter 6: To apply the GMU in multimodal and information fusion tasks is a reasonable strategy to boost results of single classifiers and standard early and late fusion strategies.

7.7 Conclusions

In this chapter we propose new features for predicting the success of books. We used two main feature categories: hand-crafted and RNN-learned features. Hand-crafted features included typed character n-grams and sentic concepts. For the learned features we proposed two different strategies based on neural networks. The first extends Word2Vec-type representations to work in large documents such as books, and the second one uses an RNN to capture sequential patterns in large texts. Finally, we used the GMU architecture con effectively combine both representations. We evaluated the methods in the Goodreads dataset, whose classes are not based on download counts, but rather are a function of average star ratings and number of reviewers. Our results outperform state-of-the-art methods. We conclude that instead of having either deep-learning or hand-crafted features outperform the other, both methods capture complementary information, which upon combination gives better performance. An interesting research direction would be to explore features that capture plot-related aspects, such as character profiles and interaction through social network analysis, historical setting, and other feature-learning strategies.

8 Summary and conclusions

Machine learning methods have received a lot of attention because of its successful application addressing complex problems. These promising results are mainly due to the development of methods that automatically learn the representation of complex objects directly from large amounts of training data. These methods are an evolution of neural networks and the main topic of this research, known as deep learning. Deep learning leads the state-of-the-art in different areas with success cases in object recognition, scene image labeling, autonomous car driving and speech recognition among others. Despite huge advances during last decade, representation learning faces unique challenges. One of them is how to take advantage of different types of information. This research was conducted to address such challenge: to use deep learning models for combining multiple modalities and jointly learn an unified representation for supervised tasks.

A novel strategy was proposed to learn fusion transformations from multimodal sources. Similarly to the way recurrent models control the information flow, the proposed model is based on multiplicative gates. The Gated Multimodal Unit (GMU) receives two or more input sources and learns to determine how much each input modality affects the unit activation. This contrasts the traditional fusion methods that adjust weights for each modality and are fixed for all instances, while the GMU weights are determined by the input. In synthetic experiments the GMU was able to learn hidden latent variables. A key property of the GMU is that, being a differentiable operation, it is easily coupled in different neural network architectures and trained with standard gradient-based optimization algorithms. In the following, the four supervised tasks used to validate the proposed model are described.

Medical image analysis Before the deep learning era, traditional approaches for pattern recognition problems used to engineer a set of feature extractors that performed the best for a particular task. For instance, in medical image classification, the bag-of-visual-words approach was a common strategy to represent the medical image content. This strategy splits the image in square patches and finds a dictionary with the most common ones. Then, the image is represented as a frequency histogram of occurrence of those patches in the image. The feature engineering approach requires high specialization for each problem. With this kind of methods, automatic analysis systems can take advantage of the prior knowledge given by the experts. On the other hand, researches integrated the feature extraction stage in the learning process. This allows the algorithms to automatically learn what transformations are required to extract the meaningful content of the input. We initially explored simple concatenation of engineered and learned representation to improve performance in breast mass lesion classification [14]. However, when the learned features went better, the classifier did not benefit from the handcrafted features. We extended such work with the integration of GMU in the architecture and the model was able to better fuse handcrafted and learned features to outperform previous state-of-the-art results.

Multimodal genre prediction In recent years, multimodal tasks have received attention by the representation learning community. Strategies for visual question answering, or image captioning have developed interesting ways of combining different representation learning architectures. Most of these models are focused on mapping from one modality to another or solving an auxiliary task to create a common representation with the information of all modalities. This research addressed the task of predicting the target variable using two or more modalities as input. We built the MMImdb dataset which, to the best of our knowledge, is the largest publicly available multimodal dataset for genre prediction on movies. Then, the GMU was evaluated on a multilabel scenario for genre classification of movies using the plot and the poster. The GMU improved the macro f-score performance of single-modality approaches and outperformed other fusion strategies, including early, late and mixture of experts models.

Multimodal image segmentation The GMU has been integrated with convolutional networks for addressing natural image segmentation where it outperformed the single-modality, early and late fusion approaches. The gated multimodal network involved an end-to-end convolutional architecture taking as input the RGB and depth images and output the segmented image with 6 semantic concepts. Likewise, the model outperformed other single and multimodal approaches measuring the Intersection-over-union score. The activations of the GMU layer were mapped to the output concepts finding correlations between input modalities and output concepts, e.g. depth information was more correlated with "sky" and "tree" while RGB is more correlated with "grass" and "vegetation". It should be noted that even though the model is capable of combining information, the content representation is critical to correctly take advantage of the different modalities.

In contrast to other weighted fusion or feature selection strategies, the GMU is able to give independent weights for each sample, and those weights are automatically assigned depending on the values for each modality. The evaluation went beyond multiple modalities and was applied to perform feature combination showing that the gated approach can be also applied to unimodal escenarios, provided that there are more than one representation. We also explored ways to understand how the model gives importance to each input. The analysis associated input modalities with the output categories. Interesting findings in genre prediction show, for instance, that the model associates the visual information with animation movies while textual information is more associated with Drama or Romance movies. Similar behavior appeared in image segmentation where the closest and farthest concepts in an image where associated with depth information. while more complex structures like trees are more related to RGB representation. It is also noteworthy that the GMU is easily adapatable with different neural network architectures provided that the function used to guide the learning be differentiable. In consequence, we integrated the GMU with convolutional and fully connected networks. We also observed that the GMU does not present learning vanishing or exploding gradient, problems that are common in mixture-of-experts approaches. Finally, this dissertation supported reproducible research by contributing three public datasets, all of them created in this project.

A future research direction in this area is to build a more general unit that can learn more complex transformations and interactions between input modalities, as well as dealing with scenarios where, at inference time, some modalities are absent.

Bibliography

[1] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep learning". In: *Nature* 521.7553 (May 2015), pp. 436–444. DOI: 10.1038/nature14539. URL: http://dx.doi.org/10.1038/nature14539.

[2] Yoshua Bengio, Aaron Courville, and Pascal Vincent. "Representation Learning: A Review and New Perspectives". In: *IEEE Trans. Pattern Anal. Mach. Intell.* 35.8 (Aug. 2013), pp. 1798–1828. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2013.50.

[3] Mohammad Norouzi et al. "Zero-Shot Learning by Convex Combination of Semantic Embeddings". In: *CoRR* abs/1312.5 (Dec. 2014). arXiv: 1312.5650. URL: http://arxiv.org/abs/1312.5650.

[4] Andrea Frome et al. "DeViSE: A Deep Visual-Semantic Embedding Model". In: *Advances in Neural Information Processing Systems 26*. Ed. by C J C Burges et al. Curran Associates, Inc., 2013, pp. 2121–2129. URL: http://papers.nips.cc/paper/5204-devise-a-deep-visual-semantic-embedding-model.pdf.

[5] Chidansh Bhatt and Mohan Kankanhalli. "Multimedia data mining: state of the art and challenges". In: *Multimedia Tools and Applications* 51.1 (2011), pp. 35–76. ISSN: 1380-7501. DOI: 10.1007/s11042-010-0645-5.

[6] Pradeep K. Atrey et al. "Multimodal fusion for multimedia analysis: a survey". In: *Multimedia Systems* 16.6 (Apr. 2010), pp. 345–379. ISSN: 0942-4962. DOI: 10.1007/s00530-010-0182-0.

[7] Li Deng. "A tutorial survey of architectures, algorithms, and applications for deep learning". In: *APSIPA Transactions on Signal and Information Processing* 3 (2014). ISSN: 2048-7703. DOI: 10.1017/atsip.2013.9. URL: http://journals.cambridge.org/article%7B%5C_%7DS2048770313000097.

[8] Dong Yu Li Deng. *Deep Learning: Methods and Applications*. NOW Publishers, May 2014, pp. 1–134. ISBN: 9781405161251. URL: https://www.microsoft.com/en-us/research/publication/deep-learning-methods-and-applications/.

[9] Yoshua Bengio. "Learning Deep Architectures for AI". In: *Found. Trends Mach. Learn.* 2.1 (Jan. 2009), pp. 1–127. ISSN: 1935-8237. DOI: 10.1561/2200000006. URL: http://dx.doi.org/10.1561/2200000006.

[10] Jürgen Schmidhuber. "Deep learning in neural networks: An overview". In: *Neural networks* 61 (2015), pp. 85–117.

[11] Nitish Srivastava and Ruslan R Salakhutdinov. "Multimodal learning with deep boltzmann machines". In: *Advances in neural information processing systems*. 2012, pp. 2222–2230.

[12] Jiquan Ngiam et al. "Multimodal Deep Learning". In: *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*. Ed. by Lise Getoor and Tobias Scheffer. ICML '11. Bellevue, Washington, USA: ACM, June 2011, pp. 689–696. ISBN: 978-1-4503-0619-5. URL: http://ai.stanford.edu/%7B~%7Dang/papers/icml11-MultimodalDeepLearning.pdf.

[13] John Arevalo et al. "An unsupervised feature learning framework for basal cell carcinoma image analysis". In: *Artificial intelligence in medicine* 64.2 (2015), pp. 131–145.

[14] John Arevalo et al. "Representation learning for mammography mass lesion classification with convolutional neural networks". In: *Computer methods and programs in biomedicine* 127 (2016), pp. 248–257.

[15] John Arevalo, Angel Cruz-Roa, et al. "Histopathology image representation for automatic analysis: A state-of-the-art review". In: *Revista Med* 22.2 (2014), pp. 79–91.

[16] Jorge A Vanegas et al. "MindLab at ImageCLEF 2014: Scalable Concept Image Annotation". In: *CLEF 2014 Evaluation Labs and Workshop, Online Working Notes. Sheffield, UK (September 15-18 2014)*. 2014.

[17] Jorge A Vanegas, John Arevalo, and Fabio A Gonzalez. "Unsupervised feature learning for content-based histopathology image retrieval". In: *Content-Based Multimedia Indexing (CBMI), 2014 12th International Workshop on*. IEEE. 2014, pp. 1–6.

[18] John Arevalo et al. "Convolutional neural networks for mammography mass lesion classification". In: *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE*. IEEE. 2015, pp. 797–800.

[19] Angel Cruz-Roa et al. "A comparative evaluation of supervised and unsupervised representation learning approaches for anaplastic medulloblastoma differentiation". In: *Proc. SPIE*. Vol. 9287. 2015, 92870G–92870G.

[20] Angel Cruz-Roa et al. "A method for medulloblastoma tumor differentiation based on convolutional neural networks and transfer learning". In: *11th International Symposium on Medical Information Processing and Analysis (SIPAIM 2015)*. International Society for Optics and Photonics. 2015, pp. 968103–968103.

[21] Sebastian Otálora et al. "Combining Unsupervised Feature Learning and Riesz Wavelets for Histopathology Image Representation: Application to Identifying Anaplastic Medulloblastoma". In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part I*. Ed. by Nassir Navab et al. Cham: Springer International Publishing, 2015, pp. 581–588. ISBN: 978-3-319-24553-9. DOI: 10.1007/978-3-319-24553-9_71. URL: https://doi.org/10.1007/978-3-319-24553-9_71.

[22] Luis Pellegrin et al. "INAOE-UNAL at ImageCLEF 2015: Scalable Concept Image Annotation." In: *CLEF (Working Notes)*. 2015.

[23] Luis Pellegrin et al. "A Two-Step Retrieval Method for Image Captioning". In: *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer International Publishing. 2016, pp. 150–161.

[24] John Arevalo et al. “Gated Multimodal Units for Information Fusion”. In: *5th International conference on learning representations 2017 workshop*. 2017.

[25] Suraj Maharjan et al. “A Multi-task Approach to Predict Likability of Books”. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Vol. 1. 2017, pp. 1217–1227.

[26] John Arevalo, Raúl Ramos-Pollan, and Fabio A González. “Distributed Cache Strategies for Machine Learning Classification Tasks over Cluster Computing Resources”. In: *High Performance Computing*. Springer Berlin Heidelberg. 2014, pp. 43–53.

[27] Oscar Perdomo et al. “A Novel Machine Learning Model Based on Exudate Localization to Detect Diabetic Macular Edema”. In: (2016).

[28] Oscar Perdomo, John Arevalo, and Fabio A González. “Convolutional network to detect exudates in eye fundus images of diabetic subjects”. In: *12th International Symposium on Medical Information Processing and Analysis*. International Society for Optics and Photonics. 2017, 101600T–101600T.

[29] Mehrdad J Gangeh et al. “Supervised dictionary learning and sparse representation-a review”. In: *arXiv preprint arXiv:1502.05928* (2015).

[30] Yu-Xiong Wang and Yu-Jin Zhang. “Nonnegative matrix factorization: A comprehensive review”. In: *IEEE Transactions on Knowledge and Data Engineering* 25.6 (2013), pp. 1336–1353.

[31] Frank Seide, Gang Li, and Dong Yu. “Conversational Speech Transcription Using Context-Dependent Deep Neural Networks.” In: *INTERSPEECH*. 2011, pp. 437–440.

[32] Yoshua Bengio et al. “Greedy layer-wise training of deep networks”. In: *Advances in neural information processing systems* 19 (2007), p. 153.

[33] Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. “A fast learning algorithm for deep belief nets”. In: *Neural computation* 18.7 (2006), pp. 1527–1554.

[34] Kaiming He et al. “Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification”. In: *CoRR* abs/1502.0 (2015).

[35] Kelvin Xu et al. “Show, attend and tell: Neural image caption generation with visual attention”. In: *arXiv preprint arXiv:1502.03044* 2.3 (2015), p. 5.

[36] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. “Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models”. In: *arXiv preprint arXiv:1411.2539* (2014).

[37] Oriol Vinyals et al. “Show and tell: A neural image caption generator”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 3156–3164.

[38] Junhua Mao et al. “Explain images with multimodal recurrent neural networks”. In: *arXiv preprint arXiv:1410.1090* (2014), pp. 1–9.

[39] Richard Socher et al. "Grounded Compositional Semantics for Finding and Describing Images with Sentences". In: *Transactions of the Association for Computational Linguistics (TACL)* 2.April (2014), pp. 207–218. URL: http://nlp.stanford.edu/%7B~%7Dsocherr/SocherLeManningNg%7B%5C_%7DnipsDeepWorkshop2013.pdf.

[40] Ryan Kiros, Richard S Zemel, and Ruslan Salakhutdinov. "Multimodal Neural Language Models". In: *NIPS 2013 Deep Learning Workshop* (2013).

[41] Douwe Kiela and Léon Bottou. "Learning Image Embeddings using Convolutional Neural Networks for Improved Multi-Modal Semantics." In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-14)*. Doha, Qatar, 2014, pp. 36–45.

[42] Zeynep Akata, Honglak Lee, and Bernt Schiele. "Zero-Shot Learning with Structured Embeddings". In: *CoRR* abs/1409.8 (2014). URL: http://arxiv.org/abs/1409.8403.

[43] Richard Socher et al. "Zero-Shot Learning Through Cross-Modal Transfer". In: *Advances in Neural Information Processing Systems 26*. Ed. by C J C Burges et al. Curran Associates, Inc., 2013, pp. 935–943. URL: http://papers.nips.cc/paper/5027-zero-shot-learning-through-cross-modal-transfer.pdf.

[44] Yoonseop Kang, Saehoon Kim, and Seungjin Choi. "Deep Learning to Hash with Multiple Representations". In: *Data Mining (ICDM), 2012 IEEE 12th International Conference on*. IEEE. IEEE, 2012, pp. 930–935. ISBN: 9780769549057. DOI: 10.1109/ICDM.2012.24.

[45] Devendra Singh Sachan, Umesh Tekwani, and Amit Sethi. "Sports Video Classification from Multimodal Information Using Deep Neural Networks". In: *2013 AAAI Fall Symposium Series*. 2013.

[46] Eric H Huang et al. "Improving word representations via global context and multiple word prototypes". In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*. Association for Computational Linguistics. 2012, pp. 873–882.

[47] Tomas Mikolov et al. "Distributed representations of words and phrases and their compositionality". In: *Advances in Neural Information Processing Systems*. 2013, pp. 3111–3119.

[48] Zhaoquan Yuan et al. "A Unified Framework of Latent Feature Learning in Social Media". In: *IEEE Transactions on Multimedia* 9210.c (Oct. 2014). ISSN: 1520-9210. DOI: 10.1109/TMM.2014.2322338.

[49] Wei Wang et al. "Effective Multi-Modal Retrieval based on Stacked". In: *Proceedings of the VLDB Endowment* 7.8 (2014), pp. 649–660. ISSN: 21508097.

[50] Nitish Srivastava and Ruslan Salakhutdinov. "Learning representations for multimodal data with deep belief nets". In: *International Conference on Machine Learning Workshop*. 2012.

[51] Pengcheng Wu et al. "Online multimodal deep similarity learning with application to image retrieval". In: *Proceedings of the 21st ACM international conference on Multimedia - MM '13*. MM '13. ACM Press, 2013, pp. 153–162. ISBN: 9781450324045. DOI: 10.1145/2502081.2502112.

[52] Maurizio Calo Caligaris. "Unsupervised Learning of Multimodal Features: Images and Text". In: *http://cs229.stanford.edu/* (2010).

[53] Fangxiang Feng, Ruifan Li, and Xiaojie Wang. "Constructing hierarchical image-tags bimodal representations for word tags alternative choice". In: *arXiv preprint arXiv:1307.1275* (2013).

[54] Xinyan Lu et al. "Learning multimodal neural network with ranking examples". In: *Proceedings of the 22nd ACM international conference on Multimedia*. MM '14. ACM. New York, NY, USA: ACM, 2014, pp. 985–988. ISBN: 9781450330633. DOI: 10.1145/2647868.2655001.

[55] Yin Zheng, Yu-Jin Zhang, and Hugo Larochelle. "Topic modeling of multimodal data: an autoregressive approach". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2014, pp. 1370–1377. ISBN: 2011000211.

[56] Jing Huang and Brian Kingsbury. "Audio-visual deep learning for noise robust speech recognition". In: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. ICASSP. IEEE, May 2013, pp. 7596–7599. ISBN: 978-1-4799-0356-6. DOI: 10.1109/ICASSP.2013.6639140.

[57] Jian Tu et al. "Challenge Huawei challenge: Fusing multimodal features with deep neural networks for Mobile Video Annotation". In: *Multimedia and Expo Workshops (ICMEW), 2014 IEEE International Conference on*. 2014, pp. 1–6. DOI: 10.1109/ICMEW.2014.6890609.

[58] Yelin Kim, Honglak Lee, and Emily Mower Provost. "Deep learning for robust feature generation in audiovisual emotion recognition". In: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, May 2013, pp. 3687–3691. ISBN: 978-1-4799-0356-6. DOI: 10.1109/ICASSP.2013.6638346.

[59] Yuki Yamaguchi et al. "Learning and association of synaesthesia phenomenon using deep neural networks". In: *Proceedings of the 2013 IEEE/SICE International Symposium on System Integration*. IEEE, Dec. 2013, pp. 659–664. ISBN: 978-1-4799-2625-1. DOI: 10.1109/SII.2013.6776750.

[60] Deli Pei et al. "Unsupervised multimodal feature learning for semantic image segmentation". In: *The 2013 International Joint Conference on Neural Networks (IJCNN)*. IEEE, Aug. 2013, pp. 1–6. ISBN: 978-1-4673-6129-3. DOI: 10.1109/IJCNN.2013.6706748. URL: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6706748.

[61] Heung Il Suk and Dinggang Shen. "Deep learning-based feature representation for AD/MCI classification". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 8150 LNCS. 2013, pp. 583–590. ISBN: 9783642407628. DOI: 10.1007/978-3-642-40763-5_72.

[62] Galen Andrew et al. "Deep canonical correlation analysis". In: *International Conference on Machine Learning*. 2013, pp. 1247–1255.

[63] Dong Yi et al. "Shared Representation Learning for Heterogeneous Face Recognition". In: *CoRR* abs/1406.1 (2014).

[64] Heung-Il Suk, Seong-Whan Lee, and Dinggang Shen. "Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis". In: *NeuroImage* 101 (2014), pp. 569–582. ISSN: 1053-8119. DOI: http://dx.doi.org/10.1016/j.neuroimage.2014.06.077.

[65] Liang Ge et al. "Multi-source deep learning for information trustworthiness estimation". In: *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '13*. KDD '13. Chicago, Illinois, USA: ACM Press, 2013, p. 766. ISBN: 9781450321747. DOI: 10.1145/2487575.2487612. URL: http://doi.acm.org/10.1145/2487575.2487612.

[66] Haifeng Hu et al. "Multimodal DBN for Predicting High-Quality Answers in cQA portals". In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, 2013, pp. 843–847.

[67] Zhaoquan Yuan, Jitao Sang, and Changsheng Xu. "Tag-aware image classification via Nested Deep Belief nets". In: *2013 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, July 2013, pp. 1–6. ISBN: 978-1-4799-0015-2. DOI: 10.1109/ICME.2013.6607503.

[68] Angeliki Lazaridou, Elia Bruni, and Marco Baroni. "Is this a wampimuk? Cross-modal mapping between distributional semantics and the visual world." In: *ACL (1)*. 2014, pp. 1403–1414.

[69] Ruslan Salakhutdinov, Joshua B Tenenbaum, and Antonio Torralba. "Learning with hierarchical-deep models." In: *IEEE transactions on pattern analysis and machine intelligence* 35.8 (Aug. 2013), pp. 1958–71. ISSN: 1939-3539. DOI: 10.1109/TPAMI.2012.269.

[70] Yue Zhao, Hui Wang, and Qiang Ji. "Audio-Visual Tibetan Speech Recognition Based on a Deep Dynamic Bayesian Network for Natural Human Robot Interaction". In: *International Journal of Advanced Robotic Systems* (2012), p. 1. ISSN: 1729-8806. DOI: 10.5772/54000.

[71] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems*. Ed. by F. Pereira et al. Curran Associates, Inc., 2012, pp. 1106–1114.

[72] Adam Coates and Andrew Ng. "The importance of encoding versus training with sparse coding and vector quantization". In: *Proceedings of the 28th International Conference on Machine Learning (ICML-11)* (2011), pp. 921–928.

[73] Sinno Jialin Pan and Qiang Yang. "A Survey on Transfer Learning". In: *IEEE Transactions on Knowledge and Data Engineering* 22.10 (Oct. 2010), pp. 1345–1359. ISSN: 1041-4347. DOI: 10.1109/TKDE.2009.191.

[74] Tomer Michaeli, Yonina C Eldar, and Guillermo Sapiro. "Semi-supervised multi-domain regression with distinct training sets". In: *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Mar. 2012, pp. 2145–2148. ISBN: 978-1-4673-0046-9. DOI: 10.1109/ICASSP.2012.6288336.

[75] Carina Silberer and Mirella Lapata. "Learning Grounded Meaning Representations with Autoencoders". In: *The 52nd Annual Meeting of the Association for Computational Linguistics* (2014).

[76] Ian Lenz, Honglak Lee, and Ashutosh Saxena. "Deep Learning for Detecting Robotic Grasps". In: *CoRR* abs/1301.3 (2013).

[77] Y Bengio. "Deep learning of representations for unsupervised and transfer learning". In: *Workshop on Unsupervised and Transfer Learning, ...* (2011).

[78] G Mesnil, Y Dauphin, and X Glorot. "Unsupervised and transfer learning challenge: a deep learning approach". In: *Journal of Machine Learning Research-Proceedings Track* 27 (2011), pp. 97–110.

[79] Jeffrey Dean et al. "Large scale distributed deep networks". In: *Advances in neural information processing systems*. 2012, pp. 1223–1231.

[80] A Karpathy et al. "Large-Scale Video Classification with Convolutional Neural Networks". In: *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. June 2014, pp. 1725–1732. DOI: 10.1109/CVPR.2014.223.

[81] Alex Krizhevsky. "One weird trick for parallelizing convolutional neural networks". In: *CoRR* abs/1404.5997 (2014).

[82] John Arevalo et al. "Gated Multimodal Networks". Submitted. 2018.

[83] Stanislaw Antol et al. "Vqa: Visual question answering". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2015, pp. 2425–2433.

[84] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. "Densecap: Fully convolutional localization networks for dense captioning". In: *arXiv preprint arXiv:1511.07571* (2015).

[85] Ioannis Kanaris and Efstathios Stamatatos. "Learning to recognize webpage genres". In: *Information Processing and Management* 45.5 (2009), pp. 499–512. ISSN: 03064573. DOI: 10.1016/j.ipm.2009.05.003. URL: http://dx.doi.org/10.1016/j.ipm.2009.05.003.

[86] Eric Makita and Artem Lenskiy. "A movie genre prediction based on Multivariate Bernoulli model and genre correlations". In: *arXiv preprint arXiv:1604.08608* (Mar. 2016). arXiv: 1604.08608. URL: http://arxiv.org/abs/1604.08608.

[87] Michael Treml et al. "Speeding up semantic segmentation for autonomous driving". In: *NIPSW* 1.7 (2016), p. 8.

[88] Andrew Janowczyk and Anant Madabhushi. "Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases". In: *Journal of pathology informatics* 7 (2016).

[89] Robert A Jacobs et al. "Adaptive mixtures of local experts". In: *Neural computation* 3.1 (1991), pp. 79–87.

[90] Seniha Esen Yuksel, Joseph N Wilson, and Paul D Gader. "Twenty years of mixture of experts". In: *IEEE transactions on neural networks and learning systems* 23.8 (2012), pp. 1177–1193.

[91] Ricardo F Alvear-Sandoval and Aníbal R Figueiras-Vidal. "On building ensembles of stacked denoising auto-encoding classifiers and their further improvement". In: *Information Fusion* 39 (2018), pp. 41–52.

[92] Jing Zhao et al. "Multi-view learning overview: Recent progress and new challenges". In: *Information Fusion* 38 (2017), pp. 43–54.

[93] Kyunghyun Cho et al. "Learning phrase representations using RNN encoder-decoder for statistical machine translation". In: *arXiv preprint arXiv:1406.1078* (2014).

[94] Sepp Hochreiter and Jürgen Schmidhuber. "Long Short-Term Memory". In: *Neural Computation* 9.8 (1997), pp. 1735–1780. DOI: 10.1162/neco.1997.9.8.1735. eprint: http://dx.doi.org/10.1162/neco.1997.9.8.1735. URL: http://dx.doi.org/10.1162/neco.1997.9.8.1735.

[95] Deepa Anand. "Evaluating folksonomy information sources for genre prediction". In: *Advance Computing Conference (IACC), 2014 IEEE International*. Feb. 2014, pp. 887–892. DOI: 10.1109/IAdCC.2014.6779440.

[96] Marina Ivasic-Kos, Miran Pobar, and Luka Mikec. "Movie posters classification into genres based on low-level features". In: *2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. Vol. i. IEEE, May 2014, pp. 1198–1203. ISBN: 978-953-233-077-9. DOI: 10.1109/MIPRO.2014.6859750. URL: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6859750.

[97] Marina Ivasic-Kos, Miran Pobar, and Ivo Ipsic. "Automatic Movie Posters Classification into Genres". In: *ICT Innovations 2014: World of Data*. Ed. by Madevska Ana Bogdanova and Dejan Gjorgjevikj. Cham: Springer International Publishing, 2015, pp. 319–328. ISBN: 978-3-319-09879-1. DOI: 10.1007/978-3-319-09879-1_32. URL: http://dx.doi.org/10.1007/978-3-319-09879-1%7B%5C_%7D32.

[98] Eric Makita and Artem Lenskiy. "A multinomial probabilistic model for movie genre predictions". In: *arXiv preprint arXiv:1603.07849* (2016).

[99] Ian Goodfellow et al. "Maxout Networks". In: *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*. Ed. by Sanjoy Dasgupta and David Mcallester. Vol. 28. JMLR Workshop and Conference Proceedings, May 2013, pp. 1319–1327.

[100] Yoshua Bengio et al. "A Neural Probabilistic Language Model". In: *J. Mach. Learn. Res.* 3 (Mar. 2003), pp. 1137–1155. ISSN: 1532-4435.

[101] Tomas Mikolov et al. "Efficient estimation of word representations in vector space". In: *arXiv preprint arXiv:1301.3781* (2013).

[102] Karen Simonyan and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition". In: *arXiv preprint arXiv:1409.1556* (2014).

[103] Gjorgji Madjarov et al. "An extensive experimental comparison of methods for multi-label learning". In: *Pattern Recognition* 45.9 (2012), pp. 3084–3104. ISSN: 0031-3203. DOI: http://dx.doi.org/10.1016/j.patcog.2012.03.004. URL: http://www.sciencedirect.com/science/article/pii/S0031320312001203.

[104] Sergey Ioffe and Christian Szegedy. "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift". In: *Proceedings of The 32nd International Conference on Machine Learning*. 2015, pp. 448–456.

[105] Diederik Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014).

[106] James Bergstra and Yoshua Bengio. "Random search for hyper-parameter optimization". In: *Journal of Machine Learning Research* 13.Feb (2012), pp. 281–305.

[107] Bart Van Merriënboer et al. "Blocks and fuel: Frameworks for deep learning". In: *arXiv preprint arXiv:1506.00619* (2015).

[108] Abhinav Valada, Ankit Dhall, and Wolfram Burgard. "Convoluted Mixture of Deep Experts for Robust Semantic Segmentation". In: *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) Workshop, State Estimation and Terrain Perception for All Terrain Mobile Robots*. 2016.

[109] Abhinav Valada et al. "Deep Multispectral Semantic Scene Understanding of Forested Environments using Multimodal Fusion". In: *The 2016 International Symposium on Experimental Robotics (ISER 2016)*. Tokyo, Japan, Oct. 2016. URL: http://ais.informatik.uni-freiburg.de/publications/papers/valada16iser.pdf.

[110] Alfredo Huete, Chris Justice, and Wim Van Leeuwen. "MODIS vegetation index (MOD13)". In: *Algorithm theoretical basis document* 3 (1999), p. 213.

[111] Fayao Liu, Chunhua Shen, and Guosheng Lin. "Deep convolutional neural fields for depth estimation from a single image". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 5162–5170.

[112] László Tabár et al. "Swedish two-county trial: impact of mammographic screening on breast cancer mortality during 3 decades". In: *Radiology* 260.3 (2011), pp. 658–663. DOI: 10.1148/radiol.11110469.

[113] Turgay Ayer et al. "Computer-aided diagnostic models in breast cancer screening". In: *Imaging in medicine* 2.3 (2010), pp. 313–323.

[114] Daniel C. Moura and Miguel A. Guevara López. "An evaluation of image descriptors combined with clinical data for breast cancer diagnosis". English. In: *International Journal of Computer Assisted Radiology and Surgery* 8.4 (2013), pp. 561–574. ISSN: 1861-6410. DOI: 10.1007/s11548-013-0838-2.

[115] Raúl Ramos-Pollán et al. "Discovering Mammography-based Machine Learning Classifiers for Breast Cancer Diagnosis". English. In: *Journal of Medical Systems* 36.4 (2012), pp. 2259–2269. ISSN: 0148-5598. DOI: 10.1007/s10916-011-9693-2.

[116] Raúl Ramos-Pollán, Miguel Ángel Guevara-López, and Eugénio Oliveira. "A Software Framework for Building Biomedical Machine Learning Classifiers through Grid Computing Resources". English. In: *Journal of Medical Systems* 36.4 (2012), pp. 2245–2257. ISSN: 0148-5598. DOI: 10.1007/s10916-011-9692-3.

[117] Xiaoming Liu and Jinshan Tang. "Mass Classification in Mammograms Using Selected Geometry and Texture Features, and a New SVM-Based Feature Selection Method". In: *Systems Journal, IEEE* 8.3 (Sept. 2014), pp. 910–920. ISSN: 1932-8184. DOI: 10.1109/JSYST.2013.2286539.

[118] Min Dong et al. "An Efficient Approach for Automated Mass Segmentation and Classification in Mammograms". English. In: *Journal of Digital Imaging* 28.5 (2015), pp. 613–625. ISSN: 0897-1889. DOI: 10.1007/s10278-015-9778-4. URL: http://dx.doi.org/10.1007/s10278-015-9778-4.

[119] John Arevalo, Angel Cruz-Roa, and Fabio A. González. "Hybrid image representation learning model with invariant features for basal cell carcinoma detection". In: *IX International Seminar on Medical Information Processing and Analysis*. Ed. by Jorge Brieva and Boris Escalante-Ramírez. Vol. 8922. SPIE, Nov. 2013. DOI: 10.1117/12.2035530. URL: http://dx.doi.org/10.1117/12.2035530.

[120] Angel Cruz-Roa et al. "A Deep Learning Architecture for Image Representation, Visual Interpretability and Automated Basal-Cell Carcinoma Cancer Detection". English. In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013*. Ed. by Kensaku Mori et al. Vol. 8150. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2013, pp. 403–410. ISBN: 978-3-642-40762-8. DOI: 10.1007/978-3-642-40763-5_50. URL: http://dx.doi.org/10.1007/978-3-642-40763-5%5C_50.

[121] Heung-Il Suk, Seong-Whan Lee, and Dinggang Shen. "Latent feature representation with stacked auto-encoder for AD/MCI diagnosis." In: *Brain structure & function* (Dec. 2013), pp. 1–19. ISSN: 1863-2661. DOI: 10.1007/s00429-013-0687-3.

[122] Feng Li et al. "Robust Deep Learning for Improved Classification of AD/MCI Patients". In: *Machine Learning in Medical Imaging*. Ed. by Guorong Wu, Daoqiang Zhang, and Luping Zhou. Vol. 8679. Lecture Notes in Computer Science. Springer International Publishing, 2014, pp. 240–247. ISBN: 978-3-319-10580-2. DOI: 10.1007/978-3-319-10581-9_30.

[123] Adhish Prasoon et al. "Deep Feature Learning for Knee Cartilage Segmentation Using a Triplanar Convolutional Neural Network". In: *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2013*. Ed. by Kensaku Mori et al. Vol. 8150. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2013, pp. 246–253. ISBN: 978-3-642-40762-8. DOI: 10.1007/978-3-642-40763-5_31.

[124] Afsaneh Jalalian et al. "Computer-aided detection/diagnosis of breast cancer in mammography and ultrasound: a review". In: *Clinical Imaging* 37.3 (2013), pp. 420–426. ISSN: 0899-7071. DOI: 10.1016/j.clinimag.2012.09.024.

[125] Kersten Petersen et al. "Breast Density Scoring with Multiscale Denoising Autoencoders". In: *STMI workshop at MICCAI 2012 (15th International Conference on Medical Image Computing and Computer Assisted Intervention)*. 2012.

[126] Kersten Petersen et al. "Breast Tissue Segmentation and Mammographic Risk Scoring Using Deep Learning". In: *Breast Imaging*. Ed. by Hiroshi Fujita, Takeshi Hara, and Chisako Muramatsu. Vol. 8539. Lecture Notes in Computer Science. Springer International Publishing, 2014, pp. 88–94. ISBN: 978-3-319-07886-1. DOI: 10.1007/978-3-319-07887-8_13.

[127] Xin-Sheng Zhang. "A new approach for clustered MCs classification with sparse features learning and TWSVM." In: *The Scientific World Journal* (Jan. 2014). ISSN: 1537-744X. DOI: 10.1155/2014/970287.

[128] Jun Ge et al. "Computer aided detection of clusters of microcalcifications on full field digital mammograms". In: *Medical Physics* 33.8 (2006).

[129] Andrew R. Jamieson, Karen Drukker, and Maryellen L. Giger. "Breast image feature learning with adaptive deconvolutional networks". In: *Proc. SPIE* 8315 (2012). DOI: 10.1117/12.910710.

[130] GHEONEA IOANA Andreea et al. "The Role of Imaging Techniques in Diagnosis of Breast Cancer". In: *Journal of Current Health Sciences* 37.2 (2011), pp. 241–248.

[131] Noel Pérez Pérez et al. "Improving the Mann–Whitney statistical test for feature selection: An approach in breast cancer diagnosis on mammography". In: *Artificial Intelligence in Medicine* (2014). ISSN: 0933-3657. DOI: 10.1016/j.artmed.2014.12.004.

[132] K. Jarrett et al. "What is the best multi-stage architecture for object recognition?" In: *Computer Vision, 2009 IEEE 12th International Conference on*. Sept. 2009, pp. 2146–2153. DOI: 10.1109/ICCV.2009.5459469.

[133] Nicolas Pinto, David D Cox, and James J DiCarlo. "Why is real-world visual object recognition hard?" In: *PLoS computational biology* 4.1 (2008), e27. DOI: 10.1371/journal.pcbi.0040027.

[134] Siwei Lyu and E.P. Simoncelli. "Nonlinear image representation using divisive normalization". In: *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. June 2008, pp. 1–8. DOI: 10.1109/CVPR.2008.4587821.

[135] Yann LeCun. "Learning invariant feature hierarchies". In: *Computer Vision–ECCV 2012. Workshops and Demonstrations*. Springer. 2012, pp. 496–505.

[136] Alex Krizhevsky. *Learning multiple layers of features from tiny images*. Tech. rep. Toronto: University of Toronto, 2009, p. 58.

[137] Wouter JH Veldkamp and Nico Karssemeijer. "Normalization of local contrast in mammograms". In: *IEEE Transactions on Medical Imaging* 19.7 (July 2000), pp. 731–738. ISSN: 0278-0062. DOI: 10.1109/42.875197.

[138] Qiuhong Ke and Yi Li. "Is Rotation a Nuisance in Shape Recognition?" In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014*. June 2014, pp. 4146–4153. DOI: 10.1109/CVPR.2014.528.

[139] Nitish Srivastava et al. "Dropout: A Simple Way to Prevent Neural Networks from Overfitting". In: *Journal of Machine Learning Research* 15 (2014), pp. 1929–1958.

[140] Ian J Goodfellow et al. "Pylearn2: a machine learning research library". In: *arXiv preprint arXiv:1308.4214* (2013).

[141] The Theano Development Team et al. "Theano: A Python framework for fast computation of mathematical expressions". In: *arXiv preprint arXiv:1605.02688* (2016).

[142] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[143] Jeff Donahue et al. "Decaf: A deep convolutional activation feature for generic visual recognition". In: *International conference on machine learning*. 2014, pp. 647–655.

[144] Olga Russakovsky et al. "Detecting avocados to zucchinis: what have we done, and where are we going?" In: *Proceedings of the IEEE International Conference on Computer Vision*. 2013, pp. 2064–2071.

[145] Diego Rueda-Plata, Raúl Ramos-Pollán, and Fabio A González. "Supervised Greedy Layer-Wise Training for Deep Convolutional Networks with Small Datasets". English. In: *Computational Collective Intelligence*. Ed. by Manuel Núñez et al. Vol. 9329. Lecture Notes in Computer Science. Springer International Publishing, 2015, pp. 275–284. ISBN: 978-3-319-24068-8. DOI: 10.1007/978-3-319-24069-5_26.

[146] Vikas Ganjigunte Ashok, Song Feng, and Yejin Choi. "Success with Style: Using Writing Style to Predict the Success of Novels". In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Seattle, Washington, USA: Association for Computational Linguistics, Oct. 2013, pp. 1753–1764. URL: http://www.aclweb.org/anthology/D13-1181.

[147] Emily Pitler and Ani Nenkova. "Revisiting Readability: A Unified Framework for Predicting Text Quality". In: *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. Honolulu, Hawaii: Association for Computational Linguistics, Oct. 2008, pp. 186–195. URL: http://www.aclweb.org/anthology/D08-1020.

[148] Annie Louis and Ani Nenkova. "What Makes Writing Great? First Experiments on Article Quality Prediction in the Science Journalism Domain". In: *Transactions of the Association for Computational Linguistics* 1 (2013), pp. 341–352.

[149] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate". In: *CoRR* abs/1409.0473 (2014).

[150] Xiaoqing Zheng, Hanyang Chen, and Tianyu Xu. "Deep Learning for Chinese Word Segmentation and POS Tagging". In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Seattle, Washington, USA: Association for Computational Linguistics, Oct. 2013, pp. 647–657. URL: http://www.aclweb.org/anthology/D13-1061.

[151] Jianfeng Gao et al. "Learning Continuous Phrase Representations for Translation Modeling". In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Baltimore, Maryland: Association for Computational Linguistics, June 2014, pp. 699–709. URL: http://www.aclweb.org/anthology/P14-1066.

[152] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. "Domain adaptation for large-scale sentiment classification: A deep learning approach". In: *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*. 2011, pp. 513–520.

[153] Younes Samih et al. "Multilingual Code-switching Identification via LSTM Recurrent Neural Networks". In: *Proceedings of the Second Workshop on Computational Approaches to Code Switching*. Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 50–59. URL: http://aclweb.org/anthology/W16-5806.

[154] Ryan Kiros et al. "Skip-Thought Vectors". In: *Advances in Neural Information Processing Systems 28*. Ed. by C. Cortes et al. Curran Associates, Inc., 2015, pp. 3294–3302. URL: http://papers.nips.cc/paper/5950-skip-thought-vectors.pdf.

[155] Quoc V. Le and Tomas Mikolov. "Distributed Representations of Sentences and Documents." In: *ICML*. Vol. 14. 2014, pp. 1188–1196.

[156] Ronan Collobert and Jason Weston. "A unified architecture for natural language processing: Deep neural networks with multitask learning". In: *Proceedings of the 25th international conference on Machine learning*. ACM. 2008, pp. 160–167.

[157] Anders Søgaard and Yoav Goldberg. "Deep multi-task learning with low level tasks supervised at lower layers". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 231–235. URL: http://anthology.aclweb.org/P16-2038.

[158] Mohammed Attia et al. "CogALex-V Shared Task: GHHH - Detecting Semantic Relations via Word Embeddings". In: *Proceedings of the 5th Workshop on Cognitive Aspects of the Lexicon (CogALex - V)*. Osaka, Japan: The COLING 2016 Organizing Committee, Dec. 2016, pp. 86–91. URL: http://aclweb.org/anthology/W16-5311.

[159] Alice Payne Hackett. *Seventy years of best sellers, 1895-1965*. RR Bowker Co., 1967.

[160] Mariona Coll Ardanuy and Caroline Sporleder. "Structure-based Clustering of Novels". In: *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL)*. Gothenburg, Sweden: Association for Computational Linguistics, Apr. 2014, pp. 31–39. URL: http://www.aclweb.org/anthology/W14-0905.

[161] Amit Goyal, Ellen Riloff, and Hal Daume III. "Automatically Producing Plot Unit Representations for Narrative Text". In: *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Cambridge, MA: Association for Computational Linguistics, Oct. 2010, pp. 77–86. URL: http://www.aclweb.org/anthology/D10-1008.

[162] Andreas van Cranenburgh and Corina Koolen. "Identifying Literary Texts with Bigrams". In: *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*. Denver, Colorado, USA: Association for Computational Linguistics, June 2015, pp. 58–67. URL: http://www.aclweb.org/anthology/W15-0707.

[163] Upendra Sapkota et al. "Not All Character N-grams Are Created Equal: A Study in Authorship Attribution". In: *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Denver, Colorado: Association for Computational Linguistics, May 2015, pp. 93–102. URL: http://www.aclweb.org/anthology/N15-1010.

[164] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. "SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining." In: *LREC*. Vol. 10. 2010, pp. 2200–2204.

[165] Erik Cambria and Amir Hussain. *Sentic Computing: A Common-Sense-Based Framework for Concept-Level Sentiment Analysis*. Vol. 1. Springer, 2015.

[166] Robert Gunning. "The Technique of Clear Writing. McGraw-Hill". In: (1952).

[167] Rudolph Flesch. "A new readability yardstick". In: *Journal of Applied Psychology* 32.3 (1948), pp. 221–223.

[168] J. Peter Kincaid et al. *Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel*. Tech. rep. 1975.

[169] Jonathan Anderson. "LIX and RIX: Variations on a little-known readability index". In: *Journal of Reading* 26.6 (1983), pp. 490–496.

[170] RJ Senter and Edgar A Smith. *Automated readability index*. Tech. rep. CINCINNATI UNIV OH, 1967.

[171] G Harry Mc Laughlin. "SMOG grading-a new readability formula". In: *Journal of reading* 12.8 (1969), pp. 639–646.

[172] Radim Řehůřek and Petr Sojka. "Software Framework for Topic Modelling with Large Corpora". English. In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. http://is.muni.cz/publication/884893/en. Valletta, Malta: ELRA, May 2010, pp. 45–50.

[173] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. "On the difficulty of training recurrent neural networks". In: *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*. 2013, pp. 1310–1318. URL: http://jmlr.org/proceedings/papers/v28/pascanu13.html.

[174] Kyunghyun Cho et al. "On the Properties of Neural Machine Translation: Encoder–Decoder Approaches". In: *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 103–111. URL: http://www.aclweb.org/anthology/W14-4012.

[175] Richard Socher et al. "Parsing with Compositional Vector Grammars". In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Sofia, Bulgaria: Association for Computational Linguistics, Aug. 2013, pp. 455–465. URL: http://www.aclweb.org/anthology/P13-1045.

[176] Xiaojiang Lei, Xueming Qian, and Guoshuai Zhao. "Rating prediction based on social sentiment from textual reviews". In: *IEEE Transactions on Multimedia* 18.9 (2016), pp. 1910–1921.

[177] Fangtao Li et al. "Incorporating Reviewer and Product Information for Review Rating Prediction". In: *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Three*. IJCAI'11. Barcelona, Catalonia, Spain: AAAI Press, 2011, pp. 1820–1825. ISBN: 978-1-57735-515-1. DOI: 10.5591/978-1-57735-516-8/IJCAI11-305. URL: http://dx.doi.org/10.5591/978-1-57735-516-8/IJCAI11-305.

[178] Susan M. Mudambi, David Schuff, and Zhewei Zhang. "Why aren't the stars aligned? An analysis of online review content and star ratings". In: *System Sciences (HICSS), 2014 47th Hawaii International Conference on*. IEEE. 2014, pp. 3139–3147.