# Don't only Feel Read: Using Scene text to understand advertisements

Arka Ujjal Dey
IIT Jodhpur
dey.1@iitj.ac.in

Suman K. Ghosh and Ernest Valveny
Computer Vision Center, Dept. Ciències de la Computació
Universitat Autònoma de Barcelona
08193 Bellaterra (Barcelona), Spain
Email: sghosh,ernest@cvc.uab.es

## Abstract

*We propose a framework for automated classification of Advertisement Images, using not just Visual features but also Textual cues extracted from embedded text. Our approach takes inspiration from the assumption that Ad images contain meaningful textual content, that can provide discriminative semantic interpretetion, and can thus aid in classifcation tasks. To this end, we develop a framework using off-the-shelf components, and demonstrate the effectiveness of Textual cues in semantic Classfication tasks.*

## 1. Introduction

In the recent past deep CNNs have generated state of the art results in various computer vision tasks. While it started from character recognition [11], the architecture has been successfully adapted to a whole range of allied task involving natural images, word images [16], as well as scene text images [6]. Alongside this with the advent of big data, deep learning has been applied in various Natutal Language Processing tasks, long term sequence learners like LSTM [14], and context encoders like CBOW [13] are being explored for language modeling tasks. Sequential nature of text data allows it to be modeled by such encoders to provide semantic understanding of text data [10].

This understanding text and images can be used to solve more general AI problems like Image Captioning [7], Image Annotation, Visual Question Answering and Feature grounding [9]. Thus the interplay between text and image data is necessary for all aforementioned tasks,

In this context it is worth noting that images around us, apart from the visual semantic content, also contain a lot of embedded text (usually called scene text), which are provided for better human understanding of the images, whenever the image itself can not make the idea explicit. For example consider the image in Figure 1. The content of the images is not clear without explicitly reading the text.

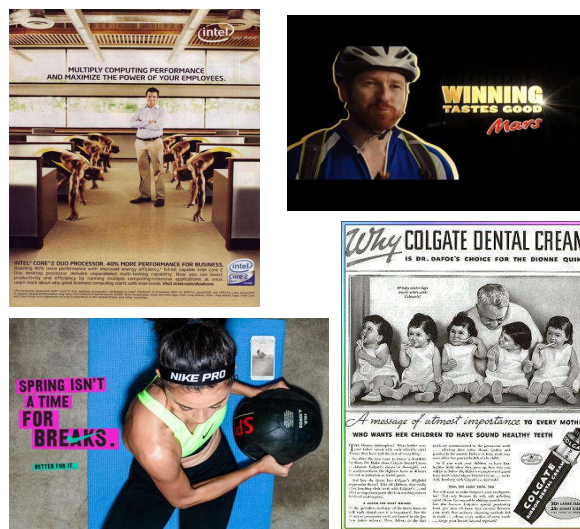In this work we build on this hypothesis that scene text



Figure 1. Example Ad images, illustrating the complementary nature of text and visual cues. In some cases the visuals can be symbolic, but embedded text gives away the context[top-left, top-right], in other cases the visuals can be simple to understand but the text can be obtuse[bottom-left]. Further, the amount of text content can vary widely [top-right, bottom-right]

(whenever available) plays very important role in understanding the image and thus we will use scene text as an extra cue and analyze its impact in in semantic image understanding.

## 2. Related Work

In this section we briefly analyze the works which use two modalities to solve computer vision or AI problems. Dealing with multimodal data poses two challenges, firstly on how to represent each of them, to effectively model the interplay and second modeling the interaction between them. The first step towards that is the generation of Global feature representation explictly or implicitly from local features. In case of images, Deep CNNs generate robust global features as the last fully connected layer activations, im-

plictly from local CNN feature maps. For textual content on the other hand its global representation mainly depends on the presence or absence of structure or sequence. For structured text, like question/answers about an image, or sentence captions, LSTM-RNN language models are the obvious choice due to their superior ability in modeling the sequential data. However for unstructured text like tags, or other Meta data, which is not sequential, the global feature is usually an aggregation (e.g. mean or averaging) of local semantic features. In some cases people have even used bag-of-words [3] and fisher vectors [8] for data aggregation to model the first order and second order statistics respectively. Once these Global textual and Visual Feature representations are computed, one has to model the interplay between them and this, depending on the end goal, one can have various ways.

For multimodal retrieval tasks usually the aim is to project both modalities into a common subspace where they are comparable, CCA (Canonical Correlation Analysis) [5] and its deep learning variant (DCCA) [1] are the most useful techniques often used by these algorithms. However this does not generate any unified third representation, encoding both the text and image features and lacks interoperability between them.

To effectively model the inter-relationship between them, generating one unified representation which can represent the data in higher semantic hierarchy, a feature fusion scheme is needed. Early methods use simple techniques like element wise sum/product, concatenation of features etc. However this is not expressive enough to capture the rich and non-linear relation between features from two different views/modalities. Outer product based bilinear interaction schemes like MCB [4], Mutan [2] allow for multiplicative interaction between the features, and are thus more suited. These schemes are used for image caption generation by leveraging the co-relation between text and image space at various level of granularity. In this work we will analyze the efficacy of both simple fusion schemes like average and concatenation and more generic outer join based schemes.

## 3. Methodology

Though textual content in images is ubiquitous in our everyday life, in the form of newspaper, magazines, print ads, store fronts, street scenes, their use in solving general problem other that text understanding itself has not been studied much in the literature. In this work our aim is to use these textual data in order to understand the world around us. In particular we deal with the problem of image classification into semantic topics, our framework is shown in Figure 2. We also apply our framework to visual question answering task from advertisement images. Our basic network is composed of three parts namely the scene text understanding part, image feature extraction and data fusion.

### 3.1. Scene Text Understanding

Though embedded text (from images) can be a rich source of information about semantic understanding this text is not there in usual text (ASCII) format. Rather it is embedded within image pixels, thus to leverage this knowledge the first task is to extract this texts from the raw image.

Thus, the first task in such a pipeline is to extract the text from the images. However this is not simple task and in reality is an active area of research [6].

In this work, as our goal is not to effectively extract text but to analyze the efficacy of text in understanding the images, we use an standard off-the-self model for text detection [12] and recognition [6] pipeline to detect and transcribe text.

Once we obtain a list of text extracted from an image with corresponding confidence measures we embed these texts into a semantic vector space such that words with similar semantics are have similar vector representation. In our current experiments we have used the word2vec [13] semantic embedding, as this has been successfully used in different semantic understanding pipeline.

Since the number of detected text varies widely from image to image in case of advertisement images, we limit ourselves to use only the top $k$ most discriminatory words according to tf-idf score. Experiment with different values of $k$ is presented in experimental results section 1.

We aggregate the $k$ corresponding word2vec vectors to generate a global text feature for a given image. The vector structure of the word2vec space, allows for simple vector sum to be a meaningful aggregation scheme. In experimental section we will show that using only text feature in this manner is comparable to image feature in understanding the semantics of the advertisement.

### 3.2. Image Feature Extraction

As the focus of our work is not to study the semantics of image features, which is a well studied topic we restrict our investigation only to standard CNN based feature extractors. In particular we use [15] to extract global feature vector from every image.

### 3.3. Combining Text and Image Feature, Data fusion

In our current set of experiments, we have explored one simple and one generic data fusion scheme. For baseline we use a simple concatenation and then we learned outer product approximation as fusion schemes3. We used a similar formulation like MCB [4]. In particular we use $1024$ dimensional feature from last fully connected layer of googlenet [15] and text feature as desribed in Sec. 3.1. Now Outer product between these two views is approximated. A low rank approximation is achieved by using a count sketch transforation.
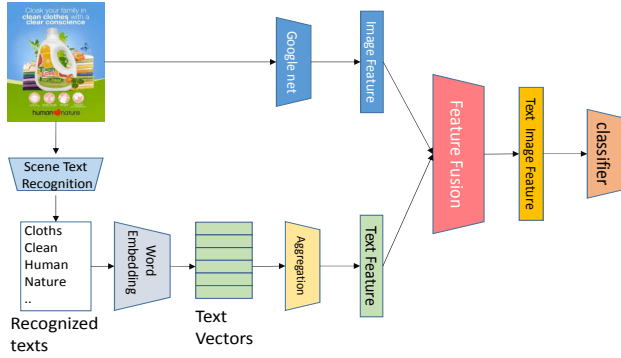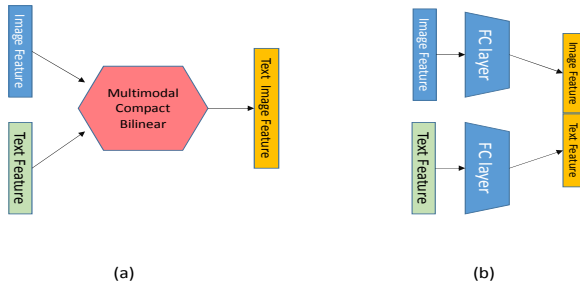
Figure 2. Topic Classification framework



Figure 3. Fusion schemes , (a) MCB , (b) Concat

In our experiments, we found outer product based on MCB [4] leads to better accuracy.

# 4. Experimental Results

We evaluate the benefit of using text features, when attempting to understand Ad images. We argue, that careful selection of relevant texts, imply rich semantic information, which can aid the visual cues, and can result in better classification results. We use the Jaderberg [6] scene text recognition engine on top of bounding boxes generated by Textboxes [12] to generate this transcription. However the nature of the Ad dataset, with varying amount of legible text, and Vocabulary limitations of our recognition engine, implied that we were not able to generate meaningful text features for every Ad image. To this end, we generated a cleaner dataset, consisting of 47000 images for which we are able to generate accurate transcriptions with 70% confidence. Our experiments are conducted on this dataset, where every image has some legible text leading to a meaningful text feature.

## 4.1. Topic Classification Task

In the Topic classification task, the objective is to classify an Ad image into 1 of 40 Topic classes. As shown in 1, sometimes the visual content can be symbolic or metaphorical, while the text content is more straight forward to inter-

pret or in other cases it could be that the text is misleading (eg. 'spring' , 'time' in the nike Ad), but the visual content is fairly straight. In conclusion, for Ad images both visual and textual cues and generate power features 1, but can lead to complementary interpretations. This lead us to try fusion schemes, whereby we learn a feature set using both, the text and visual cues. As is demonstrated in 2, fused features lead to better classification accuracy. Further, the multiplicative nature of MCB, allows for more interaction between the features than simple concatenation, and thus leads to a better performance.

Table 1. Topic Classification accuracy using only Image Features and only embedded text features from k most significant text words .

| Image | Text k=5 | Text k=10 | Text k=35 | Text k=100 |
|---|---|---|---|---|
| 45 | 41 | 41 | 40 | 40 |

Table 2. Topic Classification accuracy using fusion of Image Features and embedded text features.

| Fusion | Image Text k=5 | Image Text k=35 | Image Text k=100 |
|---|---|---|---|
| Concat | 53 | 52 | 52 |
| MCB | 58 | 57 | 57 |

## 4.2. VQA Task

Table 3. Classification accuracy using Fused text and Image Features and question features on VQA task.

| Fusion | Question | Question Image | Question Image Text |
|---|---|---|---|
| Concat | 10.93 | 11.9 | 12.44 |

For the VQA task, we observe that using only the question features we can obtain surprising results, but the incorporation of image and text features thereafter, does lead to improved performance.

# 5. Conclusion

Text based semantic embedding, originated from using meta-data, or annotation, when applied to text content from images, are dependant on robust and accurate transcription generation. Thus transcription is a weak link in this pipeline leading to significant performance drops due to vocabulary misses. These vocabulary misses can occur at two levels, misses by the wordspotting engine, and the misses by the word2vec lexicon during vector embedding. To address these issues, we are working towards an end to end embedding scheme, that generates semantic vectors from raw image pixels, without requiring any recognition and transcription.

# References

[1] G. Andrew, R. Arora, J. Bilmes, and K. Livescu. Deep canonical correlation analysis. In *International Conference on Machine Learning*, pages 1247–1255, 2013.

[2] H. Ben-younes, R. Cadene, M. Cord, and N. Thome. Mutan: Multimodal tucker fusion for visual question answering. In *The IEEE International Conference on Computer Vision (ICCV)*, volume 1, page 3, 2017.

[3] S. Clinchant and F. Perronnin. Textual similarity with a bag-of-embedded-words model. In *Proceedings of the 2013 Conference on the Theory of Information Retrieval*, page 25. ACM, 2013.

[4] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *Conference on Empirical Methods in Natural Language Processing*, pages 457–468. ACL, 2016.

[5] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 16(12):2639–2664, 2004.

[6] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman. Reading text in the wild with convolutional neural networks. *International Journal of Computer Vision*, 116(1):1–20, 2016.

[7] R. Kiros, R. Salakhutdinov, and R. S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014.

[8] B. Klein, G. Lev, G. Sadeh, and L. Wolf. Associating neural word embeddings with deep image representations using fisher vectors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4437–4446, 2015.

[9] S. Kottur, R. Vedantam, J. M. Moura, and D. Parikh. Visualword2vec (vis-w2v): Learning visually grounded word embeddings using abstract scenes. In *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*, pages 4985–4994. IEEE, 2016.

[10] Q. Le and T. Mikolov. Distributed representations of sentences and documents. In *International Conference on Machine Learning*, pages 1188–1196, 2014.

[11] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[12] M. Liao, B. Shi, X. Bai, X. Wang, and W. Liu. Textboxes: A fast text detector with a single deep neural network. In *AAAI*, pages 4161–4167, 2017.

[13] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[14] M. Sundermeyer, R. Schlüter, and H. Ney. Lstm neural networks for language modeling. In *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[15] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, et al. Going deeper with convolutions. Cvpr, 2015.

[16] T. Wilkinson and A. Brun. Semantic and verbatim word spotting using deep neural networks. In *Frontiers in Handwriting Recognition (ICFHR), 2016 15th International Conference on*, pages 307–312. IEEE, 2016.