Beyond Visual Semantics: Exploring the Role of Scene Text in Image Understanding

Arka Ujjal Dey[1], Suman K. Ghosh[2], Ernest Valveny[3], and Gaurav Harit[1]

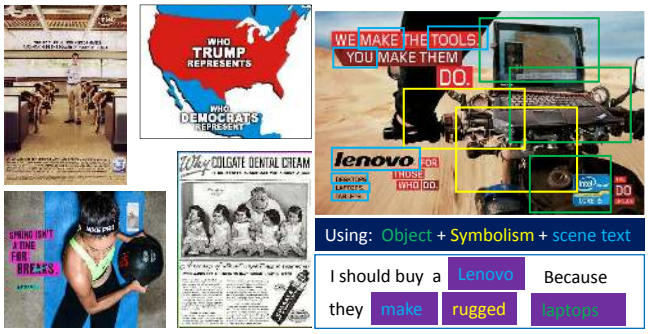
[1]IIT Jodhpur, Rajasthan, India
[3]Computer Vision Center, Universitat Autònoma de Barcelona, Bellaterra (Barcelona)
[2]University of Rouen, France

***Abstract*—Images with visual and scene text content are ubiquitous in everyday life. However, current image interpretation systems are mostly limited to using only the visual features, neglecting to leverage the scene text content. In this paper, we propose to jointly use scene text and visual channels for robust semantic interpretation of images. We do not only extract and encode visual and scene text cues, but also model their interplay to generate a contextual joint embedding with richer semantics. The contextual embedding thus generated is applied to retrieval and classification tasks on multimedia images, with scene text content, to demonstrate its effectiveness. In the retrieval framework, we augment our learned text-visual semantic representation with scene text cues, to mitigate vocabulary misses that may have occurred during the semantic embedding. To deal with irrelevant or erroneous recognition of scene text, we also apply query-based attention to our text channel. We show how the multi-channel approach, involving visual semantics and scene text, improves upon state of the art.**

I. Introduction

Images are the prevalent choice of expression these days, as they are often more engaging and less intrusive than other media. Often images use embedded scene text, in addition to visual elements, to express ideas more lucidly. Such images with visual and embedded scene text are ubiquitous in everyday life, in the form of printed advertisements, posters, propaganda bills, storefront views, and similar variants. The scene text content in such images is often crucial in the interpretation of the image. More importantly the scene text along with the visual contents often provide useful context to understand these media.

Text detection and recognition frameworks have matured in recent times, providing appealing results [17], [25] while handling real life scenarios like complex backgrounds [24], [18], irregular font sizes or arbitrarily oriented text [25]. With these advances, the underlying scene text in images, which has been inaccessible until now in most image understanding tasks, can now be leveraged to interpret images in a more generalized way. However, the use of scene text in image understanding thus far, has been scarce and constrained, basically to the realm of fine-grained classification tasks [4], [21], [20], [22] and more recently to Visual Question Answering (VQA) [30], [6]. However, these works treat visual and text features as separate channels and do not model the semantic relationships between them.



(a) (b)

Fig. 1: (a) 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] (b) Basic idea: Use detected visual symbolism and objects, together with scene text to reason about images

In this work, we go beyond the detection of text and visual objects by learning a joint contextual semantic embedding that aims at capturing the inter-object dynamics. This interaction is modelled using a Text-Visual graph and a Graph Attention network [32] to generate the final embedding. The inter-object relationships, along with the encoded features, augments the ability to reason about images. In order to show how the contextual semantic embedding can be adapted to different scenarios we apply the model in two datasets where context plays a critical role: advertisement images[16] and tweets[13]. We address two different tasks on the Ad dataset[16] (retrieval of relevant statements and topic classification), as well as a binary sentiment classification task (hate speech detection) on the tweet dataset[13]. Both datasets contain images where text, as well as visual elements, are purposefully used to propagate an agenda, a marketing strategy, or a hateful message as illustrated in Fig.1a. They may also contain socio-cultural references, symbolism[33], [16] along with wit and humor. Reasoning about such images involves understanding the context and the relationship between all the elements in that context[34].

In summary, the main contributions of this work are the following: first, we model the interplay between the detected text and the visual cues to generate a contextual embedding

that encodes the inter-object dynamics using an attentional relationship graph. Second, we show through experiments that this model can be applied successfully, improving state of the art, to different image understanding tasks, such as semantic retrieval of relevant statements, topic classification or sentiment classification. Third, we make additional contributions to better leverage the use of scene text for the specific task of retrieving relevant statements in the ad dataset: we propose a novel use of scene text by using both semantic and lexical information and we leverage the language structure of the statement by partitioning it into an action-reason pair to better model the relation between the semantics of the query and the image.

II. Related works

A. Use of scene text

Using scene text for image understanding has been attempted mainly in the context of fine-grained image classification tasks [4], [21], [20], [22]. Leveraging scene text present in an image may lead to better classification accuracy for certain types of images, e.g., storefront images [21], [20]. While in [22], the authors use a spatial encoding of n-grams as text features, in [20] they argue for word-level features and use a vocabulary based Bag of Words (BOW) representation. In both works, the final representation is a combination of visual and text cues, without using any semantic information or modelling of the interaction between the text and visual cues. Only in [4], the authors proposed encoding the text using a semantic embedding[26] improving upon the previous results.

Recently, motivated for last advances in scene text extraction, there is a surge of interest in systems and datasets that leverage scene text along with traditional visual cues, for instance in advertisement understanding[16], hate speech detection[13] or VQA[30], [6]. In these cases it is observed that visual features alone are not enough and extracting the scene text and encoding the context is critical for successful interpretation. Scene text features have been shown to be quite discriminative by themselves for advertisement understanding, as noted by the CVPR AD Workshop Challenge winners. We will rely on these results and we will also use a separate text channel encoding scene text semantics for ad retrieval. In the case of VQA the proposed baselines, built upon traditional VQA systems, combine the text channels along with the standard visual channel, but without trying to model the relationship between them. In our work, we will show that modeling such relationships generates rich contextual features leading to improved semantics.

B. Text and vision

Language and vision are the two most important ways we communicate. Thus, their combination poses important challenges like image captioning [3], text-based image retrieval (e.g., google image search) and Visual Question Answering [12] among others. In most cases, the semantic encoding of the text, is used either in conjunction with visual features through fusion[12], [5] for VQA tasks, or it is used to define a common subspace[10] to project the visual representation into for retrieval tasks. In [33], an embedding scheme projects images and statements into a common subspace, where retrieval is feasible. The embedding scheme used features from salient regions proposed by symbol detectors and automated captions generated by Densecap[19]. The generated caption acted as external knowledge and was encoded with word-embedding[26]. The success of such methods[33], [11] in embedding visual features into a common semantic subspace can be largely attributed to the discriminative nature of text semantics[26], [29] facilitated by availability of huge text corpus.

In these works, the text originates from an external source (question, caption, annotation) and not in the form of scene text present within the image. While images may usually contain visual objects, symbolism, and motifs the advertisement and tweet images that we analyse in this work often use scene text content to drive home a clear message. Thus, while we find several related work exploring the visual symbolism present[33], [9], or attending[1] to different visual components, we also explore the role of scene text in conveying that take away message.

C. Contextual Encoding

Given the nature of high-level tasks like VQA or captioning, both textual and visual cues convey essential contextual information to be leveraged. While feature fusion[12], [5] is a standard scheme for image representation encoding different modalities, it is preceded by feature aggregation of respective modalities. However, simple aggregation of local features from different modalities leads to loss of fine-grained spatial and contextual information that can be beneficial for high-level downstream tasks. Here is where attention[27], or relevance of the different detected components, comes into play. In the case of advertisement images, for instance, the need to attend to relevant information has found expression in various works[1], [33], [9]. In these works we see examples of top-down attention, guided by the final task[33] or linguistic cues, in the form of the text statement [9], to attend to the visual features corresponding to symbolism and objects. Recently, attention on graphs describing the relations among different components has been proposed with the Graph Attentional Layer (GAT)[32], [31], [23] to encode the context. We will leverage GAT in our contextual encoder as a way to explore the interplay between the detected textual and visual local features, and encode their contextual information, to generate rich semantics.

III. Method

Fig.2 gives a detailed illustration of our proposed model, which extracts and encodes visual and scene text cues to generate a contextual encoding, applicable to different tasks, viz semantic embedding, classification. The basic stages of our pipeline consist of a Visual Encoder, a Scene Text Encoder, and a Multi-Modal Contextual encoder.

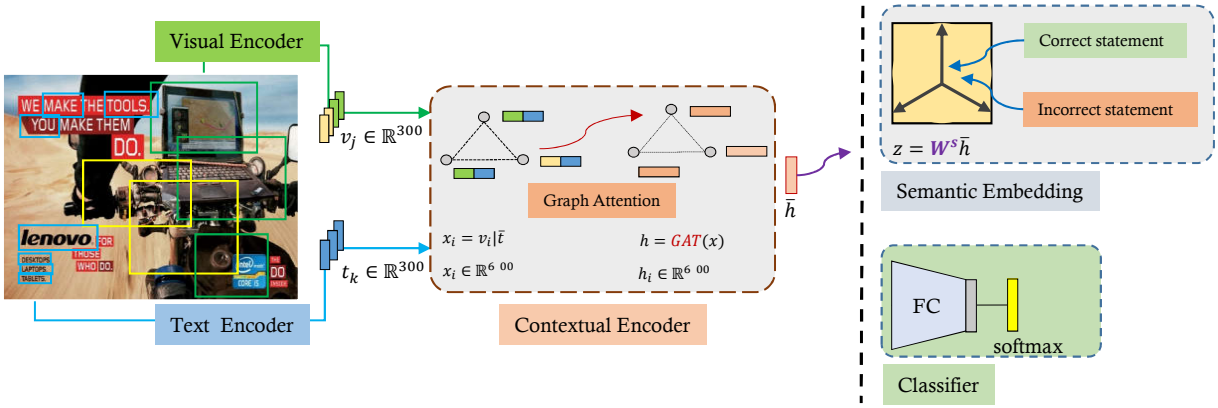


Fig. 2: Model architecture of the Proposed Contextual Embedding applied to the separate tasks of semantic embedding and classification

A. Visual Encoder

Images often consists of multiple visual elements providing different semantic information. Thus we argue that local region patches are better suited[20] to this task than global image-level features. We use two different channels to generate meaningful local patches. We use a pre-trained standard object detector[15], to detect salient objects in the image, which can convey relevant semantic information. Recently it has been shown that symbolism associated with the local visual patches in the image (for instance, concepts like danger, cool, freedom) can play a significant role in semantic understanding [33]. Thus, we also leverage symbol annotations in the dataset[16] and use a pre-trained [33] symbolism detector to generate an additional set of local region patches. A pre-trained deep network [14] is used to extract visual features $v_i \in R^{300}$, corresponding to both detected object and symbolism local patches.

B. Scene Text Encoder

While scene text (present in the image) may be a rich source of information for semantic understanding of the image, extracting that text involves dealing with complexities like cluttered background, orientation, or uneven lighting. Considering that the accuracy of text extraction is a critical factor for later image reasoning, we have analyzed different OCR alternatives[24], [18] to evaluate their impact on the final system. Finally, we settled on using Google Vision API, as it leads to improved text extraction, generating legible scene text for about 94% images. The extracted text is embedded [26] into a word embedding space that encodes the semantics of the text.

a) Anchor based Text Attention: The number of recognized scene text words varies widely and besides, not all of the words are relevant to the given task. Therefore, we propose to encode the detected words in terms of a fixed number of anchors (or clusters) specific for every task. In the case of the tasks in the Ad dataset, the anchors are the 15 statements associated with each image (see section IV-A for details). For the Tweet dataset, we use the same from hatebase.org, as used for dataset collection[13].

Thus given n recognized scene text words $[t_1^{'}..t_n^{'}]$, we encode them as $[t_1..t_k]$, in terms of the k task dependent anchors A_k:

$$t_k = \sum_{i=0}^{n} r_{i,k} t_i^{'}, \text{ where } r_{i,k} = \sum_{j} \frac{1}{1 + d(t_i^{'}, A_{k,j})} \quad (1)$$

$r_{i,k}$ gives the similarity between a scene text word $t_i^{'}$ and anchor A_k, based on distance measure d. When an anchor has multiple words (statements in the ad dataset), $r_{i,k}$ is the sum of similarities with all anchor words $A_{k,j}$. Thus, given a variable number of detected words, only those that are similar to the anchors are considered relevant.

C. Contextual VT (Visual Text) Encoder

One of our main contributions is a representation that captures the rich interplay between the text and visual cues. Such a representation entails a) defining a compositional strategy encoding the contextual relationships among the co-occurring text and visual features, and b) capturing their interaction.

a) Compositional Strategy: Text Visual Graph: In the case of text semantics, the strategy employed to encode context is usually sequential [26] characterized by a sliding window. Our recognized text and visual objects do not have any particular sequence ordering, and thus we take inspiration from the recently proposed graph-structured context[31].

The top 10 visual objects v_j detected by the visual encoder and the k task dependent text anchors t_k provided by the scene text encoder are represented as nodes to construct a fully connected graph $G = (V, E)$, with $V = v \cup t$. However, the text and visual nodes have features from different domains and are not directly comparable. Thus, we augment visual nodes with the mean of their adjacent text nodes, and similarly for text nodes we augment them with the mean of their adjacent visual nodes:

$$x_i = v_i||\bar{t} \quad \text{or} \quad x_i = \bar{v}||t_i \quad (2)$$

We assume the graph is fully connected based on our earlier hypothesis of relatedness between all objects in the image

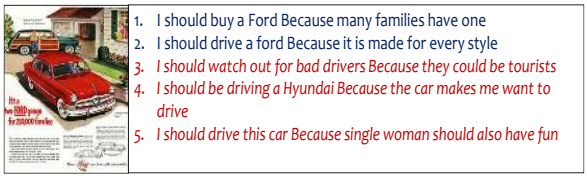


Fig. 3: A sample Ad image, with relevant sentences in blue and irrelevant sentences red. The task is to rank the relevant sentences ahead of the irrelevant ones, given an example Ad image. Showing only 5 of 15 statements for brevity

(text and visual). The edge weights representing the relevance between two nodes are implicitly learned through the Graph Attention Layer.

b) Interaction scheme: Relation Encoder: We model the text-visual interplay in our relationship encoder by allowing attentional interaction amongst the nodes of the graph through a Graph Attention Layer [32]. We allow nodes in a similar context, in this case an image, to influence each other's representation. Our interaction scheme is similar to the *Implicit Relational Encoder* proposed in [23], but we differ in our design of the attention mechanism and also allow for multimodal text-visual interaction. Given the input features x_i of a node, we learn a shared projection matrix W, and perform self-attention on the adjacent nodes to generate the output feature h_i for that node

$$h_i = \sum_j \alpha_{ij}.Wx_i \quad \text{with} \quad \alpha_{ij} = softmax(e_{i,j}) \qquad (3)$$

where α_{ij} is the attention weight defined using $e_{i,j}$ representing the importance of node j to node i. It is computed by a single layer feed forward network akin to [32]. We define the final aggregated contextual feature as $\bar{h}$, the mean of all the nodes.

IV. Application of the Model

A. Task 1: Image-Statement Relevance

For this task we will use the Ad image dataset introduced in [16]. In a later work [33], a retrieval task was proposed, where the goal is to match an Ad image against relevant sentences. For each image 3 relevant and 12 non-relevant sentences are provided. See Fig.3 for an example of this task.

Learning of the contextual semantic embedding: The Image-Statement Relevance task entails matching statements against images. Thus, we need the image and statement representations to be comparable by a distance. As explained in section III-C, images are are represented by the aggregate of their contextual embedding given by $\bar{h}$. Statements are encoded as the aggregate of their word2vec word embeddings. To make them comparable we project the aggregated contextual vector $\bar{h}$ into a semantic space z, where matching against relevant statements is feasible, as depicted in Fig.2. The weights W^s of the projection matrix are learned through triplet training

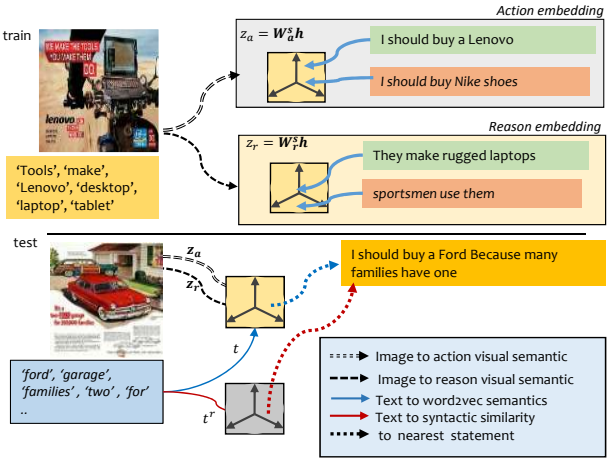


Fig. 4: Sentence Relevance Task: Training, Testing

with relevant and irrelevant statements, minimizing this triplet loss

$$l(z, s, \theta) = \sum_{i=1}^{B} \sum_{j \in ns(i)} \|z_i - s_i\| - \|z_i - s_j\| + \beta \qquad (4)$$

where B is the batch size, β is the margin of triplet loss and z is the semantic embedding, with s_i and s_j being the randomly sampled positive and negative statement semantic features respectively. Given the i^{th} image, $ns(i)$ denotes the set of irrelevant statement s_j is sampled from.

Retrieval framework: We build upon the semantic contextual embedding learnes as explained in the previous section to define the complete framework for the Image-Statement Relevance task shown in Fig.4. We integrate some specific model components that leverage certain properties of the advertisement images to boost performance in this task. More specifically, in training we learn separate semantics based on statement action-reason partitioning. During testing, we integrate additional text channels component to mitigate vocabulary misses.

a) Partitioning: Analogous to some recent works [7], [2] that combine natural language and vision we take advantage of the linguistic structure of the statements. The statements in this task can be *partitioned* into a couplet of action and reason : "I should $< action >$ because $< reason >$." e.g., "I should buy a Lenovo because they are rugged laptops", as can be seen in Fig. 1b. We exploit this structure and learn, using the triplet loss defined in eq.4, two separate semantic embeddings, viz one related to actions z_a, and another one to reasons z_r. Thus, given an image, we now evaluate its relevance separately for the action and the reason. Such partitioning allows not just exploiting fine-grained intermediate data, but it also enables to mitigate long term dependencies associated with long sentences [28].

b) Scene Text Semantic Channel: During test time, while matching images against statements, we augment the trained contextual semantic embedding with an independent scene text semantics channel. This text semantics uses the same anchor based encoding of the scene text as described in eq. 1, but in this case we only use the query statement as a single anchor.

c) Lexical similarity scoring: Often brand names, or brand-related terms, like 'googling, Mcchicken', are not present in the pretrained word2vec vocabulary used for text semantics. This can be further aggravated by erroneous word recognition. In particular, for the Ad dataset, a total of 15% of the 3 million recognized words could not be mapped to any semantic vector. One way to cope with vocabulary misses in word embedding is to use Lexical Similarity. It provides us a way to check for the similarity between the raw scene text and the query statement without the need for any further embedding. This enables us to use all the extracted scene text words, taking advantage of any word correspondence. We measure the lexical distance $d(s_j{}^r, t^r)$, as the cosine distance between tf-idf vectors of the raw scene text words $t_i{}^r$ and the query statement words $s_{j_k}{}^r$.

d) Final matching: Combining contextual semantics with text scoring: The final distance measure used for the ranking of the query statement tries to capture the semantic distance between the statement s_j and the image taking into account the contextual semantic features z (with action-reason partitioning z_a and z_r), the semantic text features t, and the lexical distance from the scene text features t^r. It is given by:

$$\arg\min_{j \in Q} d(z_a, s_{j_a}) + d(z_r, s_{j_r}) + d(t, s_j) + d(t^r, s_j{}^r) \quad (5)$$

where we have Q query statements to rank against an image.

B. Task 2: Classification

We also apply the contextual semantic embedding to two different classification tasks: topic classification on the Ad Image dataset[16] and tweet classification on the MMHS150K dataset[13]. While the Topic classification[16] task consists of categorizing the Ad image under one of 38 different product heads viz, 'car','beauty','coffee', the tweet classification involves marking tweets as hateful or benign. In both cases the classifier is built upon the contextual encoding framework as depicted in Fig.2. In particular, the contextual representation is fed to a softmax classifier and trained end-to-end with cross-entropy loss.

V. Experimental Results

A. Task1: Image-Statement Relevance

For the Image-Statement Relevance task we follow the protocol introduced in the CVPR Workshop [1], and rank 15 statements (3 relevant, 12 non relevant) based on their relevance or similarity to the image.

a) Metrics: We compute 3 different metrics: 1) Accuracy, which records a hit whenever any of the 3 relevant statements is picked 2) Rank Average, which is the average rank of the highest-ranked relevant statement and 3) Recall at 3, which denotes the number of correct statements ranked in top 3. For a good model, we expect high accuracy and recall, with a low average rank.

1) Comparison with the state-of-the-art: In this section, we compare our results with the current state of the art. We first give a brief description of the methods used for comparison. VSE++ [10] is one of the major visual semantic embedding schemes, but it does not incorporate the symbolism or scene text content present in the Ad image. ADVISE [33] played the crucial role of leveraging the symbol annotation[16] present in the dataset, and use the symbol channel in the visual semantic embedding. While these schemes do use external knowledge, in the form of automatically generated captions[19], to augment their visual understanding, we are the first ones to formally introduce scene text in the context of visual understanding. Both VSE++ [10] and ADVISE [33] had also participated in the CVPR 2018 Workshop Challenge, organised on this dataset. In the results we can clearly see the improvement brought upon by our complete framework using contextual semantics trained on visual and text features, augmented with text scoring and statement partitioning.

TABLE I: Comparison with state-of the-art. Results marked with * do not use the exactly our same partitions for training and test.

Model	RankAvg ↓	Accuracy ↑
VSE++ [10]	3.85	66.6 *
ADVISE [33]	3.55	72.84 *
CVPRW winner	-	82 *
Our full system	3.09	**90.9**

2) Ablation Study:

a) Training of Semantic Contextual Embedding: In Tab.II we analyze the contribution of the different channels and components involved in training the semantic contextual embedding. Visual and text baselines are proposed to show the contribution of each individual channel in the contextual embedding (columns 1 and 2 in Tab.II). The visual baseline only uses the ResNet visual features and excludes the use of scene text or GAT in the pipeline. For the text baseline, only the scene text word embeddings are used. In both cases, the local features are aggregated to generate a semantic vector trained with triplet loss. In columns 3 and 4 textual and visual features are fused with simple concatenation while the full model using the contextual VT encoder is shown in column 5.

b) Contribution of the different channels: For the sentence relevance task, as eq.5 shows, the ranking involves, not just the visual semantic features, but also semantic and lexical text features. In Tab.III, we detail the contribution of each of these separate channels. We also show the improvement due to using statement based attention when aggregating scene text features.

TABLE II: Semantic Embedding : Role of Text and Visual channels, partitioning and Contextual VT encoder in semantic embedding

Visual	✓	×	✓	✓	✓
Text	×	✓	✓	✓	✓
Partitioning	×	×	×	✓	✓
Contextual VT Encoder	×	×	×	×	✓
Accuracy ↑	55.6	74.4	82.4	83.5	85.7
RankAvg ↓	4.77	4.31	3.34	3.29	3.2
Recall@3 ↑	1.4	1.7	2.12	2.14	2.19

[1]https://people.cs.pitt.edu/ kovashka/ads_workshop/

TABLE III: Sentence Relevance: Role of components Semantic and Text channel in Sentence Relevance task

Text Semantic	×	✓	×	×
Text Semantic w/ attention	×	×	✓	✓
Lexical	×	×	×	✓
Semantic Embedding	✓	×	×	✓
Accuracy ↑	85.76	72	74.4	90.9
RankAvg ↓	3.2	4.52	4.31	3.08
Recall@3 ↑	2.19	1.6	1.7	2.3

Fig. 5: Semantic Retrieval. The semantic features of the query image, bounded in red, is used to find its top 3 matches among the other test images. The Top row lists images, that were retrieved using Visual cues based semantic feature. The bottom row uses our contextual encoder for semantic features, using both scene text and visual cues. Our improvements leads to retrieval of images not just pertaining to cars, but also gets the type and brand right.

3) Qualitative Results: Fig.5 shows examples of query by image, i.e. the semantic features of an image are used to find similar images. We show that the proposed scheme can encode the visual and text cues, and generate a holistic semantic feature. Comparison with the baseline that uses only visual features shows the effectiveness of scene text in generating more fine-grained results.

In Fig.7a, we display instances where the visual features by themselves were not able to map the image to the correct statements, and we had to incorporate scene text in the semantic representation. This can be attributed to the co-occurrence of certain semantically related words in both the scene text and the relevant statements. However, the simple co-occurrence of semantically related words does not suffice for all examples, as is illustrated in Figs.7b and 7c. In particular, in Fig.7c, we show test instances that were only correctly mapped to their statements when we incorporated the relationship encoder, going beyond simple visual or text similarity and exploiting non-literal relationships. For example, in the second example, it had to relate that getting the service amounted to bridging the challenge between being anxious and excited.

B. Task2: Classification

a) Topic Classification: In the Topic classification task the objective is to classify an Ad image into 1 of 38 Topic

Fig. 6: Qualitative Results on Classification task. Correct class labels are marked in green, and incorrect ones are marked with red

TABLE IV: Classification results

Ads Dataset		MMHS150K	
Dey et al[8]	58		
Hussain et al [16]	64.34	SCM[13]	68.5
pretrained on Task1	66.35	TKM[13]	68.2
Our Model trained	**69.23**	**Our Model trained**	**67.44**

classes. Topic classification was initially attempted by [16] by training 152-layer ResNet using the visual features only. In Tab.IV row 1, we see another scheme[8] which uses both text and visual features, but uses simple concatenation to aggregate them. Thus, we stress that simple use of text is not enough; we have to find ways to capture the multimodal interaction, as in our contextual encoding. In row 3, we observe that using features from the network trained on the sentence relevance task, we can already improve upon the previous results. In row 4, we present the results of our final end to end trained topic classifier.

b) Tweet Classification: In the recently proposed Hate Speech dataset MMHS150K[13], the binary classification task of marking tweets, containing both visual and text content, as "hateful" and "benign" was proposed. We address the class imbalance problem in the original dataset by training on equal number of random samples from each class. Application of our contextual encoder, leads to results comparable with the multimodal models proposed by the authors

VI. Conclusion

We proposed a framework for interpreting images by leveraging both visual and text contents present in the images. Visual cues, both symbols and objects, together with scene text are extracted and embedded in a semantic space trained with triplet loss. Our embedding also incorporates text-visual

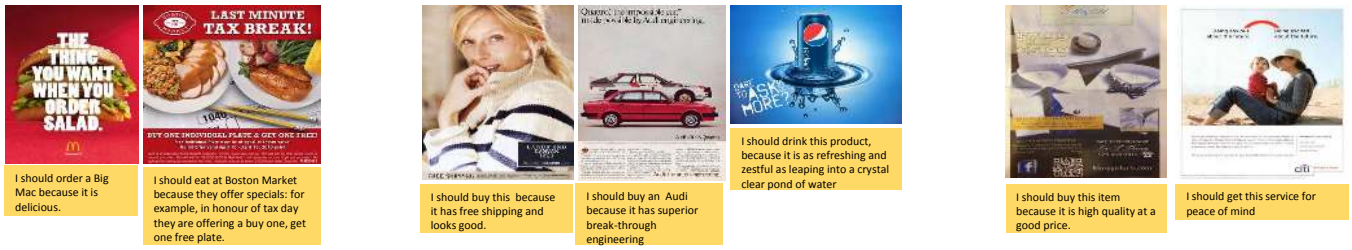


Fig. 7: (a) Examples where just visual features were not enough to generate robust semantics, and only with the incorporating of Scene Text content in Semantic Embedding were we able to map the images to their relevant statements. (b)Examples where semantic features with visual and text, still fell short of retrieving the correct relevant statements and Partitioning had to help. (c)Examples that required exploiting Non-Literal relationships using our visual text relationship encoder.

inter-object dynamics encoding, which leads to capturing non-literal relationships between the detected objects. This idea of extracting and encoding, followed by embedding in semantic space, finds application in semantic retrieval and classification tasks.

In addition, we leverage the linguistic structure, training separate branches of the network for action-reason parts of the statement. We augment the visual-text semantic representation of the image with the lexical similarity between scene text and the query statement. Results confirm our initial hypothesis that scene text plays an important role in semantic understanding of images. These results encourage us to extend the application of our framework to more generic domains, for instance, the recently released datasets[30], [6] for VQA using scene texts.

References

[1] K. Ahuja, K. Sikka, A. Roy, and A. Divakaran. Understanding visual ads by aligning symbols and objects using co-attention. *arXiv preprint arXiv:1807.01448*, 2018. 2

[2] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein. Learning to compose neural networks for question answering. *CoRR*, abs/1601.01705, 2016. 4

[3] S. Bai and S. An. A survey on automatic image caption generation. *Neurocomputing*, 311:291–304, 2018. 2

[4] X. Bai, M. Yang, P. Lyu, and Y. Xu. Integrating scene text and visual appearance for fine-grained image classification with convolutional neural networks. *arXiv preprint arXiv:1704.04613*, 2017. 1, 2

[5] H. Ben-younes, R. Cadène, M. Cord, and N. Thome. Mutan: Multimodal tucker fusion for visual question answering. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2631–2639, 2017. 2

[6] A. F. Biten, R. Tito, A. Mafla, L. Gomez, M. Rusiol, E. Valveny, C. V. Jawahar, and D. Karatzas. Scene text visual question answering, 2019. 1, 2, 7

[7] K. R. Chandu, M. A. Pyreddy, M. Felix, and N. N. Joshi. Textually enriched neural module networks for visual question answering. *CoRR*, abs/1809.08697, 2018. 4

[8] A. U. Dey, S. K. Ghosh, and E. Valveny. Don't only feel read: Using scene text to understand advertisements. *CoRR*, abs/1806.08279, 2018. 6

[9] R. Doshi and W. Hinthorn. Symbolic vqa on visual advertisements with symvise networks. *p*, 2018. 2

[10] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler. Vse++: Improved visual-semantic embeddings. *arXiv preprint arXiv:1707.05612*, 2017. 2, 5

[11] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov, et al. Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*, 2013. 2

[12] 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*. ACL, 2016. 2

[13] R. Gomez, J. Gibert, L. Gomez, and D. Karatzas. Exploring hate speech detection in multimodal publications. *arXiv preprint arXiv:1910.03814*, 2019. 1, 2, 3, 5, 6

[14] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016. 3

[15] J. Huang, V. Rathod, C. Sun, M. Zhu, A. Korattikara, A. Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama, et al. Speed/accuracy trade-offs for modern convolutional object detectors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017. 3

[16] Z. Hussain, M. Zhang, X. Zhang, K. Ye, C. Thomas, Z. Agha, N. Ong, and A. Kovashka. Automatic understanding of image and video advertisements. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017. 1, 2, 3, 4, 5, 6

[17] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman. Synthetic data and artificial neural networks for natural scene text recognition. *arXiv preprint arXiv:1406.2227*, 2014. 1

[18] 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), 2016. 1, 3

[19] J. Johnson, A. Karpathy, and L. Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 2, 5

[20] S. Karaoglu, R. Tao, T. Gevers, and A. W. Smeulders. Words matter: Scene text for image classification and retrieval. *IEEE Transactions on Multimedia*, 19(5), 2017. 1, 2, 3

[21] S. Karaoglu, R. Tao, J. C. van Gemert, and T. Gevers. Con-text: Text detection for fine-grained object classification. *IEEE Transactions on Image Processing*, 26(8), 2017. 1, 2

[22] S. Karaoglu, J. C. van Gemert, and T. Gevers. Con-text: text detection using background connectivity for fine-grained object classification. In *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 2013. 1, 2

[23] L. Li, Z. Gan, Y. Cheng, and J. Liu. Relation-aware graph attention network for visual question answering. *arXiv preprint arXiv:1903.12314*, 2019. 2, 4

[24] M. Liao, B. Shi, X. Bai, X. Wang, and W. Liu. Textboxes: A fast text detector with a single deep neural network. In *AAAI*, 2017. 1, 3

[25] X. Liu, D. Liang, S. Yan, D. Chen, Y. Qiao, and J. Yan. Fots: Fast oriented text spotting with a unified network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018. 1

[26] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013. 2, 3

[27] V. Mnih, N. Heess, A. Graves, and K. Kavukcuoglu. Recurrent models of visual attention. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pages 2204–2212, Cambridge, MA, USA, 2014. MIT Press. 2

[28] Z. Niu, M. Zhou, L. Wang, X. Gao, and G. Hua. Hierarchical multimodal lstm for dense visual-semantic embedding. In *Computer Vision (ICCV), 2017 IEEE International Conference on*. IEEE, 2017. 4

[29] J. Pennington, R. Socher, and C. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014. 2

[30] A. Singh, V. Natarajan, M. Shah, Y. Jiang, X. Chen, D. Batra, D. Parikh, and M. Rohrbach. Towards vqa models that can read. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

1, 2, 7

[31] S. Vashishth, M. Bhandari, P. Yadav, P. Rai, C. Bhattacharyya, and P. Talukdar. Incorporating syntactic and semantic information in word embeddings using graph convolutional networks. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019. 2, 3

[32] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017. 1, 2, 4

[33] K. Ye and A. Kovashka. Advise: Symbolism and external knowledge for decoding advertisements. *arXiv preprint arXiv:1711.06666*, 2017. 1, 2, 3, 4, 5

[34] M. Zhang, R. Hwa, and A. Kovashka. Equal but not the same: Understanding the implicit relationship between persuasive images and text. *CoRR*, abs/1807.08205, 2018. 1