An Interpretable Model for Scene Graph Generation

Ji Zhang[1,2], Kevin Shih[1], Andrew Tao[1], Bryan Catanzaro[1], Ahmed Elgammal[2]

[1]Nvidia Research
[2]Department of Computer Science, Rutgers University

Abstract

We propose an efficient and interpretable scene graph generator. We consider three types of features: visual, spatial and semantic, and we use a late fusion strategy such that each feature's contribution can be explicitly investigated. We study the key factors about these features that have the most impact on the performance, and also visualize the learned visual features for relationships and investigate the efficacy of our model. We won the champion of the OpenImages Visual Relationship Detection Challenge on Kaggle, where we outperform the 2nd place by 5% (20% relatively). We believe an accurate scene graph generator is a fundamental stepping stone for higher-level vision-language tasks such as image captioning and visual QA, since it provides a semantic, structured comprehension of an image that is beyond pixels and objects.

1 Introduction

Scene graph generation is a fundamental task that bridges low-level vision such as scene parsing and object detection and high-level vision-language problems such as image captioning [10, 16] and visual QA [1, 4]. It produces a structured semantic understanding of an image from individual objects, and provides rich information for those high-level tasks. Current state-of-the-art methods [9, 18, 22, 11, 2, 20, 15, 8, 13, 19, 17, 21] use three types of features to represent relationships: 1) *visual features*: the CNN features of the two objects or their combination; 2) *spatial features*: coordinates of the two objects which encodes their spatial layouts; 3) *semantic features*: class labels of the two objects which provide a strong prior of the predicate. Most of them, if not all, combine the three features in an early stage to learn a compositional feature for relationship prediction. The contribution of each feature is thus implicit and probably not optimized. In this paper we propose a structure that instead explicitly builds three branches for the three features, each contributing to the output in an interpretable way, and we fuse their outputs in the final stage to get optimized predictions.

Our contributions are: 1) we propose a new model that efficiently combines three features and show explicitly what each feature contributes to the final prediction and how much the contribution is. 2) we demonstrate the efficacy of our model on three datasets: OpenImages (OI) [5], Visual Genome (VG) [7] and Visual Relationship Detection (VRD) [9]. We won the 1st place in the OpenImages Challenge, and we outperform state-of-the-art methods on VG and VRD by significant margins.

2 Model Description

The task of visual relationship detection can be defined as a mapping f from image I to 3 labels and 2 boxes l_S, l_P, l_O, b_S, b_O

$$I \xrightarrow{f} l_S, l_P, l_O, b_S, b_O \tag{1}$$

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

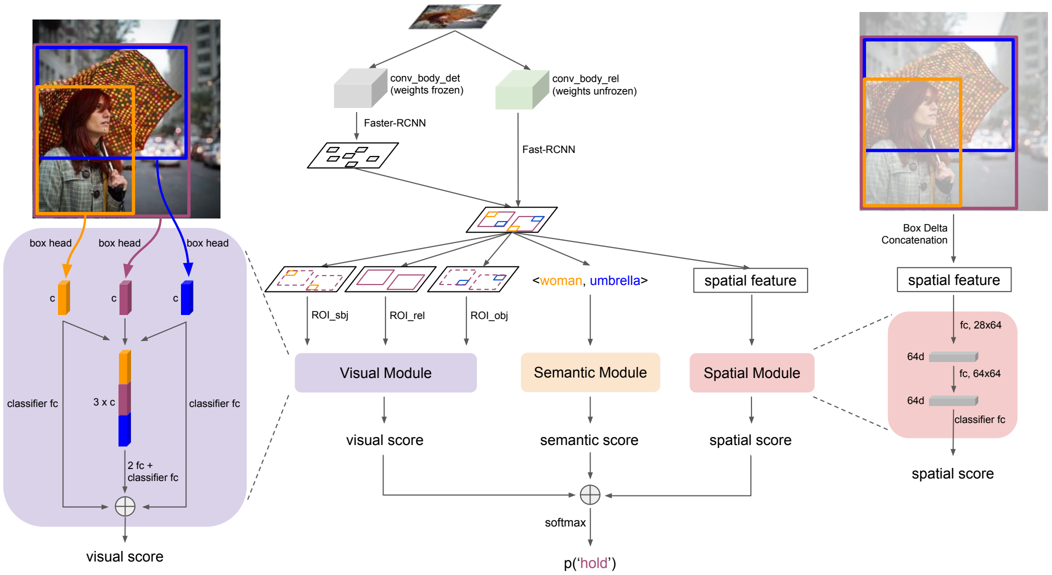


Figure 1: Model Architecture

where l, b stand for labels and boxes, S, P, O stand for subject, predicate, object. We decompose f into object detector f_{det} and relationship classifier f_{rel}:

$$I \xrightarrow{f_{det}} l_S, l_O, b_S, b_O, v_S, v_O \xrightarrow{f_{rel}} l_P \tag{2}$$

The decomposition means that we can run an object detector on the input image to obtain labels, boxes and visual features for subject and object, then use these as input features to the relationship classifier which only needs to output a label. There are two obvious advantages in this model: 1) learning complexity is dramatically reduced, since we can simply use an off-the-shelf object detector as f_{det} without the need for re-training, hence the learn-able weights exist only in the small subnet f_{rel}; 2) We have much richer features for relationships, i.e., $l_S, l_O, b_S, b_O, v_S, v_O$ for f_{rel}, instead of only the image I for f.

We further assume that the semantic feature l_S, l_O, spatial feature b_S, b_O and visual feature v_S, v_O are independent from each other. So we can build 3 separate branches of sub-networks for them. This is the basic work flow of our model.

Figure 1 shows our model in details. The network takes an input image and outputs the 6 aforementioned features, then each branch uses its corresponding feature to produce a confidence score for predicates, then all scores are added up and normalized by softmax. We now introduce each module's design and their motivation.

2.1 Relationship Proposal

A relationship proposal is defined as a pair of objects that is very likely related[21]. In our model we first detect all meaningful objects by running an object detector, then we simply consider each pair of objects is a relationship proposal. The following modules learn to classify each pair as either "no relationship" or one of the 9 predicates, not including the "is" relationship.

2.2 Semantic Module

Zeller, et al.[19] introduced a frequency baseline that performs reasonably well on Visual Genome dataset[14] by counting frequencies of predicates given subject and object. Its motivation is that in general cases, the types of relationships between two objects are usually limited, e.g., given the subject being person and object being horse, their relationship is highly likely to be "ride", "walk", "feed", but less likely to be "stand on", "carry", "wear", etc. In short, the $\langle subject, predicate, object \rangle$ composition is usually biased. Furthermore, any specific relationship detection dataset can only contain a limited number of them, making the bias even stronger. This is a factor that we find every useful to leverage.

We improved this baseline by removing the background class of subject and object. Specifically, for each training image we count the occurrence of l_P given l_S, l_O in the ground truth annotations, and we end up with an empirical distribution $p(P|S, O)$ for the whole training set. We do this under the assumption that the test set is also drawn from the same distribution. We then build the remaining modules to learn a complementary residual on top of the output of this baseline.

2.3 Spatial Module

In the challenge dataset, the three predicates "on", "under", "inside_of" indicate purely spatial relationships i.e., the relative locations of subject and object are sufficient to tell the relationship. A common solution, as applied in Faster-RCNN[12], is to learn a mapping from visual features to location offsets. However, the learning becomes significantly hard when the distance of two objects are very far[3], which is often the case for relationships. We capture spatial information by encoding the box coordinates of subjects and objects using box delta[12] and normalized coordinates:

$$\langle \Delta(b_S, b_O), \Delta(b_S, b_P), \Delta(b_P, b_O), c(b_S), c(b_O) \rangle \tag{3}$$

where $\Delta(b_1, b_2)$ are box delta of two boxes b_1, b_2, and $c(b)$ are normalized coordinates of box b, which are defined as:

$$\Delta(b_1, b_2) = \langle \frac{x_1 - x_2}{w_2}, \frac{y_1 - y_2}{h_2}, \log \frac{w_1}{w_2}, \log h_1 h_2 \rangle \tag{4}$$

$$c(b) = \langle \frac{x_{min}}{w}, \frac{y_{min}}{h}, \frac{x_{max}}{w}, \frac{y_{max}}{h}, \frac{a_{box}}{a_{img}} \rangle \tag{5}$$

where $b_1 = (x_1, y_1, w_1, h_1)$ and $b_2 = (x_2, y_2, w_2, h_2)$, w, h are width and height of the image, a_{box} and a_{img} are areas of the box and image.

2.4 Visual Module

Visual Module is useful mainly for three reasons: 1) it accounts for all other types of relationships that spatial features can hardly predict, e.g., interactions such as "man play guitar" and "woman wear handbag"; 2) it solves relationship reference problems[6], i.e., when there are multiple subjects or objects that belong to a same category, we need to know which subject is related to which object; 3) for some specific interactions, e.g., "throw", "eat" "ride", the visual appearance of the subject or object alone is very informative about the predicate. With these motivations, we feed subject, predicate, object ROIs into the backbone and get the feature vectors from its last fc layer as our visual features, then we concatenate these three features and feed them into 2 additional randomly initialized fc layers followed by an extra fc layer to get a logit, i.e., unnormalized score. We also add one fc layer on top of the subject feature and another fc layer on top of the object feature to get two scores. These two scores are the predictions made solely by the subject/object feature according to the third reason mentioned above.

2.5 The "is" Relationship

In the OpenImages challenge, "$\langle object \rangle$ is $\langle attribute \rangle$" is also considered as relationships, where there is only one object involved. We achieve this sub-task by using a completely separate, single-branch, Fast-RCNN based model. We use the same object detector to get proposals for this model, then for each proposal the model produces a probability distribution over all attributes with the Fast-RCNN pipeline.

3 Experiments

We present quantitative and qualitative results on OpenImages (OI). We show ablation study on each component of our model. We also show results on Visual Genome (VG) and Visual Relationship Detection (VRD) datasets in the supplementary material.

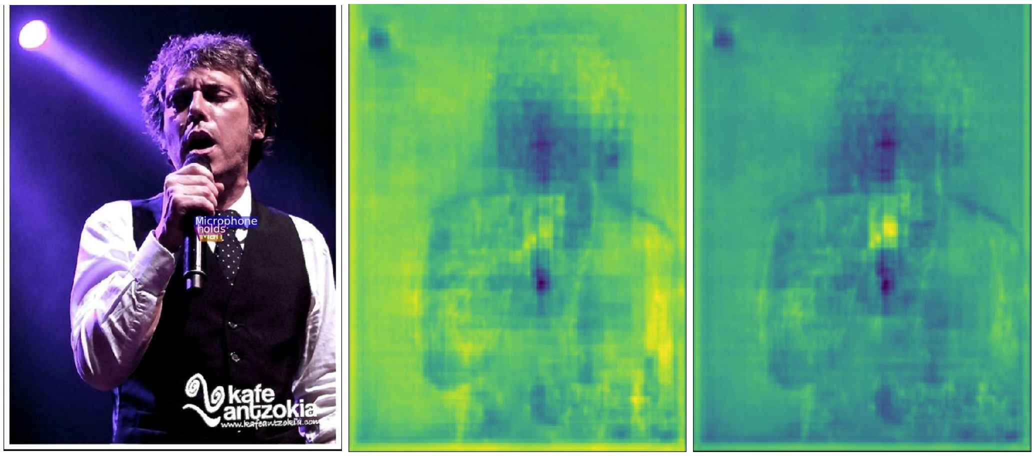
Team ID	Public
VRD_NN (8th)	0.20643
anokas (7th)	0.21573
MIL (6th)	0.21774
tito (5th)	0.25571
toshif (4th)	0.25621
Kyle (3rd)	0.28043
radek (2nd)	0.28886
Seiji (Ours)	**0.33213**

Team ID	Private
[ods.ai] ZFTurbo (8th)	0.17621
anokas (7th)	0.17960
MIL (6th)	0.19666
radek (5th)	0.20113
toshif (4th)	0.22832
Kyle (3rd)	0.23491
tito (2nd)	0.23709
Seiji (Ours)	**0.28544**

Table 1: Kaggle leader boards.

	R@50	mAP_rel	mAP_phr	score
Baseline	72.98	26.54	32.77	38.32
$\langle S, P, O\rangle$	74.13	32.41	39.55	43.61
$\langle S, P, O\rangle + S + O$	**74.46**	34.16	39.59	44.39
$\langle S, P, O\rangle + S + O + spt$	74.40	**34.96**	**40.70**	**45.14**

Table 2: Ablation Study on OI.



(a) image with gt relationship (b) feature from conv_body_det (c) feature from conv_body_rel

Figure 2: Qualitative results. (a) shows the image with $\langle man, holds, microphone\rangle$, (b) shows the convolution feature from the object detector backbone, and (c) shows the feature from the predicate backbone that we train along with the whole model.

Quantitative Results: In Table 1 we show the competition results on both the public and private leader board. The score is computed by weight average of three metrics: recall of top 50 predictions (R@50), mean average precision of relationships (mAP_rel), mean average precision of phrases (mAP_phr). The weights for them are 0.2, 0.4, 0.4, respectively. Our model surpasses the 2nd place by 15% relatively on the public dataset and 20% relatively on the private dataset.

Visualization Results: In Figure 2 we show convolution feature maps from the two backbones described in Figure 1 given an image with a ground-truth relationship $\langle man, holds, microphone\rangle$. It is very clear that the object detector focuses mostly on the contour of the person, while the predicate branch accurately learns to capture the most informative region that represents "holds", i.e., the intersection of the microphone and the fingers that are holding it. This is the most critical reason why our model performs well.

Ablation Study: We show evaluation results on the validation set of four models with the following settings: 1) **baseline**: only the semantic module. 2) $\langle$**S,P,O**$\rangle$: using semantic module and visual module without the direct predictions from subject/object. 3) $\langle$**S,P,O**$\rangle$**+ S + O**: using semantic module and the complete visual module 4) $\langle$**S,P,O**$\rangle$**+ S + O + spt**: our complete model.

References

[1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[2] B. Dai, Y. Zhang, and D. Lin. Detecting visual relationships with deep relational networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3298–3308. IEEE, 2017.

[3] G. Gkioxari, R. Girshick, P. Dollár, and K. He. Detecting and recognizing human-object intaractions. *CVPR*, 2018.

[4] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. L. Zitnick, and R. Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 1988–1997. IEEE, 2017.

[5] I. Krasin, T. Duerig, N. Alldrin, V. Ferrari, S. Abu-El-Haija, A. Kuznetsova, H. Rom, J. Uijlings, S. Popov, S. Kamali, M. Malloci, J. Pont-Tuset, A. Veit, S. Belongie, V. Gomes, A. Gupta, C. Sun, G. Chechik, D. Cai, Z. Feng, D. Narayanan, and K. Murphy. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://storage.googleapis.com/openimages/web/index.html*, 2017.

[6] R. Krishna, I. Chami, M. Bernstein, and L. Fei-Fei. Referring relationships. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[7] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. Bernstein, and L. Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. 2016.

[8] Y. Li, W. Ouyang, and X. Wang. Vip-cnn: A visual phrase reasoning convolutional neural network for visual relationship detection. *CVPR*, 2017.

[9] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei. Visual relationship detection with language priors. In *ECCV*, 2016.

[10] J. Lu, J. Yang, D. Batra, and D. Parikh. Neural baby talk. In *CVPR*, 2018.

[11] B. A. Plummer, A. Mallya, C. M. Cervantes, J. Hockenmaier, and S. Lazebnik. Phrase localization and visual relationship detection with comprehensive image-language cues. In *ICCV*, 2017.

[12] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[13] D. Xu, Y. Zhu, C. Choy, and L. Fei-Fei. Scene graph generation by iterative message passing. In *Computer Vision and Pattern Recognition (CVPR)*, 2017.

[14] D. Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei. Scene graph generation by iterative message passing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, 2017.

[15] X. Yang, H. Zhang, and J. Cai. Shuffle-then-assemble: Learning object-agnostic visual relationship features. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[16] T. Yao, Y. Pan, Y. Li, and T. Mei. Exploring visual relationship for image captioning. In *ECCV*, 2018.

[17] G. Yin, L. Sheng, B. Liu, N. Yu, X. Wang, J. Shao, and C. Change Loy. Zoom-net: Mining deep feature interactions for visual relationship recognition. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[18] R. Yu, A. Li, V. I. Morariu, and L. S. Davis. Visual relationship detection with internal and external linguistic knowledge distillation. In *The IEEE International Conference on Computer Vision (ICCV)*, 2017.

[19] R. Zellers, M. Yatskar, S. Thomson, and Y. Choi. Neural motifs: Scene graph parsing with global context.

[20] H. Zhang, Z. Kyaw, S.-F. Chang, and T.-S. Chua. Visual translation embedding network for visual relation detection. *arXiv preprint arXiv:1702.08319*, 2017.

[21] J. Zhang, M. Elhoseiny, S. Cohen, W. Chang, and A. Elgammal. Relationship proposal networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5678–5686, 2017.

[22] B. Zhuang, L. Liu, C. Shen, and I. Reid. Towards context-aware interaction recognition for visual relationship detection. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.