Collecting, Analyzing and Predicting Socially-Driven Image Interestingness

Eloïse Berson, Ngoc Duong, Claire-Hélène Demarty

Collecting, Analyzing and Predicting Socially-Driven Image Interestingness

Eloïse Berson
Supélec
Cesson Sévigné, France
eloise.berson@gmail.com

Ngoc Q. K. Duong
InterDigital R&D France
Cesson Sévigné, France
quang-khanh-ngoc.duong@technicolor.com

Claire-Hélène Demarty
InterDigital R&D Francer
Cesson Sévigné, France
claire-helene.demarty@technicolor.com

***Abstract*—*Interestingness* has recently become an emerging concept for visual content assessment. However, understanding and predicting image interestingness remains challenging as its judgment is highly subjective and usually context-dependent. In addition, existing datasets are quite small for in-depth analysis. To push forward research in this topic, a large-scale interestingness dataset (images and their associated metadata) is described in this paper and released for public use. We then propose computational models based on deep learning to predict image interestingness. We show that exploiting relevant contextual information derived from social metadata could greatly improve the prediction results. Finally we discuss some key findings and potential research directions for this emerging topic.**

***Index Terms*—Image interestingness; content and social interestingness; Flickr; LaFin dataset; contextual information; deep learning.**

I. Introduction

There has been a recent surge of interest in understanding and predicting user perceptions on visual content. Such perceptions have moved toward challenging high level subjective concepts such as emotion [1], [2], popularity [3], [4], memorability [5]–[8], and interestingness [9]–[12]. One reason for this surge is the exponential increase of sharing of images and videos, that raises an essential need for filtering the content in each large-scale retrieval and recommendation system. Such a filtering process is impossible without a clear understanding of the content's subjective meaning. Focusing solely on interestingness and how contextual information may help its prediction, this paper presents our contributions to this emerging topic in multimedia assessment.

Interestingness usually refers to arousing interest, curiosity, as well as the *ability of holding or catching attention* [13]. This important notion has been intensively investigated for years in psychology and vision research as it involves human perception [14], [15]. These studies revealed that interest is determined by certain factors like *novelty*, *uncertainty*, *conflict*, *complexity*, and their combinations. This finding was also supported in the appraisal theory presented in [16], where the author explained that appraisals like the novelty, the comprehensibility, and the complexity of an event are likely to arouse interest in this event.

Following the existing literature, we distinguish two different notions, namely *socially-driven* interestingness and *content-driven* interestingness. The latter refers to human annotations that assess interestingness solely on the perceived content. Content-driven interestingness has been extensively addressed in the MediaEval benchmark on Predicting Media Interestingness[1] [17]. In this paper, we focus on *socially-driven* interestingness for which the definition is derived from media sharing websites such as Flickr and Pinterest where posted images are assigned with interestingness labels. Such labels are usually inferred based on social aspects such as number of views, tags, comments, user reputations and viewer's profiles[2] but also on the characteristics of the content itself. In that sense, socially-driven interestingness differs from *popularity* which solely corresponds to the number of likes, favorites, reshares or views attached to a given content[3] [3], [4], even though the two notions might be linked. The assumption that socially-driven interestingness also derives from the content itself was in part proven in [18] where the authors achieved good performance for the prediction of Flickr interestingness scores with high level aesthetics attributes extracted from the content as input to an SVM classifier.

Moving toward computational aspects, visual interestingness have recently been explored in multimedia and computer vision communities [11], [19]–[21]. As an example, Gygli *et al.* [9] investigated the use of various features (RGB values, GIST, spatial pyramids of SIFT histograms, colorfulness, complexity, contrast and edge distributions, arousal and composition of parts) that computationally capture unusualness, aesthetics, and general preferences. These features were used to train a Support Vector Regression (SVR) model for interestingness prediction. Authors in [22] investigated the correlation of color, texture, edge and saliency features with both content-driven and socially-driven interestingness. Recently, deep learning has also been investigated for the task, but with small available datasets, its power was limited [23]. Readers are referred to [11], [12] for a more comprehensive survey on the existing work.

In the literature, some work also investigated the use of context for the prediction of interestingness. Chu *et al.* [24] investigated the effect of familiarity (facial familiarity and

The work was done while the authors were with Technicolor.

[1]http://www.multimediaeval.org/mediaeval2017/mediainterestingness/
[2]https://www.flickr.com/explore/interesting/
[3]http://www.acmmm.org/2017/challenge/social-media-prediction/

familiarity with image context) in the perceived interestingness of images. Liu *et al.* [25] built a computational model based on viewer data and viewer's profiles to estimate the interestingness of images. Rajani *et al.* [26] predicted interestingness of fashion products for online shopping, where they assumed that all pins related to fashion on Pinterest are interesting [26]. The authors used Word2Vec [27] to transform textual data from pins to machine learning features. Closer to our work, although the authors are targeting popularity, Gelli *et al.* [28] tried to use visual sentiments conveyed by the images together with contextual features extracted from tags, descriptions and titles of Flickr images, to predict popularity of images.

This paper states several contributions to the emerging topic of image interestingness as follows. In Section II we first describe a large-scale socially-driven interestingness dataset, which consists of more than 123k Flickr images together with their metadata and interestingness labels. We then present some insight on the associated metadata, that may help for the understanding of the socially-driven interestingness concept. The dataset will be made publicly available for the community so as to push forward research in the domain[4]. In Section III, we propose computational models to predict interestingness of images. From the obtained results, we highlight the importance of contextual information and discuss some key findings in Section IV.

II. LaFin: Large-scale Flickr interestingness dataset

Apart from some very recent dataset[5], no large-scale socially-driven interestingness dataset with additional contextual information was existing up to now. Some datasets exist that come from Flickr's images but either they are not publicly released [18], or they are not associated with interestingness scores[6], or they contain only images with positive interestingness labels, and no negative samples [29]. This motivated the construction of the LaFin dataset, which contains binary interestingness label for images. Note that defining a relevant human annotation protocol to collect such social labels is far from being easy due to the subjectivity of the task, especially when one wants to build annotations at large. Thus, we rely on available Flickr interestingness labels for our dataset. We believe that this initiative is already providing valuable information and can be considered as a first step toward building a new dataset with human labels in the future.

A. Image dataset collection

The LaFin dataset consists of more than 123k images equally balanced between interesting and non interesting samples and collected by following a two-step process: 1/ a first set of 200k images was collected from Flickr[7], 2/ followed by a filtering of the images based on their associated metadata.

A first set of 100k interesting images was gathered through Flickr Interestingness API[8], at a rate of 500 images per day (Flickr maximal limit) for several months. When ever possible, each image was downloaded at the highest available resolution. No clear description of the algorithm used to compute Flickr interestingness labels is available but refering to [30], we may infer that Flickr interestingness label is related to at least the number of views, the number of comments and who comments on a specific picture, tags applied to the picture, Flickr discussion groups in which the picture appears, favorites, a.k.a Flickr bookmarking, of the picture, existing relationship between the owner and a viewer of the picture and time varying behavior of the above factors. From this list of probable factors used in the computation of interestingness labels, it appears clearly that what Flickr provides is an estimation of a socially-driven interestingness level, more than of a content-driven interestingness level, as defined in Section I. It should also be noted that most of the images returned by Flickr API are owned by semi professional or professional photographers, leading to the conclusion that the subset of interesting images might be biased towards aesthetics images.

In parallel, 100k additional Flickr images were collected from the same days, to build a subset of non interesting images, by following the simple rule that one image is non interesting as long as it does not belong to the list of interesting images provided by Flickr. Note that we processed so because Flickr is not providing any API to collect non interesting images and it does not also provide any interestingness scores with its interesting images. With the assumption that Flickr will provide the first 500 most interesting images per day through its API, we can only assume that these non interesting samples are not ranked among the first 500 interesting images.

B. Contextual information from the metadata

Together with the images, classified into interesting/non interesting, we also collected all additional socially-driven metadata available on Flickr's web site in an attempt to provide the associated contextual information that might influence the appraisal of users. Such metadata include title, description, associated tags, owners information, license, upload date, comments, number of views, original format, location, *etc*. Although some metadata are equally present in both classes, *e.g., locations*, *tags*, *titles*, *views*, others have an unbalanced distribution over the two classes, *e.g., comments* and *descriptions*. Indeed a single image in the interesting class may have several *comments* and tens of *views* attached to it, whereas most of non interesting images will have no comments, nor views. Indeed, interesting images got 97.2% of the total number of *views*.

With the aim of further modeling contextual features, we proceeded to a qualitative study of *descriptions*, *titles* and *tags* attached to LaFin images. *Descriptions*, as it could have been expected, often only correspond to a plain description of the visual content. It does not really bring new information regarding context, nor does it really directly relate with the

[4]https://www.technicolor.com/dream/research-innovation/lafin-dataset. Please note that because of copyright issues, we do not release the images but their links to Flickr's web site.

[5]https://github.com/gyglim/personalized-highlights-dataset

[6]http://www.acmmm.org/2017/challenge/social-media-prediction/

[7]https://www.flickr.com/

[8]https://www.flickr.com/services/api/flickr.interestingness.getList.html

social aspect of interestingness. This is also true for *titles* for a non negligible part of the images, although they might also contain other additional information than pure description (*e.g., May be next time*, *Deep thoughts*). As for *tags*, although some of them also correspond to descriptive attributes of the images (*sunset*, *street*), they seem to bring other information relative to context such as the location (*Japan*, *London*), the quality and type of the image (*canon*, *black and white*), its topic (*art*, *portrait*, *architecture*), *etc*. We believe that this additional contextual information, different from a pure description of the image content might help in the task of predicting the interestingness of content. Thus in the following section, we further filter our dataset thanks to the tags and we use them in Section III as input to our computational models.

C. Filtering and additional features

With the target of keeping the largest possible number of images from the initial set of 200k images, while retaining the balance between the two classes interesting/non interesting in terms of contextual information, we further filtered the original image set to keep only images with at least one *tag*. Although, from the above study, *titles* also seem to bring contextual information, filtering on *tags* was maximizing the size of LaFin dataset. LaFin dataset finally contains roughly 123k images, equally spread between interesting and non interesting samples. Additionally to the images, their binary labels, and associated metadata, some precomputed features are provided: CNNs, semantic features that derived from image captioning and Word2Vec representations of Flickr tags. See Section III-A for a complete description of these features.

III. Image interestingness prediction

The global workflow of our investigated computational models for image interestingness prediction is depicted in Figure 1. The system takes as input images and their associated *tags* and outputs the corresponding interestingness labels. As far as the modeling is concerned, we targeted to: 1) Find what different features bring to the prediction performance. 2) Study the influence of context information coming from the *tags* on the prediction. 3) Prove that social interestingness is also related to the content itself as content-based features help improving the prediction when compared to tag features alone. 4) Finally we would like to note that, as far as we know, high level semantic image captioning features are used for the first time for the task. For each model, the LaFin dataset was split into 64% for training, 16% for validation, and 20% for testing. Details of each step are given below.

A. Feature extraction

CNN features: As CNN provides a powerful feature representation useful for many different tasks, we used the well-known VGG16 network [31] pre-trained on the ImageNet dataset for the extraction of a first image feature. The feature is extracted from the last fully-connected layer before the softmax and has a dimension of 4096.

Image captioning-based features (IC): As scene semantics and high-level visual attributes (such as emotions, actions, movements, appearance of objects, *etc*.) may also derive human interest in an image, we further investigated the use of semantic features as computed by an image captioning (IC) system [32]. Such an IC model builds an encoder comprising a CNN and a long short-term memory recurrent network (LSTM) for learning a joint image-text embedding. The CNN image feature and the Word2Vec representation [27] of the image caption are projected on a 2D embedding space which enforces the alignment between an image and its corresponding semantic caption. We extracted the projected CNN feature for each image, as logically it should contain some semantic information expected to be helpful for the prediction of image interestingness. This feature has a dimension of 1024.

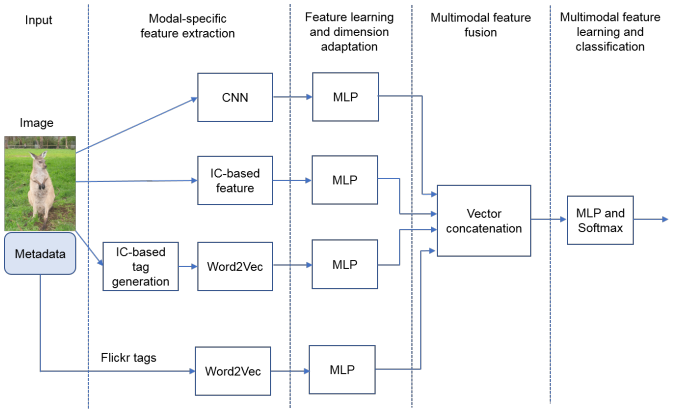

Fig. 1. General workflow of the proposed computational models.

Word2Vec features from Flickr tags: With the target of investigating the use of contextual information, we used textual information coming from the *tags* associated to LaFin images. We are especially interested in discovering how IC features compare with those contextual features, as both convey semantic information. We extracted a 300-dimensional word embedding vector [27] for each tag associated with one image. Those Word2Vec features are then averaged to obtain a single feature vector per image. Note that we also filtered all tags containing the string *explor* as potentially referring to Explore[9].

Word2Vec features from IC-based generated tags: For each image and its projection in the IC embedding space, we extract the 10 nearest neighbors. These neighbors in the feature space allow us to artificially build 10 tags related to the input image, called *generated tags*. As the COCO[10] dataset used in the training of the IC system is large enough, we were able to get relevant tags for each image. From those *generated tags*, we compute Word2Vec features similarly as above.

B. Feature learning, fusion, and classification

Each modal-specific feature vector is passed through a multilayer perceptron (MLP), which contains up to two fully connected layers, for higher-level feature learning. Another

[9]https://www.flickr.com/explore
[10]http://mscoco.org/

important benefit of these MLP is that they allow to control the output feature dimension (*i.e.,* by choosing a relevant number of hidden units in the last layer of each MLP). Thus, even if the original feature dimensions are highly unbalanced: 4096 for CNNs, 1024 for IC features, and 300 for Word2Vec, they have all balanced dimensions after this first learning phase (features with higher dimension are reduced to match the lowest feature dimension). These balanced modal-specific features are then combined by vector concatenation to form a multimodal feature vector. Note that we also tried direct concatenation of all original features without dimension reduction, meaning that no modal-specific learning was performed, but performances were lower.

The fused multimodal feature is fed to an MLP with two fully connected layers for the higher-level multimodal feature learning step. A final classification layer is added (*i.e.,* softmax [33] followed by quantization) to produce the binary prediction results. For each MLP in the global workflow, we retain a single set of parameters after some investigation on the performances (RMSProp, lr=0.0001, batchsize=512, ReLu, dropout=0.5). Optimization was conducted on validation loss.

IV. Experiment result and discussion

We performed experiments with the use of different input features and their combinations as summarized in Table I. Results are presented in terms of accuracy for the training, validation and test sets. It clearly appears that performance remains quite similar for the three sets, showing that no over-fitting occurred and that the investigated models generalize well to new data. From models based on individual features (first four rows), it is shown that IC features do bring valuable information for the task as they perform quite as good as classic VGG16 features. Furthermore, both are complementary (see 5th row) and reach a high accuracy of more than 83% on the test set, proving that IC features bring more semantic information to the task. Nevertheless, social information collected from LaFin *tags* alone outperforms all other specific features with a final accuracy of 89.63% on the test set. Note that we did also experiment with textual features extracted from LaFin images' *titles*, by similarly generating Word2Vec features. Nevertheless models with such textual features led to lower performances than with *tags*. Thus we did not consider them in the remaining experiments. Contrary to *Flickr tags*, *generated tags* were not able to capture high predictive information: used alone they reach a low accuracy of 65% on both the train and test sets and seem to be redundant with classic VGG16 features as the combination does not really bring any benefit (see 6th row). Going further with the combination of features, we then tested VGG166 and Word2Vec features together and, in the case of *Flickr tags*, got an improvement of accuracy (91%) that shows that both features contain significantly different information. A full combination of all three VGG16, IC features and Flickr tags further slightly improves the overall accuracy (92.6% on the test set). As a conclusion, this last model successfully predicts LaFin interestingness labels, thanks to a combination of both content and social information. These results are to be compared with the work of [18] which obtained a best recall value of 70% for a precision value of roughtly 85% on a dataset of 40k Flickr images.

Inputs and *features*	Accuracy (%)		
	Train	Validation	Test
Image *VGG16*	78.07	75.54	76.45
Image *IC*	76.3	75.53	76.35
Flickr tags *Word2Vec*	89.96	89.68	89.63
Generated tags *Word2Vec*	65.42	63.47	65.12
Image *VGG16+IC*	85.27	83.34	83.59
Image + Generated tags *VGG16+Word2Vec*	78.82	76.48	75.65
Image+Flickr tags *VGG16+Word2Vec*	**92.76**	**90.99**	**91.08**
Image+Flickr tags *VGG16+IC+Word2Vec*	**93.72**	**92.46**	**92.59**

TABLE I
PREDICTION RESULTS IN TERMS OF ACCURACY OBTAINED BY MODELS WITH DIFFERENT SETS OF INPUT FEATURES ON LAFIN.

We have shown that the use of contextual information in the form of image tags enables to increase the performances of *socially-driven* interestingness prediction. Nevertheless, it must be noted that the single use of this information alone despite all the other social metadata available did perform really well, although it seems that Flickr's assessment of interestingness was based on more than only the image tags. One interpretation for this might be that contextual information is redundant in the available metadata, *e.g.,* the number of views is partially correlated with the tags, comments, *etc*. The high performance of our classifiers for LaFin dataset also proves that the two classes, non interesting and interesting, are separable and that our strategy of collecting the non interesting images was correct, even though we did not have access to the description of Flickr's interestingness prediction. This reinforces the quality of LaFin dataset that is publicly released to the community.

V. Conclusion

In this paper we disclosed a new large-scale socially-driven interestingness dataset (LaFin) collected from Flickr website and presented an analysis of the associated metadata. We built computational models for interestingness prediction where different types of features derived from different sources of information were taken into account. The prediction results revealed the correlation between contextual information and socially-driven interestingess. Future work may be devoted to re-annotate the LaFin dataset with a protocol for content-driven interestingness inspired from [17] and compare the dataset's social labels with content-driven labels. Then a subsequent in-depth study of the differences between these two notions of interestingness would be of great interest.

REFERENCES

[1] U. Rimmele, L. Davachi, R. Petrov, S. Dougal, and E. A. Phelps, "Emotion enhances the subjective feeling of remembering, despite lower accuracy for contextual details," *Psychology Association*, 2011.

[2] M. Soleymani, G. Chanel, J. J. M. Kierkels, and T. Pun, "Affective characterization of movie scenes based on content analysis and physiological changes," *Int. J. Semantic Computing*, vol. 3, pp. 235–254, 2009.

[3] A. Khosla, A. Sarma, and R. Hamid, "What makes an image popular?" in *Proceedings of the International conference on World Wide Web*, 2013, pp. 867–876.

[4] S. Cappallo, T. Mensink, and C. G. M. Snoek, "Latent factors of visual popularity prediction," in *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval (ICMR)*, 2015, pp. 195–202.

[5] P. Isola, J. Xiao, A. Torralba, and A. Oliva, "What makes an image memorable?" in *Proceedings of the 2011 IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 145–152.

[6] H. Squalli-Houssaini, N. Q. K. Duong, G. Marquant, and C.-H. Demarty, "Deep learning for predicting image memorability," in *Proceedings of the IEEE International Conference on Audio, Speech and Language Processing (ICASSP)*, 2018.

[7] A. Siarohin, G. Zen, C. Majtanovic, X. Alameda-Pineda, E. Ricci, and N. Sebe, "How to make an image more memorable?: A deep style transfer approach," in *Proceedings of the ACM on International Conference on Multimedia Retrieval (ICMR)*, 2017, pp. 322–329.

[8] R. Cohendet, C.-H. Demarty, N. Q. K. Duong, M. Sjöberg, B. Ionescu, and T. T. Do, "Mediaeval 2018 predicting media memorability task," in *Proceedings of the MediaEval Workshop*, 2018.

[9] M. Gygli, H. Grabner, H. Riemenschneider, F. Fabian, and L. V. Gool, "The interestingness of images," in *Proceedings of the IEEE International Conference on Computer Vision (CVPR)*, 2013, pp. 1633–1640.

[10] G. Zen, P. de Juan, Y. Song, and A. Jaimes, "Mouse activity as an indicator of interestingness in video," in *Proceedings of the ACM on International Conference on Multimedia Retrieval (ICMR)*, 2016, pp. 47–54.

[11] C.-H. Demarty, M. Sjöberg, G. Constantin, N. Q. K. Duong, B. Ionescu, T. T. Do, and H. Wang, "Predicting interestingness of visual content," in *Visual Content Indexing and Retrieval with Psycho-Visual Models*. Springer, 2017.

[12] M. Constantin, M. Redi, G. Zen, and B. Ionescu, "Computational understanding of visual interestingness beyond semantics: literature survey and analysis of covariates," *ACM Computing Surveys*, 2019.

[13] D. Berlyne, *Conflict, arousal and curiosity*. Mc-Graw-Hill, 1960.

[14] A. Chen, P. W. Darst, and R. P. Pangrazi, "An examination of situational interest and its sources," *British Journal of Educational Psychology*, vol. 71, no. 3, pp. 383–400, 2001.

[15] L. Elazary and L. Itti, "Interesting objects are visually salient," *Journal of vision*, vol. 8, no. 3, pp. 3–3, 2008.

[16] P. J. Silvia, "What is interesting? exploring the appraisal structure of interest." *Emotion*, vol. 5, no. 1, p. 89, 2005.

[17] C.-H. Demarty, M. Sjöberg, B. Ionescu, T. T. Do, M. Gygli, and N. Q. K. Duong, "Mediaeval 2017 predicting media interestingness task," in *Proceedings of the MediaEval Workshop*, 2017.

[18] S. Dhar, V. Ordonez, and T. L. Berg, "High level describable attributes for predicting aesthetics and interestingness," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 1657–1664.

[19] X. Amengual, A. Bosch, and J. L. Rosa, *Review of Methods to Predict Social Image Interestingness and Memorability*. Springer, 2015, pp. 64–76. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-23192-1_6

[20] M. Soleymani, "The quest for visual interest," in *Proceedings of the 23rd ACM International Conference on Multimedia*, 2015, pp. 919–922.

[21] H. Grabner, F. Nater, M. Druey, and L. Van Gool, "Visual interestingness in image sequences," in *Proceedings of the 21st ACM International Conference on Multimedia*, New York, USA, 2013, pp. 1017–1026.

[22] L.-C. Hsieh, W. H. Hsu, and H.-C. Wang, "Investigating and predicting social and visual image interestingness on social media by crowdsourcing," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 4309–4313.

[23] Y. Shen, C.-H. Demarty, and N. Q. K. Duong, "Technicolor@MediaEval 2016 Predicting Media Interestingness Task," in *Proceedings of the MediaEval Workshop*, Hilversum, Netherlands, October 2016.

[24] S. L. Chu, E. A. Fedorovskaya, F. K. Quek, and J. Snyder, "The effect of familiarity on perceived interestingness of images." in *Human Vision and Electronic Imaging*, 2013.

[25] M. B. Liu, P. Kato, and K. Tanaka, "Estimating interestingness of images based on viewer data," in *Proceedings of the 7th Forum on Data Engineering and Information Management (DEIM)*, 2015.

[26] N. Rajani, K. Rohanimanesh, and E. Oliveira, "Identifying interestingness in fashion e-commerce using pinterest data," https://www.cs.utexas.edu/ nrajani/srs.pdf, 2015.

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

[28] F. Gelli, T. Uricchio, M. Bertini, A. Del Bimbo, and S.-F. Chang, "Image popularity prediction in social media using sentiment and context features," in *Proceedings of the 23rd ACM International Conference on Multimedia*, 2015, pp. 907–910.

[29] M. J. Huiskes, B. Thomee, and M. S. Lew, "New trends and ideas in visual concept detection: The mir flickr retrieval evaluation initiative," in *Proceedings of the ACM International Conference on Multimedia Information Retrieval (ICMR)*, New York, USA, 2010, pp. 527–536.

[30] D. Butterfield, C. FakeCallum, H. Mourachov, and S. Mourachov, "Interestingness ranking of media objects," February 2006.

[31] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014. [Online]. Available: http://arxiv.org/abs/1409.1556

[32] R. Kiros, R. Salakhutdinov, and R. Zemel, "Unifying visual-semantic embeddings with multimodal neural language models," *CoRR*, vol. abs/1411.2539, 2014. [Online]. Available: http://arxiv.org/abs/1411.2539

[33] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., 2006.