Multi-Modal Reasoning Graph for Text-based Fine-Grained Image Classification and Retrieval

Andrés Mafla Sounak Dey Ali Furkan Biten Lluís Gómez Dimosthenis Karatzas
Computer Vision Center, UAB, Spain
{amafla, sdaey, abiten, lgomez, dimos}@cvc.uab.es

Abstract

Scene text instances found in natural images carry explicit semantic information that can provide important cues to solve a wide array of computer vision problems. In this paper, we focus on leveraging multi-modal content in the form of visual and textual cues to tackle the task of fine-grained image classification and retrieval. First, we obtain the text instances from images by employing a text reading system. Then, we combine textual features with salient image regions to exploit the complementary information carried by the two sources. Specifically, we employ a Graph Convolutional Network to perform multi-modal reasoning and obtain relationship-enhanced features by learning a common semantic space between salient objects and text found in an image. By obtaining an enhanced set of visual and textual features, the proposed model greatly outperforms previous state-of-the-art in two different tasks, fine-grained classification and image retrieval in the ConText[24] and Drink Bottle[4] datasets.

Figure 1. The proposed model employs a Graph-based Multi-Modal Reasoning (MMR) module to enrich location-based visual and textual features in a combined semantic representation. The network learns at the output of the MMR to map strong complementary regions of visual (red) and text (green) instances to obtain discriminative features to perform fine-grained image classification and retrieval.

1. Introduction

Since the advent of written text to represent ideas, humans have employed it to communicate non-trivial and semantically rich information. Nowadays, text can be found in an ubiquitous manner in images and video, especially in urban and man-made environments[54, 25]. Extracting and analyzing such textual information in images jointly with the visual content, is indispensable to achieve full scene understanding. In this work, we explore the role of such multi-modal cues, specifically in the form of visual and textual features to solve the task of fine-grained image classification and retrieval.

The task of fine-grained image classification (FGIC) consists of labeling a set of images that are visually alike. A lot of research on this problem has been oriented to differentiate visually similar objects such as birds[16], aircrafts [41], and dog breeds[27] among others, which more often than not require domain specific knowledge. However, differentiating objects by leveraging available textual instances in the scene is an omnipresent practice in daily life. A seminal work on leveraging textual cues was presented by Movshovitz *et al*. [43], who showcased that in order to classify store-fronts, a trained Convolutional Neural Network (CNN) had automatically learned to attend to scene text instances as the sole way to solve the given task. In this manner, scene text found in an image serves as an additional discriminative signal that a model should incorporate into its design. Additional research devoted to leveraging textual cues in the task of FGIC has been explored. Similar to our work, Karaoglu *et al*. [24, 23] introduces a simple pipeline to perform fine-grained classification using scene text and extending the previous work, an attention mechanism is proposed by Bai *et al*. [4] to learn a common semantic space. In a different approach, Mafla *et al*. [40] learns a morphological space by using textual instances as discrimi-

native features rather than semantics to solve this task.

Departing from previous approaches, we exploit a structural representation between the studied modalities. Our work summarized in Figure 1 with publicly available code at [1], focuses on learning an enhanced visual representation that incorporates reasoning between salient regions of an image and scene text to construct a semantic space over which fine-grained classification is performed. In this example, we can observe that relevant regions such as the text "Bakery" and "Bread" are associated with a visual region that depicts pastry, both important cues to classify the given image. Additionally, we show experiments of fine-grained image retrieval, using the same multi-modal representation, in the two evaluated datasets. Overall, the main contributions can be summarized as following:

- We propose a novel architecture that surpasses previous state-of-the-art results by more than 5% on fine-grained classification and 10% on image retrieval by considering text and visual features of an image.
- We design a fully end-to-end trainable pipeline that incorporates a Multi-Modal Reasoning module that does not rely on ensemble models or pre-computed features.
- We provide exhaustive experiments in which we analyze the effectiveness of different modules in our model architecture and the importance of scene text towards comprehensive models of image understanding.

2. Related Work

2.1. Scene Text Detection and Recognition

Localizing and recognizing text instances found in a natural image is a challenging problem due to the variability, orientation, occlusion and background noise among other factors [10]. Initial approaches on text detection and recognition in the wild relied on the usage of hand-crafted features such as histogram of oriented gradients [55] descriptors and connected components [45]. However, those approaches were outperformed by Deep Learning based methods that implicitly learn features that are representative for a given task [31]. Deep learning based methods began with the work proposed by [22] which focused on a sliding window and a CNN to filter the proposals. The proposals were used as input into another CNN that posed the task as a classification problem over a large fixed dictionary of words. Later works take object detection pipelines such as YOLO [48] used by [19] to obtain a Fully Convolutional Neural Network along with a focus on generating synthetic training data, which later became the go-to data to train text detectors and recognizers. Along these lines, a variation of SSD [38] is presented by [35, 34] to develop a text detector which easily integrates with a module trained for recognition. Methods that focus on an end-to-end recognition have been explored by [7] based on Faster R-CNN [49], which performs text detection and incorporates a Connectionist Temporal Classification (CTC) [18] to recognize a given text instance. Similarly, [21] presents a CNN as a region-based feature extractor, which are fed to two attention-based Long-Short Term Memories (LSTM) to predict bounding boxes and recognize the textual proposals. Multi-lingual models have been proposed as in the case of [8], work that uses a CNN as an encoder and a CTC to decode the characters from a set of different languages.

On a different approach, the Pyramidal Histogram of Characters (PHOC) [1] is used to represent words and it has been amply used in text spotting in documents [52] and in text retrieval in natural images [17]. Despite all the progress done in scene text detection and recognition, it remains as an open problem in the computer vision community, with a special focus placed lately on multi-oriented text localization and recognition.

2.2. Fine-Grained Classification

The task of Fine-Grained Image Classification (FGIC) focuses on finding discriminative visual regions that often require domain specific knowledge to correctly perform the labeling task [56]. Different to visual based FGIC methods, there has been growing interest to use textual cues to achieve this task by incorporating two modalities. Closely related to this work, the initial approach taken by [23] was to extract scene text and construct a bag of words, while the visual features where obtained by employing a pre-trained GoogLeNet [53]. Soon after, [4] proposes the usage of textboxes [35] to read text instances in an image, a CNN to obtain visual features along with an attention mechanism and a concatenation of the final features to learn a semantic space suitable for FGIC. Later work performed by [40] employs a CNN as a visual feature extractor and uses the PHOC representation of a word along with the Fisher Vector [47] to learn a space based on the morphology of text instances to overcome Optical Character Recognition (OCR) errors. Several fusion methods are explored in the work by [40] but finally a concatenation of features is performed to solve the task of image classification and retrieval.

2.3. Multi-Modal Fusion and Reasoning

Several fusion-based techniques such as MCB [15, 13], MLB [28] and Block [5] have been explored to model relationships between language and vision. To model these interaction, attention-based [3] approaches also have been have been proposed [2, 57, 26]. With the aim of designing models capable of reasoning, the intrinsic synergy between visual and textual features has been explored. Work such as [60, 32] employ variations of an LSTM and a Gated Recurrent Unit (GRU) to perform reasoning in a sequential manner. However, significant advances have been made by

[1] `http://www.uponAcceptance.com`

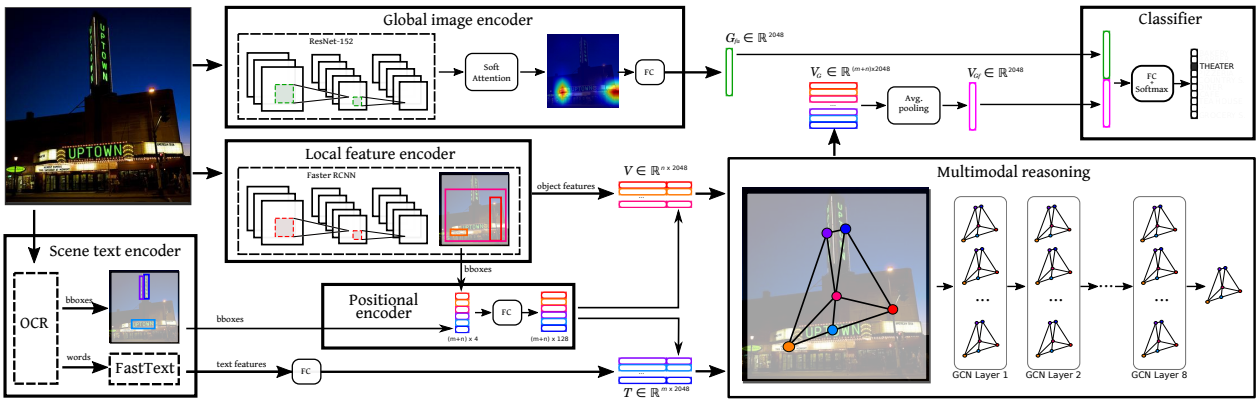

Figure 2. Detailed model architecture. The proposed model combines features of regions of scene text and visual salient objects by employing a graph-based Multi-Modal Reasoning (MMR) module. The MMR module enhances semantic relations between the visual regions and uses the enriched nodes along with features from the Global Encoder to obtain a set of discriminatory signals for fine-grained classification and retrieval.

the usage of Graph Convolutional Networks (GCN) [29], due to the proven capability of modelling relationships [50] between nodes in a given graph. Along this road, GCNs have been successfully used in VQA [44, 51, 14], image captioning [33, 58] and image-sentence retrieval [32, 36].

In this work, we propose a method to learn a richer set of visual features and model a more discriminative semantic space by employing a GCN. To the best of our knowledge, this is the first approach that integrates multimodal sources which come in the form of visual along textual features jointly with positional encoding into a GCN pipeline that performs reasoning for the task of fine-grained image classification and retrieval.

3. Method

In this section, we detail each of the components that comprise the proposed architecture. Figure 2 depicts the overall scheme of the proposed model, which is formed by 6 different modules: global image encoder, local feature encoder, text encoder, positional encoder, multi-modal reasoning graph and classification module. The local feature encoder employs features extracted based on the regions of interest obtained by a Faster R-CNN [49] in a similar manner as the bottom-up attention model [2]. The scene text encoder uses an OCR model to obtain scene text and further embed it into a common space. The goal is to obtain multi-modal node representations that leverage the semantic relationships found between salient objects and text instances within an image that are discriminative enough to perform fine-grained classification.

3.1. Global Image Encoder

We employ a CNN as an encoder, which in our case is a ResNet-152 [20] pre-trained on ImageNet [11] to acquire global image features. Particularly, given an image I we take the output features before the last average pooling layer and the output is denoted as $G_f = \psi(I)$. In order to obtain a more descriptive set of global features and due to its differentiable properties, we compute a soft attention mechanism on top of the global features in a similar way as [12]. This self-attention mechanism provides an attention mask, $attn_{mask}$, that assigns weights on different regions of the input image. The attention weights are learned in an end-to-end manner by convolving 1×1 kernels that are projected into a single dimensional filter and later followed by a Softmax function. In order to preserve information from the original global features, a residual connection is employed after multiplying the attention mask with the global features and then we pass the features through a Fully-Connected, FC, layer in the form of:

$$G_{fa} = FC(G_f + (G_f \times attn_{mask})) \quad (1)$$

where $G_{fa} \in \mathbb{R}^{1\times D}, G_f \in \mathbb{R}^{HxWxD}, attn_{mask} \in \mathbb{R}^{HxW}$ stands for the final encoded global features, where $D = 2048$, $H = 7$ and $W = 7$.

3.2. Local Feature Encoder

Following [2], we employ a Faster R-CNN [49] pre-trained on Visual Genome [30] as the extractor of local visual features. This approach allows us to obtain salient image regions that are potentially discriminative for our task. By using an IOU threshold of 0.7 and a confidence threshold of 0.3, we sort the obtained predictions before the last average pooling layer and use the n most confident regions of interest. Thus, we can represent the output of an image I with a set of local features $R_f = \{(r_1, bbox_{r_1})..., (r_n, bbox_{r_n})\}, r_i \in \mathbb{R}^d$, where r_i is i^{th} region of interest and $bbox_{r_i}$ is the r_i's corresponding bounding box coordinates normalized with respect to the image.

In our experiments, we set $n = 36$ and the obtained features have a dimension of $d = 2048$. In order to encode the local visual features, we project the features through a fully-connected layer using the equation:

$$v_i = W_r \times r_i + b_r \tag{2}$$

where W_r is the weight matrix and b_r is the bias. In this manner we obtain the final encoded local features that will serve as input to the multi-modal GCN in the form of $V_f = \{v_1, ..., v_n\}, v_i \in \mathbb{R}^D$, where $D = 1920$ is the dimension of the final embedding space. The bounding boxes obtained to represent these regions are later used as input into the positional encoder module. If there are less than $n = 36$ regions in an image, a zero padding scheme is adopted.

3.3. Text Encoder

In order to extract text contained in an image, we ran several public state of the art text recognizers as well as a commercial OCR model provided by Google[2]. We extract the transcriptions of each word, denoted as w_i, as well as the corresponding bounding boxes, $bbox_{w_i}$. In particular, we extract the top m most confident textual instances found in an image. The transcriptions are embedded using Fasttext [6] and the bounding boxes will be used as input in the positional encoder branch. We employ the FastText embedding due to its capability of encoding word morphology in the form of n-grams as well as preserving a semantic space similar to Word2Vec [42] while at the same time dealing with out of vocabulary words.

We project the obtained embedded textual features by passing them through a fully-connected layer represented by the following equation:

$$t_i = W_t \times w_i + b_t \tag{3}$$

where W_t, b_t stands for the learnable weights and biases, respectively. The final textual features $T_f = \{t_1, ..., t_m\}, t_i \in \mathbb{R}^D$, where $D = 1920$ is the dimension of the final embedding space and $m = 15$ is the number of text proposals extracted from an image. In the case that there is no text found in a given image, similarly to the local encoder module, zero padding is employed.

3.4. Positional Encoder

Encoding the position of objects and text instances within an image can provide important relational information about the scene. For example text found on top of a building often refers to its class in a explicit manner contrary to text found in any other location in the image. To meet this end, we design a positional encoding that takes as input a predicted bounding box of an object or text instance. The input to the positional encoder describes the top left (x_1, y_1), and bottom right (x_2, y_2) coordinates normalized according to the image size, and is a concatenation of the bounding boxes of the local and text regions of interest. The bbox matrix is given by: $bboxes_{input} = \{bbox_{r_1}, .., bbox_{r_n}, bbox_{t_1}, ..., bbox_{t_m}\}$ where $bbox_i = (x_1, y_1, x_2, y_2)$. In order to encode them, we pass the bounding boxes over a fully-connected in a similar way as the same as previous sections. The final encoded representation can be described as: $bboxes = \{bbox_{r_1}, .., bbox_{r_n}, bbox_{t_1}, ..., bbox_{t_m}\}, bbox_i \in \mathbb{R}^b$, in which the dimension $b = 128$ represents the final encoded bounding boxes.

3.5. Multi-modal Reasoning Graph

Due to the showcased capability of graphs to describe complex relations between objects [51, 58, 14, 36], we construct a richer set of region-based visual descriptors that exploit the semantic correlation between visual and textual features. In order to do so, we initialize the node features as local visual features and textual features concatenated with their respective positional encoding of bounding boxes. We can describe the node features as:

$$V = \{(v_1, bbox_{r_1}), ..(v_n, bbox_{r_n}), (t_1, bbox_{t_1}), ... \\ ..., (t_m, bbox_{t_m})\}, v_i, t_j \in \mathbb{R}^{(n+m)\times D}$$

where n, m is the number of visual and textual features, respectively. In our case, $n+m = 51$ and $D = 1920+128 = 2048$. Furthermore, we construct the affinity matrix R which measures the degree of correlation of between two visual regions. The construction of the affinity matrix is given by:

$$R_{ij} = \phi(k_i)^T \gamma(k_j) \tag{4}$$

where $k_i, k_j \in V$, $\phi(.)$ and $\gamma(.)$ are two fully connected layers that are learned end-to-end by back propagation at training time. If we define $k = n + m$, then the obtained affinity matrix consists of a shape $k \times k$. Once R is calculated, we can define our graph by $G = (V, R)$, in which the nodes are represented by the local and textual features V and the edges are described by R. The obtained graph describes through the affinity matrix R the degree of semantic and spatial correlation between two nodes. We use the formulation of Graph Convolutional Networks given by [29] to obtain a reasoning over the nodes and edges. Particularly, we use residual connections in the GCN formulation as it is presented by [32]. We can write the equation that describes a single Graph Convolution layer performed as:

$$V_g^l = W_r^l(R^l V^{l-1} W_g^l) + V^{l-1} \tag{5}$$

where $R \in \mathbb{R}^{k\times k}$ is the affinity matrix , $V \in \mathbb{R}^{k\times D}$ the local visual features, $W_g \in \mathbb{R}^{D\times D}$ is a learnable weights

[2] https://cloud.google.com/vision/

matrix of the GCN, $W_r \in \mathbb{R}^{k \times k}$ corresponds to the residual weights matrix and l is the number of GCN layer. Notice that passing V through the GCN layer, a richer set of multi-modal features is obtained. In order to find an enhanced representation of the visual features we apply $l = 8$ GCN layers in total, which finally yields a set of enriched nodes that represent the visual features V_G such that:

$$V_G = \{v_{g1}, .., v_{gk}\}, V_G \in \mathbb{R}^{k \times D}$$

3.6. Classification

In order to combine the global G_{fa} and the enriched local and textual V_G visual features, firstly we perform an average pooling of the V_G tensor. Specifically, we can rewrite the final local feature vector V_{Gf} as:

$$V_{Gf} = \frac{1}{k} \sum_{n=1}^{k} V_{gi} \tag{6}$$

Lastly, we simply concatenate the two obtained vectors V_{Gf} and G_{fa}, to obtain the final vector F that is used as input for the final fully-connected layer for classification denoted by: $F = [G_{fa}, V_{Gf}]$

By applying a softmax to the output of the final layer, we obtain a probability distribution of a class label given an input image. The model is trained in an end-to-end fashion optimized with the cross entropy loss function described by:

$$J(\theta) = -\frac{1}{N} \sum_{n=1}^{N} \sum_{i=1}^{\mathcal{C}} y_i^n log(p_i^n) \tag{7}$$

Where, C is the number of classes, N the dataset samples such that each pair contains an annotation $\{x^{(n)}, y^{(n)}\} | n = 1, 2, ..., N$, and p^n is the predicted output label.

4. Experiments and Results

This section presents an introduction to the datasets employed in this work, as well as the implementation details, ablation studies performed, and a thorough analysis of the results obtained in the experiments conducted.

4.1. Datasets

The Con-Text dataset was introduced by Karaoglu *et al*. [24] and is a subset of ImageNet [11], constructed by selecting the sub-categories of "building" and "place of business". This dataset contains $24,255$ images in total divided into three-folds to divide training and testing sets. This dataset introduces 28 visually similar categories of images such as Cafe, Pizzeria, and Pharmacy in which in order to perform fine-grained classification, text is a necessary cue to solve otherwise a very difficult task even for humans. This dataset closely resembles natural circumstances due to the fact that the images are taken without considering scene text instances, thus some images do not have text present in them.

The Drink Bottle dataset was presented by Bai *et al*. [4] and as the Con-Text dataset, it is a subset of images of ImageNet [11], specifically taken from the sub-categories of soft drink and alcoholic drink. The dataset is divided in three-folds as well and contains $18,488$ images. There are 20 image categories which include visually similar instances such as Coca Cola, Pepsi Cola and Cream Soda.

4.2. Implementation Details

In our experiments in order to extract salient visual regions of an image, we use the same settings as [2]. We take the top $n = 36$ ROIs and encode them along with their bounding boxes into a common space of 2048-d. The transcribed text is sorted by confidence score and we take the top $m = 15$ confident predictions. We embed the textual instances by a using a pre-trained FastText model with 1 million 300_d word vectors, trained with sub-word information on Wikipedia2017, UMBC webbase corpus and statmt.org news dataset. The obtained 300-d textual vectors, and following the case of visual features, are projected with the corresponding bounding boxes into a 2048-d space. The Faster R-CNN [49] from [2] and the OCR models, both employed as initial feature extractor modules use pre-trained weights and are not updated at training stage. The rest of the weights of each module in the model are learned in an end-to-end manner during training. The graph-based multimodal reasoning module employs 8 multi-modal GCN layers to obtain the final enriched visual features. In the last full-connected layer before classification, we employ a dropout rate of 0.3 to avoid over-fitting on the evaluated datasets. In general, we employ Leaky ReLU as an activation function in all layers except in the last one, in which we use a Softmax to compute the class label probabilities. The proposed model is trained for 45 epochs, but an early stop condition is employed. We use a combination of optimizers comprised by RAdam [37] and Lookahead [61]. The batch size employed in all our experiments is 64, with a starting learning rate of 0.001 that decays by a factor of 0.1 on the epochs 15, 30 and 45. The momentum value used on the optimizers is 0.9 and the weight decay is 0.0005.

4.3. Comparison with the State-of-the-Art

We show the experimental results of our method compared to previous state-of-the-art on Table 1. We can note that the performance obtained in the Con-Text significantly surpasses the previous best performing method by 5.9%. The improvement in the Drink-Bottle dataset is more modest, of about 1.98%, however it is still significant. We believe the improvement is greater in Con-Text due to the text instances found in it, which refer mostly to business places

without particular out of vocabulary words, therefore a semantic space for classification is more discriminative when compared to the Drink-Bottle dataset. To provide further insights, we conducted experiments by employing the final model along with different OCRs and word embeddings in both datasets. It is essential to note that state-of-the-art results are achieved by the usage of other OCRs as well, proving the proposed pipeline outperforms previous methods. We should point out that when learning a semantic space, the performance of the models remains highly dependant on the accuracy of the employed OCR. Results showing the classification scores of each evaluated class and a further analysis are shown in the Supplementary Material section.

Method	OCR	Emb.	Context	Bottles
Karao.[24]	Custom	BoB[1]	39.0	–
Karao.[23]	Jaderberg	Probs[2]	77.3	–
Bai[4]	Textboxes	GloVe	78.9	–
Bai[4]†	Textboxes	GloVe	79.6	72.8
Bai[4]†	Google OCR	GloVe	80.5	74.5
Mafla[40]	SSTR-PHOC	FV	80.2	77.4
Proposed	E2E-MLT	Fasttext	82.36	78.14
Proposed	SSTR-PHOC	PHOC	82.77	78.27
Proposed	SSTR-PHOC	FV	83.15	77.86
Final	**Google OCR**	**Fasttext**	**85.81**	**79.87**

Table 1. Classification performance of state-of-the art methods on the Con-Text and Bottles dataset. The results depicted with † are based on an ensemble model. The embeddings labeled as: [1] refer to a Bag of Bigrams, [2] is a probability vector along a dictionary. The acronym FV stands for Fisher Vector. The metric depicted is the mean Average Precision (mAP in %).

When comparing to previous methods it is worth revisiting previous approaches. The results reported by [4] use an ensemble of classifiers in order to reach the reported performance. As an additional experiment to showcase the effect of using the same OCR as our proposed model is included. On the other side, the work done by [40] requires offline pre-computation of the Fisher Vector by training a Gaussian Mixture Model and tuning the hyper-parameters involved. In this manner, the method proposed in this work do not require an ensemble and the features used are learned in an end-to-end manner at training time. We clearly show that the proposed pipeline surpasses other approaches even when employing a set of different scene-text OCRs.

4.4. Importance of Textual Features

In order to assess the importance of the scene text found in images, we follow the previous works [23, 4, 40] by defining two different evaluation baselines, the visual features based and the textual features based. Moreover, due to the fact that the evaluated datasets do not contain text transcriptions as ground truth, we evaluated the effectiveness of the OCR employed in the fine-grained classification task.

	Model	Con-Text	Bottles
Visual	CNN	62.11	65.15
	CNN + Self Attention	63.78	66.62
Textual	Texspotter+w2v†	35.09	50.68
	Texspotter+glove†	34.52	50.26
	Texspotter+fasttext†	36.71	51.93
	E2E_MLT+w2v†	44.36	43.98
	E2E_MLT+glove†	44.25	42.64
	E2E_MLT+fasttext†	45.07	44.31
	FOTS+w2v	43.22	41.33
	FOTS+glove	43.71	41.85
	FOTS+fasttext	44.19	42.69
	Google OCR+w2v	53.87	53.47
	Google OCR+glove	54.48	54.39
	Google OCR+fasttext	**55.61**	**55.16**
	PHOC†	49.18	52.39
	Fisher Vector (PHOC)†	**63.93**	**62.41**

Table 2. Visual only and Textual only results. The textual only results were performed on the subset of images that contained spotted text. The results with † were reported by [40]. The metric depicted is the mean Average Precision (mAP in %).

The visual only evaluates all the test set images that belong to a specific split by only employing the global encoder features G_f in the first scenario and the global encoder along with the self attention features G_{fa} in the second scenario. In both cases the output of the global encoder, a 2048-d feature vector, is directly passed through a fully connected layer to obtain the final classification prediction. In the textual only, the baselines are evaluated only in the subset of images which contained spotted scene text. The results of each baseline by employing visual only, different OCRs and word embeddings are shown in Table 2. To provide with comparable results and following the approach given by [40], we employ $m = 15$ text instances on each experiment and pre-trained word embeddings are used, which yield a 300-d vector in the case of Word2Vec [42], GloVe [46] and FastText [6]. The textual tensor obtained is used as input to a fully connected layer, which output is used for classification purposes. In our experiments we evaluate two additional state-of-the-art scene text recognizers, FOTS [39] and the commercially used Google OCR Cloud Vision based on an API. We note that the embedding that performs the best is Fasttext due to the capability of embedding out of vocabulary words by using character n-grams. Regarding the results, it was found that the best performing standard recognizer is the Google OCR, which employs a more compact (300-d) vector compared to a PHOC or a Fisher Vector. The PHOC embedding employs a 604-d feature vector along with $m = 15$ and the Fisher Vector is a single 38400-d vector in our experiments. Overall, by using only textual features, the Fisher Vector based on PHOCs remains as the best performing descriptor. However, be-

Input	MMR	Projection	Late Fusion	Con-Text	Bottles
G_f	-	-	-	62.11	65.15
G_{fa}	-	-	-	63.70	66.56
G_{fa} + V	-	Avg Pooling	Concat	70.48	73.21
G_{fa} + V + T	-	Avg Pooling	Concat	78.72	76.43
G_{fa}	V	Avg Pooling	Concat	72.88	74.96
G_{fa} + T	V	Avg Pooling	Concat	82.51	77.46
G_{fa}	V+T	Avg Pooling	Concat	84.92	79.31
G_{fa}	V+T	Attention	Concat	80.39	78.15
G_{fa}	V+T	Attention	Hadamard	80.62	78.23
G_{fa}	V+T	GRU	Concat	83.32	78.28
G_{fa}	V+T	GRU	Hadamard	83.44	78.05
G_{fa}	V+T	Avg Pooling	MLB [28]	84.51	78.55
G_{fa}	V+T	Avg Pooling	Block [5]	85.02	77.68
G_{fa}	V+T+BBox	Avg Pooling	Hadamard	**86.15**	79.38
G_{fa}	**V+T+BBox**	**Avg Pooling**	**Concat**	85.81	**79.87**

Table 3. Results obtained on each incremental step of our model. *Input* depicts the features that the model is using that are not an input to the multi-modal reasoning module (MMR). *MMR* refers to the features being used by itself. *Projection* is the method used to reduce the dimensionality of the output coming from the MMR as V_G. *Late Fusion* showcases the method employed to combine the features coming from V_{Gf} and G_{fa}. Results are shown in terms of the mAP(%).

sides the high dimensional vector employed, extensive offline pre-computation is required to obtain such descriptor. Nonetheless, as it can be seen in Table 1, the FV descriptor does not achieve the best results in our model.

4.5. Ablation studies

In this section, we present the incremental improvements and the effects obtained by the addition of each module that comprises the final pipeline in the method proposed. Table 3 shows the quantitative results obtained on each variation of the model. Initially by using self attention on the global image encoder CNN an improvement of 1.59% on Con-Text and 1.41% on the Drink-Bottle dataset is reached. In order to assess the effectiveness of the multi-modal reasoning graph module, we compare a model that uses the Faster R-CNN ROIs without the usage of the MMR. It is observed that solely by using the Faster R-CNN features, a significant boost is achieved. Nonetheless, the improvement is accentuated by the usage of the MMR module, which produces as output richer local features coming from the graph nodes. One of the biggest improvements is achieved by the usage of scene text, which enforces the idea that textual information is essential to successfully discriminate between visually similar classes. By the incorporation of scene text, an improvement of 9.7% is gained in Con-Text and 2.5% in the Drink-Bottle datasets. However, another increase in performance is gained by the incorporation of the textual instances into the MMR along with the local features. Due to several works showing performance gains by the usage of attention [59, 57] and Recurrent Neural Networks [32, 9] as reasoning modules, we explored those alternatives as well without a significant improvement. In the same manner, as it is presented by [40], we explored two fusion mechanisms, MLB [28] and Block [5] but no gains were found. By adding the positional encoder module into the MMR, another increase in the results is achieved. This encourages us to think that the MMR module learns relationships coming from semantic and spatial information. When exploring late fusion methods, we obtained a model that by using the Hadamard product between V_{Gf} and G_{fa} performed the best in Con-Text and a model that by concatenating the features achieved a highest score in the Drink-Bottle dataset. Lastly, we opted for the model that on average performed the best as the final proposed model in this work. Insights into the attention masks learned and the correlation in visual regions can be found in the Supplementary Material section.

4.6. Qualitative Results

Qualitative results of the fine-grained image classification task are shown in Figure 3. By reviewing the samples obtained, we can note that our model is capable of learning a semantic space which combines successfully visual and textual signals coming from a single image. Classified samples such as "Pizzeria", "Tea House" and "Diner" often contain similar classes ranked on second and third positions. Images belonging to the Drink Bottle dataset on the second row, are correctly classified even though text instances belong to specific brands, thus showing generalization capability of our method. The seventh image on the first row is wrongly classified as "Theatre" due to OCR recognition errors and a lack of strong enough visual cues. The remaining wrongly classified images are very challenging and contain

Figure 3. Classification predictions. The top-3 probabilities of a class are shown as well as the Ground Truth label performed on the test set. Without recognizing textual instances some images are extremely hard to classify even for humans. Text in blue and red is used to show correct and incorrect predictions respectively. Best viewed in color.

some degree of ambiguity even for humans.

4.7. Fine-Grained Image Retrieval

As an additional experiment that gives us insights about the capabilities of the proposed model, we show the results obtained in Table 4 by performing query-by-example (QbE) image retrieval. In QbE, a system must return images in the form of a ranked list that belong to the same class as the image used as a query. To provide comparable results and following the work from [4, 40], we use the final classification vector as the image descriptor without using a specific metric-learning method. This vector is used to retrieve the nearest samples computed by the usage of the cosine similarity as a distance metric. In our experiments, the query, as well as the database is formed by unseen samples at training time. The results demonstrate that a very significant

Method	Con-Text	Drink-Bottle
Bai[4]	62.87	60.80
Mafla[40]	64.52	62.91
Proposed	**75.50**	**65.39**

Table 4. Retrieval results on the evaluated datasets. The retrieval scores are depicted in terms of the mAP(%).

boost of 10.78% and 2.39% in Con-Text and Drink-Bottle is achieved respectively. The lower boost in the Drink-Bottle dataset directly depends on the harder to recognize text instances, as well as the low image quality of several samples. Qualitative results that show the robustness of the model, as well as experiments addressing the importance of text can be found in the Supplementary Material section.

5. Conclusions

In this paper, we have presented a simple end-to-end model that employs a multi-modal reasoning graph to encounter semantic and positional relationships between text and salient visual regions. The learned space is composed by enriched features obtained from nodes in a graph, module that acts as an appropriate reasoning scheme. Exhaustive experiments in two datasets and two different tasks validate the robustness of the presented model which achieve state-of-the-art results by a significant margin over previous methods. Moreover, our end-to-end pipeline does not require pre-computed handcrafted features or a collection of ensemble models as earlier works. In the future we expect to explore the effectiveness of this approach into other vision and language related tasks.

References

[1] Jon Almazán, Albert Gordo, Alicia Fornés, and Ernest Valveny. Word spotting and recognition with embedded attributes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(12):2552–2566, 2014.

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018.

[3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[4] Xiang Bai, Mingkun Yang, Pengyuan Lyu, Yongchao Xu, and Jiebo Luo. Integrating scene text and visual appearance for fine-grained image classification. *IEEE Access*, 6:66322–66335, 2018.

[5] Hedi Ben-Younes, Rémi Cadene, Nicolas Thome, and Matthieu Cord. Block: Bilinear superdiagonal fusion for visual question answering and visual relationship detection. *arXiv preprint arXiv:1902.00038*, 2019.

[6] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

[7] Fedor Borisyuk, Albert Gordo, and Viswanath Sivakumar. Rosetta: Large scale system for text detection and recognition in images. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 71–79, 2018.

[8] Michal Bušta, Yash Patel, and Jiri Matas. E2e-mlt-an unconstrained end-to-end method for multi-language scene text. In *Asian Conference on Computer Vision*, pages 127–143. Springer, 2018.

[9] Charles Chen, Ruiyi Zhang, Eunyee Koh, Sungchul Kim, Scott Cohen, Tong Yu, Ryan Rossi, and Razvan Bunescu. Figure captioning with reasoning and sequence-level training. *arXiv preprint arXiv:1906.02850*, 2019.

[10] Xiaoxue Chen, Lianwen Jin, Yuanzhi Zhu, Canjie Luo, and Tianwei Wang. Text recognition in the wild: A survey. *arXiv preprint arXiv:2005.03492*, 2020.

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. Ieee, 2009.

[12] Sounak Dey, Pau Riba, Anjan Dutta, Josep Llados, and Yi-Zhe Song. Doodle to search: Practical zero-shot sketch-based image retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2179–2188, 2019.

[13] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*, 2016.

[14] Difei Gao, Ke Li, Ruiping Wang, Shiguang Shan, and Xilin Chen. Multi-modal graph neural network for joint reasoning on vision and scene text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12746–12756, 2020.

[15] Yang Gao, Oscar Beijbom, Ning Zhang, and Trevor Darrell. Compact bilinear pooling. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 317–326, 2016.

[16] ZongYuan Ge, Chris McCool, Conrad Sanderson, Peng Wang, Lingqiao Liu, Ian Reid, and Peter Corke. Exploiting temporal information for DCNN-based fine-grained object classification. In *International Conference on Digital Image Computing: Techniques and Applications*, 2016.

[17] Lluis Gomez, Andres Mafla, Marcal Rusinol, and Dimosthenis Karatzas. Single shot scene text retrieval. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[18] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376. ACM, 2006.

[19] Ankush Gupta, Andrea Vedaldi, and Andrew Zisserman. Synthetic data for text localisation in natural images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2315–2324, 2016.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[21] Tong He, Zhi Tian, Weilin Huang, Chunhua Shen, Yu Qiao, and Changming Sun. An end-to-end textspotter with explicit alignment and attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5020–5029, 2018.

[22] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Reading text in the wild with convolutional neural networks. *International Journal of Computer Vision*, 116(1):1–20, 2016.

[23] Sezer Karaoglu, Ran Tao, Theo Gevers, and Arnold WM Smeulders. Words matter: Scene text for image classification and retrieval. *IEEE Transactions on Multimedia*, 19(5):1063–1076, 2017.

[24] Sezer Karaoglu, Jan C van Gemert, and Theo Gevers. Context: text detection using background connectivity for fine-grained object classification. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 757–760. ACM, 2013.

[25] Dimosthenis Karatzas, Lluis Gomez-Bigorda, Anguelos Nicolaou, Suman Ghosh, Andrew Bagdanov, Masakazu Iwamura, Jiri Matas, Lukas Neumann, Vijay Ramaseshan Chandrasekhar, Shijian Lu, et al. ICDAR 2015 competition on robust reading. In *Proc. of the IEEE International Conference on Document Analysis and Recognition*, pages 1156–1160, 2015.

[26] Vahid Kazemi and Ali Elqursh. Show, ask, attend, and answer: A strong baseline for visual question answering. *arXiv preprint arXiv:1704.03162*, 2017.

[27] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. Novel dataset for fine-grained image categorization: Stanford dogs. In *Proc. CVPR Workshop on Fine-Grained Visual Categorization (FGVC)*, volume 2, page 1, 2011.

[28] Jin-Hwa Kim, Kyoung-Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. Hadamard product for low-rank bilinear pooling. *arXiv preprint arXiv:1610.04325*, 2016.

[29] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[30] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017.

[31] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[32] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. Visual semantic reasoning for image-text matching. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4654–4662, 2019.

[33] Xiangyang Li and Shuqiang Jiang. Know more say less: Image captioning based on scene graphs. *IEEE Transactions on Multimedia*, 21(8):2117–2130, 2019.

[34] Minghui Liao, Baoguang Shi, and Xiang Bai. Textboxes++: A single-shot oriented scene text detector. *IEEE Transactions on Image Processing*, 27(8):3676–3690, 2018.

[35] Minghui Liao, Baoguang Shi, Xiang Bai, Xinggang Wang, and Wenyu Liu. Textboxes: A fast text detector with a single deep neural network. In *AAAI*, pages 4161–4167, 2017.

[36] Chunxiao Liu, Zhendong Mao, Tianzhu Zhang, Hongtao Xie, Bin Wang, and Yongdong Zhang. Graph structured network for image-text matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10921–10930, 2020.

[37] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the variance of the adaptive learning rate and beyond. *arXiv preprint arXiv:1908.03265*, 2019.

[38] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

[39] Xuebo Liu, Ding Liang, Shi Yan, Dagui Chen, Yu Qiao, and Junjie Yan. Fots: Fast oriented text spotting with a unified network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5676–5685, 2018.

[40] Andres Mafla, Sounak Dey, Ali Furkan Biten, Lluis Gomez, and Dimosthenis Karatzas. Fine-grained image classification and retrieval by combining visual and locally pooled textual features. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 2950–2959, 2020.

[41] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

[42] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[43] Yair Movshovitz-Attias, Qian Yu, Martin C Stumpe, Vinay Shet, Sacha Arnoud, and Liron Yatziv. Ontological supervision for fine grained classification of street view storefronts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1693–1702, 2015.

[44] Medhini Narasimhan, Svetlana Lazebnik, and Alexander Schwing. Out of the box: Reasoning with graph convolution nets for factual visual question answering. In *Advances in neural information processing systems*, pages 2654–2665, 2018.

[45] Lukáš Neumann and Jiří Matas. Real-time scene text localization and recognition. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3538–3545. IEEE, 2012.

[46] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[47] Florent Perronnin and Christopher Dance. Fisher kernels on visual vocabularies for image categorization. In *2007 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2007.

[48] Joseph Redmon and Ali Farhadi. YOLO9000: better, faster, stronger. *arXiv preprint arXiv:1612.08242*, 2016.

[49] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[50] Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In *European Semantic Web Conference*, pages 593–607. Springer, 2018.

[51] Ajeet Kumar Singh, Anand Mishra, Shashank Shekhar, and Anirban Chakraborty. From strings to things: Knowledge-enabled vqa model that can read and reason. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4602–4612, 2019.

[52] Sebastian Sudholt, Neha Gurjar, and Gernot A Fink. Learning deep representations for word spotting under weak supervision. *arXiv preprint arXiv:1712.00250*, 2017.

[53] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[54] Andreas Veit, Tomas Matera, Lukas Neumann, Jiri Matas, and Serge Belongie. Coco-text: Dataset and benchmark for text detection and recognition in natural images. *arXiv preprint arXiv:1601.07140*, 2016.

[55] Tao Wang, David J Wu, Adam Coates, and Andrew Y Ng. End-to-end text recognition with convolutional neural networks. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, pages 3304–3308. IEEE, 2012.

[56] Xiu-Shen Wei, Jianxin Wu, and Quan Cui. Deep learning for fine-grained image analysis: A survey. *arXiv preprint arXiv:1907.03069*, 2019.

[57] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua

Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015.

[58] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. Auto-encoding scene graphs for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10685–10694, 2019.

[59] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4651–4659, 2016.

[60] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6720–6731, 2019.

[61] Michael R Zhang, James Lucas, Geoffrey Hinton, and Jimmy Ba. Lookahead optimizer: k steps forward, 1 step back. *arXiv preprint arXiv:1907.08610*, 2019.