1-1-2018

Representation Learning With Convolutional Neural Networks

Haotian Xu
Wayne State University,

REPRESENTATION LEARNING WITH CONVOLUTIONAL NEURAL NETWORKS

by

HAOTIAN XU

DISSERTATION

Submitted to the Graduate School

of Wayne State University,

Detroit, Michigan

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

2018

MAJOR: COMPUTER SCIENCE

Approved By:

_____

Advisor Date

_____

_____

_____

_____

DEDICATION

To my family for the unconditional love and support

ACKNOWLEDGEMENTS

There are many great people that have been there along this long journey and I am honored to express my sincere appreciations here.

First, I want to thank my parents. I wouldn't be here without their unconditional love and support. They encouraged me to pursue my dream and their selfless love and trust make me brave. I also appreciate my wife, Ye. I am so grateful that she is here with me all the way along. I will always remember those sleepless nights and busy weekends working together with her.

I would like to express my sincere gratitude to my advisor, Dr. Ming Dong, for taking me as his Ph.D. student and persistent help throughout my graduate study. This dissertation cannot be completed without his tremendous support. I really appreciate that he gives me an opportunity to see a world of science that I was never exposed to before.

I also want to say thank you to my committee members, Dr. Alexander Kotov, Dr. Dongxiao Zhu, and Dr. Ratna Chinnam for their time and efforts on this dissertation. Their scientific insights and helpful suggestions are valuable for my research study.

I've had the great fortune to collaborate with Dr. Dongxiao Zhu, Dr. Zichun Zhong, and Dr. Justin Jeong during my graduate study. They helped me a lot and shared their knowledge and experience with me. Their support and encouragements are very important for my graduate life.

I also appreciate all Dong Lab members, especially Shixing Chen, Canjin Zhang, Hajar Emami, and Dr. Raed Almomani. The time going through this long journey and solving problems together is memorable to me.

Finally, I am thankful to the financial support from the National Science Foundation (CNS-1637312 and ACI-1657364), the Ford Motor Company University Research Program (2015-9186R), and the National Institute of Neurological Disorders and Stroke (R01-NS089659).

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1 INTRODUCTION

Representation learning, also known as feature learning, is a set of methods that takes raw data as input and discovers the intrinsic structure of data for specific tasks. It is motivated by the fact that machine learning tasks such as classification often require input that is mathematically and computationally convenient to process. However, real-world data such as images and videos are usually complex. Thus, it is necessary to discover useful features or representations from raw data. As a critical step to facilitate the subsequent classification, detection, retrieval, and other tasks, many representation learning approaches have been proposed in the past 100 years (some are shown in Fig. 1.1).

In this chapter, we will review the data representation learning algorithms. Specifically, both conventional feature learning methods and recent deep learning frameworks are included.

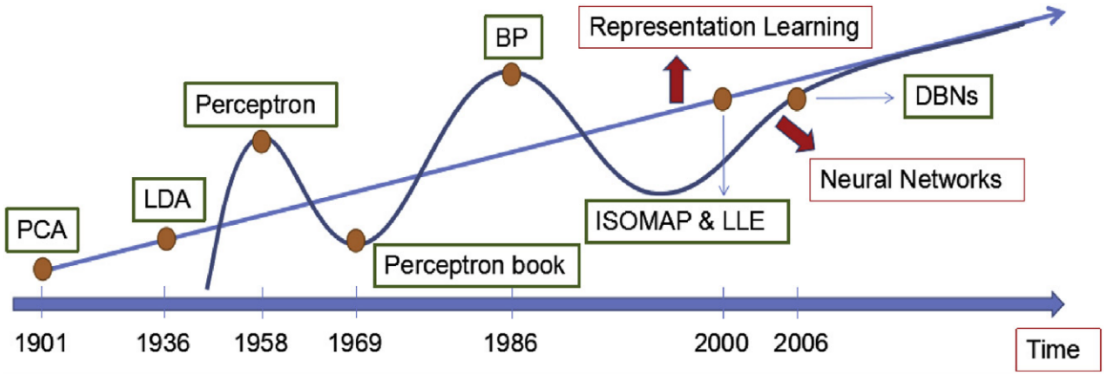


Figure 1.1: The development of representation learning [131].

Data Representation

In the perspective of pattern recognition, data contains the information of a set of objects or patterns that can be processed by computers [65]. Data representation is what would help us differentiate between different concepts, and in turn would also help us find out similarities between them. More specifically, a data point is represented by an n-dimensional vector, which is called a *feature vector*. Each feature vector describes the measurement results and various properties of the corresponding data point. The same data can be represented in

different ways and the choice of data representation has significant impact on the performance of sequent machine learning approaches.

Conventional Representation Learning

In this section, we discuss the conventional feature learning approaches which aim to learn transformations of raw data input to effective representations that can be exploited in subsequent machine learning tasks. An algorithm can be generally categorized into linear or nonlinear, supervised or unsupervised, global or local. For example, Principle Component Analysis (PCA) is a linear, unsupervised, global representation learning method. Linear Discriminant Analysis (LDA) [34] is a linear, supervised, global approach. In this dissertation, we consider representation learning approaches as global methods or local ones. Generally, global algorithms aim to preserve the global relationship and information of data points in the learned feature space, while local approaches aim to preserve the local similarity between each raw data points and its neighbors.

Global Representation Learning

As mentioned above, PCA is one of the earliest representation learning approaches which has been widely used for dimensionality reduction. It applies an orthogonal transformation to convert a set of (possibly) correlated variables into linearly uncorrelated ones which are also known as *principle components*. More formally, the transformation is defined in such a way that the first principle component has the largest variance and explains the most of data variability, and every succeeding component in turn explains the most of left data variability under the constraint that each component should be orthogonal to preceding components. Eigenvalue decomposition is applied for optimization.

LDA is a supervised, linear representation learning algorithm, which encourages data points belonging to the same class to be close to each other and that belonging to different classes to be far away in the learned feature space. It has been successfully used for face recognition, and the learned features are named *Fisherfaces* [9]. Similar to *Eigenfaces* [102], which is obtained by PCA, Fisherfaces is also extracted from face images and a nearest

neighbor classifier can be applied for the subsequent face recognition. However, comparing with Eigenfaces, Fisherfaces has intra-class compactness and inter-class dispersion even under severe variation in lighting and facial expressions.

Local Representation Learning

Local representation learning, also known as *manifold learning*, focuses on mining the locality-based feature relationship. Although most of the manifold learning methods are nonlinear dimensionality reduction approaches, some are linear ones, such as Locality Preserving Projections (LPP) [42]. Some manifold learning approaches are shown in Fig. 1.2. Meanwhile, some nonlinear dimensionality reduction algorithms are not manifold learning approaches, as they are not aimed to discover the intrinsic structure of high dimensional data, such as Kernel PCA [97].

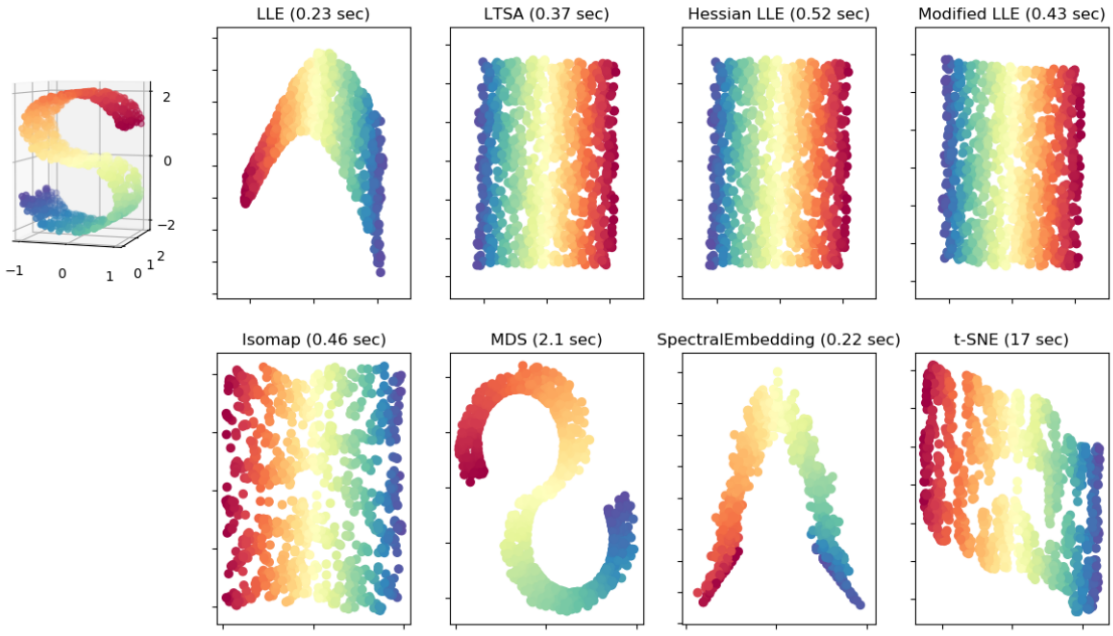


Figure 1.2: An illustration of various manifold learning methods [86].

Locally Linear Embedding (LLE) [94] encodes the locality information at each point into the reconstruction weights of its neighbors. Following the idea of LLE, Local Tangent Space Alignment (LTSA) [130] was proposed to represent the local geometry of the manifold in the tangent space. For the Isometric Feature Mapping (Isomap) [107], it combines the

Floyd-Warshall algorithm [35] with classic MDS [94]. Isomap computes the pair-wise geodesic distances between local neighbors of data samples using the Floyd-Warshall algorithm, which is utilized to find the shortest distance between each pair of samples, and then learns the data embeddings with MDS on the precomputed pair-wise distances [131].

Deep Representation Learning

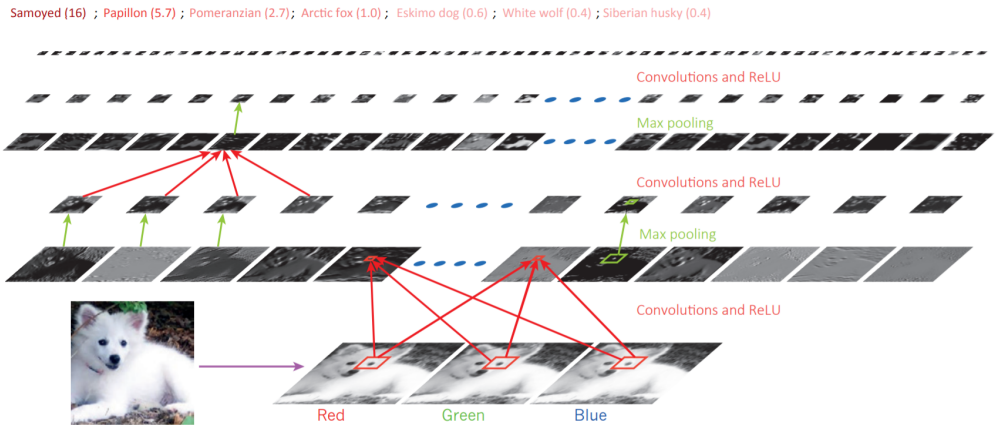


Figure 1.3: Typical CNNs used in computer vision applications [18].

As opposed to conventional machine learning methods, deep learning frameworks require little manual feature engineering and can easily take advantage of the increasing amount of data and computational ability. An example Convolutional Neural Network (CNN) architecture is shown in Fig. 1.3.

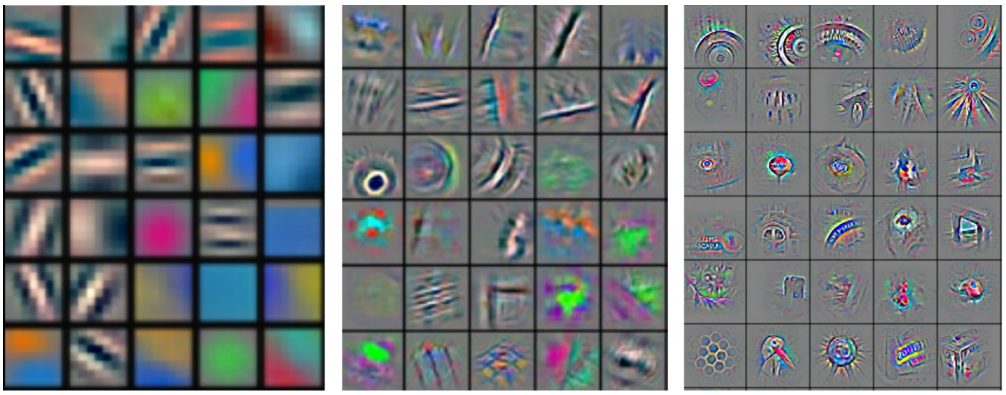


Figure 1.4: The learned representations are from coarse to fine [127].

Deep learning is part of the boarder family of representation learning which learns multiple levels of data representations by stacking non-linear layers that each transforms the

representation from a lower and coarser level into one at higher and fine level. For a specific task, high level layers of representation amplify aspects of the input that are important for discrimination and suppress irrelevant variations [63]. As shown in Fig. 1.4, the CNN takes an image of a Samoyed as input which is represented as three channels of pixel arrays, and the learned low level features typically describe edges at particular orientations and locations in the image. For high level layers, abstract of the whole images is extracted and then used for specific tasks. The most significant advantage of deep learning is that data representation is learned with a general purpose learning framework instead of being designed by human engineers [63].

Deep learning frameworks have turned out to be very good at discovering intrinsic structures in high dimensional data and are therefore applicable to many domains of science. In addition to overperforming state-of-the-arts in image recognition [58, 31, 21] and speech recognition [96, 83], deep learning also produced extremely promising results for various tasks in natural language understanding [54], question answering [4] and machine translation [7, 106].

Our Contributions

In this dissertation, we focus on representation learning with deep neural networks. Specifically, we propose the following research topics with significant intellectual merit and novelty. We summarize below the three research projects that we accomplished as part of this dissertation.

Learning topic-based word embedding for text analysis

We propose a novel word embedding learning approach, which provides topic-based semantic word embeddings and two CNN architectures, that can utilize multiple word representations simultaneously for text classification. Specifically, the main contributions are summarized as follows:

• We develop a word embedding learning model, Topic-based Skip-gram, which captures word semantic relationship with topic models, e.g., LDA, and then integrate it into distributed word embedding learning with a novel objective function.

• We introduce two complementary multimodal CNN architectures that simultaneously take multiple kinds of word embeddings as inputs for text classification.

• We combine the proposed topic-based word embedding and other state-of-the-art word embeddings as inputs to the proposed multimodal CNN architectures. Our experiments conducted on several real-world datasets show that combination of the proposed topic-based word representations and our multimodal CNNs outperforms state-of-the-art word representations in various text classification tasks, including indexing of biomedical articles.

Learning 3D polygon representation for shape segmentation

We propose Directionally Convolutional Network that extends convolution operations from images to the surface mesh in the spatial domain. Furthermore, we introduce a two-stream framework combining proposed Directionally Convolutional Network and a neural network for segmentation of 3D shapes. Instead of fusing the two streams by a simple concatenation, we take our framework as an approximation of a directed graph and combine the probabilities inferred by the two streams with an element-wise product. Finally, Conditional Random Field is applied to optimize the surface mesh segmentation. The main contributions are summarized as follows:

• By defining rotation-invariant convolution and pooling operations on the surface of 3D shapes, we learn effective shape representations from raw geometric features, *i.e.*, face normals and distances, to achieve robust segmentation of 3D shapes.

• Based on the proposed Directionally Convolutional Network, we introduce a two-stream framework to classify each face of a given mesh into predefined semantic parts. Our approach achieves state-of-the-art segmentation results on a large variety of 3D shapes.

Learning deep brain fiber representation for classification

We propose a CNN-based end-to-end learning framework with direct representation learning to differentially delineate diffusion tensor imaging (DTI)-based eloquent axonal pathways by incorporating functional MRI (fMRI) and electrical stimulation mapping (ESM) observations. The main contributions of this work are as follows:

• Two CNN architectures with different depth were investigated in this study: a shallow CNN model with 3 layers from our previous work [119]; inspired by the great success of very deep CNNs [41, 92], we also adapted the shallow CNN into a deep model with 21 layers. The proposed CNN models generate different feature maps of the input data (*i.e.*, 3D spatial coordinates of individual fiber streamlines) by using a sequence of convolutional and pooling layers before classifying input data using fully connected layers.

• A couple of novel CNN loss functions [68, 116] were introduced for pathway classification task. First, since our dataset is highly unbalanced, which cannot be handled well by CNN with the conventional cross-entropy loss, we introduced focal loss into the proposed CNN models. Focal loss applies a modulating term to the cross-entropy loss to help focus on hard examples and down-weight the numerous easy examples. Second, to further improve the classification performance and generalization ability of proposed CNN models, the learned fiber representations need to be not only separable but also discriminative. Center loss was introduced which adds a cluster-based loss term to the cross-entropy loss to ensure the learned representations have both compact intra-class variations and large inter-class margins.

• Although CNNs have led to breakthroughs of state-of-the-arts, the end-to-end learning strategy makes the entire CNN model a black box. This weakness is highlighted in the biomedical imaging: if we do not know how the trained CNNs classify each fiber, we cannot fully trust the classification results given by the CNN models. In this study, we applied attention mechanism [120] in the proposed CNNs, which highlights the most useful segments of a fiber for classification. In this study, we will demonstrate that the attention mecha-

nism provides a machine perspective on how the CNNs classify functionally-important white matter pathways.

Organization

The rest of this dissertation is organized as follows: In Chapter 2, we introduce our method on learning topic-based word embedding for text analysis. In Chapter 3, we describe the proposed approach to learning 3D polygon representation for shape segmentation using a two-stream deep neural network framework. In Chapter 4, we introduce our framework to learn effective, discriminative, and interpretable brain fiber representations for classification in detail. In Chapter 5, we conclude and review future research directions.

CHAPTER 2 Learning Word Embedding for Text Analysis

In this chapter, we propose a novel neural language model, Topic-based Skip-gram, to learn topic-based word representation for text analysis with CNNs. Topic-based Skip-gram leverages textual content with topic models, *e.g.*, Latent Dirichlet Allocation, to capture precise topic-based word relationship and then integrate it into distributed word embedding learning. We then describe two multimodal CNN architectures, which are able to employ different kinds of word embeddings at the same time for text classification.

Introduction

As the amount of biomedical textual data in MEDLINE of the US National Library of Medicine (NLM) is growing exponentially, the indexing of biomedical articles is becoming a much more difficult task. Medical Text Indexer (MTI)[1] [5] has been assigned to this task as a support tool which produces (semi-)automated recommendation indexing based on predefined Medical Subject Headings (MESH)[2]. Meanwhile, biomedical literature indexing can also be viewed as a classification over textual data into a set of predefined classes. However, as discussed in [93, 124], traditional machine learning algorithms, including Naive Bayes, Support Vector Machine and Logistic Regression, cannot outperform MTI system without ensemble.

Recently, CNN models have achieved remarkably strong performance in natural language processing and become commonly used architectures for text classification [49, 54, 60, 128]. As input features of CNNs, various types of word vector representations have been proposed. Generally speaking, there are two model families to represent words with real-valued vectors: 1)matrix factorization methods, such as [28, 71] and 2)local window-based methods, such as [12, 25, 78]. Both families have their own pros and cons. Although matrix factorization methods do not require much domain expertise of word embedding and efficiently leverage statistical information of corpora, their main problem is that most frequent words (or characters) have a large negative impact on word similarity measure, which leads

[1] `http://ii.nlm.nih.gov/MTI/index.shtml`

[2] `https://www.nlm.nih.gov/pubs/factsheets/mesh.html`

to poor performance on word analogy tasks. Local window-based methods perform better on analogy tasks, but they poorly utilize statistical information about corpus because these models are trained on separate local windows of content.

In the presented work, we propose a novel word embedding learning approach, which provides topic-based semantic word embeddings and two CNN architectures, which can utilize multiple word representations simultaneously for text classification. Specifically, our framework first leverages the whole text corpus with topic models to capture semantic relationship between words and then take it as the input for word representation learning using Topic-based Skip-gram with a novel objective function. Then, these topic-based word representations are used together with other state-of-the-art word embeddings for text classification in multimodal CNN models. Specifically, the main contributions of this work are summarized as follows:

• We propose a learning-based word embedding model, Topic-based Skip-gram, which captures word semantic relationship with topic models, e.g., LDA, and then integrate it into distributed word embedding learning with a novel objective function.

•We introduce two complementary multimodal CNN architectures that are able to simultaneously take multiple kinds of word embeddings as inputs for text classification.

• We combine the proposed topic-based word embedding and other state-of-the-art word embeddings as inputs to the proposed multimodal CNN architectures. Our experiments conducted on several real-world datasets show that combination of the proposed topic-based word representations and our multimodal CNNs outperforms state-of-the-art word representations in various text classification tasks, including indexing of biomedical articles.

The rest of this chapter is organized as follows. In Section , we review related work in biomedical literature indexing and word representation learning. The details of our word embedding learning approach and multimodal CNN models are introduced in Section . In Section , we demonstrate that our topic-based word embedding produces competitive results

with CNN architecture and outperforms state-of-the-art approaches with our multimodal CNN models in three case studies. At last, we conclude in Section .

Related Work

Indexing of Biomedical Literature

Our work shares the high-level goal of biomedical literature indexing with many previous works, such as USI [33], MeSHLabeler [69], MeSH Now [73] and Atypon [84]. Several other works [93, 124] tried to improve the MTI system with automatic machine learning methods. Among them, Yepes et al. [124] pointed out that ensemble of classic machine learning methods can outperform indexing performance of MTI. Rios and Kavuluru [93] surpassed MTI performance by utilizing CNNs for sentence-level textual classification [54] with word embeddings trained by the Skip-gram model [78], which is more closely related to our work. However, these works focus on utilizing classic machine learning methods for biomedical literature indexing, while we propose a novel Topic-based Skip-gram for learning topic-based semantic word representations and obtain state-of-the-art classification performance with deep learning architectures.

Topic Models

Topic models are probabilistic generative models to discover main themes of documents. These models share the same assumptions: 1) they posit there are a set of latent topics, which are multinomial distributions over vocabulary; 2) each document is a mixture of these topics. Recently, topic models have become a popular tool for text classification [75, 90], image classification [32, 91], transfer learning [20, 121] and unsupervised analysis of textual data [14, 15]. As one of the most commonly used unsupervised topic models, Latent Dirichlet Allocation (LDA) [15] can extract semantic information from corpora. The basic assumption of LDA is that each document is a mixture of topic proportions and each topic is a distribution over fixed vocabulary. In this work, we employ LDA to identify topic-based semantic relationships between words in each corpus.

Word Embedding Learning Methods

Recently, Mikolov et al. introduced an algorithm for learning fixed length distributed representations of words in a vector space, the Skip-gram model [77], which is a single-layer neural network based on inner products between word vectors. As one of the local window-based methods, Skip-gram's objective is to learn word embeddings that can predict the textual content of a word given the word itself. Through experiments on word and phrase analogy tasks, this model demonstrated its capacity to capture linguistic relationships between word vectors. However, Skip-gram model suffers from the disadvantage that it does not utilize the co-occurrence statistics of the corpus. Instead, Skip-gram scans textual corpus with local context windows, which fails to make use of statistical information of the whole corpus. Pennington et al. [87] took the advantages of both global matrix factorization and local content window-based methods by training their model only on nonzero elements in the word co-occurrence matrix. Different from their approach, Topic-based Skip-gram leverages global statistical information of the whole corpus with LDA and learns the semantic information with local content windows.

CNNs for Text Classification

A number of CNN architectures have been developed for text classification [49, 54, 60, 128]. Kalchbrenner et al. [49] focused on sentence modeling with a CNN-based model for word-level input. Zhang and LeCun [128] concentrated on character-level input with a very deep CNN architecture which requires a large amount of training data and training time. Lai et al. [60] proposed a model which combines Recurrent Neural Networks (RNN) with CNN. Kim [54] proposed a two-layer CNN model for sentence-level text classification with single kind of word embeddings. This model is simple but very effective for text classification. Our multimodal approach is inspired by this model. In contrast to the architecture described by Kim, our multimodal approaches are able to simultaneously take multiple kinds of word representations as inputs.

The Proposed Approach

In this section, we first present technical details of Topic-based Skip-gram for learning topic-based semantic word embeddings and then introduce two multimodal CNN architectures which employ multiple kinds of word embeddings as inputs for text classification.

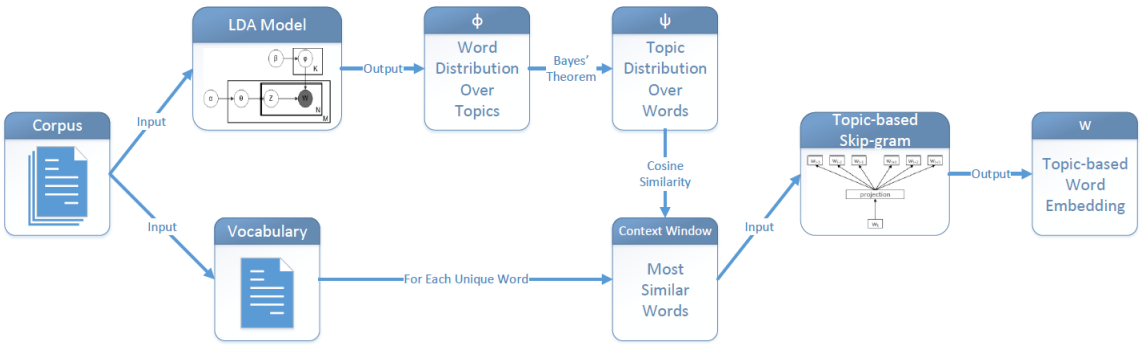


Figure 2.1: Workflow of Topic-based Skip-gram.

Topic-based Skip-gram

Topic-based Skip-gram identify semantic relationship between words from corpus using LDA and then integrate it into word representation learning with a novel objective function. The workflow is shown in Fig. 2.1 and we will introduce the details in this subsection.

Leveraging Topic-based Semantic Information with LDA

LDA. The basic idea of LDA is that each document d is a distribution over K latent topics and each topic is a distribution over V unique words in the dictionary. Given a corpus of M documents and each document has N_m words, the generative process of LDA is as follows:

1. Choose $\boldsymbol{\theta} \sim \text{Dirichlet}(\boldsymbol{\alpha})$

   $\boldsymbol{\theta}$ denotes topic distribution over documents. Each document has its own $\boldsymbol{\theta}$, which needs to be estimated during the training stage. Each $\boldsymbol{\theta}$ is a vector of length K, where K is the number of topics and chosen manually at the beginning of training. $\boldsymbol{\alpha}$ is the hyperparameter of document-topic distribution.

2. Choose $\boldsymbol{\phi} \sim \text{Dirichlet}(\boldsymbol{\beta})$

   $\boldsymbol{\phi}$ is word distribution over topics, also known as topic in [15], which is a matrix of K rows and V columns. Element $\phi_{i,j}$ equals $p(w_j|z_i)$, which is the probability of generating word w_j given this word belonging to topic z_i. $\boldsymbol{\beta}$ is the hyperparameter of topic-word distribution.

3. For each of the N words w_n in each document d_m of the M documents in the corpus:

   (a) Choose a topic $z_n \sim \text{Multinomial}(\boldsymbol{\theta})$

   The topic indicator z_n is the topic k assigned to word w_n.

   (b) Generate a word $w_n \sim \text{Multinomial}(z_n, \boldsymbol{\beta})$

   Generate a word as w_n, which is the nth unique word in the dictionary, from Multinomial distribution $p(w_n|z_n, \boldsymbol{\beta})$.

Topic-based Semantic Information of Corpus. In this chapter, we treat the topic distribution over words $\boldsymbol{\psi}$ as topic-based semantic information of corpus for learning word embeddings. $\boldsymbol{\psi}$ is a $V \times K$ matrix. Its element $\psi_{i,j}$ is equal to $p(z_i|w_j)$, which is the probability for word w_j to be assigned to topic z_i. It can be approximated with word distribution over topics $\boldsymbol{\phi}$ based on Bayes' theorem:

$$p(z_i|w_j) = \frac{p(w_j|z_i) \cdot p(z_i)}{p(w_j)}, \tag{2.1}$$

where $p(z_i)$ is the marginal probability of topic z_i and $p(w_j)$ denotes the marginal probability of word w_j in the dictionary. $p(z_i)$ and $p(w_j)$ can be calculated as follows:

$$p(z_i) = \frac{\sum_{m=1}^{M} z_i^m}{M}, \tag{2.2}$$

$$p(w_j) = \frac{\sum_{m=1}^{M} N_m^j}{\sum_{m=1}^{M} N_m}, \tag{2.3}$$

where z_i^m is the topic proportion of z_i in document d_m and N_m^j is the count of word w_j in the document d_m. The topic-based semantic information matrix $\boldsymbol{\psi}$ is then used as training data in the word embedding learning step.

Learning Topic-based Word Embeddings

Skip-gram. The training objective of Skip-gram [78] is to learn distributed word representations which aim at predicting the surrounding words in the documents. Given a training corpus of T words $w_1, w_2, w_3, \cdots, w_T$, the learning objective of the Skip-gram model is to maximize the average log probability

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leqslant j\leqslant c, j\neq 0}\log p(w_{t+j}|w_t), \tag{2.4}$$

where c is the size of training content. In other words, given a local window of size $2\cdot c+1$, the objective of Skip-gram model is to maximize prediction log probability of the $2\cdot c$ words $w_{t-c}, w_{t-c+1}, \cdots, w_{t-1}, w_{t+1}, \cdots, w_{t+c-1}, w_{t+c}$ given the word w_t in the center.

Learning Semantic Word Embeddings. We propose a novel training objective function for Topic-based Skip-gram that is to learn distributed word embeddings which are useful to predict words with similar topic-based semantic information. The basic assumption of Topic-based Skip-gram is that if topic distributions of two words $\boldsymbol{\psi_i}$ and $\boldsymbol{\psi_j}$ have a large cosine similarity between each other, then these two words share similar topic-based semantic information. Given a dictionary of N unique words $w_1, w_2, w_3, \cdots, w_N$ of a corpus, the objective of Topic-based Skip-gram model is to maximize the average log probability

$$\frac{1}{N}\sum_{n=1}^{N}\sum_{-c\leqslant j\leqslant c, j\neq 0}\log p(w_{n+j}|w_n). \tag{2.5}$$

In other words, given half window size c (s.t. window size is $2c+1$) and a word in the dictionary w_n, the training objective of Topic-based Skip-gram is to maximize prediction log probability of the top $2c$ words similar to w_n. The probability $p(w_{n+j}|w_n)$ is defined using

$softmax$ function

$$p(w_{n+j}|w_n) = \frac{\exp(v_{w_{n+j}}^\top v_{w_n})}{\sum_{1 \leqslant i \leqslant N, i \neq n} \exp(v_{w_i}^\top v_{w_n})}, \tag{2.6}$$

where v_{w_n} is the vector representation of word w_n. In practice, the cost of computing $\nabla \log p(w_{n+j}|w_n) \propto N$, where N can be very large ($10^6 - 10^8$ unique words).

Optimization. Same with Skip-gram, we use Negative Sampling [78] to optimize the objective function of Topic-based Skip-gram. In Negative Sampling, $p(w_{n+j}|w_n)$ is replaced as

$$\log \sigma(w_{n+j}|w_n) + \sum_{i=i}^{k} E_{w_i \sim P_n(w)}[\log \sigma(-v_{w_i}^\top v_{w_n})]. \tag{2.7}$$

The idea is to distinguish target word w_{n+j} from k noise words which are drawn from noise distribution $P_n(w)$ using logistic regression by maximizing the probability of target word (first item) and minimizing the probability of noise words (second term). According to results reported in [78], we choose $k = 15$ and $P_n(w) \sim \frac{U(w)^{0.75}}{Z}$,where $U(w)$ is unigram distribution.

Time efficiency. Given a dataset of N unique words and L words in total, proposed Topic-based Skip-gram optimizes N word windows and Skip-gram optimizes L windows. Note that $N \ll L$ in most cases. Furthermore, Topic-based Skip-gram can also work with other semantic indexing models in addition to LDA, which may significantly expedite the training process.

We summarize the learning procedure for topic-based semantic word embedding in Algorithm 1.

Multimodal CNN Architectures

In this part, we first introduce a single channel CNN model [54], which is used as baseline architecture in the experiments. Then we will describe the two proposed multimodal CNN architectures which can take multiple types of word embeddings with different length.

Algorithm 1 Topic-based Skip-gram

Input: Raw training textual corpus $\boldsymbol{\mathcal{D}}$; Topic number K, Hyperparameters $\boldsymbol{\alpha}, \boldsymbol{\beta}$ for LDA, Half window size c
Output: Topic-based semantic word embedding $\mathcal{W}$
procedure GETWORDEMBEDDING
    $\boldsymbol{\phi} = LDA(\boldsymbol{\mathcal{D}}, \boldsymbol{\alpha}, \boldsymbol{\beta}, K)$ ▷ *Train LDA model on the corpus $\boldsymbol{\mathcal{D}}$ and get word distribution over topics $\boldsymbol{\phi}$*
    for Each topic z_i **do**
        Compute marginal probability of each topic $p(z_i)$ with Eq. (2.2)
    for Each word w_j **do**
        Compute marginal probability of each word $p(w_j)$ with Eq. (2.3)
    Compute topic distribution over words $\boldsymbol{\psi}$ based on Eq. (2.1), (2.2) and (2.3)
    for Each word w_j **do**
        Find $2c$ words with most similar topic distribution over words to w_j according to cosine similarity ▷ *These $2c+1$ words are then used as an input window win_j for Topic-based Skip-gram*
    $\mathcal{W}$=Topic-based-Skip-gram($\boldsymbol{win}$) ▷ *Take all word windows $\boldsymbol{win}$ as input of Topic-based Skip-gram to learn topic-based word embedding $\mathcal{W}$ based on the objective function in Eq. (2.5)*

Baseline CNN

The baseline CNN contains one input layer, one convolution layer, one maxpooling layer and one fully connected layer. Although one output neuron with *sigmoid* or *tanh* function is sufficient for binary classification, we choose multiple neurons with $softmax$ function to make it easier to adopt CNN models for multi-class classification. The details of each layer are described as follows.

Input layer. Formally, we denote $\mathbf{x_i} \in \mathbb{R}^k$ as the k-dimensional word representation for the ith word in a sentence. A sentence of length n is denoted as

$$\mathbf{X_{1:n}} = \mathbf{x_1} \oplus \mathbf{x_2} \oplus \cdots \oplus \mathbf{x_n}, \tag{2.8}$$

where $\oplus$ is the concatenation operator. By this, each input sentence is represented as a $n \times k$ matrix. In practice, short sentences are padded with zeros to same length, such that, each matrix shares the same size.

Convolution layer. A convolution filter $\mathbf{w} \in \mathbb{R}^{h \times k}$, which is applied to a window of h words of k-dimensional embeddings, produces a new feature. For instance, given a window of words $\mathbf{X_{i:i+h-1}}$ and a bias term $b \in \mathbb{R}$, a new feature c_i is generated by

$$c_i = f(\mathbf{w} \cdot \mathbf{X_{i:i+h-1}} + b), \tag{2.9}$$

where f is a non-linear function. In our case, we apply the element-wise function Rectified Linear Unit (ReLU) to the input matrices:

$$ReLU(x) = \begin{cases} x, & \text{if } x > 0 \\ 0, & \text{otherwise} \end{cases} \tag{2.10}$$

Each filter produces a feature map $\mathbf{c} = [c_1, c_2, \cdots, c_{n-h+1}]$ from every possible window $\{\mathbf{X_{1:h}}, \mathbf{X_{2:h+1}}, \cdots, \mathbf{n-h+1:n}\}$ of a sentence of length n. In [54], multiple layers of various sizes are applied in the convolution layer, and multiple feature maps are generated.

Sub-sampling layer. There are several sub-sampling methods, such as average pooling, median pooling and max pooling. In this case, we apply max pooling over each feature map produced by the convolution layer and take the maximum element $\hat{c} = max\{\mathbf{c}\}$. Let's denote features generated by this max pooling layer as

$$\hat{\mathbf{c}} = \hat{c}_1 \oplus \hat{c}_2 \oplus \cdots \oplus \hat{c}_m, \tag{2.11}$$

where m is the number of feature maps.

Fully connected layer. Given $\hat{\mathbf{c}}$ as the input, the fully connected layer produces

$$P(Y = i|\hat{\mathbf{c}}, \boldsymbol{\theta}) = softmax_i(\mathbf{W} \cdot (\hat{\mathbf{c}} \circ \mathbf{r}) + b), \tag{2.12}$$

where Y is the prediction, $\boldsymbol{\theta}$ denotes parameters $\{W, b\}$, $\mathbf{W}$ denotes weights, $\circ$ denotes the element-wise multiplication operator and $\mathbf{r} \in \mathbb{R}^m$ is a dropout mask vector of Bernoulli variables with probability p of being zero. During the back propagation stage, only unmarked elements in $\hat{\mathbf{c}}$ are involved in the computation. l_2-norm [44] is also applied to weight matrices W. If $\|W\|_2 > s$ after gradient descent step, we rescale W, such that $\|W\|_2 = s$. Here, s is a manually defined parameter. By applying dropout and l_2-norm, we prevent the overfitting problem.

Optimization. A reasonable training objective is to minimize categorical (or binary) cross-entropy loss. The average loss for each sample is

$$\begin{aligned} Q(\boldsymbol{\theta}) &= \frac{1}{|\boldsymbol{\mathcal{D}}|}\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\mathcal{D}}) \\ &= -\frac{1}{|\boldsymbol{\mathcal{D}}|}\sum_{i=1}^{|\boldsymbol{\mathcal{D}}|} \log P(Y = y^i | x^i, \boldsymbol{\theta}), \end{aligned} \tag{2.13}$$

where x^i is the ith sample in the dataset and y^i is the prediction for it. In the proposed framework, we update the parameters $\boldsymbol{\theta}$ by Adadelta [126], which is an adaptive learning rate approach for classic SGD.

Multi-channel CNN (CNN-channel)

As shown in the top panel of Fig. 2.2, CNN-channel model combines two baseline CNN models. More formally, we denote two kinds of word embeddings $\mathbf{x_i^1} \in \mathbb{R}^{k_1}$ and $\mathbf{x_i^2} \in \mathbb{R}^{k_2}$ as k_1- and k_2-dimensional word representations for the ith word in a sentence. So, a sentence of length n can be represented in two ways

$$\mathbf{X_{1:n}^1} = \mathbf{x_1^1} \oplus \mathbf{x_2^1} \oplus \cdots \oplus \mathbf{x_n^1} \tag{2.14}$$

and

$$\mathbf{X_{1:n}^2} = \mathbf{x_1^2} \oplus \mathbf{x_2^2} \oplus \cdots \oplus \mathbf{x_n^2}, \tag{2.15}$$

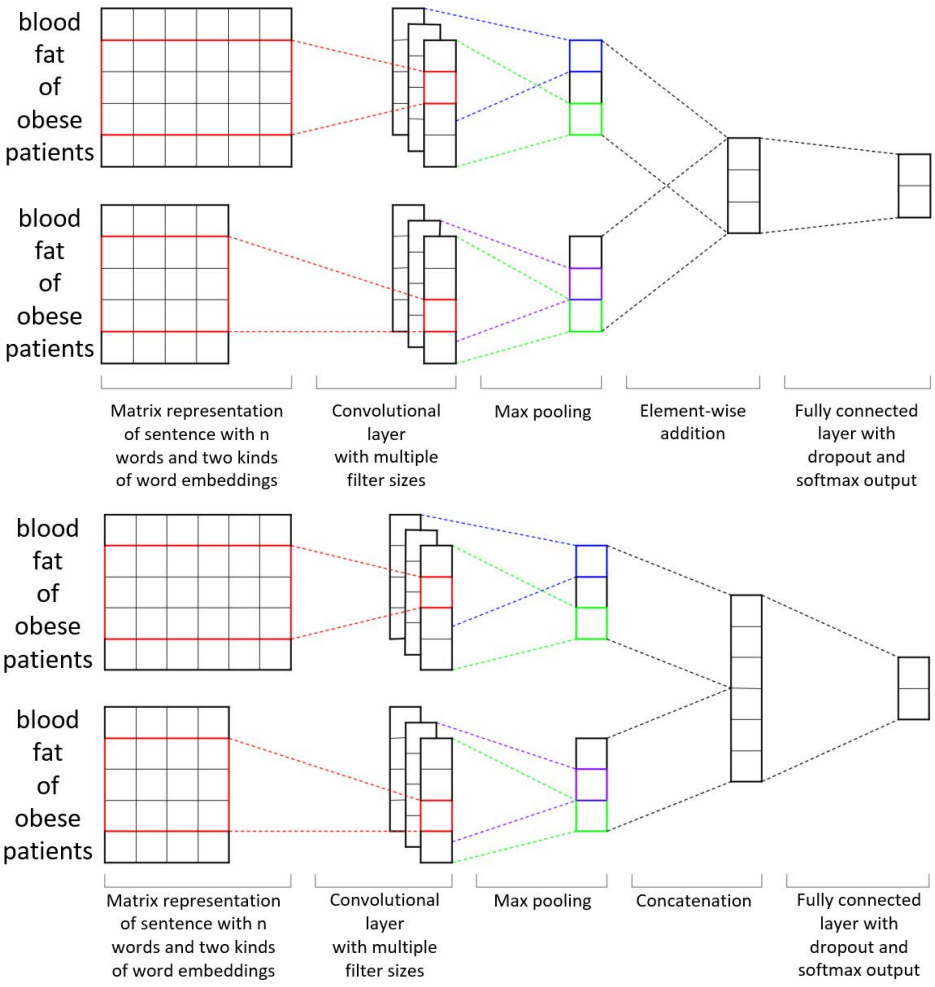


Figure 2.2: Architecture of CNN-channel (top) and CNN-concat (bottom).

where $\mathbf{X_{1:n}^1}$ is used as the input matrix for the 'top channel' of CNN-channel and $\mathbf{X_{1:n}^2}$ is the input for the 'bottom channel' of CNN-channel. Similarly, after applying convolution and max-pooling layers, $\hat{\mathbf{c}}^{\mathbf{1}}$ and $\hat{\mathbf{c}}^{\mathbf{2}}$ are generated. In CNN-channel, they are merged by element-wise addition:

$$\hat{\mathbf{c}} = \hat{\mathbf{c}}^{\mathbf{1}} + \hat{\mathbf{c}}^{\mathbf{2}} \tag{2.16}$$

Here, + denotes element-wise addition. Then we apply the fully connected layer with dropout and softmax output and l_2 regularization as in the baseline CNN model.

Concatenation CNN (CNN-concat)

As shown in the bottom panel of Fig. 2.2, CNN-concat is also built on top of the baseline CNN model. Different from CNN-channel, $\hat{\mathbf{c}}^{\mathbf{1}}$ and $\hat{\mathbf{c}}^{\mathbf{2}}$ are merged by concatenation

$$\hat{\mathbf{c}} = \hat{\mathbf{c}}^{\mathbf{1}} \oplus \hat{\mathbf{c}}^{\mathbf{2}} \tag{2.17}$$

Then $\hat{\mathbf{c}}$ is taken as the input of fully connected layer as in the baseline CNN model. Although CNN-channel and CNN-concat models can be expended to utilize as many types of word embeddings as needed, we only employ two kinds of word representations in our experiments.

Deep Understanding of Multimodal CNNs

Multimodal CNNs vs. original CNN model. Original CNN architecture, which was proposed in [54], can only take one kind of word embedding as input. Meanwhile, our proposed multimodal CNNs are able to simultaneously take multiple types of word embeddings as inputs, which means that multimodal CNNs have stronger learning ability than the original CNN model. Specifically, by combing the topic-based word embedding and local window-based word embeddings, the multimodal CNNs are able to utilize both topic-based semantic relationship and local content information and outperform the original CNN model.

CNN-channel vs. CNN-concat. CNN-channel combines the two kinds of word representations by element-wise addition, commonly used for multi-channel image classification. On the other hand, CNN-concat concatenates two parts together, which introduces more parameters to fit. In other words, CNN-concat has stronger learning ability but needs more training data to preserve from overfitting than CNN-channel. When the training set has enough positive samples for binary classification task or is balanced for multi-class classification problem, CNN-concat is a better choice than CNN-channel.

Experiments

We evaluate our framework by three tasks: 1)indexing of biomedical articles; 2)annotation of clinical text fragments with behavior codes; and 3)classification of benchmark newsgroups. Baselines and state-of-the-art algorithms are compared with our method in these experiments. In our experiments, we used the same code[3] and parameter settings as in [93] for the baseline CNN model. We make implementation of proposed multimodal CNNs publicly available[4].

Datasets

Indexing of Biomedical Articles

MEDLINE citations. A public dataset[5] of MEDLINE citations from November 2012 to February 2013 is used in this work. The dataset contains 143,853 citations in total, from which 94,942 citations were selected for training and 48,911 were selected for testing. As in [93], we categorize 29 MeSH terms into three groups according to MTI's performance: check tags, low precision terms and low recall terms. The check tags group is a common set of top 12 MeSH headings routinely considered for almost all articles (e.g. Humans, Female and Male), the low precision group contains 10 MeSH headings with the lowest precision performance using MTI and the low recall group contains 7 MeSH headings with the lowest recall performance using MTI. We build CNN models as binary classifiers for each MeSH to

[3] https://github.com/yoonkim/CNN_sentence

[4] https://github.com/HaotianMXu/Multimodal-CNNs

[5] http://ii.nlm.nih.gov/MTI_ML/index.shtml

Table 2.1: Description of five behavior code annotations.

Behavior	Definition	Sample Quote
Positive Commitment Language	Statement describing intentions, plans for, and action steps toward changing the current behavior pattern	Well, I've been trying to lose weight,but it really never goes anywhere.
Negative Commitment Language	Statement describing intentions, plans for, and action steps toward maintaining the current behavior pattern	I eat a lot of junk food, like cake and cookies, stuff like that.
Positive Change Talk	Statement describing the desire, ability, reason, or need for changing the current behavior pattern	Hmmm, I guess I need to lose some weight.
Negative Change Talk	Statement describing the desire, ability, reason, or need for maintaining the current behavior pattern	I just don't feel like I want to eat before. I'm just not hungry at all.
Ambivalence	Statements that combine positive and negative commitment language and/or change talk	Fried foods may taste good, but it's not good for your health.

classify if a document belongs to this MeSH term. Note that although only 29 terms are used in this experiment, our framework works for arbitrary number of MeSH terms.

Annotation of Clinical Text Fragments with Behavior Codes

Clinical interview fragments. As discussed in [57], behavior code annotation can be treated as a classification problem which assigns a code to each utterance. We use a collection of motivational interviewing-based weight loss sessions, which consists of 11,353 utterances that were manually annotated by two human coders as a golden standard. On top of this dataset, we conduct three behavior code annotation tasks: A) Positive, Negative and Ambivalence; B) Commitment Language, Change Talk and Ambivalence; C) Positive Commitment Language, Negative Commitment Language, Positive Change Talk, Negative Change Talk and Ambivalence. The description of behavior code is listed in Table 2.1.

Classification of News Groups

20 Newsgroups. This publicly available[6] dataset[61] has been widely used to evaluate text classification algorithms. The 20 Newsgroups dataset is a collection of approximately

[6] `http://qwone.com/~jason/20Newsgroups/`

20,000 newsgroup documents across six categories, i.e., computers, recreation, science, politics, religion and forsale. In this work, we use four most common classes, which are computers, recreation, science and politics, as a four-class classification task to evaluate our framework.

Methods Compared

Baseline Approaches

The following non-CNN models are used as our baseline:

• **MTI.** Medical Text Indexer, which is commonly used in biomedical literature indexing. We only compare our method with MTI in the indexing task of biomedical articles.

• **Prior-best.** Prior-best is the best-performed method in the experiments of several classic machine learning methods, including Naive Bayes(NB), Logistic Regression(LR) and Support Vector Machine(SVM). For indexing of biomedical articles, Support Vector Machine with Huber Loss (SVM HL) [123] is also compared.

CNN-based Methods

In our experiments, we compared Topic-based Skip-gram with several baseline and state-of-the-art distributed word embedding learning methods, including:

• **CNN-rand.** Each word embedding is initialized with values drawn from continuous uniform distribution $U \sim [-0.25, 0.25]$. CNN-rand is used as a baseline of CNN-based methods.

• **CNN-gn.** These word vectors were trained by Mikolov et al. [78] on Google News and are publicly available[7]. • **CNN-glove.** The word embeddings used in this work were trained by Pennington et al. [87] [8].

• **CNN-local.** The word representations are trained by Skip-gram on the datasets to classify. The implementation of Skip-gram is publicly available[9].

[7] `https://code.google.com/archive/p/word2vec/`
[8] `http://nlp.stanford.edu/projects/glove/`
[9] `http://word2vec.googlecode.com/svn/trunk/`

• **CNN-topic.** These word embeddings are learned by our Topic-based Skip-gram on the datasets to categorize.

These kinds of word embeddings are compared under the baseline CNN architecture. Our two multimodal CNN architectures are also compared in this work:

• **CNN-channel.** We utilize two kinds of word embeddings for CNN-channel, CNN-local and CNN-topic.

• **CNN-concat.** CNN-local and CNN-topic are employed for CNN-concat.

We also tried to combine CNN-gn and CNN-glove with CNN-topic for multimodal CNN models, but their classification performance is not as good as combination of CNN-local and CNN-topic. The reason is that there are quite a few appeared words not in the CNN-gn and CNN-glove vocabulary, and embeddings for these words need to be randomly initialized. For example, more than 60% of words in the vocabulary of MEDLINE citations are not in the pre-trained CNN-glove vocabulary and need to be randomly initialized. This significantly and negatively impacts the performance of CNN-gn and CNN-glove.

Metrics

In this work, we use F_1 score to evaluate the performance of binary classifiers and macro-averaged F_1 score for multi-class classifiers.

F_1 score

F_1 score is a measure of binary classification accuracy, which is robust to unbalanced data distribution:

$$F_1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}, \tag{2.18}$$

where precision is ratio of instances which are classified as positive are correct and recall is the ratio of positive instances that are correctly classified.

Table 2.2: F_1 scores for check tags group.

MeSH Term	Positive	Prior best	CNN rand	CNN gn	CNN glove	CNN local	CNN topic	CNN channel	CNN concat
Adolescent	3824	0.4144	0.4321	0.4311	0.2677	0.4382	0.4104	0.4321	**0.4437**
Adult	8792	0.5700	0.6095	0.6192	0.5389	0.6159	0.6121	**0.6354**	0.6278
Aged	6151	0.5614	0.5695	0.5705	0.4378	0.5568	0.5645	**0.5841**	0.5737
Aged, 80 and over	2328	0.3227	0.321	0.3406	0.0642	0.3231	0.3316	0.3428	**0.3639**
Child, Preschool	1573	0.4954	0.4998	0.5126	0.4270	0.4944	0.4909	**0.5363**	0.5289
Female	16483	0.7517	0.7644	0.7761	0.7169	0.7761	0.7784	0.7810	**0.7840**
Humans	35967	0.9269	0.9307	0.9360	0.9113	0.9365	0.9351	**0.9366**	0.9361
Infant	1281	0.4441	0.4642	0.5032	0.1296	0.4923	0.4957	**0.5262**	0.5206
Male	15530	0.7294	0.7469	0.7477	0.6822	**0.7631**	0.7561	0.7543	0.7545
Middle Aged	8392	0.6377	0.6558	0.6665	0.6076	0.6692	0.6784	**0.6803**	0.6759
Swine	285	0.7071	0.7190	0.7332	0.6252	0.7406	0.7444	**0.7539**	0.7496
Young Adult	3807	0.3371	0.3125	0.3238	0.0499	0.3389	0.3128	0.3229	**0.3652**

Macro-averaged F_1 score

For multi-class classifiers, we employ macro-averaged F_1 score to evaluate their performance, which is an arithmetic average of F_1 score for each class:

$$\text{Macro-averaged } F_1 = \frac{1}{n}\sum_{i=1}^{n} F_1^i, \tag{2.19}$$

where n is total number of classes and F_1^i is F_1 score for ith class.

Experimental Results

In this section, we report the experimental results of baselines, state-of-the-art methods and our topic-based word embedding and multimodal CNN models. Best results are marked in bold.

Results of Indexing of Biomedical Articles

F_1 scores of each method over the check tags group, the low precision group and the low recall group are listed in Table 2.2, 2.3 and 2.4, respectively. The Positive column shows the number of positive samples for each MeSH. The results of MTI and Prior-best were reported in [93]. Although no single method outperforms all of the other approaches, the following observations can be made.

Table 2.3: F_1 scores for low precision MeSH group.

MeSH Term	Positive	MTI	Prior best	CNN rand	CNN gn	CNN glove	CNN local	CNN topic	CNN channel	CNN concat
Age Factors	889	0.0844	0.1450	0.2150	0.2212	0.0001	0.2142	0.2233	0.2206	**0.2429**
Brain	823	**0.5201**	0.4182	0.4300	0.4596	0.1902	0.4226	0.4571	0.4697	0.4821
Cell Line	781	0.2876	0.2265	0.2277	0.2139	0.0721	0.3009	0.2389	0.2704	**0.3212**
Cells, Cultured	1079	0.3046	0.2784	0.2457	0.2936	0.0841	0.2807	0.2723	**0.3350**	0.2739
Models, Molecular	851	0.4292	0.3734	0.3769	0.4283	0.2282	0.3893	0.4138	0.4209	**0.4307**
Molecular Sequence Data	1527	**0.5495**	0.4094	0.3863	0.4035	0.2141	0.4140	0.3532	0.4211	0.4024
RNA, Messenger	628	0.4477	0.4385	0.4421	0.4397	0.3110	0.3918	0.4374	**0.4576**	0.4486
Severity of Illness Index	751	0.1824	0.1924	0.1598	0.2106	0.0372	0.1588	0.2106	0.1927	**0.2237**
Time Factors	2153	0.098	0.1393	0.091	0.1188	0.0221	0.1123	0.1179	**0.1401**	0.1364
United States	2658	0.3585	0.3655	0.4128	0.4599	0.1081	0.4213	0.4292	**0.4791**	0.4653

Table 2.4: F_1 scores for low recall MeSH group.

MeSH Term	Positive	MTI	Prior best	CNN rand	CNN gn	CNN glove	CNN local	CNN topic	CNN channel	CNN concat
Child	2780	0.5863	0.5723	0.6015	0.6099	0.5488	0.6102	0.6040	0.6180	**0.6192**
Follow-Up Studies	1470	0.0407	0.2300	0.2189	0.2368	0.1187	0.2247	0.2284	**0.2514**	0.2264
Reproducibility of Results	1206	0.3191	0.3138	0.2963	0.3220	0.1921	0.3261	0.3110	0.3147	**0.3274**
Retrospective Studies	2183	0.6608	0.6580	0.6647	0.6578	0.6346	0.6585	0.6617	**0.6754**	0.6589
Risk Assessment	1014	**0.2556**	0.1610	0.2063	0.1854	0.1411	0.2145	0.1979	0.2100	0.2298
Risk Factors	2365	0.4989	0.3778	0.4438	0.4510	0.3446	0.4711	0.4514	0.4654	**0.5003**
Treatment Outcome	2999	**0.4202**	0.3859	0.3635	0.3590	0.2274	0.3752	0.3592	0.3831	0.3876

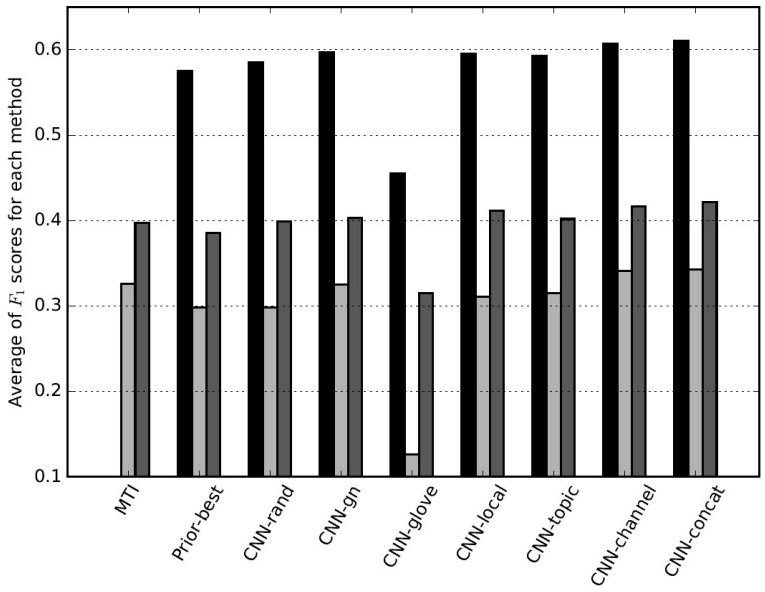


Figure 2.3: Macro-averaged F_1 scores of each method from the three groups.

First, CNN-channel and CNN-concat give the best performance for more than 82.7% selected MeSH terms. Only for four MeSH terms: Brain, Molecular Sequence Data, Risk Assessment and Treatment Outcome, MTI system demonstrates better results than the proposed multimodal CNNs.

Second, our multimodal CNN architectures outperform baseline CNN models with a single type of word embedding. This is mainly because multimodal CNNs utilize topic-based semantic word embedding as well as local content-based embedding. According to results shown in Table 2.2, 2.3 and 2.4, introducing topic-based semantic information improves indexing results.

Third, CNN-concat gives better results than CNN-channel for 15 terms among 29 terms and CNN-concat performs better than CNN-channel for more balanced MeSH terms. Considering there are 94,942 training samples in total, most MeSH terms are highly imbalanced. Among the 13 more balanced terms (Positive samples : Negative samples $> 0.025 : 1$), CNN-concat performs better than CNN-channel for eight MeSH terms and the average F_1 score of CNN-concat is 0.0063 higher than CNN-channel for the 13 terms.

Fourth, baseline CNN model with our proposed topic-based word embedding produces competitive results with CNN-gn and CNN-local. Word vectors used in CNN-gn and CNN-local are both trained with Skip-gram, which is the state-of-the-art word representation learning approach.

Fifth, CNN-glove demonstrates poor performance. The reason is that more than 60% of unique words in MEDLINE are not in the CNN-glove vocabulary and need to be randomly initialized. CNN-glove is pre-trained on Wikipedia2014 and Gigaword5 which do not contain many technical terms in biomedical domain. We can see that CNN-glove gives better performance on clinical text fragments and newsgroups datasets because more unique words are contained in the pre-trained vocabulary.

Sixth, the Prior-best columns refer to the best F_1 scores for traditional machine learning algorithms which give worse performance than CNN-based models. It indicates that CNN-based approaches are more effective for indexing problems than NB, LR, SVM and SVM HL.

Finally, we summarize average of F_1 scores for each method in all of the three MeSH term groups in Fig. 2.3. Although there is no model outperforming all of the other models, CNN-concat demonstrates the best overall performance and CNN-channel gives very competitive average F_1 scores. Further, word embedding learned by our proposed Topic-based Skip-gram produces state-of-the-art results with baseline CNN model.

Results of Behavior Code Annotation of Clinical Text Fragments

Three cases of multi-class behavior code annotation are conducted for this task: case 1, annotation over positive, negative and ambivalence, with sample ratio $1 : 0.014 : 0.150$; case 2, annotation over commitment language, change talk and ambivalence, with sample ratio $0.527 : 1 : 0.094$; case 3, annotation over positive commitment language, negative commitment language, positive change talk, negative change talk and ambivalence, with sample ratio $0.067 : 0.573 : 0.214 : 1 : 0.114$. Clearly, all of these three data splits are highly imbalanced. For each case, we conduct 5-fold cross validation and report average

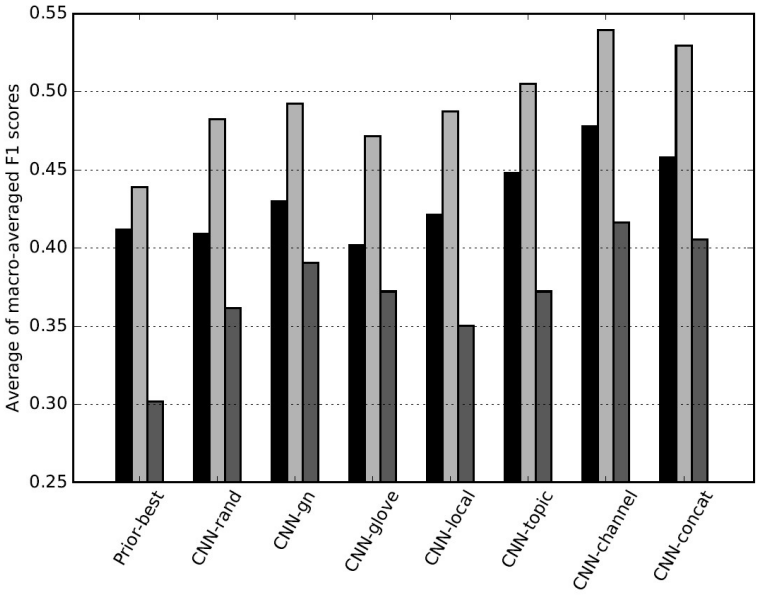


Figure 2.4: Macro-averaged F_1 scores for clinical text fragments.

macro-averaged F_1 scores for all methods over five folds. As shown in Fig. 2.4, CNN-channel gives the best F_1 scores among all the compared methods in the three cases and CNN-concat produces comparable results, which shows that CNN-channel performs better than CNN-concat for classification on highly imbalanced datasets. For word representation learned with baseline CNN models, CNN-topic, which is trained with proposed Topic-based Skip-gram, demonstrates better performance than other state-of-the-art word embeddings. Prior-best, which includes NB, LR and SVM in this task, is less effective than all CNN-based models.

Results of Classification of Newsgroups

This task is a 4-class classification problem over computers, recreation, science and politics. The sample ratio of the four categories is $1 : 0.876 : 0.811 : 0.668$, which is roughly balanced. 5-fold cross validation is applied to the whole dataset and the average macro-averaged F_1 scores over the five folds are reported in Fig. 2.5. First, CNN-channel and CNN-concat outperform other baselines and state-of-the-art methods. Second, CNN-concat demonstrates better performance than CNN-channel on this balanced dataset. Third, CNN-topic with baseline CNN model produces a comparable F_1 score with other state-of-the-

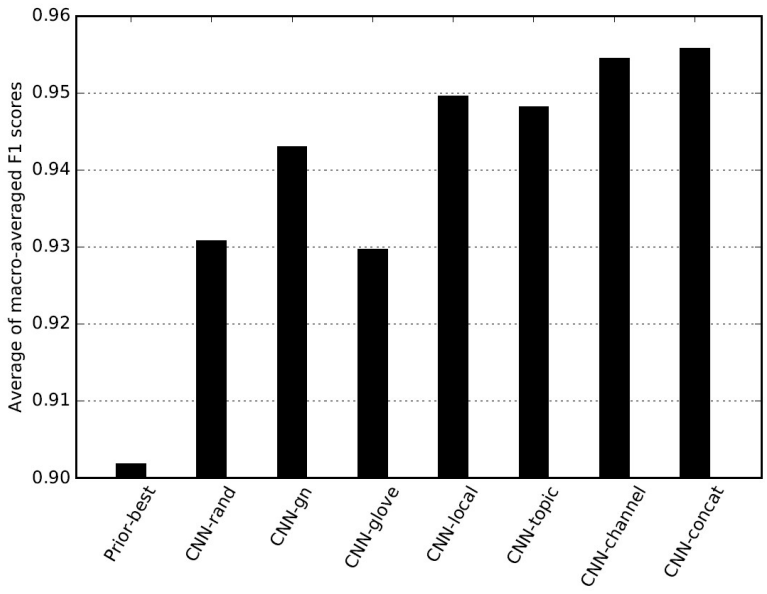


Figure 2.5: Macro-averaged F_1 scores for news groups.

art word embeddings. Furthermore, CNN-based models significantly outperform non-CNN models (NB, LR and SVM).

Conclusion

In this chapter, we proposed a novel framework,Topic-based Skip-gram, for learning topic-based semantic word embeddings for text classification with CNNs and achieved highly competitive results with word embeddings learned by Skip-gram. While Skip-gram focuses on context information from local word windows, the proposed Topic-based Skip-gram leverages semantic information from documents.

We also described two multimodal CNN architectures, CNN-channel and CNN-concat, which can ensemble different kinds of word embeddings. CNN-channel has a better imbalanced data resistance than CNN-concat, while CNN-concat has stronger learning ability and performs better on more balanced datasets.

Through experiments on indexing biomedical literature, annotation of clinical text fragments with behavior codes and text classification of a textual benchmark, we showed that our topic-based semantic word embeddings with multimodal CNNs outperform state-of-the-art word representations in text classification.

CHAPTER 3 Learning 3D Representation for Segmentation

Previous approaches on 3D shape segmentation mostly rely on heuristic processing and hand-tuned geometric descriptors. In this chapter, we propose a novel 3D shape representation learning approach, Directionally Convolutional Network (DCN), to solve the shape segmentation problem. DCN extends convolution operations from images to the surface mesh of 3D shapes. With DCN, we learn effective shape representations from raw geometric features, *i.e.*, face normals and distances, to achieve robust segmentation. More specifically, a two-stream segmentation framework is proposed: one stream is made up by the proposed DCN with the face normals as the input, and the other stream is implemented by a neural network with the face distance histogram as the input. The learned shape representations from the two streams are fused by an element-wise product. Finally, Conditional Random Field (CRF) is applied to optimize the segmentation. Through extensive experiments conducted on benchmark datasets, we demonstrate that our approach outperforms the current state-of-the-arts (both classic and deep learning-based) on a large variety of 3D shapes.

Introduction

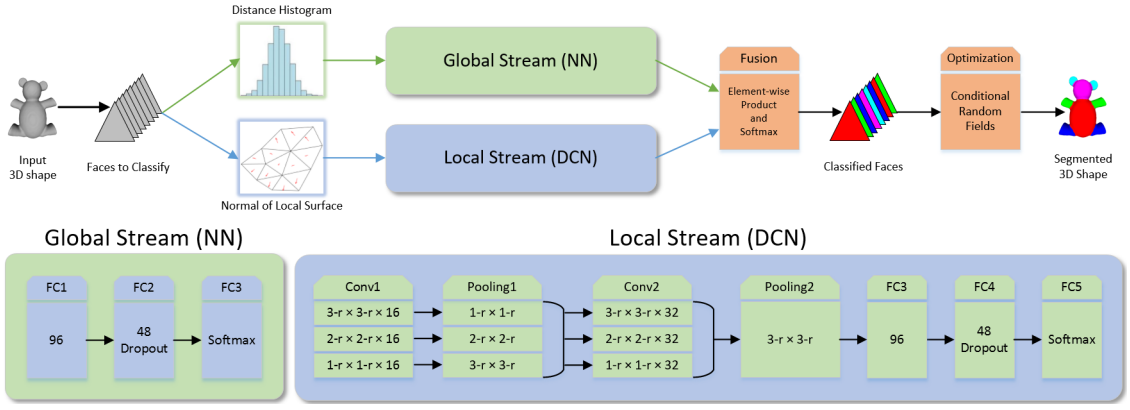


Figure 3.1: Workflow of our two-stream framework for 3D shape segmentation.

Segmentation over 3D shapes, also known as compositional part-based reasoning on 3D shapes, plays an important role in computer graphics and computer vision. It has been applied to various applications, such as 3D modeling [122], 3D object detection [66, 104], 3D scene understanding [52], and human pose estimation [99]. In the past few years, many

methods have been proposed to segment 3D shapes into semantic parts. Among these approaches, they either rely on heuristic processing and hand-engineering geometric features [10, 70], or apply co-segmentation schemes based on geometric characteristics of 3D shapes [101, 53]. More recently, (convolutional) neural networks have been applied to 3D shape segmentation [118, 38].

Inspired by the remarkable success of applying Convolutional Neural Network (CNN) in image recognition tasks, a few approaches have been proposed to extent convolution to graphs [19, 30, 29], most of which operate convolutions in the spectral domain - taking convolution as a linear operator in the Fourier space of a graph. However, [29] pointed out that a convolution filter defined in the spectral domain is not naturally localized and the translations are very costly. For approaches that define convolution in the spatial domain, they require relatively weak regularity assumptions on the graph and utilize the advantage of graphs, *i.e.*, having localized neighborhoods. However, the method in [19] only works for a given domain as eigenbases vary arbitrarily from shape to shape.

In this chapter, we propose Directionally Convolutional Network (DCN) that extends convolution operations from images to the surface mesh in the spatial domain. As a special case of graphs, polygon meshes inherit the advantage of being natural to define localized neighborhoods. Furthermore, we introduce a two-stream framework combining proposed DCN and a neural network (NN) for segmentation of 3D shapes. Instead of fusing the two streams by a simple concatenation, we take our framework as an approximation of a directed graph and combine the probabilities inferred by the two streams with an element-wise product. Finally, Conditional Random Field (CRF) is applied to optimize the surface mesh segmentation. The main contributions of our work are as follows:

• By defining rotation-invariant convolution and pooling operation on the surface of 3D shapes, we learn effective shape representations from raw geometric features, *i.e.*, face normals and distances, to achieve robust segmentation of 3D shapes.

• Based on the proposed DCN, we introduce a two-stream framework (shown in Fig. 3.1) to classify each face of a given mesh into predefined semantic parts. Our approach achieves state-of-the-art segmentation results on a large variety of 3D shapes.

In the rest of the chapter, we first review related work in Section 3.2. Then, we describe details of DCN in Section 3.3. In Section 3.4, we show our two-stream segmentation framework. In Section 3.5, we compare our approach with the current state-of-the-arts on benchmark datasets. At last, we conclude in Section 3.6.

Related Work

3D Shape Segmentation

Nowadays, 3D shape segmentation and labeling are widely used in computer graphics and computer vision fields. We can generally divide existing methods into the following categories.

Traditional feature-based approaches. Some early studies aim to manually design a single geometric descriptor that is effective in mesh segmentation [10, 70]. However, a single descriptor is insufficient to deal with various kinds of 3D shapes.

Co-segmentation approaches. In order to address the aforementioned limitation, data-driven approaches are utilized to extract common geometric features, including unsupervised co-segmentation methods [101, 53], and (semi-)supervised methods [51, 114]. These learning-based approaches generally outperform single geometric features. However, the simple combinations of geometric features are still not robust enough to describe complicated 3D shapes in many cases.

CNN-based approaches. Recently, neural networks have been popularly employed in 3D model analysis, due to their capabilities in extracting effective representations from low-level features. Xie et al. [118] proposed a shallow network to learn high-level features for segmentation, but this approach does not offer better performance than standard shallow classifiers [50]. Guo et al. [38] and Shu et al. [100] utilized CNNs to learn high-level

features from hand-engineering descriptors. These approaches simply concatenate hand-tuned features and lack geometric spatial coherence.

Kalogerakis et al. [50] proposed a view-based deep architecture for 3D shape segmentation and achieved state-of-the-art performance. However, their approach suffers from strong requirements on view selection, *i.e.*, minimizing occlusions, covering shape surface, and guaranteeing each part of shapes is visible in at least three views.

Comparing to the aforementioned methods, our approach only relies on the two most fundamental geometric descriptors for 3D shape representation learning: one preserves local precision and the other preserves global spatial consistency. Instead of combining the same kind of feature at different scales as in [80], we combine two different kinds of features. Furthermore, DCN defines convolution and pooling operations on 3D shape surfaces directly and thus has clearer geometric coherence than previous methods.

Convolutional Networks on Graphs

Inspired by the great success of CNNs in computer vision tasks, several approaches have been proposed to extend convolutional networks from images to arbitrarily structured graphs [19, 43, 74, 17, 29, 18].

Bruna et al. [19], Henaff et al. [43] and Defferrard et al. [29] proposed spectral networks which utilize spectral graph theory to define graph convolution as multiplication of a filter and graph node values in the Fourier space. The spatial network proposed in [19] is based on a hierarchical clustering of a graph. However, this approach does not have an efficient strategy of weight sharing across different locations of the graph [19]. By contrast, the proposed DCN operates convolution and pooling on the surface mesh of 3D shapes in the spatial domain and thus does not require strong regularity assumptions on the input shape structure [19]. Moreover, it is natural to define a face and its neighbors in the mesh as a cluster for convolution filters, which provides efficient weight sharing.

Some works also defined convolution on the surface mesh of 3D shapes for shape correspondence. Masci et al. [74] took geometry vector as input and had to compute the

convolution for all possible filter rotations due to angular coordinate ambiguity. Boscaini et al. [17] took local SHOT descriptor as input. In contrast to these earlier works, our method learns the shape representation (layer by layer and coarse to fine) from the most fundamental low-level geometric features. Furthermore, the proposed DCN is rotation-invariant and does not have the angular coordinate ambiguity problem in [74].

Directional Convolution and Pooling

In this section, we present the details of directional convolution and pooling methods defined on surface meshes of 3D shapes, and how to generalize to cloud points.

Mesh Face Normal and Curvature

In geometry, curvatures can effectively represent the local shape variations. The local directions of minimum and maximum curvatures indicate the slowest and steepest variation of the surface normal, respectively. In this subsection, we define the fundamental low-level geometric features on each face based on surface normal and curvature.

Face normal: For each mesh face f_i, let $\{\mathbf{v}_{f_{i_1}}, \mathbf{v}_{f_{i_2}}, \mathbf{v}_{f_{i_3}}\}$ denote its vertices. The face f_i's normal can be represented using the cross product of two edge vectors as follows:

$$\mathbf{n}_{f_i} = (\mathbf{v}_{f_{i_2}} - \mathbf{v}_{f_{i_1}}) \times (\mathbf{v}_{f_{i_3}} - \mathbf{v}_{f_{i_1}}). \tag{3.1}$$

As mentioned in [95], we can estimate the local shape variations and properties over a local region by using the differences in face normals, which is similar to estimate the curvatures of a face.

Face curvature: The mesh vertex curvature magnitudes and directions are computed based on [2]. We can use the average value of three vertex's curvatures (including magnitude and direction) to approximate the minimum and maximum curvatures on mesh face f_i as follows:

$$k_i = \frac{(k_{i_1} + k_{i_2} + k_{i_3})}{3} \text{ and } \mathbf{d}_i = \frac{(\mathbf{d}_{i_1} + \mathbf{d}_{i_2} + \mathbf{d}_{i_3})}{3}, \tag{3.2}$$

where magnitude k_i and direction $\mathbf{d}_i$ can be used to represent the minimum and maximum curvatures on face f_i, respectively.

Directional N-ring Face Neighbors

In order to define a convolution on the surface mesh, we need to define a robust rotation-invariant face neighboring mechanism at first. There are two aspects needed to be defined in the concept of *directional n-ring face neighbors*: (1) The set of n-ring face neighbors: the nth ring of a face i is the set of faces that are at distance $n-1$ from f_i in the given mesh, where the distance n is the minimum number of edges between two faces. (2) The order of n-ring face neighbors: the order of neighbors is important in convolutions since filter weights are adaptive according to their significance.

As mentioned above, local directions of curvatures indicate the local shape variation. No matter how the model rotates, the local shape geometry is invariant. So, we choose the curvature direction as a guidance to define the order of face neighbors. For face f_i, we traverse the neighbors ring by ring based on the direction of maximum curvature counter-clockwise, and the first face neighbor for each ring is the one having the minimal angle difference between two vectors, *i.e.*, the maximum curvature direction and the vector $\mathbf{c}_{ij}$ defined by the centroids of faces f_i and f_j. The angle between two vectors can be computed by using the geometric definition of dot product, *i.e.*, $\theta_{ij} = \cos^{-1}\left(\frac{\mathbf{d}_i^{max} \cdot \mathbf{c}_{ij}}{\|\mathbf{d}_i^{max}\|\|\mathbf{c}_{ij}\|}\right)$. Fig. 3.2 illustrates the first nth rings of neighbors of face i under the defined order on a 3D hand mesh model.

Directional Convolution on Mesh

Once directional n-ring face neighbors are defined, we can present the definition of *directional convolution* of the feature ϕ with a kernel w on mesh face f_i as follows:

$$(\phi * w)(i) = \frac{1}{K} \sum_{j \in N_n(i)} w(j)\phi(j), \tag{3.3}$$

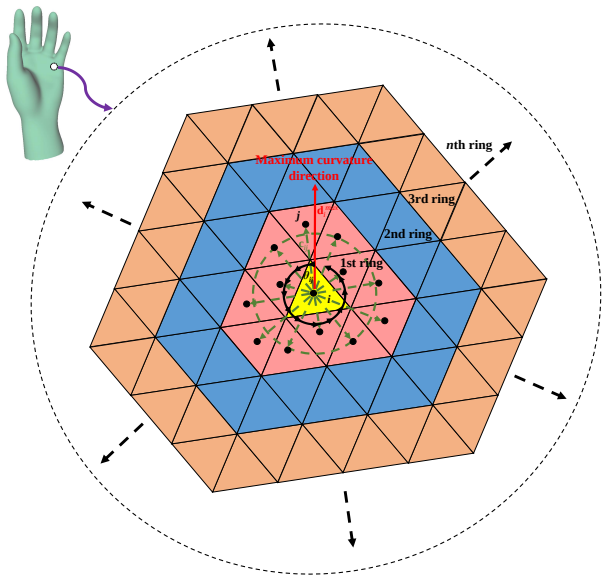


Figure 3.2: The illustration of the first nth rings of neighbors.

where ϕ can be a scale or vector function based on the mesh face features, such as normal, curvature, shape diameter, etc. In this work, we only use the face normal vectors as the feature. Face normals are computed in Eq. (3.1). The kernel w weighs the participation of neighbouring faces f_j, which will be learned during the optimization of DCN. K is the normalization factor, *i.e.*, $K = \sum_{j \in N_n(i)} w(j)$. $N_n(i)$ is the set of neighbors of face f_i. n is the index of ring for face neighbors. The order of neighbors is computed as in Section .

We define the filter size as n-r$\times n$-r, which means that a face and its first n rings of neighbors are convolved by the filter. If $n = 0$, then only one face is convolved by the convolution filter. Since neighboring face number in n-ring of different faces varies, we choose the average neighbor number as filter size for n-ring, and pad zeros for faces without enough neighbors (or omit redundant neighbors).

Pooling on Mesh

Classic pooling layers in CNN make use of the natural multi-scale clustering of grid: they input all the feature maps over a cluster, and output a single feature for that cluster [19]. On surface mesh, we define a cluster as a face and its 1- to n-ring neighbors. Thus, given such a cluster, the pooling is manipulated by a downsampling strategy of *a cluster of*

faces to 1 and denoted as n-r×n-r pooling. For max pooling, the maximum value of feature maps in the cluster is taken as output.

Generalization to Cloud Points

Although meshes and point clouds are two different representations of 3D objects, we can easily modify the proposed method to segment 3D point clouds. Specifically, we can first use principal component analysis to compute the local point normal and curvatures[10]. Then, we define the point neighbors by finding the k nearest points. Finally, we can employ the proposed directional convolution on point clouds, same as on meshes.

3D Segmentation with DCN

In this section, we describe our shape segmentation approach in detail (see Fig. 3.1). First, we compute normals and distances for faces in a given 3D shape. Second, we feed these raw features to the proposed two-stream framework with DCN and NN, and then fuse the two streams by an element-wise product and softmax. Finally, the segmentation is optimized by CRF.

Input Features

In our approach, we aim to learn an effective 3D shape representation, robust for a large variety of shapes. Two types of input geometric features, face normals and distance histogram, are utilized in order to ensure local precision and global spatial consistency, respectively.

Face normal as local features

Normal is one of the most fundamental geometric features to describe the shape of a surface mesh. We select face normals to ensure the local precision of the segmentation. To capture the local shape information of a surface at a higher level of details, we extract a patch of the target face and its first n rings of neighbors. Generally, taking a very small patch as the input is insufficient to accurately describe the local geometry. A larger patch will help but typically leads to inefficiency in computing. Empirically, we choose $n = 6$.

[10] `http://pointclouds.org/documentation/tutorials/normal_estimation.php`

The normals of faces in the patch are used as the local input features of the surface patch centered on the target face.

Distance histogram as global features

For segmentation over the same category of 3D shapes, semantic parts consistently preserve the same relative positions in all the models. Thus, including global information is likely to yield improvements. Although simply increasing the size of local patches would cover larger part of a 3D shape, it is computationally inefficient. Another strategy is to use the coordinates of face centroids, but this scheme is not shift-, scale-, or rotation-invariant. In this work, we use normalized histograms of the pairwise face distances to ensure the global spatial consistency.

To define pairwise distances, we first denote an input 3D shape dataset of M models as $D = \{S_1, S_2, \ldots, S_M\}$. For a 3D shape S_m with N_m faces to segment, we denote each face as $f_i^{S_m}$, where $m \in [1, M]$ and $i \in [1, N_m]$. We build a dual graph $S_m^{'}$ with N_m vertices, in which each vertex corresponds to a face of S_m and two vertices are connected by an edge if and only if the two corresponding faces share at least one vertex in S_m. The pairwise distance between two faces $f_i^{S_m}$ and $f_j^{S_m}$ is denoted as $d_{i,j}$, which is the shortest distance between corresponding vertices $\mathbf{v}_i^{S_m^{'}}$ and $\mathbf{v}_j^{S_m^{'}}$ in the dual graph $S_m^{'}$. Since our input is "water-tight" polygon meshes, every two vertices in the same 3D shape are connected by one or more edges. Thus, the existence of the pairwise distance between every two faces in the same shape is guaranteed.

In this way, we can get $N_m - 1$ pairwise distances for face $f_i^{S_m}$. Then, a histogram is computed based on these distances. Empirically, we choose a 50-bin histogram. Finally, we perform L_2 normalization on the 50 elements of each histogram, making the distance histogram insensitive to the total number of faces in a 3D shape. Unlike coordinates, normalized distance histograms are robust to scaling and invariant to shifting and rotation.

Two-stream Framework with DCN and NN

In the proposed two-stream segmentation framework, the local stream is a DCN with the face normal as the input, and the global stream is a neural network with the distance histogram as the input. Then we fuse the two streams with an element-wise product and softmax.

Local stream DCN. The architecture of the proposed DCN is design based on the idea of multi-scale and multi-level feature ensembling. Limited by the size of experimental datasets, the network consists of only two convolution and pooling layers and three fully connected layers.

Specifically, an input patch to DCN contains a center face and its 1- to 6-ring neighbors. To get multi-scale features, we first employ three sizes of convolution filters, *i.e.*, 3-r×3-r, 2-r×2-r, and 1-r×1-r, of stride 1 in layer Conv1. A sliding max pooling of stride 1 [67] is separately applied for the three feature maps in layer Pooling1. Since neighbors that are closer to the target face carry more information and less noise, we only keep the center and first 3 neighboring rings of three feature maps and concatenate them together. In Conv2 and Pooling2, we employ three sizes of filters of stride 1 and average pooling to flatten the feature map. Finally, three fully connected layers with dropout and softmax are used and output classification probabilities P_{local}. P_{local} is a real-valued vector of size C, where C is the number of predefined classes.

Global stream NN. The global stream is implemented by a three-layer neural network with dropout and softmax, which takes the distance histogram as input. We denote the output softmax scores of the global stream as P_{global}, which represents the probabilities of an input face belonging to segmentation classes. Similar to P_{local}, P_{global} is the vector of probabilities inferred based on distance histograms.

Fusion in a graphical-model style. Assuming that the local feature and global feature of the same face are independent, we take our two-stream segmentation framework as a directed graphical model. That is, the probability of segmentation classes P_{seg} can be

computed by an Hadamard product:

$$P_{seg} = \sigma(P_{local} \circ P_{global}), \tag{3.4}$$

where $\sigma(\cdot)$ denotes the softmax function.

Mesh Label Optimization with CRF

Given P_{seg} for each triangle in a test 3D shape S_m with N_m faces, we employ CRF [59] to refine the labels by incorporating the constraint on label consistency. The energy function is

$$E(\mathbf{x}) = \sum_i \xi_i(x_i) + \lambda \sum_{i,j} \xi_{ij}(x_i, x_j), \tag{3.5}$$

where x is the label assignment for faces and λ is a non-negative constant. In particular, we set $\lambda = 50$ as in [38]. Here, we define the unary item $\xi_i(x_i)$ as

$$\xi_i(x_i) = -\log P(x_i), \tag{3.6}$$

where $P(x_i)$ is the ith element in P_{seg} of face f_i. Thus, assigning f_i to a class with low probability will result in a high penalty.

We define the pairwise item $\xi_{ij}(x_i, x_j)$ as

$$\xi_{ij}(x_i, x_j) = \begin{cases} 0 & \text{if } i = j \\ -\log(\theta_{ij}/\pi)d_{ij} & \text{otherwise} \end{cases}, \tag{3.7}$$

where θ_{ij} and d_{ij} are the dihedral angle and distance between triangle face f_i and f_j, respectively. With this pairwise item, we penalize the smoothness between the labels of adjutant face pairs.

Experimental Results

Datasets and Experimental Setups

In this section, we compare our approach with current state-of-the-arts on the segmentation of a large variety of shapes. The following benchmark datasets are employed to evaluate our approach: Princeton Segmentation Benchmark (PSB) [22] (19 categories, 20 meshes per category) and four categories from the Shape COSEG Dataset [113], including Iron (18 meshes), Lamp (20 meshes), Candelabra (28 meshes), and Guitar (44 meshes). We also perform our experiments on two large categories, Vases (300 meshes) and Chairs (400 meshes).

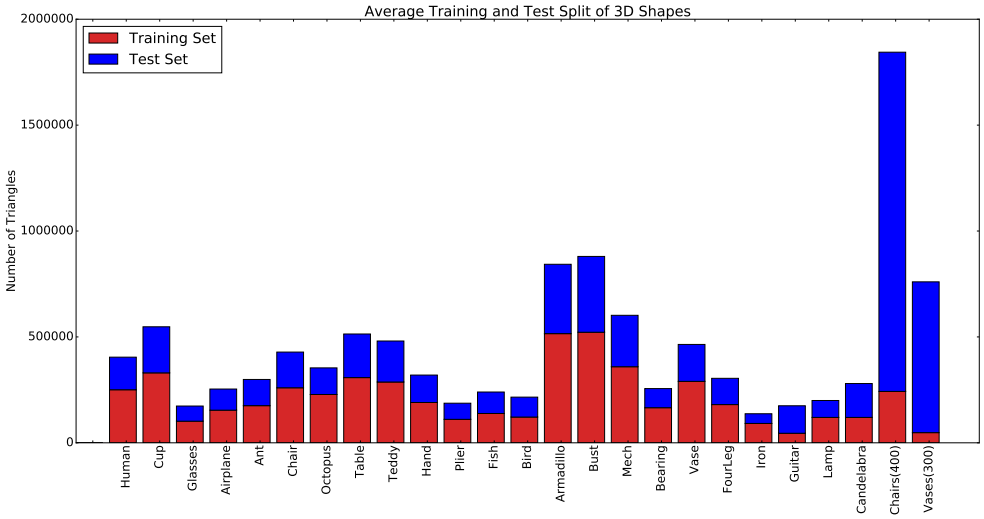


Figure 3.3: The triangle face numbers of training and testing split.

We followed the experimental setup of [38]. For small categories from PSB and the Shape COSEG, we take 12 meshes for training and the remaining for testing. For the two large categories, we take 20 meshes as the training set for Vases and 50 meshes for Chairs. For each category, we repeat our approach three times and report the average accuracy and standard deviation. The average number of faces used in the training and testing dataset is shown in Fig. 3.3.

In our experiments, we compared our approach with several state-of-the-arts (both classic and deep learning-based), including Sidi et al. [101], Kalogerakis et al. [51], Wang et al. [114], Guo et al. [38], and shapePFCN [50]. Besides, we also combined the two streams at penultimate layers as an alternative approach (named *Ours-early* in Tables).

Our network is implemented using Python and Theano [108]. Adam [55] with learning rate 10^{-4} is applied for optimization. Our deep learning framework runs on a GTX 980Ti GPU, and it takes about 20 minutes to train a model with 20K-30K faces, excluding the pre-processing time of computing the face normals and distance histograms.

Directional vs. Non-directional Convolutions

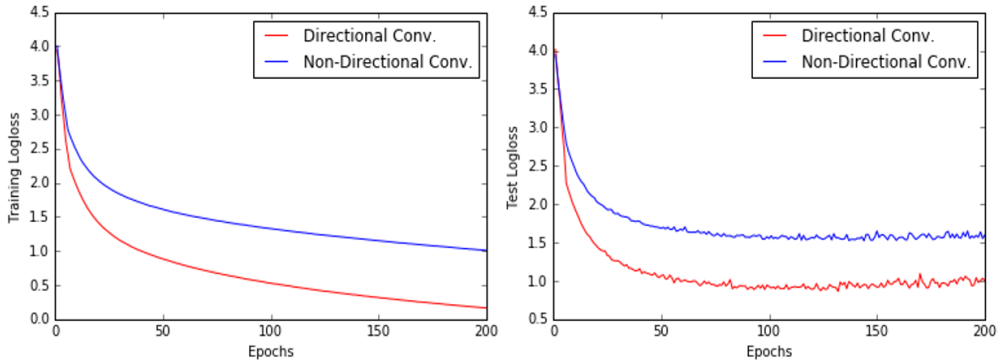


Figure 3.4: Logloss of directional (red) and non-directional convolution (blue).

First, we show that directional convolution is necessary for shape representation learning. In Fig. 3.4, we compare the training and test loss of directional convolutions with non-directional convolutions on one exemplar category: Human. For the latter one, the input triangle faces are randomly shuffled. Clearly, the network with the directional convolution converges faster and to a lower error in both the training (12 meshes) and the testing (8 meshes) datasets.

Segmentation Accuracy

We compare the segmentation accuracy between our approaches and the current state-of-the-arts. Following [38], the accuracy is computed as the percentage of area of correctly

Table 3.5: Mesh segmentation accuracy on 23 categories.

Category	Kalogerakis[51]	Wang[114]	Guo[38]	ShapePFCN[50]	Ours
Human	0.9320	0.5560	0.9122	0.9380	**0.9408**(±0.0088)
Cup	0.9960	0.9960	0.9973	0.9370	**0.9979**(±0.0003)
Glasses	0.9720	-	0.9760	0.9630	**0.9869**(±0.0020)
Airplane	0.9610	-	0.9667	0.9250	**0.9766**(±0.0078)
Ant	0.9880	-	0.9880	0.9890	**0.9898**(±0.0008)
Chair	0.9840	**0.9960**	0.9867	0.9810	0.9935(±0.0051)
Octopus	0.9840	-	0.9879	0.9810	**0.9934**(±0.0007)
Table	0.9930	**0.9960**	0.9955	0.9930	0.9959(±0.0003)
Teddy	0.9810	-	0.9824	0.9650	**0.9908**(±0.0022)
Hand	0.8870	-	**0.8871**	0.8870	0.8861(±0.0028)
Plier	0.9620	-	0.9622	0.9570	**0.9714**(±0.0054)
Fish	0.9560	-	0.9564	0.9590	**0.9705**(±0.0016)
Bird	0.8790	-	0.8835	0.8630	**0.9039**(±0.0096)
Armadillo	0.9010	-	0.9227	0.9330	**0.9382**(±0.0012)
Bust	0.6210	-	0.6984	0.6640	**0.7898**(±0.0266)
Mech	0.9050	0.9130	0.9560	**0.9790**	0.9660(±0.0012)
Bearing	0.8660	-	0.9246	0.9120	**0.9470**(±0.0036)
Vase	0.8580	**0.9050**	0.8911	0.8570	0.8931(±0.0089)
FourLeg	0.8620	0.5430	0.8702	**0.8950**	0.8742(±0.0083)
Iron	-	-	**0.9737**	0.8770	0.9714(±0.0022)
Guitar	-	-	0.9715	0.9790	**0.9932**(±0.0037)
Lamp	-	-	0.9628	0.9090	**0.9789**(±0.0007)
Candelabra	-	-	0.9447	**0.9630**	0.9546(±0.0048)
Average	0.9204	0.8436	0.9409	0.9263	**0.9522**

Table 3.6: Mesh segmentation accuracy on large datasets.

Category	Sidi[101]	Kim[53]	Guo[38]	Ours
Chairs(400)	0.8020	0.9120	0.9252	**0.9573**(±0.0013)
Vases(300)	0.6990	0.8560	0.8854	**0.9086**(±0.0060)

classified faces over area of all the surface. The performance of existing methods is based on publicly reported results in the literature. Best results are marked in bold.

The segmentation performance on the small datasets is shown in Table 3.5. Our early fusion approach (Ours-early) performs 0.0026% weaker than late fusion approach (Ours). On average, our framework gains an improvement of 1.13% against the best-performing previous approach. When it comes to each category, our approaches outperform all existing methods on 15 out of 23 objects. In the remaining 8 categories, the accuracies of our method are only a bit less than the prior-best. For the two large datasets, *i.e.*, Chairs and Vases, their segmentation accuracies are listed in Table 3.6. We can see that our approaches outperform previous ones significantly, which benefits from the high learning capacity of our two-stream framework with DCN and NN.

Visualization of DCN Kernels and Feature Maps

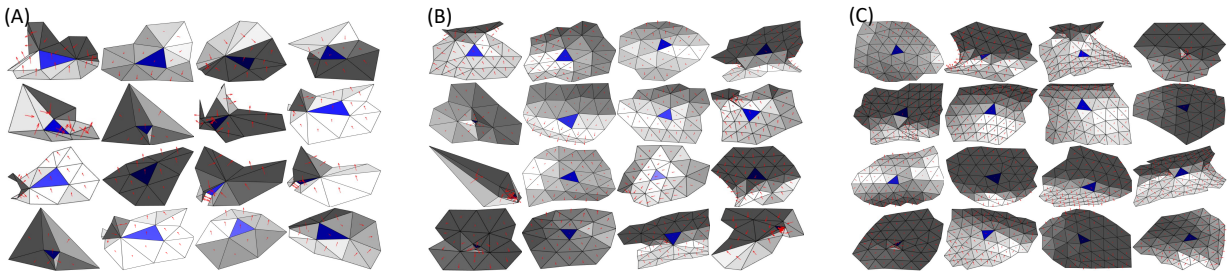


Figure 3.5: Strongest responses of convolution filters in Conv1 of DCN.

To visualize kernels learned in DCN, we convolve the trained filters in layer Conv1 over the surfaces of all 3D shapes in the corresponding category and find out the patches with strongest response for each filter (see Fig. 3.5). Clearly, 1) filters of smaller size tend to focus on the fine details of a shape, *e.g.*, steep surface changes; 2) filters of larger size tend to focus on the main trend of the shape, the matched patches are more smooth.

Moreover, to better understand what the proposed two-stream segmentation framework learns from face normals and distance histograms, we randomly select $10,000$ patches from each 3D shape category and feed them into the trained network. Then, we extract the 48-dimensional feature maps of layer FC4 in the local stream and FC2 in the global stream.

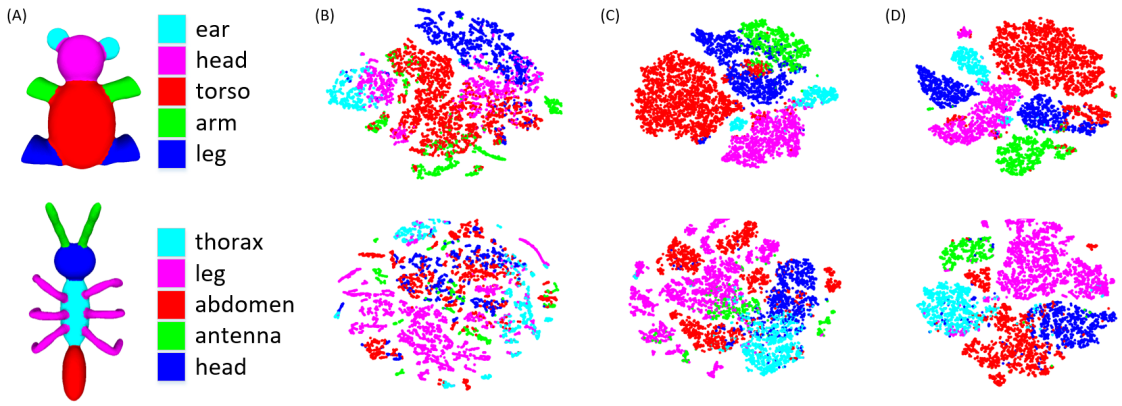


Figure 3.6: t-SNE visualization of global and local representations.

We also apply element-wise multiplication on extracted local and global feature maps and use the product as final feature maps. The three types of feature maps for category Teddy and Ant are visualized by t-SNE [72], which embeds high-dimensional feature maps in a 2D space while preserving the pairwise distance of the instances. As shown in Fig. 3.6, the clusters from different parts overlap with each other with global features, indicating a high similarity. That is mainly due to the symmetry of Teddy and Ant. By contrast, the clusters with local features are better separated. Combining global and local features together, the t-SNE clusters are best separated in both categories, which clearly indicates that our framework learned effective shape representations for segmentation.

Segmentation Examples

Fig. 3.7 demonstrates some exemplar segmentation results of the proposed framework. In this figure, our approach performs well on shapes with rotation and different poses. However, the edges between different parts are not smooth, which means that some faces locating at the edges are challengingly segmented. Apparently, if two neighboring faces with similar face normal and distance histogram are located at the edge of two separate parts, our approach has difficulty classifying them correctly.

Another way to provide an intuitive understanding of the segmentation results is to visualize probability maps of layer FC3 in the global stream and of layer FC5 in the local stream. For a better comparison, we also include segmentations from the two-stream

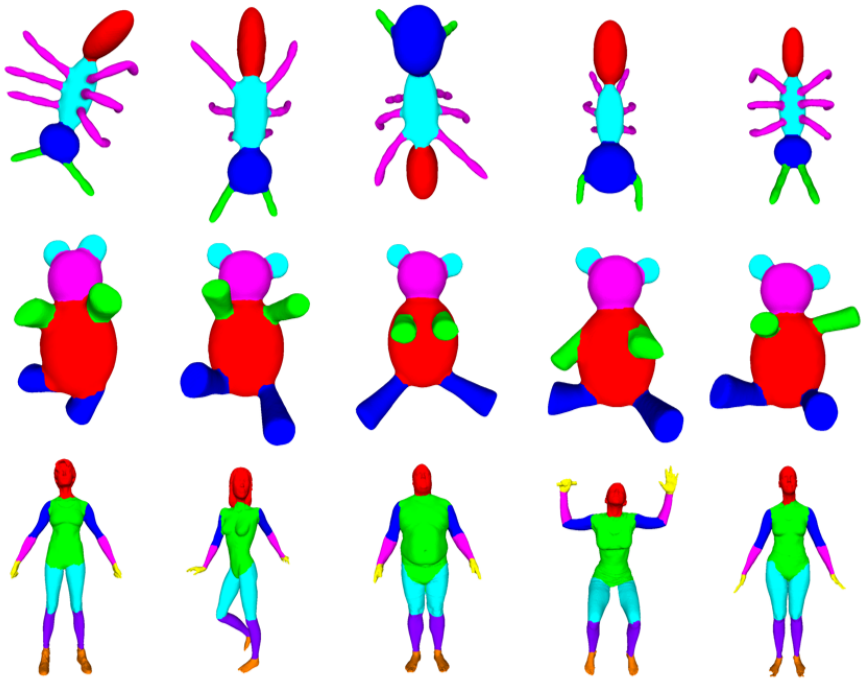


Figure 3.7: Visualization of segmentation on category Ant, Teddy, and Human.

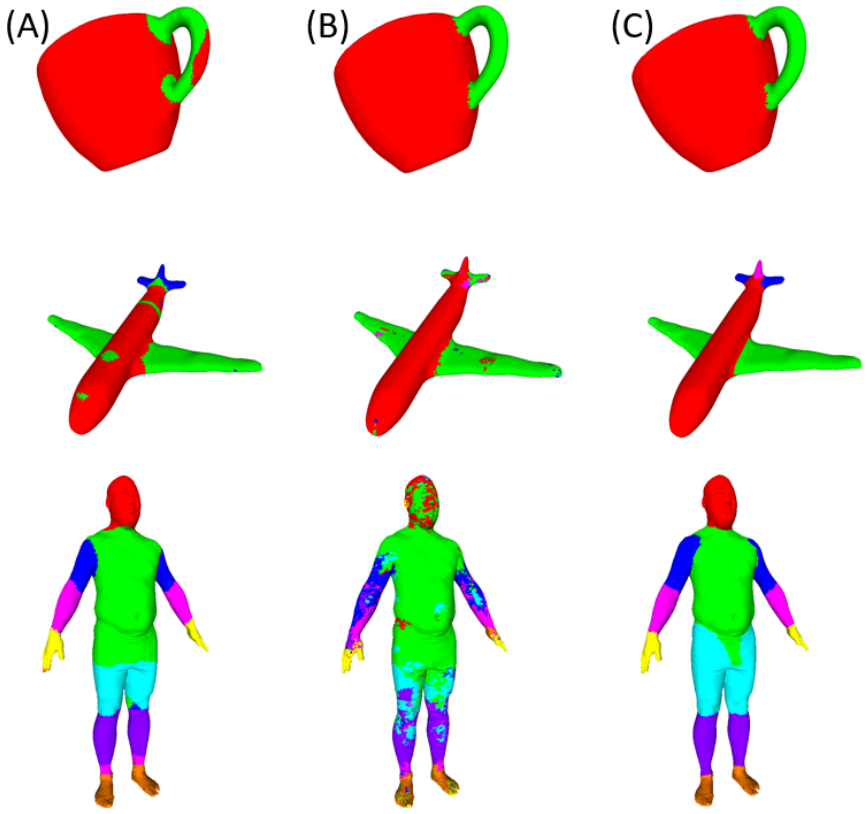


Figure 3.8: Visualization of segmentation results inferred by different streams.

framework in the figure (last column in Fig. 3.8). For the global stream, the segmentation suffers from the symmetry of the shapes. Taking shape Human as an example, faces near ankle are classified into lower arm. That is because they are all near the end of body and are not differentiable based on distance histograms. For the local stream DCN, it performs well when face normals of patches belonging to different classes are quite distinct (*e.g.*, category Cup). However, when the surfaces of different classes are similar or change smoothly, the performance is weak (*e.g.*, category Human). That is because it is difficult to tell lower arm from upper arm with a small patch of surface. In this case, the global stream plays a more important role. By fusing the two streams, we make the best use of local and global input features for segmentation.

Conclusion

In this chapter, we proposed a novel 3D shape representation learning approach, Directionally Convolutional Network (DCN). Based on a two-stream segmentation framework, we learn effective shape representations from raw geometric features, *i.e.*, face normals and distances, for robust segmentation. Our method achieved the state-of-the-art results on a large variety of 3D shapes in benchmark datasets.

Limited by the size of datasets, our approach is based on patches, which is time-consuming. In the future, we plan to integrate Fully Convolutional Networks and CRF to build an end-to-end learning framework and apply our method on larger 3D shape datasets.

CHAPTER 4 Learning Fiber Representation for Classification

Introduction

The principle of presurgical evaluation for epilepsy is to determine the spatial relationship between the epileptogenic zone and functionally important cortex, such as somatorsensory, language, auditory, and visual areas ("also known as eloquent cortex") [117]. Without accurate localization of such brain regions, one cannot achieve the ultimate goal of epilepsy surgery, which is to eliminate epileptic seizures without creating a new functional deficit. The current gold standard method employed to identify eloquent cortex often requires invasive direct electrical stimulation mapping (ESM) [64]. However, ESM is not an ideal method, since it requires implantation of intracranial electrodes, carries the inherent risk of electrically-induced seizures, and sometimes fails to identify eloquent cortex, especially in children. For instance, our previous study [40] reported that a contralateral hand movement was not elicited by electrical stimulation in 15 of 65 children. Children of average age 3.4 years belonged to this "no motor response group", suggesting that young age is a risk for failure to identify the primary motor hand areas using ESM. Also, of the 50 children with a contralateral hand movement elicited by electrical stimulation, 24 had the motor hand area in the post-central gyrus and 17 children showed the hand area in both pre- and post-central gyri, indicating that a substantial proportion of young patients with focal epilepsy had a prominent variation in the hand motor area ranging from the precentral gyrus to the postcentral gyrus. Such variations could be more prominent in lesional cases.

An alternative approach to ESM is functional MRI (fMRI) [76, 27], which is non-invasive but highly susceptible to movement artifacts and demands cooperative behaviors during scanning. Thus, it is challenging to perform fMRI studies in young patients with epilepsy (success rate $< 60\%$ at age 4-6 years [125]). Furthermore, the epileptogenic zone frequently involves the bottom of a deep sulcus [27, 13], which is in close proximity to adjacent axonal pathways. Both ESM via intracranial grid electrodes and fMRI are inherently unable to localize crucial subcortical white matter structures, which may therefore be at risk for

unwanted damage during surgery. Thus, there is an urgent need in presurgical planning to accurately identify eloquent regions of interest including both cortical areas and white matter pathways to prevent postoperative deficits in young children with intractable epilepsy.

The present study proposes a critical translational step toward the clinical application of a novel diffusion tensor imaging (DTI) tractography method, which may serve as an efficient noninvasive localizing tool supplementing (and ultimately replacing) fMRI and ESM in children with intractable epilepsy. In the last decade, DTI has been a powerful technique to visualize the whole brain white matter tracts with minimal patient cooperation [79, 8]. The central hypothesis of our study is based on the notion that, even though both partial volume and crossing fiber problems limit the feasibility of DTI by generating excessive false fibers near the cortical mantle [1, 48], DTI tractography can still provide clinically acceptable accuracy to map cortical endings of eloquent white matter tracts within a 1-cm error (i.e., spatial resolution of ESM). Thus, we investigated if DTI can serve as an alternative localizing tool to prevent the occurrence of postoperative deficits in various eloquent functions. Indeed, our previous studies [46, 47] successfully demonstrated the reliability of our central hypothesis by showing that a maximum a posterior probability (MAP) classification using a priori information of primary motor and language tract atlas could detect separate functional pathways associated with primary motor and language areas, compared with their gold standard ESM successfully performed on children with intractable epilepsy.

A major advantage of our previous DTI-MAP method is the simultaneous localization of functionally-important white and gray matter structures without using any supplementary acquisitions like fMRI and ESM. This method does not require patient cooperation, and can be ultimately extended to localize other important pathways of infants and young patients in whom functional localization cannot be done using either fMRI or ESM (about 30% of surgical cases). However, this approach was originally designed to classify a given streamline into one of the limited numbers of target classes (*i.e.*, six primary motor pathways including face, finger and leg fibers in both hemispheres and six language pathways in left hemisphere),

performing by comparing a maximum posteriori probability of individual fiber streamline where fMRI-derived white matter probability maps and equal class prior assumption were applied under the Bayesian classification framework [46, 47]. Thus, it is an ad-hoc approach typically limited by an inevitable type II error falsely inferring true positive outliers.

As one of the most powerful deep learning frameworks, CNNs have been widely used for biomedical imaging tasks with unknown priori distribution [98, 37]. In the present work, an offline, retrospective IRB-approved study was conducted to investigate the functionally-important white matter pathway detection capability of CNNs on 64 pathways of DTI streamline fibers, including somatosensory, language, auditory, and visual functions that should be preserved after epilepsy surgery. Comparing to existing approaches, our key insight is that rather than first building a tract atlas based on priori information and then feeding the input into a statistical model, we can instead utilize CNNs to provide an end-to-end learning framework which integrates white matter pathway classification with direct representation learning without any priori distribution information [62]. From a computing perspective, the novelty of present work is as follows:

• Two CNN architectures with different depth were investigated in this study: a shallow CNN model with 3 layers from our previous work [119]; inspired by the great success of very deep CNNs [41, 92], we also adapted the shallow CNN into a deep model with 21 layers. The proposed CNN models generate different feature maps of the input data (*i.e.*, 3D spatial coordinates of individual fiber streamlines) by using a sequence of convolutional and pooling layers before classifying input data using fully connected layers.

• A couple of novel CNN loss functions [68, 116] were introduced for pathway classification task. First, since our dataset is highly unbalanced, which cannot be handled well by CNN with the conventional cross-entropy loss, we introduced focal loss into the proposed CNN models. Focal loss applies a modulating term to the cross-entropy loss to help focus on hard examples and down-weight the numerous easy examples. Second, to further improve the classification performance and generalization ability of proposed CNN models, the learned

fiber representations need to be not only separable but also discriminative. Center loss was introduced which adds a cluster-based loss term to the cross-entropy loss to ensure the learned representations have both compact intra-class variations and large inter-class margins.

• Although CNNs have led to breakthroughs of state-of-the-arts, the end-to-end learning strategy makes the entire CNN model a black box. This weakness is highlighted in the biomedical imaging: if we do not know how the trained CNNs classify each fiber, we cannot fully trust the classification results given by the CNN models. In this study, we applied attention mechanism [120] in the proposed CNNs, which highlights the most useful segments of a fiber for classification. In this study, we will demonstrate that the attention mechanism provides a machine perspective on how the CNNs classify functionally-important white matter pathways.

To the best of our knowledge, we are the first to demonstrate that CNNs can benefit spatial localization of DTI tractography and further provide the spatially more complete functional brain map to minimize the risk of functional deficits following epilepsy surgery. Via intensive in-vivo comparisons with current gold standard ESM, this study demonstrates that CNN has high translational value in that the concepts derived might lead to more accurate localization for presurgical planning of epilepsy surgery by minimizing a risk to resect functionally important brain tissue during surgery.

The rest of this chapter is as follows: Section describes the detailed structures of shallow and deep CNN model with focal loss and center loss including a soft attention mechanism to provide a new perspective to interpret how fibers are classified in the framework of CNNs. Section describes the setup and results of the proposed CNN-based fiber classification in vivo experiments. At last, Section presents discussion, limitation, future application of our CNN methods.

Methodology

Subjects

To construct training and test dataset of the proposed CNN-based fiber classification, 70 healthy children (age: 12.01 ± 4.80, 36 boys) were recruited in the present study. Also, 70 children with drug-resistant epilepsy who underwent investigations for epilepsy surgery between 2009 and 2016 were retrospectively selected for the validation dataset (age: 11.60 ± 4.80 years, 36 males). Inclusion criteria were 1) drug-resistant epilepsy requiring two-stage epilepsy surgery with chronic subdural ESM mapping at the ChildrenâĂŹs Hospital of Michigan or Harper University Hospital, 2) no motor and/or language impairment, and 3) MRI abnormalities except massive brain malformation and other extensive lesions that likely destroyed the ipsilateral tracts and led to reorganization. Exclusion criteria were 1) history of prematurity or perinatal hypoxic-ischemic event, 2) hemiplegia on preoperative examination by pediatric neurologists, and 3) dysmorphic features suggestive of a clinical syndrome.

Data acquisition

All participants underwent a 3T DTI with eight channel head coil at TR = 12500 ms, TE = 88.7 ms, FOV = 24 cm, 128×128 acquisition matrix (nominal resolution = 1.89 mm), contiguous 3 mm thickness in order to cover entire axial slices of whole brain using 55 isotropic gradient directions with b= 1000 s/mm^2, and number of excitations at 1. For anatomical reference, a three-dimensional fast spoiled gradient echo sequence (FSPGR) was acquired for each participant at TR/TE/TI of 9.12/3.66/400 ms, slice thickness of 1.2 mm, and planar resolution of 0.94×0.94 mm^2.

Healthy children underwent two fMRI studies at TR = 2000 ms, TE = 30 ms, FOV = 24 cm, 64×64 acquisition matrix, 4 mm thickness in order to localize 4 somatosensory areas: face, fingers, arm, leg, 9 language regions: inferior frontal operculum/triangularis (ifop/iftr), middle frontal (mfg), precentral (prec), superior/middle/inferior temporal (stg/mtg/itg), angular/supramarginal (ang/spm), 2 auditory regions: superior/middle temporal (stg/mtg),

and 7 visual regions: inferior/middle/superior occipital (iocc/mocc/socc), calcarine(calc), lingual(ling), cuneous(cune) and fusiform (fusi). Briefly, for mapping somatosensory areas, event-related tasks to trigger single movement of face muscle, fingers, arm, and leg to each side (left/right) was presented every five seconds in a 15-second block. The block was repeated 10 times for each side, resulting in total 20 sequential movements of face, fingers, arm and leg. To map semantic language, auditory and visual areas, three different patterns (square, triangle, and circle) was randomly displayed every five seconds in a 30-second block. Subjects was instructed to tag one of two buttons (yes, no) in response to an audio question (ON 30-second block) or visual pattern comparison (OFF 30-second block). These ON-OFF blocks were repeated four times. SPM package [88] was used to process all fMRI data including motion correction, general linear modeling, and statistical analysis to identify the locations of neuronal activities responding to functional tasks at uncorrected p-value < 0.05. BOLD activation was recorded for each functional area and utilized as a binary mask to sort out relative white matter pathways from DTI data (Table 4.7).

Epilepsy children underwent subdural electrode placement as a part of the clinical management for medically-uncontrolled seizures. ESM, using the method previously established [56, 81], was performed as part of clinical care during extraoperative electrocorticography recordings. When both clinical response and after-discharges occur, another pulse-train of the same or 1 mA smaller intensity was used until either clinical response or after-discharge fails to develop. Finally, a site with a contralateral movement induced by stimulation, without after-discharges, was defined as "somatosensory area" for a given body part. Likewise, a site with speech arrest, expressive aphasia, receptive aphasia, auditory hallucination and visual perception was classified as an essential eloquent area for the comparison with the proposed CNN-based fiber classification. Table 4.8 shows 22 eloquent ESM electrode classes selected for the present study. Using the landmark based registration procedure [110], those electrodes were spatially registered to DTI brain space and used as the ground truth of the CNN-based fiber classification.

Table 4.7: 64 functionally important white matter pathways of interest.

Eloquent function	Class index	From	To
somatosensory	$C_{1,34}$	arm area	internal capsule
	$C_{4,37}$	face area	internal capsule
	$C_{5,38}$	finger area	internal capsule
	$C_{16,49}$	leg area	internal capsule
language	$C_{7,40}$	ifop	itg
	$C_{8,41}$	ifop	mtg
	$C_{9,42}$	ifop	sma
	$C_{10,43}$	ifop	spm
	$C_{11,44}$	ifop	stg
	$C_{12,45}$	iftr	itg
	$C_{13,46}$	iftr	mtg
	$C_{14,47}$	iftr	stg
	$C_{18,51}$	mdfg	ang
	$C_{19,52}$	mdfg	itg
	$C_{20,53}$	mdfg	mtg
	$C_{21,54}$	mdfg	sma
	$C_{22,55}$	mdfg	spm
	$C_{23,56}$	mdfg	stg
	$C_{26,59}$	prec	ang
	$C_{27,60}$	prec	itg
	$C_{28,61}$	prec	mtg
	$C_{29,62}$	prec	spm
	$C_{30,63}$	prec	stg
auditory	$C_{25,58}$	mtg	inferior colliculus
	$C_{32,65}$	stg	inferior colliculus
visual	$C_{2,35}$	calc	lateral geniculuate
	$C_{3,36}$	cune	lateral geniculuate
	$C_{6,39}$	fusi	lateral geniculuate
	$C_{15,48}$	iocc	lateral geniculuate
	$C_{17,50}$	ling	lateral geniculuate
	$C_{24,57}$	mocc	lateral geniculuate
	$C_{31,64}$	socc	lateral geniculuate
other	C_{33}	-	-

Table 4.8: 22 eloquent ESM electrode classes of interest.

Eloquent function	Class index	Clinical response
somatosensory motor processing of the contralateral body	$D_{1,2}$	right,left arm
	$D_{3,4}$	right,left face
	$D_{5,6}$	right,left hand
	$D_{7,8}$	right,left foot
specific types of language function	$D_{9,10}$	left,right speech arrest
	$D_{11,12}$	left,right receptive aphasia
	$D_{13,14}$	left,right expressive aphasia during naming impairment
	$D_{15,16}$	left, right expressive aphasia during visual naming
auditory	$D_{17,18}$	left,right hallucination
visual	$D_{19,20}$	left,right phosphene
	$D_{21,22}$	left,right distortion

DTI tractography analysis

NIH TORTOISE [89] and FSL eddy package [3] were used to correct motion, noise, physiological artifacts, susceptibility-induced distortion, eddy current-induced distortion in the DTI data. Whole brain streamline tractography was then reconstructed using probabilistic SIFT tractography with second-order integration over fiber orientation distributions (iFOD2) to sample up to three FOD at every voxel [109]. At every voxel of gray/white matter boundary identified by FSL FAST package [129], 100 dynamically randomized seeding points were applied in the framework of anatomically constrained tractography [103] to reconstruct biologically realistic streamlines. Then, the binary masks of the fMRI activation were applied as an inclusion mask to sort out their relative streamline pathways from whole brain tractography (Table 4.7). The resulting treamline pathways were spatially normalized into FreeSurfer average template with the advanced normalization tools [6], sampled into 100 equal-distance segmentation points and finally 3D coordinates of the 100 segmentation points were used to represent each fiber for subsequent CNN classification.

Shallow CNN model for DTI streamline classification

Fig. 4.1 presents our shallow CNN model which has one input layer, one convolution layer, one maxpooling layer and one fully connected layer with softmax function. The details of each layer are described as follows.

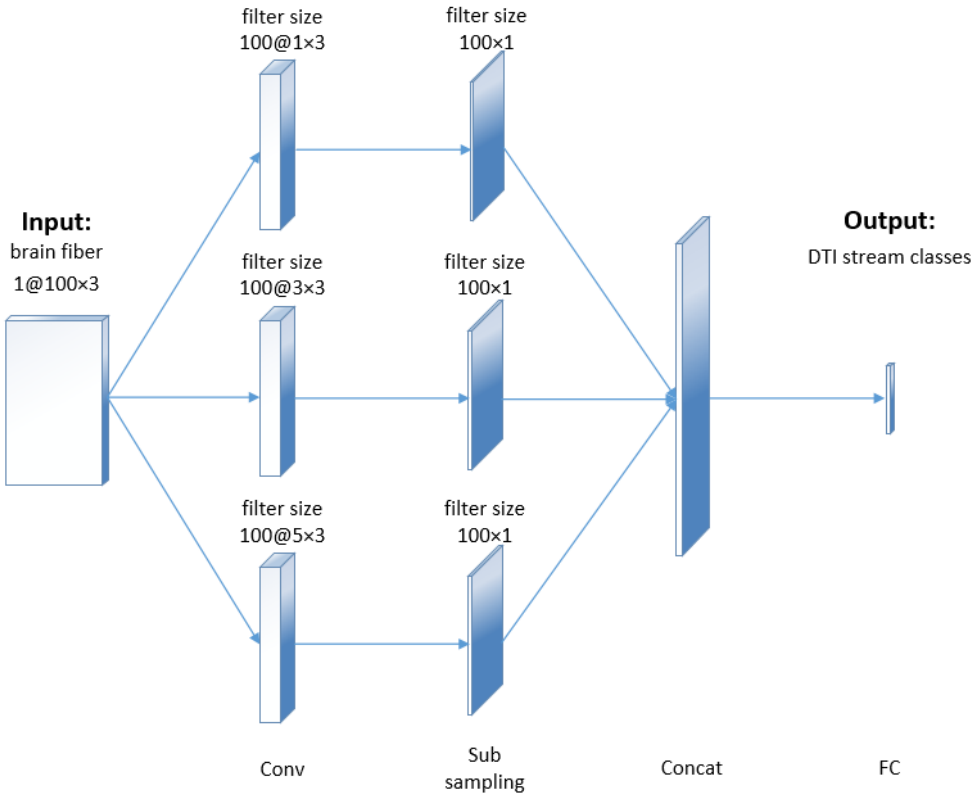


Figure 4.1: Network architecture of the proposed shallow CNN model.

Input layer

Formally, we denote $\mathbf{x_l} \in \mathbb{R}^k$ as the k-dimensional point representation for the lth point in a fiber. A fiber of length L is denoted as

$$\mathbf{X_{1:L}} = \mathbf{x_1} \oplus \mathbf{x_2} \oplus \cdots \oplus \mathbf{x_L}, \tag{4.1}$$

where $\oplus$ represents concatenation operators. By this, each input fiber is represented as a $L \times k$ matrix. In practice, we sample 100 points for each fiber and utilize coordinates of each point as its representation. Thus, each matrix has the same size 100×3.

Convolution layer

A convolution filter $\mathbf{w} \in \mathbb{R}^{h \times k}$ is applied to a window of h points of k-dimensional embeddings in the convolution layer to produce a feature map. For instance, given a window of points $\mathbf{X_{l:l+h-1}}$ and a bias term $b \in \mathbb{R}$, a feature g_i is generated by

$$g_i = f(\mathbf{w} \cdot \mathbf{X_{l:l+h-1}} + b), \tag{4.2}$$

where f is a non-linear function. In our case, we apply the element-wise Rectified Linear Unit (ReLU) to the input matrices which sets negative elements in g_i as 0. A feature map $\mathbf{g} = [g_1, g_2, \cdots, g_{L-h+1}]$ is obtained from all the possible windows of a fiber of length L. In our system, multiple filters of various sizes are applied in the convolution layer to produce multi-scale feature maps.

Sub-sampling layer

In the sub-sampling layer, we apply max pooling over each feature map produced in the convolution layer and output the maximum element $\hat{g} = max\{\mathbf{g}\}$. We denote features generated by the maxpooling layer as

$$\hat{\mathbf{G}} = \hat{\mathbf{g}}_\mathbf{1} \oplus \hat{\mathbf{g}}_\mathbf{2} \oplus \cdots \oplus \hat{\mathbf{g}}_\mathbf{M}, \tag{4.3}$$

where M denotes feature map number.

Dropout

Dropout is a technique to reduce overfitting for neural networks [105]. Given feature map $\hat{\mathbf{G}}$, we generate a dropout mask vector $\mathbf{r} \in \mathbb{R}^m$ of Bernoulli variables with probability p_d of being set as 0 and $1 - p_d$ of being set as 1. The output of dropout is

$$\hat{\mathbf{G}}_d = \hat{\mathbf{G}} \circ \mathbf{r}, \tag{4.4}$$

where $\circ$ denotes the element-wise multiplication operator. Empirically, we chose $p_d = 0.5$ in this study.

Fully connected layer

Given $\hat{\mathbf{G}}_d$ as the input, fully connected layers generate output

$$\hat{\mathbf{G}}_{fc} = \mathrm{ReLU}(\mathbf{w} \cdot \hat{\mathbf{G}}_d + b). \tag{4.5}$$

Output layer

On the output layer, we apply softmax function instead of ReLU to get the final classification probabilities

$$\mathbf{p}^i = \mathrm{sofmax}(\hat{\mathbf{G}}_{fc}), \tag{4.6}$$

where $\mathbf{p}^i$ denotes prediction probabilities of the ith fiber belonging to each class. The class with highest probability is chosen as the final classification result for corresponding fiber.

Optimization

Cross-entropy loss is selected as the training objective to minimize. The cross-entropy loss for the ith fiber is defined as

$$L^i_{CE} = -\log p^i_c, \tag{4.7}$$

where p^i_c is the prediction probability of the ith fiber in the dataset belonging to its ground truth class c. Adam [55], an adaptive learning rate approach for stochastic gradient descent, is utilized for CNN parameter updating.

Deep CNN model for DTI streamline classification

Fig. 4.2 shows the proposed deep network consisting of a series of stages. The first stage is composed of two types of layers: convolutional and pooling layers. The input fibers are passed through a set of filters followed by non-linear transformations. Then, the maximum of local patches are extracted. Second, 40 blocks of convolutional, pooling, and

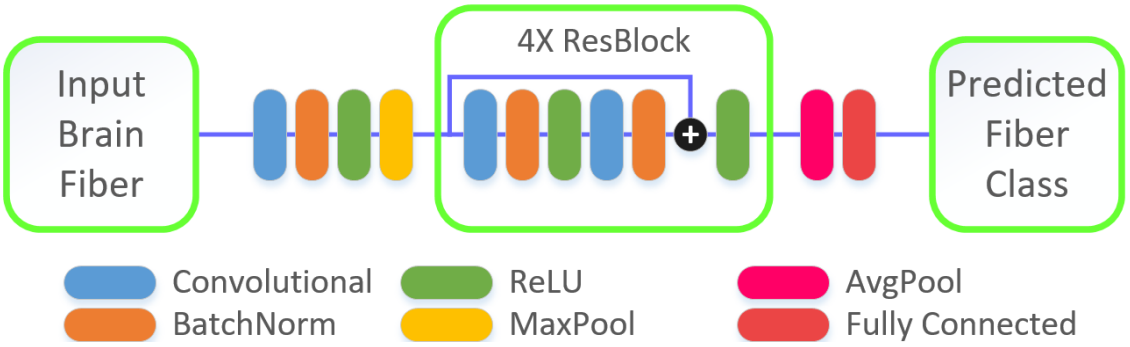


Figure 4.2: Network architecture of the proposed deep CNN model.

concatenation layers are applied to learn high-level fine features from brain fibers. For the residual units, their output is the sum of input and output of a block. The intuition behind is: rather than expecting blocks to approximate the fiber classification function, we explicitly let these layers approximate a residual function, which is easier for the optimization. Third, fully connected and softmax layers are induced to get the final prediction which contains the probabilities of the input fiber belonging to each class. The dropout units are applied to help prevent overfitting.

For optimization, the cross-entropy loss is applied here for the comparison of proposed two CNN frameworks. To further improve the classification performance, we also introduce two alternative loss functions:

Focal loss

In general, the large class unbalance encountered during training overwhelms the cross-entropy loss. Easily classified fibers comprise most of the loss and dominate the gradient [68]. In this study, we replace the conventional cross-entropy loss with focal loss in order to reduce the loss for well-classified fibers and focus on hard or mis-classified ones. We define the focal loss for the ith fiber as

$$L_F^i = -(1 - p_c^i)^\gamma \log p_c^i, \tag{4.8}$$

where γ is the focusing parameter. Empirically, we choose $\gamma = 2$. The modulating factor $(1 - p_c^i)^\gamma$ reduces the loss contribution from easy examples: a fiber classified with $p_c^i \geq 0.9$ contributes at least $100\times$ lower focal loss compared to cross-entropy loss; while hard examples with $p_c^i \leq 0.5$ would only be scaled down at most $4\times$.

Center loss

The conventional cross-entropy loss only encourages the separability of features [116]. To further improve the performance and generalization ability of proposed CNN classifier, we add a cluster-based loss item, *i.e.*, center loss, to the classification loss, which simultaneously learns a clustering centroid for CNN-learned features of each class and penalizes the distances between class centroids and the deep features. More formally, we denote the center loss of the ith fiber as

$$L_C^i = \frac{1}{2}\|f^i - c_y^i\|_2^2, \tag{4.9}$$

where $f^i \in R^d$ denotes the deep feature vector and $c_y^i \in R^d$ denotes the centroid of ground truth of the ith fiber. Thus, the overall loss to optimize is

$$L^i = L_{class}^i + \lambda L_C^i, \tag{4.10}$$

where L_{class}^i denotes the classification loss, which is the same with the proposed shallow CNN, and L_C^i is the center loss. Empirically, we choose $\lambda = 1$ in this study.

As we described in Sec. , the class with highest probability is taken as the final prediction result for each fiber.

Learning interpretable fiber representation

CNNs have achieved superior performance in supervised tasks. However, the end-to-end learning strategy makes the entire CNN model a black box. This weakness is highlighted in the biomedical imaging: if we do not know how the trained CNN classifies each fiber, we cannot fully trust the classification results given by this model. By introducing attention

mechanism into our brain fiber classifier, we are able to highlight the attention of the CNN model and understand how it makes predictions.

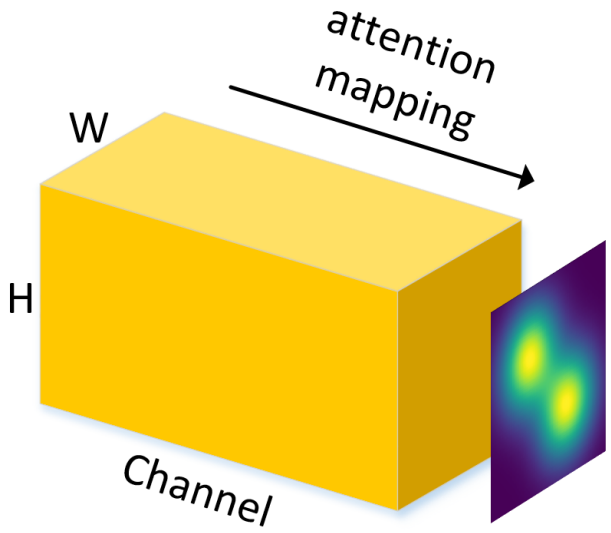


Figure 4.3: An example to conceptualize the attention map.

In this study, we apply the *soft attention* mechanism. The soft attention generates a weighted average over different locations of all the feature channels (Fig. 4.3). More formally, we denote the location variable as s, *i.e.*, where the model decides to focus on, the attention weight map as α, and the feature map of the ith fiber as $\mathbf{G}^i$. Thus, the expectation of output weighted feature map $\hat{\mathbf{G}}^i$ is

$$\mathbb{E}_s[\hat{\mathbf{G}}^i] = \alpha \mathbf{G}^i. \tag{4.11}$$

This soft attention mechanism is continuous and differentiable, so it is trivial to update the attention weight map α by using standard backpropagation during the training phase of CNN models. In particular, we insert one attention unit to the end of each deep CNN block (Fig. 4.4). The visualization of where CNNs focus on is reported in Sec. .

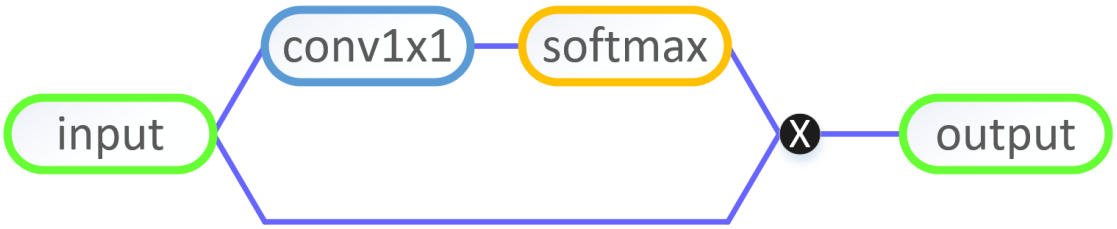


Figure 4.4: A systematic diagram of the attention mapping process.

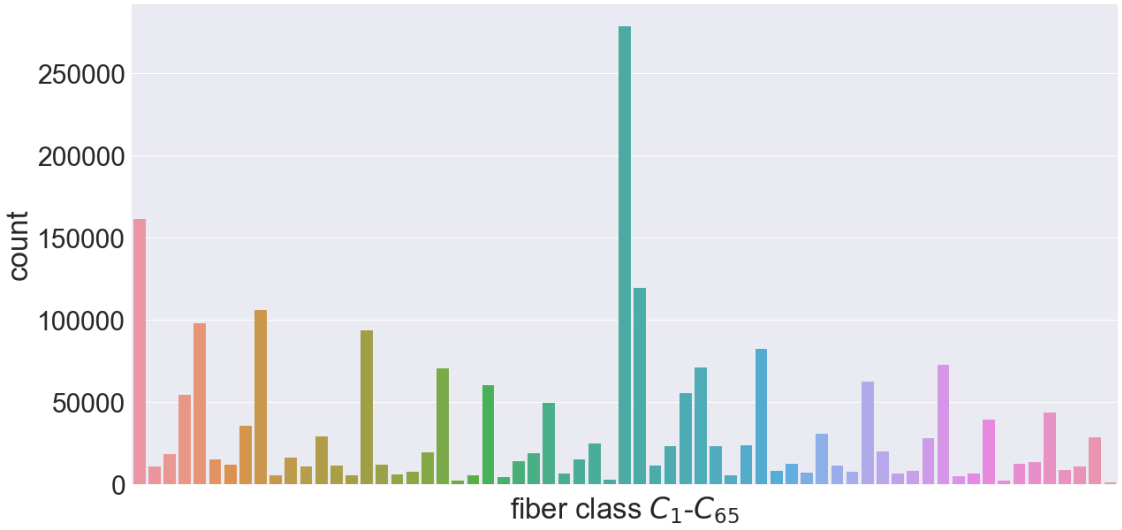


Figure 4.5: Total number of DTI fiber streamlines.

Experimental Results

Experiment setup

In this study, we utilized the fibers from 56 randomly chosen healthy subjects for training and fibers from the remaining 14 healthy subjects for testing. The number of fibers for each class in the training set was presented in Fig. 4.5. Clearly, the distribution is highly unbalanced: the most frequent classes have 40× to 220× more fibers than the least frequent classes.

To validate the performance of the proposed CNN-based fiber classifiers, the classified fibers were inversely warped to the native DTI space and then compared with the gold standard ESM.

Performance evaluation

To select the most optimal CNN-based fiber classifier, the same training and testing splits were analyzed by the following methods: linear SVM (LSVM), multiclass Logistic Regression (LR), shallow CNN with cross-entropy loss (SCNN-CE), deep CNN with cross-entropy loss (DCNN-CE), deep CNN with focal loss (DCNN-FL), and deep CNN with both focal loss and center loss (DCNN-CL). In addition, attention mechanism was combined with DCNN-CL (DCNN-CL-ATT). To evaluate the performance of each model over the highly unbalanced dateset, F_1 score was selected as the metric, which can be calculated by

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}. \tag{4.12}$$

Finally, cortical terminals of the selected CNN determined fiber classes, C_i were compared with their gold standard ESM electrode locations, D_j, where a match was considered to occur if CNN terminals contacted and overlapped the area of the gold standard. The percentage of overlap was assessed as a function of Euclidean distance. That is, the border

of the CNN area was extended by contact, 1cm, 1.5cm, and 2cm to determine whether the number of overlap varies.

The proposed CNN frameworks were implemented with pytorch [85] and trained on a GTX 1080 Ti graphics card. The CNN model with most parameters, DCNN-CL-ATT, takes about 6 hours to train through entire stages as described in Sec. and . For testing on a unseen whole brain tractography consisting of 1 million streamlines, this model takes about 15 minutes for 64-class classification, which is advantageous compared to our previous DTI-MAP analysis [46, 47], taking about 20 minutes to classify 12 classes of somatosensory and language related fibers.

Table 4.9: Mean and standard deviation of the average macro-averaged F_1 scores.

Method	Macro-averaged Score
LSVM	0.2986 ± 0.0021
LR	0.3381 ± 0.0131
SCNN-CE	0.8632 ± 0.0020
DCNN-CE	0.9211 ± 0.0098
DCNN-FL	0.9362 ± 0.0026
DCNN-CL	0.9494 ± 0.0066
DCNN-CL-ATT	$\mathbf{0.9525} \pm 0.0053$

Fiber classification results

The average classification performance over all the classes are listed in Table 4.9. For the baselines, LR performed 13.23% better than LSVM, which shows the advantage of non-linear models over linear models on the brain fiber classification.

The CNN models significantly outperformed LR and LSVM by 155.31% or more, which indicates the strong classification ability of deep learning models. DCNN-CE outperformed SCNN-CE by 6.71%, which shows the necessity of applying deep CNN architectures.

It is worth pointing out that introducing focal loss to deep CNN improved the performance by 1.64% comparing to deep CNN with conventional cross-entropy loss. This demonstrates that the focal loss function suits better for the classification of highly unbalanced datasets. Moreover, DCNN-CL achieved better performance than DCNN-FL by

1.41%, mainly attributed to the more discriminative representation learned using the center loss. Overall, DCNN-CL-ATT achieved the best performance.

Validation results

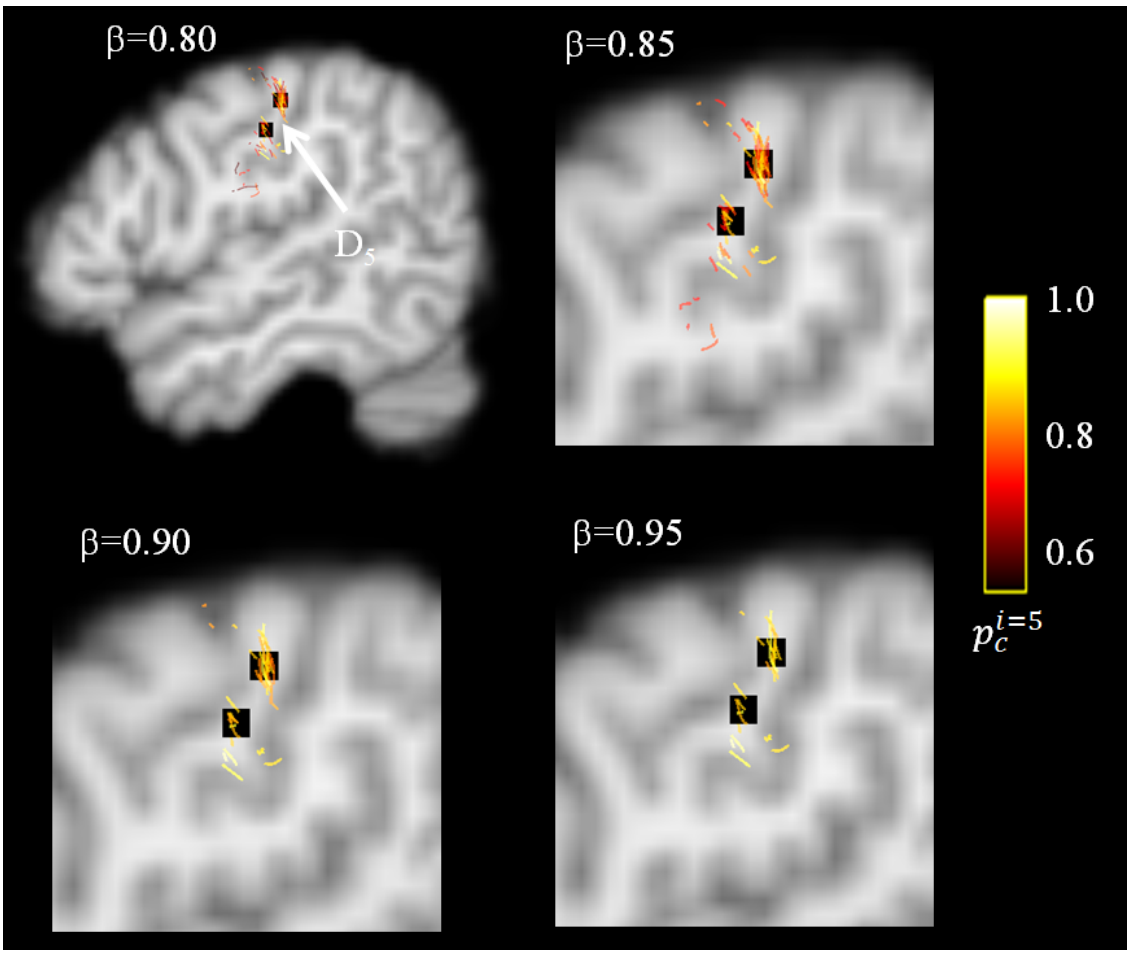


Figure 4.6: Examples of DCNN-CL-ATT determined-white matter pathway.

An illustrative example of DCNN-CL-ATT classification on white matter tracts associated with finger movement, C_5, is presented in Fig. 4.6. This example shows a clinical case where the triggering of right hand fingers was successfully induced during the ESM procedure of an 8 years old patient. Clearly, DCNN-CL-ATT successfully localized individual streamlines of cortico-spinal tracts terminating ESM finger areas in precentral gyus (two black colored boxes). False detections localized outside the electrodes were significantly reduced at $\beta = 0.95$ (β is the probability threshold to decide the final membership of a given input

fiber) without reducing true positives, suggesting high specificity of the proposed method to delineate functionally eloquent areas and pathways from individual patient. In this study, we selected $\beta = 0.95$ as an optimal value to sort out true positive fibers belonging to each C_i.

Table. 4.10 shows quantitative comparison of the DCNN-CL-ATT classification, C_i, with its gold standard ESM, D_j. Cortical terminals of class fibers, C_i, whose p_c^i were thresholded at $\beta = 0.95$ were compared with the locations of ESM results, D_j, in 70 children with focal epilepsy. The overlap match was counted if any of fiber terminals includes the measured ESM electrode within each of four Euclidean distance thresholds: contact, 1cm, 1.5cm and 2cm. The detection probability gradually increased with the distance. For instance, the average values of detection probability were 72%/83%/90%/90% (contact/1cm/1.5cm/2cm) for somatosensory areas, 74%/81%/87%/93% (contact/1cm/1.5cm/2cm) for language areas, 40%/80%/80%/90% (contact/1cm/1.5cm/2cm) for auditory areas, and 57%/85%/87%/88% (contact/1cm/1.5cm/2cm) for visual areas, respectively. We found that compared with our previous DWI-MAP analysis of somatosensory and language function[46, 47], the proposed method could improve about 9-14% of the detection probability by classifying more outliers (e.g., association fibers with higher curvatures) into correct ESM localizations.

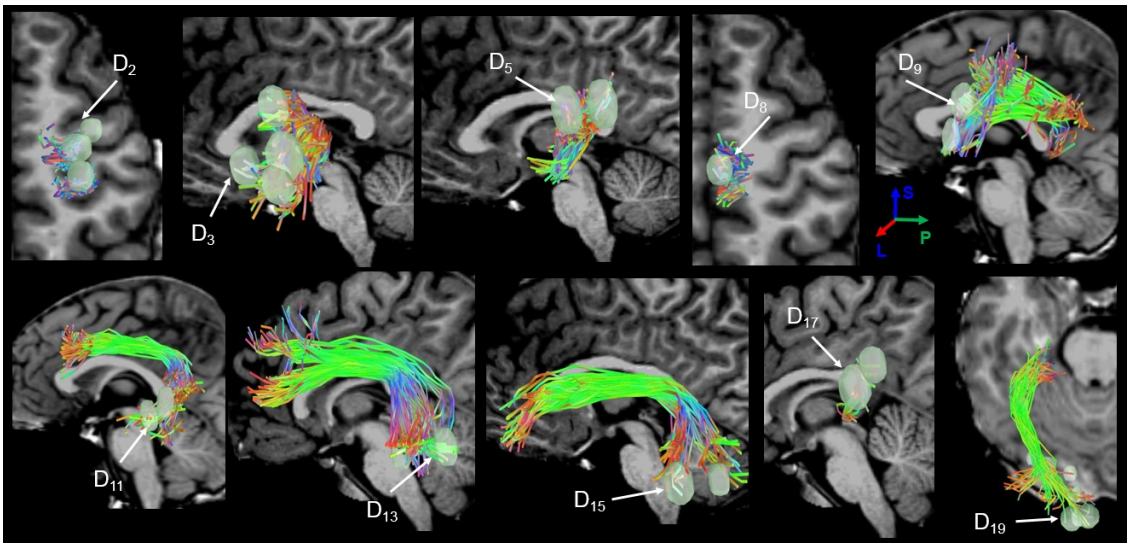


Figure 4.7: Representative examples of DCNN determined-white matter pathways.

Table 4.10: Probability of individual DTI class, C_i, to match individual ESM class.

ESM	DTI	contact	1 cm	1.5 cm	2.0 cm
D_1	C_1	0.7857	0.8571	1.0000	1.0000
D_2	C_{34}	0.6000	0.8000	0.9333	0.9333
D_3	C_4	0.6071	0.8571	0.9286	0.9286
D_4	C_{37}	0.7879	0.8182	0.8788	0.8788
D_5	C_5	0.7241	0.7931	0.8966	0.9310
D_6	C_{38}	0.6364	0.7879	0.8182	0.8182
D_7	C_{16}	0.8574	0.8574	0.8574	0.8574
D_8	C_{49}	0.7333	0.8667	0.8667	0.8667
D_9	$C_{29,30}$	0.7368	0.8947	0.8947	0.8947
D_{10}	$C_{62,63}$	0.9091	0.9091	0.9091	0.9091
D_{11}	$C_{8,11,14}$	0.6923	0.8462	0.8462	0.8462
D_{12}	N.A	N.A	N.A	N.A	N.A
D_{13}	$C_{8,13,14}$	0.6667	0.7222	1.000	1.000
D_{14}	N.A	N.A	N.A	N.A	N.A
D_{15}	$C_{7,12,19}$	0.6667	0.6667	0.6667	1.000
D_{16}	N.A	N.A	N.A	N.A	N.A
D_{17}	$C_{25,32}$	0.2000	0.8000	0.8000	0.8000
D_{18}	$C_{58,65}$	0.6000	0.8000	0.8000	1.0000
D_{19}	$C_{2,17,24}$	0.6333	0.8000	0.8333	0.8333
D_{20}	$C_{35,50,57}$	0.5625	0.7500	0.8125	0.8438
D_{21}	$C_{6,7,15}$	0.7500	1.0000	1.0000	1.0000
D_{22}	$C_{39,40,48}$	0.3333	0.8333	0.8333	0.8333

Representative examples of the proposed DCNN-CL-ATT classification derived-white matter detection of fibers, C_i, at $\beta = 0.95$ were presented at Fig. 4.7, as compared with eloquent areas determined by ESM which were obtained from four different test subjects: $D_{2,8}$ from an 8 years old boy, $D_{3,5}$ from a 12 years old girl, $D_{9,11}$ from another 8 years old boy, and $D_{13,15,17,19}$ from a 14 years old girl. It is notable that all predictions given by DCNN-CL-ATT (*i.e.*, RGB-colored fibers in Fig. 4.7) are spatially well matched to the gold standard ESM electrodes, which suggests high translational value of the proposed work as an imaging tool to improve clinical ESM procedure by guiding accurate placement of electrodes in actual eloquent areas.

Visualization of learned discriminative fiber representation

To further demonstrate the benefit of center loss, we extracted the output of penultimate layer of DCNN-FL and DCNN-CL as the representations of corresponding brain fibers for comparison. The dimensionality of extracted representation vectors were reduced to two using tSNE [72] for visualization. As shown in Fig. 4.8, the representations learned by DCNN-CL have better intra-class compactness comparing to representations learned by DCNN-FL. We further applied quantitative analysis on the advantage of center loss for discriminative feature learning by computing the intra- and inter-class distances of representation vectors learned by DCNN-FL and DCNN-CL. To make the distances comparable, the average intra-class distances were normalized to 1. The normalized average distances over all fiber classes are reported in Table 4.11. As shown in the table, the inter/intra distance ratio of fiber representations learned with DCNN-CL is 32.55 times greater than that of representations learned with DCNN-FL, which indicates that introducing center loss results in better intra-class compactness and greater inter-class variations.

Table 4.11: Normalized mean and standard deviation of intra- and inter-class distances.

Method	Intra-class Distance	Inter-class Distance
DCNN-FL	1 ± 0.5826	30.9720 ± 1.2217
DCNN-CL	1 ± 0.4958	1007.9916 ± 245.2773

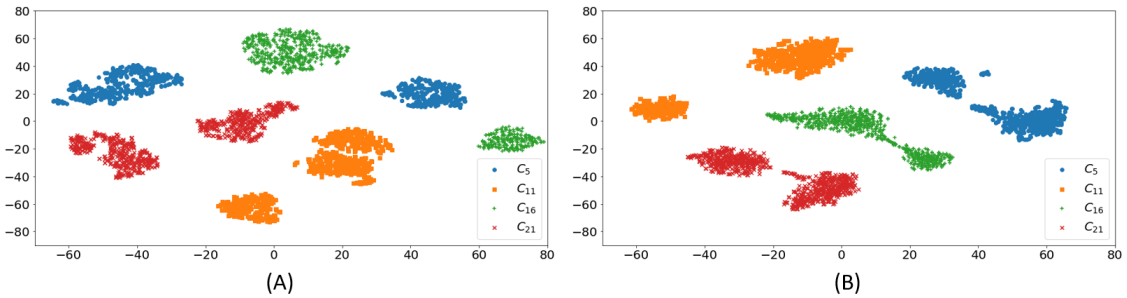


Figure 4.8: The tSNE visualization of deep fiber representations.

Visualization of interpretable fiber representation

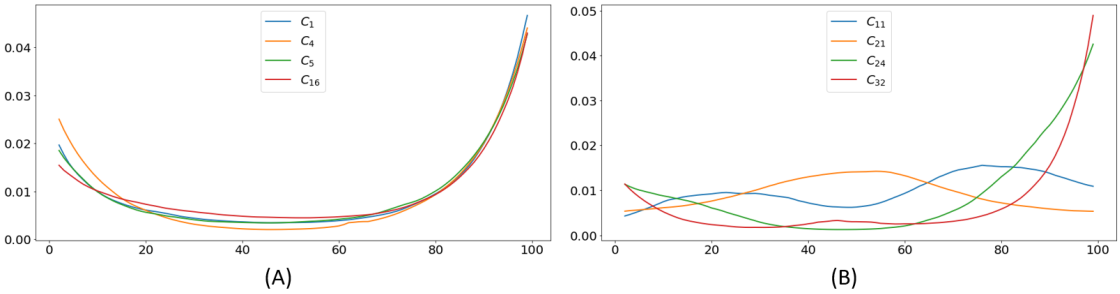


Figure 4.9: Representative examples of attention maps.

To illustrate how our CNN models classify streamlines, we visualized the attention maps for brain fibers from some example classes. First, we selected fibers with high classification confidence ($p_c^i > 0.85$). Second, the corresponding attention maps over 100 points of the selected fibers were extracted from the trained DCNN-CL-ATT model. Third, we computed the average attention weights for fibers belonging the same class and use them as the attention map for that class. The greater weights indicate the corresponding points are more important for CNNs to make predictions.

Fig. 4.9 provides a clue on how the deep CNN model makes predictions for brain fibers. Somatosensory fibers of $C_{1,4,5,16}$ showed noticeable weights of attention near both s_1 (precentral gyrus) and s_{100} (internal capsule), which indicates that DCNN-CL-ATT needs to focus on the both ends of these fibers when make classifications. These attention maps directly support the traditional homunculus representation of precentral gyrus and internal capsule in human brain [26, 39] suggesting that separate cortico-spinal tracts connect unique

segments of precentral gyrus and internal capsule resulting in multiple white matter pathway classes associated with unique smatosensory functions, $C_{1,4,5,16}$. Meanwhile, other language and auditory tracts of $C_{11,21,24,32}$ whose anatomical trajectories are terminated at different cortices (s_1, s_{100}) showed different patterns of attention weights widely spread at entire range of spatial coordinate. This example clearly demonstrated that our attention map could provide supplementary marker to localize specific functional areas of interest by identifying the most important segments of a given input tract detected by DCNN-CL-ATT.

Discussion

In this study, we proposed a novel CNN-based method, DCNN-CL-ATT, for evaluating how deep learning of in-vivo DTI trajectory can accurately detect eloquent functional areas determined by the gold starndard ESM data and selectively highlight the most important segments of DTI streamlines for the predictions. In contrast to most parametric Gaussian approaches, the proposed model makes no assumption regarding a priori probabilistic distribution of individual streamlines belong to specific eloquent white matter pathway. The proposed CNN method can process very large streamline datasets on a desktop computer in a reasonable time frame automatically with only one single user-defined parameter (*i.e.*, the probability threshold, β, to decide the final membership of a given input fiber).

In vivo visualization of neuronal connections by placing regions of interest to classify DTI streamlines is a promising but still challenging task in clinical application (*i.e.*, labor intensive and subjective to reduce reproducibility [16]). Many investigators have attempted to objectively characterize the complicated tract patterns in DTI [82, 36]. Mixed results have been reported depending on the employed geometrical features and similarity measures which do not consider functional association clinically important to know. To avoid this ambiguity, the present study has generalized the application of the-state-of-art CNN techniques to objectively learn actual spatial coordinate trajectories of function-specific-white matter pathways. However, it should be noted that the accuracy of the proposed CNN model is highly dependent on the DTI reconstruction algorithm used to generate the DTI stream-

lines. Although the utilized iFOD2 reconstruction provided clinically acceptable accuracy of 73-100% to detect eloquent functions within the spatial resolution of ESM (1 cm), other advanced methods may create higher quality data, which can further improve the performance of the proposed CNN methods.

In this study, we mainly detected fibers of major pathways of sufficient size and high coherence in the dataset. Smaller tracts or less coherent connections are currently not reliably assessable. Higher resolution employing higher field strength, stronger diffusion gradients and high angular resolution DTI could enable the delineation of such structures. More importantly, our target classes were constructed using fMRI inevitably limited by ill-posed neurovascular coupling [112]. Thus, the detection of eloquent area using our CNN methods are naturally effective and valid on the gyral level rather than the nominal voxel resolution. In the future, we plan to further investigate various attention mechanisms [111] on whether they benefit conventional connectome analysis by detecting noisy or incorrectly tracked streamlines spatially deviated from normative population (*e.g.*, wiggly false fibers). We anticipate that attention weights significantly altered in the population would indicate noisy streamlines that should be excluded from the analysis.

In conclusion, the significance of the proposed CNN framework for presurgical planning of potential surgical candidates includes: 1) no added risk or cost to identify functionally important areas at both cortical and subcortical levels; 2) no need for patient cooperation, particularly important in young children; 3) easy applicability to other types of neurosurgical procedures (*e.g.*, brain tumor resection). This study is an excellent example which translates advanced deep learning techniques to clinical practice in the pediatric population in which currently available approaches are suboptimal (*i.e.*, ESM and fMRI). Prospective investigation of the proposed CNN method will further improve presurgical planning and provide a unique opportunity to minimize or predict postsurgical functional deficits in the future.

CHAPTER 5 CONCLUSION

In this Ph.D. dissertation, we presented our research accomplishments in representation learning using convolutional neural networks with significant intellectual merit and novelty.

Summary of Contributions

In Chapter 2, we proposed a novel framework,Topic-based Skip-gram, for learning topic-based semantic word embeddings for text classification with CNNs and achieved highly competitive results with word embeddings learned by Skipgram. While Skip-gram focuses on context information from local word windows, the proposed Topic-based Skip-gram leverages semantic information from documents. We also described two multimodal CNN architectures which can ensemble different kinds of word embeddings.

In Chapter 3, we proposed Directionally Convolutional Network that extends convolution operations from images to the surface mesh in the spatial domain. Furthermore, we introduced a two-stream framework combining proposed Directionally Convolutional Network and a neural network for segmentation of 3D shapes. Instead of fusing the two streams by a simple concatenation, we take our framework as an approximation of a directed graph and combine the probabilities inferred by the two streams with an element-wise product. Finally, Conditional Random Field was applied to optimize the surface mesh segmentation.

In Chapter 4, we proposed a novel CNN-based method for evaluating how deep learning of in-vivo DTI trajectory can accurately detect eloquent functional areas determined by the gold starndard ESM data and selectively highlight the most important segments of DTI streamlines for the predictions. The proposed CNN method can process very large streamline datasets on a desktop computer in a reasonable time frame.

Future Research Directions

We believe our work will encourage new research in the area of representation learning for different data formats including text, 3D polygon, and brain fiber tracts.

The proposed work in Chapter 2 shows a promising direction of learning topic-based word embedding for text analysis. By integrating global semantic meaning and local word coherence, the learned word embedding gains competitive performance for classification while has better interpretability. Actually, the paper has been cited by recent researches: text classification algorithms [24, 45] and information retrieval works [115, 23].

For the work described in Chapter 3, we proposed a framework to learn 3D polygon representation using the most fundamental geometric features, which demonstrates a novel approach to learning 3D polygon representation for shape segmentation and pushes up the current state-of-the-art methods for future studies. Our work applied deep learning techniques for geometric feature learning on the 3D surface, which became a trend in recent years. We believe there will be more future researches in this area.

The presented work in Chapter 4 shows the proposed framework to learn discriminative and interpretable brain fiber representation for classification. Our study provided a plausible solution to explore how deep learning frameworks make decisions. From the perspective of medicine, we believe this work will encourage researchers to further investigate various attention mechanisms on whether they benefit conventional connectome analysis by detecting noisy or incorrectly tracked streamlines spatially deviated from normative population (*e.g.*, wiggly false fibers).

APPENDIX

Journal Publications Under Revision

R1. H. Xu, M. Dong, M.-H. Lee, Y. Nakai, E. Asano, and J.-W. Jeong. Objective Detection of Eloquent Axonal Pathways to Minimize Postoperative Deficits in Pediatric Epilepsy Surgery using Diffusion Tractography and Convolutional Neural Networks. *IEEE Transactions on Medical Imaging (TMI)*, 2018.

Conference Publications

C1. H. Xu, M. Dong, Y. Nakai, E. Asano, and J.-W. Jeong. Automatic detection of eloquent axonal pathways in diffusion tractography using electrical stimulation mapping and convolutional neural networks. *IEEE International Symposium on Biomedical Imaging (ISBI)*, 2018.

C2. H. Xu, M. Dong, and Z. Zhong. Directionally convolutional networks for 3D shape segmentation. *ICCV*, 2017.

C3. H. Xu, M. Dong, D. Zhu, A. Kotov, A. Carcone, and S. Naar-King. Text classification with topic-based word embedding and convolutional neural networks. *ACM Conference on Bioinformatics, Computational Biology, and Health Informatics (ACM BCB)*, 2016.

REFERENCES

[1] A. L. Alexander, K. M. Hasan, M. Lazar, J. S. Tsuruda, and D. L. Parker, "Analysis of partial volume effects in diffusion-tensor mri," *Magnetic Resonance in Medicine*, vol. 45, no. 5, pp. 770–780, 2001.

[2] P. Alliez, D. Cohen-Steiner, O. Devillers, B. Lévy, and M. Desbrun, "Anisotropic polygonal remeshing," in *ACM TOG*, vol. 22. ACM, 2003, pp. 485–493.

[3] J. L. Andersson and S. N. Sotiropoulos, "An integrated approach to correction for off-resonance effects and subject movement in diffusion mr imaging," *Neuroimage*, vol. 125, pp. 1063–1078, 2016.

[4] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh, "Vqa: Visual question answering," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2425–2433.

[5] A. R. Aronson, J. G. Mork, C. W. Gay, S. M. Humphrey, and W. J. Rogers, "The NLM indexing initiative's medical text indexer," *Medinfo*, vol. 11, no. Pt 1, pp. 268–72, 2004.

[6] B. B. Avants, N. J. Tustison, G. Song, P. A. Cook, A. Klein, and J. C. Gee, "A reproducible evaluation of ants similarity metric performance in brain image registration," *Neuroimage*, vol. 54, no. 3, pp. 2033–2044, 2011.

[7] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[8] P. J. Basser, S. Pajevic, C. Pierpaoli, J. Duda, and A. Aldroubi, "In vivo fiber tractography using dt-mri data," *Magnetic resonance in medicine*, vol. 44, no. 4, pp. 625–632, 2000.

[9] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Transactions on PAMI*, vol. 19, no. 7, pp. 711–720, 1997.

[10] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE transactions on PAMI*, vol. 24, no. 4, pp. 509–522, 2002.

[11] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on PAMI*, vol. 35, no. 8, pp. 1798–1828, 2013.

[12] Y. Bengio and R. Ducharme, "A neural probabilistic language model," *Journal of Machine Learning Research*, vol. 3, pp. 1137–1155, 2003.

[13] P. Besson, F. Andermann, F. Dubeau, and A. Bernasconi, "Small focal cortical dysplasia lesions are located at the bottom of a deep sulcus," *Brain*, vol. 131, no. 12, pp. 3246–3255, 2008.

[14] D. Blei and J. Lafferty, "Correlated topic models," *NIPS*, vol. 18, p. 147, 2006.

[15] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[16] L. Bonilha, E. Gleichgerrcht, J. Fridriksson, C. Rorden, J. L. Breedlove, T. Nesland, W. Paulus, G. Helms, and N. K. Focke, "Reproducibility of the structural brain connectome derived from diffusion tensor imaging," *PloS one*, vol. 10, no. 9, p. e0135247, 2015.

[17] D. Boscaini, J. Masci, E. Rodolà, and M. Bronstein, "Learning shape correspondence with anisotropic convolutional neural networks," in *NIPS*, 2016, pp. 3189–3197.

[18] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst, "Geometric deep learning: going beyond euclidean data," *IEEE Signal Processing Magazine*, vol. 34, no. 4, pp. 18–42, 2017.

[19] J. Bruna, W. Zaremba, A. Szlam, and Y. LeCun, "Spectral networks and locally connected networks on graphs," in *Proceedings of the 2nd International Conference on Learning Representations*, 2014.

[20] C. Chen, W. Buntine, N. Ding, L. Xie, and L. Du, "Differential topic models," *PAMI, IEEE Transactions on*, vol. 37, no. 2, pp. 230–242, 2015.

[21] S. Chen, C. Zhang, M. Dong, J. Le, and M. Rao, "Using ranking-cnn for age estimation," in *Proceedings of the IEEE CVPR*, 2017, pp. 5183–5192.

[22] X. Chen, A. Golovinskiy, and T. Funkhouser, "A benchmark for 3D mesh segmentation," in *ACM ToG*, vol. 28. ACM, 2009, p. 73.

[23] A. Cieslewicz, J. Dutkiewicz, and C. Jedrzejek, "Baseline and extensions approach to information retrieval of complex medical data: Poznan's approach to the biocaddie 2016," *Database*, vol. 2018, 2018.

[24] A. Cohan, A. Fong, R. M. Ratwani, and N. Goharian, "Identifying harm events in clinical care through medical narratives," in *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*. ACM, 2017, pp. 52–59.

[25] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proceedings of ICML*. ACM, 2008, pp. 160–167.

[26] N. Dawnay and P. Glees, "Somatotopic analysis of fibre and terminal distribution in the primate corticospinal pathway," *Developmental Brain Research*, vol. 26, no. 1, pp. 115–123, 1986.

[27] S. de Ribaupierre, M. Fohlen, C. Bulteau, G. Dorfmüller, O. Delalande, O. Dulac, C. Chiron, and L. Hertz-Pannier, "Presurgical language mapping in children with epilepsy: clinical usefulness of functional magnetic resonance imaging for the planning of cortical stimulation," *Epilepsia*, vol. 53, no. 1, pp. 67–78, 2012.

[28] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American society for information science*, vol. 41, no. 6, p. 391, 1990.

[29] M. Defferrard, X. Bresson, and P. Vandergheynst, "Convolutional neural networks on graphs with fast localized spectral filtering," in *NIPS*, 2016, pp. 3837–3845.

[30] D. K. Duvenaud, D. Maclaurin, J. Iparraguirre, R. Bombarell, T. Hirzel, A. Aspuru-Guzik, and R. P. Adams, "Convolutional networks on graphs for learning molecular fingerprints," in *NIPS*, 2015, pp. 2224–2232.

[31] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Learning hierarchical features for scene labeling," *IEEE transactions on PAMI*, vol. 35, no. 8, pp. 1915–1929, 2013.

[32] L. Fei-Fei and P. Perona, "A bayesian hierarchical model for learning natural scene categories," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 2. IEEE, 2005, pp. 524–531.

[33] N. Fiorini, S. Ranwez, S. Harispe, J. Montmain, and V. Ranwez, "USI at BioASQ 2015: a semantic similarity-based approach for semantic indexing," in *Working Notes for the Conference and Labs of the Evaluation Forum (CLEF), Toulouse, France*, 2015.

[34] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of human genetics*, vol. 7, no. 2, pp. 179–188, 1936.

[35] R. W. Floyd, "Algorithm 97: shortest path," *Communications of the ACM*, vol. 5, no. 6, p. 345, 1962.

[36] E. Garyfallidis, M. Brett, M. M. Correia, G. B. Williams, and I. Nimmo-Smith, "Quickbundles, a method for tractography simplification," *Frontiers in neuroscience*, vol. 6, p. 175, 2012.

[37] H. E. Gohari, M. Dong, S. Nejad-Davarani, and C. K. Glide-Hurst, "Generating synthetic cts from magnetic resonance images using generative adversarial networks," *Medical physics*, 2018 (to appear).

[38] K. Guo, D. Zou, and X. Chen, "3D mesh labeling via deep convolutional neural networks," *ACM TOG*, vol. 35, no. 1, p. 3, 2015.

[39] T. Hardy, G. Bertrand, and C. Thompson, "The position and organization of motor fibers in the internal capsule found during stereotactic surgery," *Stereotactic and Functional Neurosurgery*, vol. 42, no. 3, pp. 160–170, 1979.

[40] A. Haseeb, E. Asano, C. Juhász, A. Shah, S. Sood, and H. T. Chugani, "Young patients with focal seizures may have the primary motor area for the hand in the postcentral gyrus," *Epilepsy research*, vol. 76, no. 2, pp. 131–139, 2007.

[41] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE CVPR*, 2016, pp. 770–778.

[42] X. He and P. Niyogi, "Locality preserving projections," in *NIPS*, 2004, pp. 153–160.

[43] M. Henaff, J. Bruna, and Y. LeCun, "Deep convolutional networks on graph-structured data," *arXiv preprint arXiv:1506.05163*, 2015.

[44] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *arXiv preprint arXiv:1207.0580*, 2012.

[45] J. Hu, S. Li, J. Hu, and G. Yang, "A hierarchical feature extraction model for multi-label mechanical patent classification," *Sustainability*, vol. 10, no. 1, p. 219, 2018.

[46] J.-W. Jeong, E. Asano, E. C. Brown, V. N. Tiwari, D. C. Chugani, and H. T. Chugani, "Automatic detection of primary motor areas using diffusion mri tractography: comparison with functional mri and electrical stimulation mapping," *Epilepsia*, vol. 54, no. 8, pp. 1381–1390, 2013.

[47] J.-W. Jeong, E. Asano, C. Juhász, and H. T. Chugani, "Localization of specific language pathways using diffusion-weighted imaging tractography for presurgical planning of children with intractable epilepsy," *Epilepsia*, vol. 56, no. 1, pp. 49–57, 2015.

[48] D. K. Jones, T. R. Knösche, and R. Turner, "White matter integrity, fiber count, and other fallacies: the do's and don'ts of diffusion mri," *Neuroimage*, vol. 73, pp. 239–254, 2013.

[49] N. Kalchbrenner, E. Grefenstette, and P. Blunsom, "A convolutional neural network for modelling sentences," *arXiv preprint arXiv:1404.2188*, 2014.

[50] E. Kalogerakis, M. Averkiou, S. Maji, and S. Chaudhuri, "3D shape segmentation with projective convolutional networks," in *Proceedings of the IEEE CVPR*, 2017.

[51] E. Kalogerakis, A. Hertzmann, and K. Singh, "Learning 3D mesh segmentation and labeling," *ACM Transactions on Graphics (TOG)*, vol. 29, no. 4, p. 102, 2010.

[52] B.-s. Kim, P. Kohli, and S. Savarese, "3D scene understanding by voxel-CRF," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 1425–1432.

[53] V. G. Kim, W. Li, N. J. Mitra, S. Chaudhuri, S. DiVerdi, and T. Funkhouser, "Learning part-based templates from large collections of 3D shapes," *ACM TOG*, vol. 32, no. 4, p. 70, 2013.

[54] Y. Kim, "Convolutional neural networks for sentence classification," *arXiv preprint arXiv:1408.5882*, 2014.

[55] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of the 3rd International Conference on Learning Representations*, 2015.

[56] K. Kojima, E. C. Brown, R. Rothermel, A. Carlson, D. Fuerst, N. Matsuzaki, A. Shah, M. Atkinson, M. Basha, S. Mittal *et al.*, "Clinical significance and developmental changes of auditory-language-related gamma activity," *Clinical Neurophysiology*, vol. 124, no. 5, pp. 857–869, 2013.

[57] A. Kotov, M. Hasan, A. Carcone, M. Dong, S. Naar-King, and K. BroganHartlieb, "Interpretable probabilistic latent variable models for automatic annotation of clinical text," in *AMIA Annual Symposium Proceedings*, vol. 2015. American Medical Informatics Association, 2015, p. 785.

[58] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012, pp. 1097–1105.

[59] J. Lafferty, A. McCallum, F. Pereira *et al.*, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proceedings of ICML*, vol. 1, 2001, pp. 282–289.

[60] S. Lai, L. Xu, K. Liu, and J. Zhao, "Recurrent convolutional neural networks for text classification." in *AAAI*, 2015, pp. 2267–2273.

[61] K. Lang, "Newsweeder: Learning to filter netnews," in *Proceedings of ICML*, 1995, pp. 331–339.

[62] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," in *IEEE*, vol. 11, 1998, pp. 2278–2324.

[63] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[64] R. P. Lesser, N. E. Crone, and W. Webber, "Subdural electrodes," *Clinical Neurophysiology*, vol. 121, no. 9, pp. 1376–1392, 2010.

[65] Y. Li, "Localized feature selection for unsupervised learning," 2008.

[66] D. Lin, S. Fidler, and R. Urtasun, "Holistic scene understanding for 3D object detection with RGBD cameras," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 1417–1424.

[67] G. Lin, C. Shen, A. van den Hengel, and I. Reid, "Efficient piecewise training of deep structured models for semantic segmentation," in *Proceedings of the IEEE CVPR*, 2016, pp. 3194–3203.

[68] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *arXiv preprint arXiv:1708.02002*, 2017.

[69] K. Liu, S. Peng, J. Wu, C. Zhai, H. Mamitsuka, and S. Zhu, "MeSHLabeler: improving the accuracy of large-scale MeSH indexing by integrating diverse evidence," *Bioinformatics*, vol. 31, no. 12, pp. i339–i347, 2015.

[70] R. Liu, H. Zhang, A. Shamir, and D. Cohen-Or, "A part-aware surface metric for shape analysis," in *Computer Graphics Forum*, vol. 28. Wiley Online Library, 2009, pp. 397–406.

[71] K. Lund and C. Burgess, "Producing high-dimensional semantic spaces from lexical co-occurrence," *Behavior Research Methods, Instruments, & Computers*, vol. 28, no. 2, pp. 203–208, 1996.

[72] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

[73] Y. Mao, C.-H. Wei, and Z. Lu, "Ncbi at the 2014 bioasq challenge task: Large-scale biomedical semantic indexing and question answering." in *CLEF (Working Notes)*, 2014, pp. 1319–1327.

[74] J. Masci, D. Boscaini, M. Bronstein, and P. Vandergheynst, "Geodesic convolutional neural networks on riemannian manifolds," in *Proceedings of the IEEE international conference on computer vision workshops*, 2015, pp. 37–45.

[75] J. D. Mcauliffe and D. M. Blei, "Supervised topic models," in *NIPS*, 2008, pp. 121–128.

[76] L. S. Medina, B. Bernal, C. Dunoyer, L. Cervantes, M. Rodriguez, E. Pacheco, P. Jayakar, G. Morrison, J. Ragheb, and N. R. Altman, "Seizure disorders: functional mr imaging for diagnostic evaluation and surgical treatmentâĂŤprospective study," *Radiology*, vol. 236, no. 1, pp. 247–253, 2005.

[77] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *ICLR Workshop*, 2013.

[78] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *NIPS*, 2013, pp. 3111–3119.

[79] S. Mori, W. E. Kaufmann, G. D. Pearlson, B. J. Crain, B. Stieltjes, M. Solaiyappan, and P. Van Zijl, "In vivo visualization of human neural pathways by magnetic resonance imaging," *Annals of neurology*, vol. 47, no. 3, pp. 412–414, 2000.

[80] M. Mostajabi, P. Yadollahpour, and G. Shakhnarovich, "Feedforward semantic segmentation with zoom-out features," in *Proceedings of the IEEE CVPR*, 2015, pp. 3376–3385.

[81] Y. Nakai, J.-w. Jeong, E. C. Brown, R. Rothermel, K. Kojima, T. Kambara, A. Shah, S. Mittal, S. Sood, and E. Asano, "Three-and four-dimensional mapping of speech and language in patients with epilepsy," *Brain*, vol. 140, no. 5, pp. 1351–1370, 2017.

[82] L. OâĂŹDonnell, M. Kubicki, M. E. Shenton, M. H. Dreusicke, W. E. L. Grimson, and C.-F. Westin, "A method for clustering white matter fiber tracts," *American Journal of Neuroradiology*, vol. 27, no. 5, pp. 1032–1036, 2006.

[83] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[84] Y. Papanikolaou, G. Tsoumakas, M. Laliotis, N. Markantonatos, and I. Vlahavas, "AUTH-Atypon at BioASQ 3: Large-scale semantic indexing in biomedicine," in *Working Notes for the Conference and Labs of the Evaluation Forum (CLEF), Toulouse, France*, 2015.

[85] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," in *NIPS-W*, 2017.

[86] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[87] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *EMNLP*, vol. 14, 2014, pp. 1532–1543.

[88] W. D. Penny, K. J. Friston, J. T. Ashburner, S. J. Kiebel, and T. E. Nichols, *Statistical parametric mapping: the analysis of functional brain images*. Elsevier, 2011.

[89] C. Pierpaoli, L. Walker, M. Irfanoglu, A. Barnett, P. Basser, L. Chang, C. Koay, S. Pajevic, G. Rohde, J. Sarlls *et al.*, "Tortoise: an integrated software package for processing of diffusion mri data," in *ISMRM 18th annual meeting*, 2010, p. 1597.

[90] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning, "Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*. Association for Computational Linguistics, 2009, pp. 248–256.

[91] N. Rasiwasia and N. Vasconcelos, "Latent dirichlet allocation models for image classification," *PAMI, IEEE Transactions on*, vol. 35, no. 11, pp. 2665–2679, 2013.

[92] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: towards real-time object detection with region proposal networks," *IEEE transactions on PAMI*, vol. 39, no. 6, pp. 1137–1149, 2017.

[93] A. Rios and R. Kavuluru, "Convolutional neural networks for biomedical text classification: Application in indexing biomedical articles," in *Proceedings of the 6th ACM Conference on Bioinformatics, Computational Biology and Health Informatics*, ser. BCB '15. New York, NY, USA: ACM, 2015, pp. 258–267. [Online]. Available: http://doi.acm.org/10.1145/2808719.2808746

[94] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[95] S. Rusinkiewicz, "Estimating curvatures and their derivatives on triangle meshes," in *2nd International Symposium on 3D Data Processing, Visualization and Transmission (3DPVT)*. IEEE, 2004, pp. 486–493.

[96] T. N. Sainath, A.-r. Mohamed, B. Kingsbury, and B. Ramabhadran, "Deep convolutional neural networks for lvcsr," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 8614–8618.

[97] B. Schölkopf, A. Smola, and K.-R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural computation*, vol. 10, no. 5, pp. 1299–1319, 1998.

[98] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers, "Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1285–1298, 2016.

[99] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, and R. Moore, "Real-time human pose recognition in parts from single depth images," *Communications of the ACM*, vol. 56, no. 1, pp. 116–124, 2013.

[100] Z. Shu, C. Qi, S. Xin, C. Hu, L. Wang, Y. Zhang, and L. Liu, "Unsupervised 3D shape segmentation and co-segmentation via deep learning," *Computer Aided Geometric Design*, vol. 43, pp. 39–52, 2016.

[101] O. Sidi, O. van Kaick, Y. Kleiman, H. Zhang, and D. Cohen-Or, "Unsupervised co-segmentation of a set of shapes via descriptor-space spectral clustering," *ACM TOG*, vol. 30, no. 6, p. 1, 2011.

[102] L. Sirovich and M. Kirby, "Low-dimensional procedure for the characterization of human faces," *Josa a*, vol. 4, no. 3, pp. 519–524, 1987.

[103] R. E. Smith, J.-D. Tournier, F. Calamante, and A. Connelly, "Anatomically-constrained tractography: improved diffusion mri streamlines tractography through effective use of anatomical information," *Neuroimage*, vol. 62, no. 3, pp. 1924–1938, 2012.

[104] S. Song and J. Xiao, "Deep sliding shapes for amodal 3D object detection in RGB-D images," in *Proceedings of the IEEE CVPR*, 2016, pp. 808–816.

[105] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[106] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *NIPS*, 2014, pp. 3104–3112.

[107] J. B. Tenenbaum, V. De Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[108] Theano Development Team, "Theano: A Python framework for fast computation of mathematical expressions," *arXiv e-prints*, vol. abs/1605.02688, May 2016. [Online]. Available: http://arxiv.org/abs/1605.02688

[109] J. D. Tournier, F. Calamante, and A. Connelly, "Improved probabilistic streamlines tractography by 2nd order integration over fibre orientation distributions," in *Proc. 18th Annual Meeting of the Intl. Soc. Mag. Reson. Med.(ISMRM)*, 2010, p. 1670.

[110] V. L. Towle, H.-A. Yoon, M. Castelle, J. C. Edgar, N. M. Biassou, D. M. Frim, J.-P. Spire, and M. H. Kohrman, "Ecog gamma activity during a language task: differentiating expressive and receptive speech areas," *Brain*, vol. 131, no. 8, pp. 2013–2027, 2008.

[111] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *NIPS*, 2017, pp. 6000–6010.

[112] A. Viswanathan and R. D. Freeman, "Neurometabolic coupling in cerebral cortex reflects synaptic more than spiking activity," *Nature neuroscience*, vol. 10, no. 10, p. 1308, 2007.

[113] Y. Wang, S. Asafi, O. van Kaick, H. Zhang, D. Cohen-Or, and B. Chen, "Active co-analysis of a set of shapes," *ACM TOG*, vol. 31, no. 6, p. 165, 2012.

[114] Y. Wang, M. Gong, T. Wang, D. Cohen-Or, H. Zhang, and B. Chen, "Projective analysis for 3D shape segmentation," *ACM TOG*, vol. 32, no. 6, p. 192, 2013.

[115] W. Wei, *Information Retrieval in Biomedical Research: From Articles to Datasets*. University of California, San Diego, 2017.

[116] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *European Conference on Computer Vision*. Springer, 2016, pp. 499–515.

[117] E. Wyllie, "Invasive neurophysiologic techniques in the evaluation for epilepsy surgery in children," *Epilepsy surgery*, pp. 409–412, 1991.

[118] Z. Xie, K. Xu, L. Liu, and Y. Xiong, "3D shape segmentation and labeling via extreme learning machine," in *Computer graphics forum*, vol. 33. Wiley Online Library, 2014, pp. 85–95.

[119] H. Xu, M. Dong, Y. Nakai, E. Asano, and J.-W. Jeong, "Automatic detection of eloquent axonal pathways in diffusion tractography using intracanial electrical stimulation mapping and convolutional neural networks," in *IEEE International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2018.

[120] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *ICML*, 2015, pp. 2048–2057.

[121] G. Xue, W. Dai, Q. Yang, and Y. Yu, "Topic-bridged plsa for cross-domain text classification," in *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2008, pp. 627–634.

[122] T. Xue, J. Liu, and X. Tang, "Example-based 3D object reconstruction from line drawings," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 302–309.

[123] L. Yeganova, D. C. Comeau, W. Kim, and W. J. Wilbur, "Text mining techniques for leveraging positively labeled data," in *Proceedings of BioNLP 2011 Workshop*. Association for Computational Linguistics, 2011, pp. 155–163.

[124] A. J. J. Yepes, J. G. Mork, D. Demner-Fushman, and A. R. Aronson, "Comparison and combination of several mesh indexing approaches," in *AMIA annual symposium proceedings*, vol. 2013. American Medical Informatics Association, 2013, p. 709.

[125] B. E. Yerys, K. F. Jankowski, D. Shook, L. R. Rosenberger, K. A. Barnes, M. M. Berl, E. K. Ritzl, J. VanMeter, C. J. Vaidya, and W. D. Gaillard, "The fmri success rate of children and adolescents: typical development, epilepsy, attention deficit/hyperactivity disorder, and autism spectrum disorders," *Human brain mapping*, vol. 30, no. 10, pp. 3426–3435, 2009.

[126] M. D. Zeiler, "Adadelta: an adaptive learning rate method," *arXiv preprint arXiv:1212.5701*, 2012.

[127] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*. Springer, 2014, pp. 818–833.

[128] X. Zhang and Y. LeCun, "Text understanding from scratch," *arXiv preprint arXiv:1502.01710*, 2015.

[129] Y. Zhang, M. Brady, and S. Smith, "Segmentation of brain mr images through a hidden markov random field model and the expectation-maximization algorithm," *IEEE transactions on medical imaging*, vol. 20, no. 1, pp. 45–57, 2001.

[130] Z. Zhang and H. Zha, "Principal manifolds and nonlinear dimensionality reduction via tangent space alignment," *SIAM journal on scientific computing*, vol. 26, no. 1, pp. 313–338, 2004.

[131] G. Zhong, L.-N. Wang, X. Ling, and J. Dong, "An overview on data representation learning: From traditional feature learning to recent deep learning," *The Journal of Finance and Data Science*, 2017.

ABSTRACT

REPRESENTATION LEARNING WITH CONVOLUTIONAL NEURAL NETWORKS

by

HAOTIAN XU

December 2018

Advisor: Dr. Ming Dong

Major: Computer Science

Degree: Doctor of Philosophy

Deep learning methods have achieved great success in the areas of Computer Vision and Natural Language Processing. Recently, the rapidly developing field of deep learning is concerned with questions surrounding how we can learn meaningful and effective representations of data. This is because the performance of machine learning approaches is heavily dependent on the choice and quality of data representation, and different kinds of representation entangle and hide the different explanatory factors of variation behind the data [11].

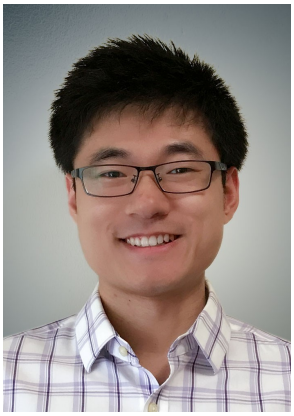
In this dissertation, we focus on representation learning with deep neural networks for different data formats including text, 3D polygon shapes, and brain fiber tracts.

First, we propose a topic-based word representation learning approach for text classification. The proposed approach takes global semantic relationship between words over the whole corpus into consideration and encodes the relationships into distributed vector representations with continuous Skip-gram model. The learned representations which capture a large number of precise syntactic and semantic word relationships are taken as input of Convolution Neural Networks for classification. Our experimental results show the effectiveness of the proposed method on indexing of biomedical articles, behavior code annotation of clinical text fragments, and classification of news groups.

Second, we present a 3D polygon shape representation learning framework for shape segmentation. We propose Directionally Convolutional Network (DCN) that extends convolution operations from images to the polygon mesh surface with rotation-invariant property. Based on the proposed DCN, we learn effective shape representations from raw geometric features and then classify each face of a given polygon into predefined semantic parts. Through extensive experiments, we demonstrate that our framework outperforms the current state-of-the-arts.

Third, we propose to learn effective and meaningful representations for brain fiber tracts using deep learning frameworks. We handle the highly unbalanced dataset by introducing asymmetrical loss function for easily classified samples and hard classified ones. The training loss avoids to be dominated by the easy samples and the training step is more efficient. In addition, we learn more effective and meaningful representations by introducing deeper network and metric learning approaches. Furthermore, we propose to improve the interpretability of our framework by inducing attention mechanism. Our experimental results show that our proposed framework outperforms current golden standard significantly on the real-world dataset.

AUTOBIOGRAPHICAL STATEMENT



Haotian Xu received his B.S. degree in Electrical Engineering from University of Science and Technology of China, P.R. China in 2012. He received his M.S. degree in Computer and Information Science from Temple University in 2013. He is currently a Ph.D. candidate in the Machine Vision and Pattern Recognition Laboratory, Department of Computer Science, Wayne State University. His research interests include computer vision, machine learning, data mining, and multimedia analysis.