JOINT EMBEDDINGS OF SCENE GRAPHS AND IMAGES

Eugene Belilovsky, Matthew B. Blaschko
KU Leuven, INRIA and University of Paris-Saclay
eugene.belilovsky@inria.fr, matthew.blaschko@esat.kuleuven.be

Jamie Ryan Kiros, Raquel Urtasun, Richard Zemel
University of Toronto
{rkiros,urtasun,zemel}@cs.toronto.edu

ABSTRACT

Multimodal representations of text and images have become popular in recent years. Text however has inherent ambiguities when describing visual scenes, leading to the recent development of datasets with detailed graphical descriptions in the form of scene graphs. We consider the task of joint representation of semantically precise scene graphs and images. We propose models for representing scene graphs and aligning them with images. We investigate methods based on bag-of-words, subpath representations, as well as neural networks. Our investigation proposes and contrasts several models which can address this task and highlights some unique challenges in both designing models and evaluation.

1 INTRODUCTION

With recent advances in perceptual tasks, attention in computer vision has been brought to problems requiring greater levels of semantic interpretation of images. Joint modeling of text and vision has led to great improvements in performance on caption generation, visual question answering, and retrieval. Text, however, often has many inherent ambiguities and, for some tasks, connecting a more precise description of image content to visual representation can be of great interest.

Compact representation of semantically precise descriptions of visual information are of great interest and can be potentially used for a variety of downstream tasks from image retrieval, generation, and visual question answering. Until recently study of multimodal embeddings of images has focused on connecting sentence level descriptions and images. One recently popular method of describing the content of images is based on the scene graph, a detailed description of the underlying image content. Recently, datasets with detailed scene graph annotations have become available (Krishna et al., 2016; Antol et al., 2015). In this work we make a first step to analyse the joint embedding of images and scene graph into a shared latent space.

We investigate several strategies for performing the embedding scene graphs. We propose as a baseline a bag of words embedding that only considers the scene objects and then consider a subpath representation as well as a graph neural network which can take advantage of the structural information within the scene, we find that, for the data we consider, subpath representations provide the best results on a retrieval task.

Related to our work (Fisher et al., 2011) proposes efficient kernels for retrieving images based on a scene graph. (Johnson et al., 2015) propose to use scene graphs for the task of image retrieval. Their model uses a probabilistic inference framework in comparing graphs. (Lu et al., 2016) consider the closely related problem of visual relationship detection. Most recently, (Teney et al., 2016), proposed to graph structured models for visual question answering. Their work considers a model similar to the Graph Neural networks (Li et al., 2015) which we also consider. The problem addressed is different as it involves graph matching instead of embedding a whole graph, furthermore visual features are used as annotations, while we consider categorical annotations.

2 Joint Representations of Scene Graphs and Images

A scene graph (Krishna et al., 2016; Johnson et al., 2015) is defined by its objects, their attributes, and relationships. Consider a set of object classes, $\mathcal{C}$, attributes $\mathcal{A}$, and relationships $\mathcal{R}$. Let a scene graph $G = (O, E)$ be a directed graph. For $o \in O$, an object in an image I, $o = (c, A)$, where $c \in \mathcal{C}$ is the class of the object and $A \subseteq \mathcal{A}$. For $t \in E$, a labeled directed edge $t = (o, s, r)$ where $r \in \mathcal{R}$ and $o, s \in O$. Define $Nbr(o)$ as the set of all (s, r) such that $(o, s, r) \in E$.

A joint representation of a scene graph, $\boldsymbol{g}$, and image,$\boldsymbol{x}$, should provide embedding functions, $\boldsymbol{f}_i(\boldsymbol{x})$ and $\boldsymbol{f}_g(\boldsymbol{g})$ which produce continuous vector representations in $\mathbb{R}^D$ for input images and scene graphs, along with a similarity metric, s (commonly the inner product in the embedding space). These vector representations should respect semantic similarities, such that for images $\boldsymbol{x}_i$, $\boldsymbol{x}_j$ it will be the case that $s(\boldsymbol{x}_i, \boldsymbol{g}) > s(\boldsymbol{x}_j, \boldsymbol{g})$ if the graph is semantically closer to the image $\boldsymbol{x}_i$ than $\boldsymbol{x}_j$. Here we consider several possible choices for embedding the graph.

Bag of Words A bag of words model takes a frequency count of nodes in the scene graphs, and does not consider the relationship information or the association of attributes. This is a natural baseline and the analogue in the text domain has shown strong performance in many joint vision and language tasks (Frome et al., 2013). Here we take the vocabulary V to be of size $|\mathcal{C}|$. Let e_o represent the one hot encoding for object class o. The embedding is defined by the matrix $\boldsymbol{W_g} \in \mathbb{R}^{D \times V}$ and given simply as $\boldsymbol{f}_g(\boldsymbol{g}_r) = \boldsymbol{W_g} \sum\limits_{o \in O} e_o$. This is then rescaled to unit norm.

SubPath Representations We consider the use of a graph path representation (Swamidass et al., 2005). Here we augment the count of node frequency by additionally considering subpaths up to length l. This allows structural information to be used in the final embedding. The final embedding is constructed as in the case of the bag of words $\boldsymbol{W_g}$ with V now the size of all unique subpaths in the dataset. Similar to the literature on text representations paths can be seen as an analog of n-grams which can still provide strong baselines in text classification (Joulin et al., 2016). In the base case of order 1 paths that only consider the nodes it reduces to the bag of words model.

Graph Neural Network Another strategy is similar in spirit to recent work proposed in (Li et al., 2015; Teney et al., 2016) which maintains a state vector for each node and uses a recurrent procedure that updates each node state based on its neighbor, progressively propagating information. Below the update sequence is defined per object.

$$h^0_{g,o} = \boldsymbol{W}_g e_o \qquad\qquad h^{i+1}_{g,o} = \tanh\left(\boldsymbol{W}_o h^i_{g,o} + \boldsymbol{W}_p \left(\sum_{(s,r) \in Nbr(o)} \boldsymbol{V}_r h^i_{g,s}\right)\right)$$

For each node we obtain its representation by performing an embedding and then updating the representation by adding a term for the neighbors based on the maximum path between any 2 nodes. In practice we will take a maximum of $i = 3$ steps. $\boldsymbol{V}_r$ is a separate term associated with each relationship, we consider only the most common relationships and fold others into one category. The final graph representation can be obtained by summing the node states and normalizing. In this work we focus on the object and their interactions but attributes can additionally be incorporated as edges in the graph.

Image embedding and loss function The image embedding we utilize are the VGG-19 fc7 10-crop features,x, For the image embedding we use VGG-19 fc7 10-crop features, as in (Kiros et al., 2014), denoted $\boldsymbol{x}$, projected as $\boldsymbol{f}_{W_m}(\boldsymbol{x}) = W_m \boldsymbol{x}$ and normalized. If we let W_G describe all parameters of the encoding model given a set of images,$\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N$ and corresponding scene graphs $\boldsymbol{g}_1, \boldsymbol{g}_2, ..., \boldsymbol{g}_N$, and the similarity measure $s(\boldsymbol{x}, \boldsymbol{g}) = \boldsymbol{f}_{W_m}(\boldsymbol{x}) \cdot \boldsymbol{f}_{W_G}(\boldsymbol{g})$ the following contrastive loss function is used to align the image and scene graph

$$\min_{W_m, W_G} \sum_{\boldsymbol{x}_k, \boldsymbol{g}_k, \boldsymbol{g}_c} \max\{0, \alpha - s(\boldsymbol{x}_k, \boldsymbol{g}_k) + s(\boldsymbol{x}_k, \boldsymbol{g}_c)\} + \sum_{\boldsymbol{g}_k, \boldsymbol{x}_k, \boldsymbol{x}_c} \max\{0, \alpha - s(\boldsymbol{x}_k, \boldsymbol{g}_k) + s(\boldsymbol{x}_c, \boldsymbol{g}_k)\}$$

Where $\alpha > 0$ is a scalar defining the size of the margin of the two hinge losses.

3 Evaluating Joint Embeddings

Evaluating joint embeddings can be challenging. A common approach in caption/image embedding is the use of a retrieval task (Vendrov et al., 2015) that involves ranking a large dataset of images by relevance for a query. However, although posed as a retrieval task the score is often only known for only one ground truth image. This problem is exacerbated for the case of highly detailed descriptions such as a paragraphs or scene graphs. Given that scene graphs can become very large (some having over 50 labeled objects) it becomes increasingly easy to match images simply based on object counts, while at the same time there can be many images in the result set which are indeed very similar to the ground truth. We thus take a different approach to the evaluation of scene graph to image retrieval than shown in Johnson et al. (2015). Since images in the test set have associated scene graphs our evaluation leverages existing graph similarity metrics to allow comparisons to all the images in the retrieval set. For the case of scene graphs we can construct a metric based on the path kernel (Borgwardt, 2007). Using this similarity metric, we can compute the Normalized Discounted Cumulative Gain (NDCG) (Wang et al., 2013) to evaluate the retrieval performance considering the returned ranking of all images in the search space to their underlying graph similarity score.

We use a dataset of images and scene graphs from (Johnson et al., 2015; Lu et al., 2016). The data consists of 5000 images with carefully curated scene graphs. We first perform a first level analysis, described in the Appendix, to determine that our visual features can indeed discriminate structural information in the scene graph.

We now evaluate the proposed joint representation approaches on the retrieval task. We use the 4000 train and 1000 test images from the splits specified in Johnson et al. (2015). We use the objective described previously with batches of size 500 and optimize using the ADAM optimizer (Kingma & Ba, 2014), ℓ_2 regularization, and $\alpha = 0.4$.

For evaluation we compute the mean Normalized Discounted Cumulative Gain (NDCG) for each test image using a path graph kernel of order 3 as our relevance metric. For each test graph we embed the graph in the joint embedding space and look for the nearest matching image out of the 999 possible remaining.

Methods	NDCG 5	NDCG 10	NDCG 20	medRank Gr2im
PathRep 3(500 latent)	**0.320**	**0.354**	**0.396**	9
PathRep 2(500 latent)	0.300	0.338	0.381	9
BOW (500 latent)	0.281	0.317	0.362	9
PathRep 3 (100 latent)	0.305	0.338	0.378	10
PathRep 2 (100 latent)	0.290	0.327	0.372	10
BOW (100 latent)	0.276	0.310	0.355	10
Graph NN (100 latent)	0.249	0.280	0.321	15
SG obj Johnson et al. (2015)	-	-	-	28
SG obj-attr-rel Johnson et al. (2015)	-	-	-	14

Table 1: Results for graph to image retrieval. NDCG is computed at the top 5,10, and 20 images. Medium rank is computed for the ground truth image retrieval

We consider results using 500 and 100 dimensional latent spaces. We also consider path representations of order 2 and 3. We observe that the use of the graph neural network under-performs the more simple linear embedding of the bag of words features. This is analogous to observations in several text based tasks (Joulin et al., 2016) and highlights the difficulty in extracting semantic information in this challenging scenario. However we can see that the path representations indeed improve substantially the performance in terms of NDCG, highlighting that graph structural information is indeed useful and can be leveraged in this task.

For reference, we also report results for the median rank of the ground truth image on the same task from Johnson et al. (2015), which uses a different model based on object detections. We note in Johnson et al. (2015) the label space is limited to the top occurring objects, which in our embedding framework is not necessary and we utilize the full set of objects provided in the dataset. Notably the results using bag of words and path representations can improve on those of Johnson et al. (2015). Additionally, it is possible to do image to graph retrieval with our model.

ACKNOWLEDGMENTS

We would like to thank MITACS for funding EB's visit to Toronto. EB and MB are partially funded by Internal Funds KU Leuven, FP7-MC-CIG 334380, an Amazon Academic Research Award, and DIGITEO 2013-0788D - SOPRANO.

REFERENCES

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2425–2433, 2015.

Karsten Michael Borgwardt. *Graph kernels*. PhD thesis, Ludwig Maximilians University Munich, Germany, 2007. URL http://edoc.ub.uni-muenchen.de/archive/00007169/.

Matthew Fisher, Manolis Savva, and Pat Hanrahan. Characterizing structural relationships in scenes using graph kernels. In *ACM Transactions on Graphics (TOG)*, volume 30, pp. 34. ACM, 2011.

Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al. Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*, pp. 2121–2129, 2013.

Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David A Shamma, Michael S Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3668–3678. IEEE, 2015.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*, 2016.

Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014.

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. *arXiv preprint arXiv:1602.07332*, 2016.

Yujia Li, Daniel Tarlow, Marc Brockschmidt, and Richard Zemel. Gated graph sequence neural networks. *arXiv preprint arXiv:1511.05493*, 2015.

Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *European Conference on Computer Vision*, pp. 852–869. Springer, 2016.

S Joshua Swamidass, Jonathan Chen, Jocelyne Bruand, Peter Phung, Liva Ralaivola, and Pierre Baldi. Kernels for small molecules and the prediction of mutagenicity, toxicity and anti-cancer activity. *Bioinformatics*, 21(suppl 1):i359–i368, 2005.

Damien Teney, Lingqiao Liu, and Anton van den Hengel. Graph-structured representations for visual question answering. *arXiv preprint arXiv:1609.05600*, 2016.

Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. Order-embeddings of images and language. *arXiv preprint arXiv:1511.06361*, 2015.

Yining Wang, Liwei Wang, Yuanzhi Li, Di He, Tie-Yan Liu, and Wei Chen. A theoretical analysis of normalized discounted cumulative gain (NDCG) ranking measures. In *Conference on Learning Theory*, 2013.

A Additional experiments

We perform a basic evaluation of whether structural information related to the scene graph is extractable from the image features (VGG fc7) we have selected to use. We consider the top occurring edges and construct binary classification problems for each of the top edges attempting to predict it's presence or absence from the visual features. We use a random forest classifier and consider the AUC. We find that 3 out of 10 have chance performance with the remaining classifiers obtaining an average AUC of $55 \pm 0.5\%$. This first order analysis indicates that there is discernible structural information in the visual features used, although expectedly the rate is rather low. We note that a given image may have a large number of interactions which can together give noticeable improvement on tasks such as retrieval.