DICTA Publication License for PID # 1222850

Paper Title: Image and Encoded Text Fusion for Multi-Modal Classification

Author: I. Gallo, A. Calefati, S. Nawaz, M.K. Janjua

I. LICENSE

Image and Encoded Text Fusion for Multi-Modal Classification

I. Gallo, A. Calefati, S. Nawaz
University of Insubria,
Dep. of Theoretical and Applied Science,
Varese, Italy
Email: ignazio.gallo@uninsubria.it

M. K. Janjua
National University of
Sciences and Technology,
Islamabad, Pakistan

Abstract—**Multi-modal approaches employ data from multiple input streams such as textual and visual domains. Deep neural networks have been successfully employed for these approaches. In this paper, we present a novel multi-modal approach that fuses images and text descriptions to improve multi-modal classification performance in real-world scenarios. The proposed approach embeds an encoded text onto an image to obtain an information enriched image. To learn feature representations of resulting images, standard Convolutional Neural Networks (CNNs) is employed for classification task. We demonstrate how a CNN based pipeline can be used to learn representations of the novel fusion approach. We compare our approach with individual sources on two large scale multi-modal classification datasets while obtaining encouraging results. Furthermore, we evaluate our approach against two famous multi-modal strategies namely early fusion and late fusion.**

II. INTRODUCTION

With the rapid rise of e-commerce, the web has increasingly become multi-modal, making the question of multi-modal strategy ever more important. However, modalities in multi-modal approach come from different input sources (text/image [1], [2], [3], audio/video [4] etc.) and are often characterized by distinct statistical properties, making it difficult to create a joint representation that uniquely captures the "concept" in the real-world applications. For example, Figure 1 shows two adverts typically available on e-commerce website, where two objects have seemingly similar text descriptions in first row but seemingly different images. On the other hand we have two different text descriptions but similar images in the second row. This leads us to create a joint representation of an image and text description for this classification problem. Multi-modal strategy can exploit such scenario to remove ambiguity and improve classification performance.

The use of multi-modal approach based on image and text features is extensively employed on a variety of tasks including modeling semantic

(a) Useful accessory for those who ride a **bike**. Size 46-52.

(b) The First **Bike** Pink Arrow dedicated to little girls.

(c) Telescopic **ladder** to partial or total opening. Ideal for any external intervention.

(d) Custom multifunction dynamic construction **scaffolding**, simple for decoration.

Fig. 1: In the top row, an example of ambiguous text descriptions that can be disambiguated with the analysis of the accompanying images. In the bottom row, an examples of ambiguous images that can be disambiguated with the analysis of the associated text descriptions.

relatedness, compositionality, classification and retrieval [5], [2], [6], [7], [3], [8]. Typically, in multi-modal approach, image features are extracted using CNNs. Whereas, to generate text features, Bag-of-Words models or Log-linear Skip-gram Models [9] 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 [10].

With this work, we present a novel strategy which combines a text encoding schema to fuse text features and image in an unified information enriched image. We merge both text encoding and image into a single source so that it can be used with a CNN. We demonstrate that by adding encoded text information in an image, multi-modal classification results can be improved compared to the best one obtained on a uni-modality (image/text).

Intuitively superimposing text descriptions onto images may not sound motivating due to several reasons. Since the idea is overlaying the encoded text description onto an image, it might affect the image perception in general. However, this is not true, the main strength of the approach is that embedded text can be overlaid onto the image with fixed width regardless of the size of text description. We experiment with different embedding sizes to verify that image perception is not affected. Figure 4 plots different embedding sizes to explain the network behavior under such scenario.

Main contributions of our paper are listed below:
- We present a novel data fusion mechanism based on encoded text description and associated image for multi-modal classification.
- We show that fused data is classified with a standard CNN based architectures, typically employed in image classification.
- We evaluate the fused multi-modal approach on two large scale datasets to show the effectiveness of our approach.

III. RELATED WORK

There are two general multi-modal fusion strategies to fuse text and images features, namely early fusion and late fusion [11], [10], each one having its advantages and disadvantages.

Early fusion is an initial attempt by the researchers towards multi-modal representation learning. Early fusion methods concatenates text and image features into a single vector which is used as input pattern for the final classifier. The technique is employed for various tasks [12], [2], [3]. The main benefit of early fusion is that it can learn to exploit the correlation and interactions between low level features of each modality.

In contrast, late fusion [13] uses decision values from each modal and fuses them using a fusion

TABLE I: Network configuration summary. k, s and p stand for kernel size, stride and padding size respectively. In the convolution layer, we use 128 filters for each of the following sizes 3,4,5 (the first one is showed below). The embedding size is $w = 128$.

Type	Configuration
Output	num. classes
Fully Connected	H_t(encoded-text-h) $\times$ W_t(encoded-text-w) $\times 3$
MaxPool-1D	h:$S - k + 1$, k:1, s:1, p:1
Convolutional-1D	w(embedding-size), k:3 $\times w$, s:1, p:1
Input	S: 100 words (sequence length)

mechanism. Multiple works [14], [15] employ different fusion mechanisms such as averaging, voting schemes, variance etc. The work in [1] showcased a comparative study of early and late fusion multi-modal methods. Late fusion produced better performance compared to early fusion method, however, it comes with the price of an increased learning effort. In addition, a strategy must be introduced to assign a weight to each classifier employed. This presents another challenge in late fusion strategy. Our method is inspired from early fusion [2], however, taking advantage of the idea of our previous work [16] we concatenate encoded text features into an image to obtain an information enriched image. In this work, we encode text features onto an image with an encoding schema similar to the one proposed in [16]. The main difference lies in the type of embedding used: our previous work [16] used the encoding extracted from Word2Vec and therefore we obtained a numeric vector for each word in a text document, while in this work we extract text features from a CNN network for text classification, trained using all the words available in a description. In the next sections, the encoding technique used to graphically represent the text above the image, will be summarized.

Multi-modal fusion methods are successfully employed to other modalities, e.g. video and audio [4], [17]. Other interesting examples of multi-modal approaches that make use of deep networks include restricted Boltzmann machines [18], autoencoders [19].

IV. THE PROPOSED APPROACH

In this work we take a cue from our previous work [16] to transform a text document onto an image to be classified with a CNN. However, instead
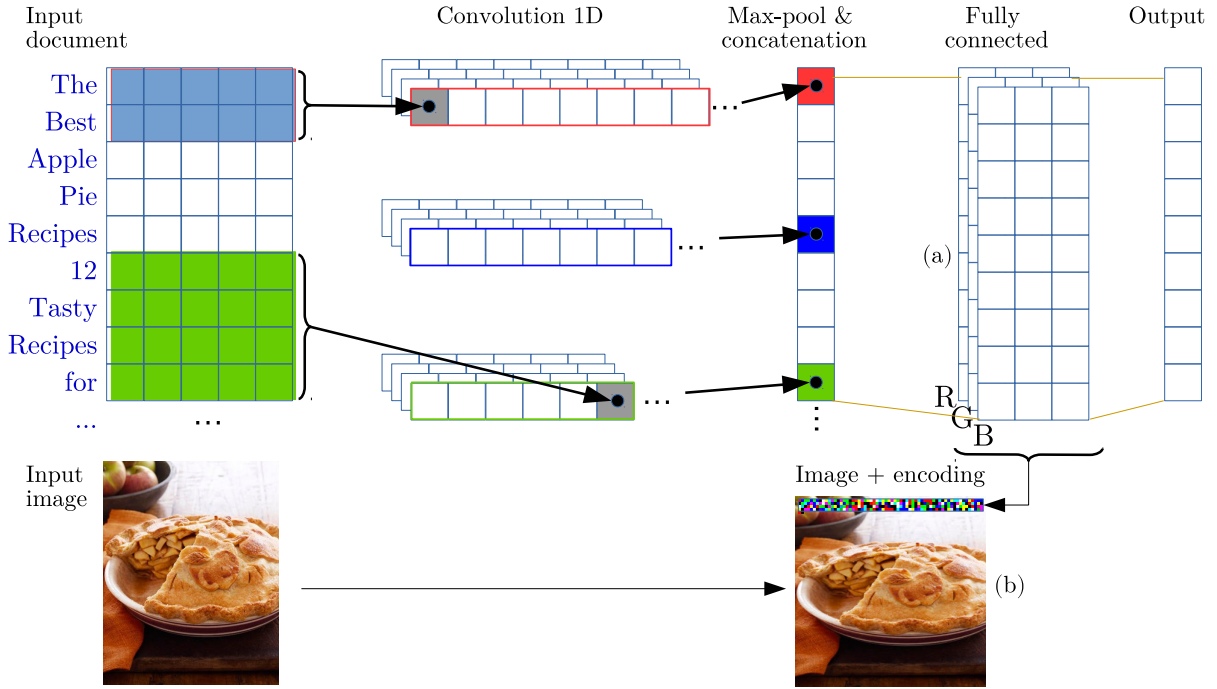
Fig. 2: The proposed text and image fusion model for deep multi-modal classification. The text encoded (a) is passed to the output layer. After the training step only the text features (a) are extracted and then drawn over the original image in order to generate a new multimodal dataset.

of using numeric values from Word2Vec model to represent a text document, we are using a new approach involving a CNN trained for text classification.

First, we transform the text document into a visual representation to construct an information enriched image containing text features and image. Finally, we solve the multi-modal problem using this image to train a CNN generally used for image classification.

We use a variant of the CNN model proposed by Kim [20] for text document classification. The input layer is a text document followed by a convolution layer with multiple filters, then a max-pooling layer followed by a fully connected layer, and finally a softmax classifier. The network configuration summary is show in Table I. Text features are extracted from the fully connected layer (Figure 2a) and transformed into an RGB encoding so that it can be overlaid onto an image associated with the text document. Figure 2 shows architectural representation of the model used to encode the text dataset into an image dataset (Figure 2b) to obtain a multimodal dataset. In the second step, resulting images are fed to any baseline CNN for classification.

The major advantage of our method is that we can cast a uni-modal into a multi-modal CNN without the need of adapting the model itself. This approach is suitable to be adopted in multi-modal methods because a CNN architecture can extract information from both the encoded text and the related image.

A. Encoding Scheme

We exploit the CNN model proposed by Kim [20] which performs text to visual features transformation within a single step. Figure 2 summarizes the encoding system used in this work, where a reshape was applied to the fully connected layer showed in Figure 2a to transform an array into an image representing the encoded text to be superimposed on the original image.

Features are extracted from the trained CNN model and transformed into a visual representation of the document. In practice, we used feature vectors showed in Figure 2a, having a size $L = 3 \cdot w \cdot h$ that is a multiple of 3 in order to be transformed into a color image. We used the same concept of *superpixel* used in [16] to represent a sequence of three values $\in L$ as an area with a uniform color of $P \times P$ dimension. In this way textual features are

TABLE II: Information on multi-modal datasets used in this work. A multi-modal dataset consists of an image and accompanying text description. The last column indicates the text description language.

Dataset	#Cls	Train	Test	Lang.
Ferramenta	52	66,141	21,869	IT
Food-101	101	67,988	22,716	EN

represented as a sequence of superpixel, drawn from left to right and from top to bottom, starting from a certain position of the scaled image (see some examples of the final multi-modal image in Figure 3 and Figure 5). Finally, we encode an entire text document within the image plane and then the next multi-modal CNN model can work simultaneously on both modalities.

This approach has an advantage to the work in [16], in fact in our work it is possible to encode long text documents because we encode the entire document in the same image area having fixed size equals to $w \times h \times 3$.

V. DATASETS

In multi-modal dataset, modalities are obtained from different input sources. Datasets used in this work consist of images and accompanying text descriptions. We select Ferramenta [1] multi-modal dataset that are created from e-commerce website. Furthermore, we select UMPC Food-101 [21] multi-modal dataset to show the applicability of our approach to other domains. Table II shows information on these datasets. The first column shows the number of class labels available in datasets. While second and third columns show split on train and test sets. The last column indicates the language of text description available for these datasets. Table III shows image and associated text description randomly selected from each multi-modal dataset.

Ferramenta multi-modal dataset [16] consists of $88,010$ adverts split in $66,141$ adverts for train set and $21,869$ adverts for test set, belonging to 52 classes. Ferramenta dataset provides a text and representative image for each commercial advertisement. It is interesting to note that text descriptions in this dataset are in Italian Language.

The second dataset that we use in our experiments is named UPMC Food-101 multi-modal dataset [21], containing about $100,000$ items of

food recipes classified in 101 classes. This dataset is collected from the web and each item consists of an image and the HTML webpage on which it was found. We have extracted the title from HTML document to use it in lieu of text description. Classes in the dataset are the 101 most popular categories from the food picture sharing website[1].

VI. EXPERIMENTS

A. Preprocessing

The proposed multi-modal approach transforms text descriptions and embeds them onto associated images to obtain information enriched images. An example of information enriched image is shown in Figure 3. In this work, the transformed text description is embedded into a RGB image with an image size of 227×227 for UMPC Food-101 and Ferramenta multi-modal datasets.

B. Detailed CNN settings

We use a standard AlexNet [22] and GoogleNet [23] on the Deep Learning GPU Training System (DIGITS) with default configuration. For fair comparison, we use same CNN settings for experiments using only images and fused images. We use standard CNN hyperparameters. The initial learning rate is set to 0.01 along with Stochastic Gradient Descent (SGD) as optimizer. The network is trained for a total of 60 epochs and/or till no further improvement is noticed to avoid over fitting. In our experiments, accuracy is used to measure classification performances. The aim of the experiment is to show that by adding encoded text information in images it is possible to obtain better classification results compared to the best one obtained using a uni-modal (Text/Image). We conducted following experiments with this aim in mind: (1) classification with CNN using only images, (2) classification with CNN using only text descriptions, (3) classification with CNN using fused images, (4) comparison with early and late fusion strategies.

The first experiment consists of extracting only text descriptions from multi-modal datasets, then we train text classification model shown in Figure 2. Results are shown in first column of Table IV. It is very important to observe how the text encoding
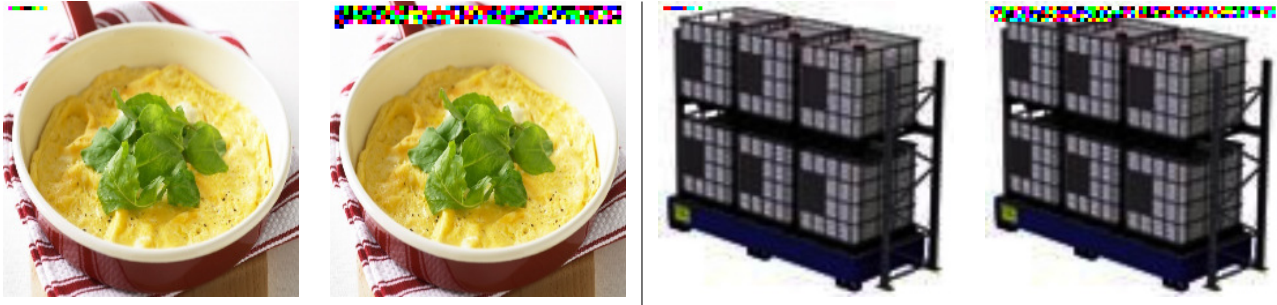
---

[1]www.foodspotting.com

Fig. 3: Two encoding examples taken from two datasets. Images on the left column show an encoding of 10 superpixel length while on the right column we have an encoding of length equal to 250 superpixel. All images are 227×227 in size having an encoding superpixel equals to 3×3 pixels for Ferramenta dataset and 4×4 for Food-101 dataset.

TABLE III: An image and an associated text description randomly taken from each multi-modal dataset. Text descriptions in Ferramenta multi-modal dataset is translated from Italian to English for readers. UMPC Food-101 multi-modal dataset contains long text descriptions for food recipes however, we include a short text description.

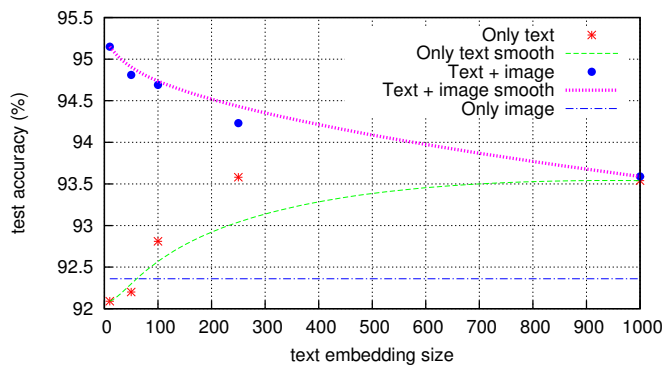| Dataset	Image	Text Description
Ferramenta [1]		saratoga chestnut brown spray paint 400 ml happy color, quick-drying bright spray enamel for interiors and exteriors for applications on furniture chairs doors frames ornaments and all surfaces in wood metal ceramic glass plaster and masonite.
UPMC Food-101 [21]		Robiola-Cheese-Filled Ravioli Recipe Pasta Recipes ...

Fig. 4: Comparison between the CNN that uses only text documents and only images with the CNN that uses fusion of image and encoded text, as the dimension of the text embedding varies. In this experiment the Ferramenta multi-modal dataset is used.

TABLE IV: Classification results comparison on only-text, only-image and fused images. There are two baseline models for images and fused-images, while we use only one baseline for text-only scores.

Dataset	Text	Image		Fusion	
		AlexNet	GoogleNet	AlexNet	GoogleNet
Ferramenta	92.09	92.36	92.47	95.15	95.45
Food-101	79.78	42.01	55.65	82.90	83.37

TABLE V: Comparison of our approach with early and late fusion strategies. The results on the Ferramenta dataset are extracted from paper [1]

Dataset	Early F.	Late F.	Proposed
Ferramenta	89.53	94.42	**95.15**
Food-101	60.83	34.43	**82.90**

extracted is similar to each other when the text description represents similar objects, even when

the text information and the images are different from each other (see the example of text encoding in Figure 5).

The second experiment consists of extracting only images from multi-modal datasets, then we train AlexNet [22] and GoogleNet [23] CNNs from scratch using DIGITS. Second and third columns of Table IV shows these results. Images in Ferramenta multi-modal dataset contain objects on a white background, this explains excellent classification results obtained on images alone. On the contrary, images in the UPMC Food-101 multi-modal dataset are with complex background and extracted from different contexts, which leads to a low classification performance on images only.

The third experiment consists of employing fused images from multi-modal datasets. We train AlexNet [22] and GoogleNet [23] CNNs from scratch using DIGITS. Results in Table IV indicate that the proposed fusion approach outperforms unimodal methods. Furthermore, the approach is language independent, Ferramenta text descriptions are in Italian. Results on UPMC Food-101 clearly indicate benefit of our proposed approach, increasing the classification performance by two folds. This performance gain is due to leveraging on multimodal representation learning.

In fourth experiment, we compare our approach with early and late fusion as shown in Table V. Experimental setting is inspired from the work [1]. In particular we use *Logarithmic Opinion Pool* [24] as a late fusion approach using Random Forest model applied to the 1000 Bag-of-Words while as early fusion we use a Support Vector Machine on the concatenation of Doc2Vec features and 4096 visual features from a trained CNN. Our proposed approach surpasses standard early and late fusion strategies which further reinforces strength of our approach.

The Figure 4 explores text embedding dimension sizes against 3 different CNN based architectures i.e. text only, image only and fused image. We see that with lower text-embedding dimension, the fused architecture has an increased performance as compared to the text only architecture. Eventually, both architectures plateau as embedding dimension increases. However, the fused image architecture always maintains the upper bound over the other.

VII. CONCLUSION

In this work, we proposed a new approach to merge image with their text description so that any CNN architecture can be employed as a multi-modal classification system. To the best of our knowledge, the proposed approach is the only one that simultaneously exploits text and image casted to a single source, making it possible to use a single classifier. We obtained promising results and the classification accuracy achieved using our approach is always higher compared to fusion strategies or 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.

REFERENCES

[1] I. Gallo, A. Calefati, and S. Nawaz, "Multimodal classification fusion in real-world scenarios," in *Document Analysis and Recognition (ICDAR)*. IEEE, 2017, pp. 36–41.

[2] D. Kiela and L. Bottou, "Learning image embeddings using convolutional neural networks for improved multi-modal semantics," in *Empirical Methods in Natural Language Processing (EMNLP)*. ACL, October 2014, pp. 36–45.

[3] D. Kiela, E. Grave, A. Joulin, and T. Mikolov, "Efficient large-scale multi-modal classification," *Proceedings of AAAI 2018*, 2018.

[4] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *International Conference on Machine Learning (ICML)*, 2011, pp. 689–696.

[5] M. Guillaumin, J. Verbeek, and C. Schmid, "Multimodal semi-supervised learning for image classification," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 902–909.

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

[7] Y. Feng and M. Lapata, "Visual information in semantic representation," in *Annual Conference of the North American Chapter of the Association for Computational Linguistics*. ACL, 2010, pp. 91–99.

[8] L. Wang, Y. Li, and S. Lazebnik, "Learning deep structure-preserving image-text embeddings," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5005–5013.

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

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

[11] 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.

[12] 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.

[13] 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.

Ferramenta		UMPC Food-101	
	bahco 9070p chiave inglese regolabile ergonomica 15 3 cm 6 pollici a becco reversibile colore nero		Cannoli Recipe - Food.com
	connex cox550110 chiave inglese regolabile 25 4 cm		homemade cannoli filling The 350 Degree Oven
	axis 28831 chiave inglese regolabile con impugnatura morbida e rullo estremamente scorrevole 200 mm . . .		Cake Boss Cannoli Cake Ideas and Designs
	sam outillage 54 c10 chiave a rullino cromata 10 lunghezza 255 mm sam		Scones* Biscotti* Cannoli on Pinterest
	faithfull chiave regolabile 150 mm		Sicilian Cannoli Recipe The Daily Meal

Fig. 5: Each column contains 5 images and associated text descriptions belonging to a particular class of Ferramenta and UPMC Food-101 datasets. Futhermore, each image contains the proposed encoded text. Note that the text encodings on each column are similar to each other even if the text and images are different from each other.

[14] E. Shutova, D. Kiela, and J. Maillard, "Black holes and white rabbits: Metaphor identification with visual features," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, pp. 160–170.

[15] G. Evangelopoulos, A. Zlatintsi, A. Potamianos, P. Maragos, K. Rapantzikos, G. Skoumas, and Y. Avrithis, "Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention," *IEEE Transactions on Multimedia*, vol. 15, no. 7, pp. 1553–1568, 2013.

[16] I. Gallo, S. Nawaz, and A. Calefati, "Semantic text encoding for text classification using convolutional neural networks," in *Document Analysis and Recognition (ICDAR)*, vol. 5. IEEE, 2017, pp. 16–21.

[17] A. Nagrani, S. Albanie, and A. Zisserman, "Seeing voices and hearing faces: Cross-modal biometric matching," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8427–8436.

[18] N. Srivastava and R. R. Salakhutdinov, "Multimodal learning with deep boltzmann machines," in *Advances in neural information processing systems*, 2012, pp. 2222–2230.

[19] P. Wu, S. C. Hoi, H. Xia, P. Zhao, D. Wang, and C. Miao, "Online multimodal deep similarity learning with application to image retrieval," in *ACM international conference on Multimedia*. ACM, 2013, pp. 153–162.

[20] Y. Kim, "Convolutional neural networks for sentence classification," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP*, 2014, pp. 1746–1751.

[21] X. Wang, D. Kumar, N. Thome, M. Cord, and F. Precioso, "Recipe recognition with large multimodal food dataset," in *Multimedia & Expo Workshops (ICMEW)*. IEEE, 2015, pp. 1–6.

[22] 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.

[23] 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.

[24] L. Rokach, "Ensemble-based classifiers," *Artificial Intelligence Review*, vol. 33, no. 1, pp. 1–39, 2010.