Learning Fused Representations for Large Scale Multi-modal Classification

Shah Nawaz[1**], Alessandro Calefati[1**], Muhammad Kamran Janjua[2], Muhammad Umer Anwaar[3] and Ignazio Gallo[1*]

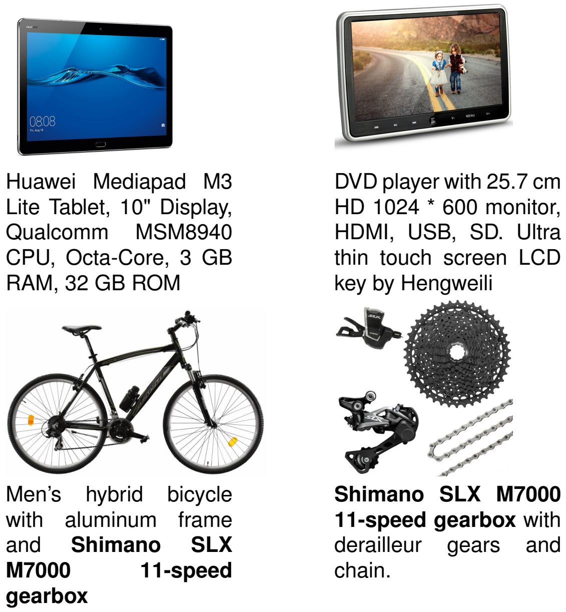
[1] Department of Theoretical and Applied Science, University of Insubria, Varese, Italy
[2] School of Electrical Engineering and Computer Science, National University of Sciences and Technology, Islamabad, Pakistan
[3] Technical University of Munich, Germany
[*] Corresponding Author
[**] Equal Contribution of these authors to this manuscript

Abstract—Multi-modal strategies combine different input sources into a joint representation that provides enhanced information than the uni-modal. In this paper, we present a novel multi-modal approach which fues image and encoded text description to obtain an information enriched image. The approach casts encoded text obtained from Word2Vec word embedding into visual embedding to be concatenated with the image. We employ standard Convolutional Neural Networks (CNNs) to learn representations of information enriched images. Finally, we compare our approach with uni-modal and their combination on three large scale multi-modal datasets. Our findings indicate that the joint representation of encoded text and image in feature space improves the multi-modal classification performance aiding the interpretability.

Index Terms—Multi-modal data fusion, Image and text fusion, Text encoding

I. INTRODUCTION

Information in real-world applications usually comes from multiple sources. Images are often associated with tags or captions; for example in the world of e-commerce products on sale are displayed using one or more images with one or more text descriptions such as product title, summary and technical details. Each source is characterized by distinct statistical properties that makes it difficult to create a joint representation that uniquely captures the "concept" in the real-world. For example, Figure 1 shows four advertisements, where, in the first row, two objects have seemingly similar images but different text descriptions, conversely, in the second row, we have two different images but similar text descriptions. For an image classification model it would not be easy to distinguish two images on the first row while it would have no difficulty in distinguishing images on the second one. Similarly, for a text classification model it would be difficult to classify two text descriptions on the second row while it would have no difficulty in classifying correctly descriptions shown on the first one. Such scenarios present a challenge to create a joint representation of an image and associated text description. This leads us to create a representation for such classification problem. This representation can exploit such scenarios to remove ambiguity and improve classification performance.



Huawei Mediapad M3 Lite Tablet, 10" Display, Qualcomm MSM8940 CPU, Octa-Core, 3 GB RAM, 32 GB ROM

DVD player with 25.7 cm HD 1024 * 600 monitor, HDMI, USB, SD. Ultra thin touch screen LCD key by Hengweili

Men's hybrid bicycle with aluminum frame and **Shimano SLX M7000 11-speed gearbox**

Shimano SLX M7000 11-speed gearbox with derailleur gears and chain.

Fig. 1. In the upper row, two examples of ambiguous images that can be disambiguated through analysis of the respective text description. In the lower row, two examples of ambiguous text description that can be disambiguated through analysis of respective images.

The use of joint representation based on image and text features is extensively employed on a variety of tasks including modeling semantic relatedness, compositionality, classification and retrieval [1]–[5]. Typically, in such approach, image features are extracted using CNNs. Whereas, to generate text features, Bag-of-Words models or Log-linear Skip-gram Models [6] are commonly employed. This represents a challenge to find relationships between features of multiple modalities along with representation, translation, alignment, and co-learning as stated in [7]. Traditionally, there are two general strategies for text and image fusion referred to as early and late fusion [7], [8]. In early fusion [3], [9], features from each modality are concatenated in a single vector and fed as input to a classification unit. In contrast, late fusion [10], [11] uses decision values from each classification unit and fuses them using a fusion mechanism employing a weight sharing strategy. The work in [10] showcases a comparative study of multi-modal fusion methods to perform multi-modal classification in real-world scenarios. Specifically, in [10], late fusion produced better performance compared to early fusion method, however, late fusion comes with the price of an increased learning effort. Recently, [5], fuses data from discrete (text) and continuous (image) domains and showcases the efficiency of fusion strategies in terms of learning and computational expense. Our approach is similar to early fusion strategy, where a single classifier is needed to perform multi-modal classification, as stated in [7], [8]. However,

Corresponding author: I. Gallo (e-mail: ignazio.gallo@uninsubria.it).
Associate Editor: Alan Smithee.
Digital Object Identifier 10.1109/LSENS.2017.0000000

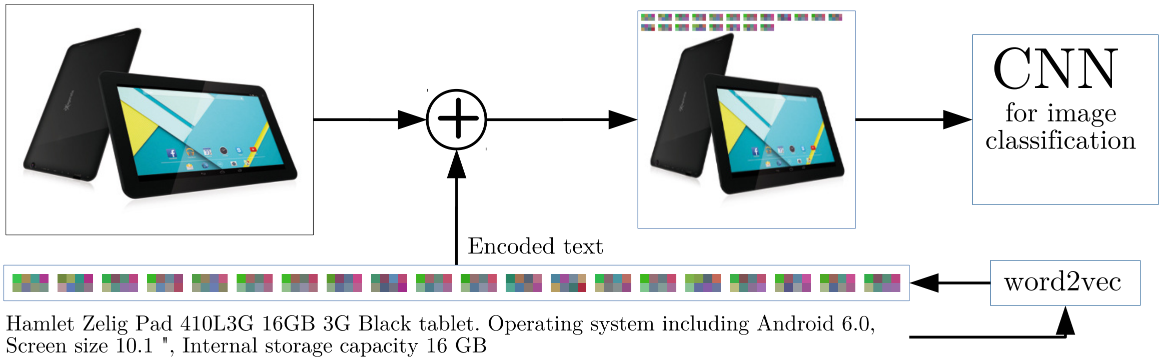

CNN
for image
classification
Encoded text
word2vec
Hamlet Zelig Pad 410L3G 16GB 3G Black tablet. Operating system including Android 6.0,
Screen size 10.1 ", Internal storage capacity 16 GB

Fig. 2. The proposed encoded text and image fusion model for deep multi-modal classification. The text is encoded within the image so that the CNN model can exploit semantics along with the information of the image.

we concatenate encoded text features into an image to obtain an information enriched image. Finally, an image classification model is trained and tested on these images. Intuitively, concatenating text descriptions onto images may not sound motivating due to several reasons. Since the idea is overlaying the encoded tex tdescription into an image, it might affect the image perception in general. However, we observed that the joint representation of encoded text and image improves the multi-modal classification.

With this work, we present a novel strategy which combines a text encoding schema to fuse image and text in a information enriched image. The encoding schema is based on Word2Vec word embedding [12] that transforms embedding to encoded text. We fuse both text description and image into a single source so that it can be used with a CNN architecture. We demonstrate that by adding encoded text information in images better classification results are obtained compared to the best one obtained using a single modality. Part of the work presented in this paper was published in [13] for timely dissemination of this work. This paper is a substantially extended version of the previous conference publication.

II. THE PROPOSED APPROACH

Multi-modal strategies fuse image and text description into integrated representation. We obtain transform Word2Vec word embedding [12] into visual embedding from our earlier work [13]. However, we fuse encoded text with associated image to obtain an information enriched image. An example of information enriched image is shown in Figure 2. This image can be fed to a CNN configured for image classification to learn representation. In other words, multi-modal classification problem is transformed into a typical image classification task. Finally, state-of-the-art image classification network can be employ. Figure 3 shows the behavior of a CNN that receives a joint representation with our approach. In the same figure we can notice how some convolutive filters of the first two layers, are activated both on the image and on the encoded text. This approach is suitable to be adopted in a multi-modal strategy because a CNN model can extract information from both sources (Text/Image).

A. Text Encoding Scheme

The encoding approach is based on Word2Vec word embeddings [12]. Our previous work [13] transforms word embedding, gathered from Word2Vec, into a visual domain keeping relationships between words. A word t_k belonging to a text document D_i is encoded into an artificial image of size $W \times H$. The approach uses a dictionary $F(t_k, v_k)$ with each word t_k associated with a feature vector $v_k(t_k)$ obtained from a trained version of Word2Vec word embedding. Given a word t_k, we obtain a visual word $\hat{t}_k$ having width V that contains values of a feature vector, called superpixels (see example in Fig. 4). A superpixel is a square area of size $P \times P$ pixels with uniform color that represents a contiguous sequence of features $(v_{k,j}, v_{k,j+1}, v_{k,j+2})$ extracted as a sub-vector of v_k. The components $v_{k,j}$ are normalized to assume values in the interval $[0 \ldots 255]$ with respect to k. We interpret triplets from the feature vector v_k as RGB sequence. In other words, the approach uses feature vector with a length multiple of 3. Another parameter associated with the approach is s, which is a blank space in pixels around each visual word $\hat{t}_k$. Each $\hat{t}_k$ is placed in the following image coordinate (x, y) as shown in Eq. 1 An example for the resulting graphical representation is given in Figure 4.

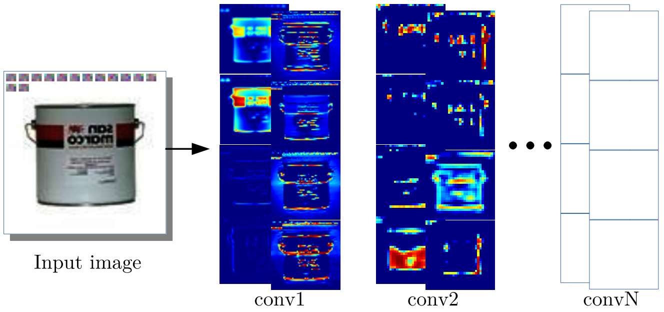


Fig. 3. An example of an input image and some feature maps of the first two convolutive layers of the CNN used for the Ferramenta dataset. These examples of feature maps show significant activations on both textual information and image details.

$$\begin{aligned} x &= k(V+s) \mod (W-V) \\ y &= (V+s)\frac{k(V+s)}{(W-V)} \end{aligned} \tag{1}$$

III. EXPERIMENTS

A. Datasets

Typically, multi-modal dataset consists of an image and associated text description. In this work, we use three large scale multi-modal datasets to show the applicability of our approach to various domains.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LSENS.2018.2880790, IEEE Sensors Letters

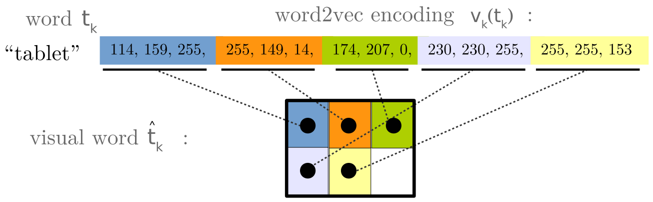


Fig. 4. In this toy example, the word "*tablet*" is encoded into a visual word $\hat{t}_k$ based on Word2Vec feature vector with vector length 15. This visual word can be transformed into different shapes, varying parameter V (in this example $V = 3$ superpixels). Note last square box is empty

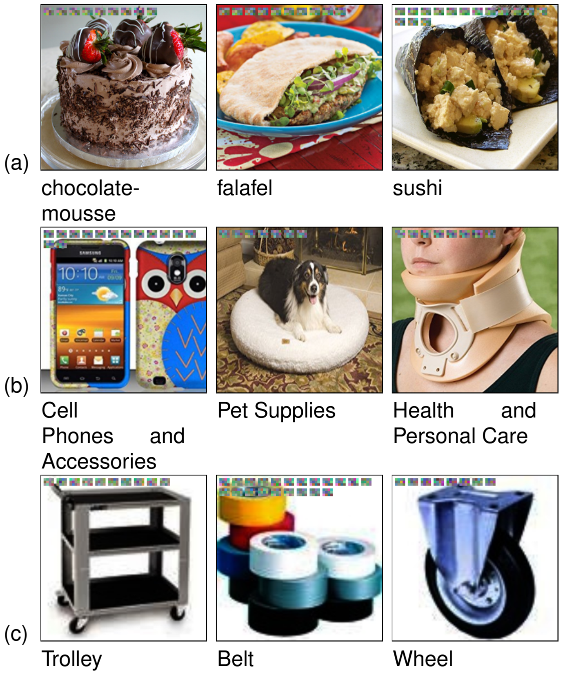


Fig. 5. Some random examples of multi-modal fusion from the datasets used. In row (a) three examples of images with the corresponding classes, extracted randomly from the UPMC Food-101 dataset, while in row (b) the three images were extracted from the Amazon Product Data dataset and in (c) three example images extracted from the Ferramenta Dataset. The size of all images is 256×256. The text associated to each image is encoded and visually represented in the upper part of the image.

The first dataset is named Ferramenta multi-modal dataset [13]. This dataset is made up of 88,010 adverts divided in 66,141 adverts for train and 21,869 adverts for test, belonging to 52 classes. Another dataset use is the UPMC Food-101 [14], a large multi-modal dataset containing about 100,000 items of food recipes classified in 101 categories. This dataset was crawled from the web and each item consists of an image and the HTML webpage on which it was found. We have only extracted the title from every HTML document.Categories in the dataset are the 101 most popular categories from the food picture sharing website[1] We used another publicly available real-world multi-modal dataset called Amazon Product Data [15]. The dataset consists of advertisements with each advertisement contain a text description and image. We randomly selected 10,000 advertisements belonging to 22 classes. Finally, we split 10,000 advertisements for each class into train and test sets with 7,500 and 2,500 advertisements respectively. We applied *mirroring* and *cropping* to these datasets. To avoid losing the semantics of the encoded text, we applied above mentioned techniques directly on images, before merging them with the encoded text descriptions.

TABLE 1. Classification results and comparisons of the CNN trained on single source (column Image) along with images and text descriptions fused using the proposed approach (column Proposed). Note that column Text shows results of a SVM trained on Word2Vec features.

Dataset	Image	Text	**Proposed**
	AlexNet		
Ferramenta	92.36	84.50	**94.84**
Amazon Product	46.07	64.37	**72.52**
Food-101	42.01	56.75	**83.04**
	GoogleNet		
Ferramenta	92.47	84.50	**95.87**
Amazon Product	51.42	64.37	**78.26**
Food-101	55.65	56.75	**85.69**

B. Implementation Details

In this work, the transformed text description is fused into original, horizontally flipped and cropped version of an image with 256×256 pixel size. Some examples are shown in Figure 5. We use a standard AlexNet [16] and GoogleNet [17] with softmax as supervision signal. We use the following hyperparameters for both networks: learning rate $lr = 0.01$, solver type = Stochastic gradient descent (SGD), training epochs = 90 and/or till no further improvement is noticed to avoid over fitting. In our experiments, accuracy is used to measure classification performance. We conducted five fold experiments on each dataset to evaluate the proposed approach.

C. Experiment Details

In the first set of experiments we extract only images from the three datasets with an image size of 256×256, then we train a standard AlexNet and GoogleNet from scratch. Results are shown in the column labelled "Image" of Table 1. In the second set of experiments, we use text descriptions and train a Word2Vec model to extract feature vectors, which then have been used as input to train a Support Vector Machine (SVM). Results are shown in the column named "Text" of Table 1. Later, we use images and text descriptions to create information enriched images with the proposed approach. Results are shown in the last column of Table 1. Analyzing results in Table 1 it can be noticed that the proposed method outperforms the accuracy of the best results obtained using only text descriptions or images on all three datasets. Images in the Ferramenta dataset contain objects on a white background, this explains the excellent classification result obtained on images alone. On the contrary, images in the UPMC Food-10 dataset are with complex background and extracted from different contexts, which leads to a low classification performance of the images without text. Results of our approach when applied to the UPMC Food-10 and Amazon Product Data datasets, highlight the strengths of our approach: the fusion of two very different information into a single image space exploits the two types of information content in the best way. From the Table 1, it is evident that the proposed approach obtains higher classification performance with GoogleNet compared with AlexNet. With this result, we expect higher multi-modal classification performance using the recent state-of-the-art CNNs for image classification.

[1] www.foodspotting.com

TABLE 2. Comparison of our approach with baseline and previous available works.

	Model	Ferramenta	UPMC Food-101	Amazon Product Data
Previous work	Wang et al [14]	–	85.10	–
	Kiela et al [5]	–	**90.8±0.1**	–
	Gallo et al [10]	94.42	60.63	–
Baseline	Image	92.47	55.65	51.42
	Text	84.50	56.75	64.37
Ours	Proposed	**95.87**	85.69	**78.26**

D. Baselines

In a multi-modal setting, text and image are two standard baselines [5], [10], [18] to which different fusion strategies are compared. In our work, we use Word2Vec word embedding as text baseline and standard image classification model as image baseline. Finally, we compare our fusion approach with these baseline strategies, results are shown in Table 2. Our approach obtains higher classification accuracy compared to the uni-modal (Text/Image).

E. Comparison

In addition, we compare our fusion approach with state-of-the-art multi-modal works available in literature, results are shown in Table 2. Our approach obtains higher or comparative classification accuracy compared to previous available works. We also include results from [14] on UPMC Food-101, where they used TF-IDF features for text and a deep convolutional neural network features for images. Furthermore, we compare our work with [5] where they explore the trade off between efficiency and multiple fusion strategies. Additionally, we compare our results with [10] where they use Word2Vec and Bag-of-Words for text and a deep convolutional neural network for images. We note that in the case of Ferramenta, our method works considerably better than previously reported results. We obtain comparable results on UPMC Food-101 dataset. We find that [10] is not scalable, whereas our work can be employed regardless of the dataset size avoiding any bottlenecks.

IV. CONCLUSION

In this work, we proposed a new approach to fuse images with their text description so that any CNN architecture can be employed as a multi-modal classification unit. To the best of our knowledge, the proposed approach is the only one that simultaneously exploits text semantics and image casted to a single source, making it possible to use a single classifier typically used in standard image classification. The classification accuracy achieved using our approach maintains an upper bound to single modalities.

Another very important contribution of this work concerns the joint representation into the same source of two heterogeneous modalities. This aspect paves the way to a still open set of problems related to the translation from one modality to another where relationships between modalities are subjective. With this aspect in mind, we are experimenting with using a CNN model to classify either text description, image or their combination. This means that we can extend this work for other cross-modal applications [1].

REFERENCES

[1] S. Nawaz, M. K. Janjua, A. Calefati, and I. Gallo, “Revisiting cross modal retrieval,” *arXiv preprint arXiv:1807.07364*, 2018.

[2] I. Gallo, A. Calefati, S. Nawaz, and M. K. Janjua, “Image and encoded text fusion for multi-modal classification,” *International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, 2018.

[3] D. Kiela and L. Bottou, “Learning image embeddings using convolutional neural networks for improved multi-modal semantics,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, October 2014, pp. 36–45.

[4] C. W. Leong and R. Mihalcea, “Going beyond text: A hybrid image-text approach for measuring word relatedness.” in *IJCNLP*, 2011, pp. 1403–1407.

[5] D. Kiela, E. Grave, A. Joulin, and T. Mikolov, “Efficient large-scale multi-modal classification,” *The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, 2018.

[6] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” *CoRR*, vol. abs/1301.3781, 2013.

[7] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, “Multimodal machine learning: A survey and taxonomy,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[8] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli, “Multimodal fusion for multimedia analysis: a survey,” *Multimedia systems*, vol. 16, no. 6, pp. 345–379, 2010.

[9] E. Bruni, G. B. Tran, and M. Baroni, “Distributional semantics from text and images,” in *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, ser. GEMS ’11, 2011, pp. 22–32.

[10] I. Gallo, A. Calefati, and S. Nawaz, “Multimodal classification fusion in real-world scenarios,” in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 2017, pp. 36–41.

[11] S. Poria, E. Cambria, and A. Gelbukh, “Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis,” in *Proceedings of the 2015 conference on empirical methods in natural language processing*, 2015, pp. 2539–2544.

[12] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[13] I. Gallo, S. Nawaz, and A. Calefati, “Semantic text encoding for text classification using convolutional neural networks,” in *Document Analysis and Recognition (ICDAR), 2017 14th IAPR International Conference on*, vol. 5. IEEE, 2017, pp. 16–21.

[14] X. Wang, D. Kumar, N. Thome, M. Cord, and F. Precioso, “Recipe recognition with large multimodal food dataset,” *2015 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pp. 1–6, 2015.

[15] R. He and J. McAuley, “Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering,” in *proceedings of the 25th international conference on world wide web*. International World Wide Web Conferences Steering Committee, 2016, pp. 507–517.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[17] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[18] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, “Multimodal deep learning,” in *Proceedings of the 28th international conference on machine learning (ICML-11)*, 2011, pp. 689–696.