Detecting Visual Relationships Using Box Attention

Alexander Kolesnikov[1]* Google Research

Alina Kuznetsova[1] Google Research

Christoph H. Lampert[2] IST Austria

Vittorio Ferrari[1] Google Research

[1]{akolesnikov,akuznetsa,vittoferrari}@google.com [2]chl@ist.ac.at

Abstract

We propose a new model for detecting visual relationships, such as "person riding motorcycle" or "bottle on table". This task is an important step towards comprehensive structured image understanding, going beyond detecting individual objects. Our main novelty is a Box Attention *mechanism that allows to model pairwise interactions between objects using standard object detection pipelines. The resulting model is conceptually clean, expressive and relies on well-justified training and prediction procedures. Moreover, unlike previously proposed approaches, our model does not introduce any additional complex components or hyperparameters on top of those already required by the underlying detection model. We conduct an experimental evaluation on two datasets,* V-COCO *and* Open Images*, demonstrating strong quantitative and qualitative results.*

1. Introduction

The task of detecting visual relationships aims at localizing all pairs of interacting objects in an input image and identifying relationships between them. The ability to recognize visual relationships is crucial for achieving comprehensive understanding of visual scenes. As a consequence, the task of detecting visual relationships has recently attracted a lot of attention in the computer vision community [1, 5, 8, 12, 14, 15, 17, 19, 24, 26].

Naturally, currently available models for detecting visual relationships [5, 17, 1] heavily rely on object detection pipelines. However, in order to enable the modeling of pairwise relationships, they augment the object detection pipelines with multiple additional components and thereby introduce additional hyperparameters. In contrast, in this paper we present a new model that almost exclusively relies on readily available detection pipelines. Crucially, our model does not require tuning any additional hyperparameters and can be implemented by adding a dozen lines of code to existing object object detection pipelines [16, 11].

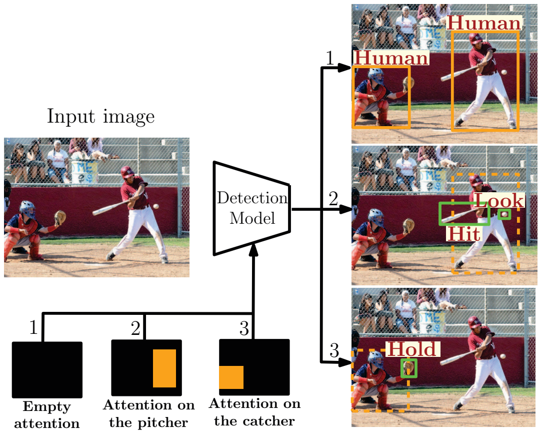

Figure 1. Schematic illustration of the proposed model. It is based on the general object detection pipeline augmented with our box attention mechanism.

We formulate the task of detecting visual relationships as a joint probabilistic model. Our key idea is to decompose the probabilistic model into two simpler sub-models using the chain rule. As a result of this decomposition the task of detecting visual relationships breaks down into two consecutive object detection tasks. A first detection model localizes all objects in an input image. Then, for each detected object, the second model detects all other objects interacting with it. Our main contribution is the *Box Attention* mechanism that augments the second model with the ability to be conditioned on objects localized by the first one.

2. Related Work

Visual relationships have been previously studied in the computer vision community. Earlier works leverage visual relationships in order to improve performance of object detection [21], action recognition and pose estimation [2], or semantic image segmentation [6]. However, [17] was the first work to formulate detection of visual relationships as a separate task. It proposes to learn a composite likelihood function that utilizes a language prior based on the word embeddings [18] for scoring visual relationships. In [24] the authors use external sources of linguistic knowledge and

*Work partially done at IST Austria.

the distillation technique [10] to improve modeling performance, while [15] formulates the task of detecting visual relationships as a reinforcement learning problem. Dai et al. [1] proposes a multistage relationship detection process, where they first run an object detector, and then apply a light-weight network for selecting promising pairs of interacting object detections. A similar approach was also leveraged by [25]. The work [14] introduces a triplet proposal mechanism and then trains a multi-stage scoring function to select the best proposal. A few works [19, 26] investigate a weakly-supervised variant of the relationship detection task. Further, several works focus on human-object relations. Earlier works exploit probabilistic graphical models [7, 23] and also investigate weakly-supervised settings [20]. Recently, high-performing models [8, 5, 4] based on deep convolutional neural networks have emerged. Finally, very recent paper [12] proposes to use an attention mechanism for solving related task of *referring relationships*.

3. Box Attention for Detecting Relationships

In this section we describe our approach for detecting visual relationships, which we call BAR-CNN (***B**ox **A**ttention **R**elational CNN*). Our overall approach is illustrated in Fig 1. Formally, the task of detecting visual relationships for a given image can be formulated as detecting all triplets in a form of $\langle$*subject* (S)*, predicate* (P)*, object* $(O)\rangle$. The subject S and object O are represented by bounding boxes b^s and b^o, and their corresponding category labels by l^s and l^o. The predicate P is represented by a label l^p.

We derive our approach for modeling visual relationships using a probabilistic interpretation of this task. The high-level idea is to model the probability $p(S, P, O|I)$ that a triplet $\langle S, P, O\rangle$ is a correct visual relationship in the input image I. It is challenging to model this joint probability distribution, as it involves multiple structured variables interacting in a complex manner. Thus, we propose to employ the chain rule in order to decompose the joint probability into simpler conditional probabilities:

$$p(S, P, O|I) = p(S|I) \cdot p(P, O|S, I). \tag{1}$$

The first factor $p(S|I)$ models the probability that a subject (b^s, l^s) is present in the image I. Thus, this factor can be modeled as a standard detection task of predicting bounding boxes and category labels for all instances in the image. The second factor, $p(P, O|S, I)$, models the probability that an object (b^o, l^o) is present in the image and is related to the subject S through a predicate l^p. Estimating $p(P, O|S, I)$ can be also seen as a detection problem. In this case the model should output bounding boxes, object labels and the corresponding predicate labels (b^o, l^o, l^p) for all objects that interact with S. We implement conditioning on S by treating it as an additional input that we call *Box Attention*. In Section 3.1 we present this *Box Attention* mechanism in detail.

Due to functional similarity of the two factors in Eq. (1) we further propose to train a single unified model for both $p(S|I)$ and $p(P, O|S, I)$. Note that our approach can be implemented within any object detection model. From now on we will refer to it as the *base detection model*.

3.1. Model details

Box attention representation. Consider an input image I. The box attention map for this image is represented as a binary image m of the same size as I, with 3 channels. The first channel represents a subject bounding box (Figure 1). Specifically, all pixels inside the subject bounding box are set to 1 and all other pixels are set to 0. An attention map can be empty: in this case the first channel is all zeros. The second and third channels are used in the following way: if the first channel is not empty, then the second channel is all zeros and the third channel is all ones. Conversely, if the first channel is empty, then the second channel is all ones and the third channel is all zeros. These two extra channels are useful because state-of-the-art detection models use deep convolutional neural networks as feature extractors [9, 22]. Neurons of these networks have limited receptive fields that might not cover the whole attention map. As a consequence, these neurons have no information whether the attention map is empty or not based only on the first channel.

Incorporating box attention maps in the base detection model. In order to incorporate the additional box attention input we use a simple and yet very effective strategy.

The proposed mechanism is illustrated in Figure 2. Consider the output u of a certain convolutional layer of the base detection model. Let's assume u has spatial resolution $H \times W$ and K channels. We condition the output u on the attention map m by performing the following steps: 1) Obtain $\hat{m}$ by resizing m to the spatial size of $H \times W$ using nearest neighbor interpolation. 2) Obtain $\tilde{m}$ by passing $\hat{m}$ through a learnable convolutional layer with K output channels and a kernel of size 3×3. 3) Update u as $u + \tilde{m}$. In principle, we can apply this procedure to every convolutional layer of the base detection model. In practice, we use ResNet-type architectures with bottleneck units [9], and apply the above conditioning procedure to the second convolution of every bottleneck unit.

The proposed conditioning procedure has several appealing properties. First, it allows to seamlessly initialize the BAR-CNN model using the pre-trained base detection model. Second, if we initialize the convolutional kernels in step 2 with all zeros, in the beginning of training our conditioning procedure does not have any effect on the outputs of the base detection model. This helps preventing disruption of the pre-trained base detection model and ensures numerical

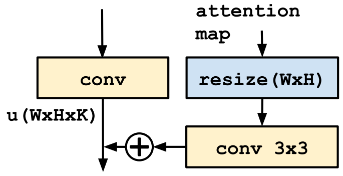

Figure 2. Our proposed attention mechanism. This procedure is applied to convolutional layers of the base detection model.

stability in the beginning of the training process.

Training. As described above, we want to learn a single model able to a) output subject predictions when the attention map is empty, and b) output object and predicate predictions when conditioned on a subject prediction through the attention map.

To achieve a) we simply add an empty attention map to each image in the training set and preserve all subject bounding boxes as ground-truth. This forms one type of training sample. To achieve b) for each subject annotated in a training image we generate a separate training sample consisting of the same image, box attention map corresponding to the subject, and all object and predicate annotations corresponding to this subject. This is the second type of training sample. Thus, given k annotated subjects in a training image, we create $k+1$ training samples. We then use this training set to train a BAR-CNN model. Note that storing multiple copies of the image is not necessary, as the training samples can be generated on the fly during the course of training.

The proposed model predicts two labels (i.e. l^o and l^p) per detection instead of one. Therefore during training we use a sigmoid multiclass loss instead of a cross-entropy loss normally used for standard object detection. During training we closely follow recommendations from the public RetinaNet implementation (see Section 4 for more details).

Predicting visual relationships. Visual relationships for a test image I are predicted using a natural two-stage procedure. First, we run the BAR-CNN model by feeding it the image I and an empty attention map. The model outputs a scored list of subject detections. Each detection has a box b^s, a class label l^s and a score s^s.

Second, we form a scored list of all detected relationships. Specifically, for every detection we apply the following two-step procedure: 1) Construct an attention map for b^s and feed it to the BAR-CNN model together with the image I. As a result BAR-CNN predicts a set of object boxes $(b^o, l^o, l^p)_i$ that interact with b^s through the relationship l^p. 2) For every detection (b^o, l^o, l^p) we first compute the score $p(P, O|S, I) = s^{p,o}$ by multiplying the scores of labels l^o and l^p obtained through multiclass prediction. The final score s of the full visual relationship detection $\langle(b^s, l^s), l^p, (b^o, l^o)\rangle$ is computed as $s = s^s s^{p,o}$.

4. Experiments

	Model C [8] impl. [5]	InteractNet [5]	BAR-CNN (proposed)
Relation	**AP$_{\text{role}}$**	**AP$_{\text{role}}$**	**AP$_{\text{role}}$**
mean AP	31.8	40.0	**43.6**

Table 1. Quantitative comparison of the proposed model with competing models on *V-COCO*.

We now present experimental evaluation of the proposed BAR-CNN model. We evaluate on the two publicly available datasets: *V-COCO* [8] and *Open Images* [13], reporting strong quantitative and qualitative results.

4.1. Implementation details

As explained in Section 3, we build BAR-CNN by combining the base detection model with the box attention input. In our experiments we use the RetinaNet [16] model with ResNet50 [9] backbone as the base detection model.

During finetuning we do not freeze any of the network's parameters. As an optimization algorithm we use stochastic gradient descent with momentum set to 0.9 and the minibatch size is set to 256 images.

Before finetuning on the *V-COCO* dataset, we initialize a BAR-CNN model from the *RetinaNet* detection model pretrained on the *MSCOCO train2014* split. For finetuning on the *Open Images* dataset, we initialize our model from the *RetinaNet* detection model pretrained on bounding boxes of the training split of the *Open Images* dataset itself. We conduct finetuning for 60 and 15 epochs for the *V-COCO* and *Open Images* datasets, respectively. The initial learning rate is always set to $8 \cdot 10^{-3}$ and is decayed twice by a factor of 10 after 50% and 75% of all optimization steps. All other hyperparameters of the *RetinaNet* model are set to their default values from the publicly available implementation[1].

4.2. Results on the VCOCO dataset [8]

Data. The V-COCO dataset contains natural images annotated by human-object relationships. There are 29 relationships (also called actions), *e.g. carry, drink, ride, cut, eat (object), eat (instrument),* etc. Overall, the dataset has $5,400$ images in the joint *train* and *val* splits, and $4,946$ images in the *test* split. On average each image has 4.5 annotated human-object relationships.

Metric. We evaluate performance on the V-COCO dataset using its official metric: "AP role" [8]. This metric computes the mean average precision (mAP) of detecting relationships triplets $\langle$*human box, action, target box*$\rangle$ by following the PASCAL VOC 2012 [3] evaluation protocol. A triplet prediction is considered as correct, if all three of its components

[1] https://github.com/tensorflow/tpu/tree/master/models/official/retinanet

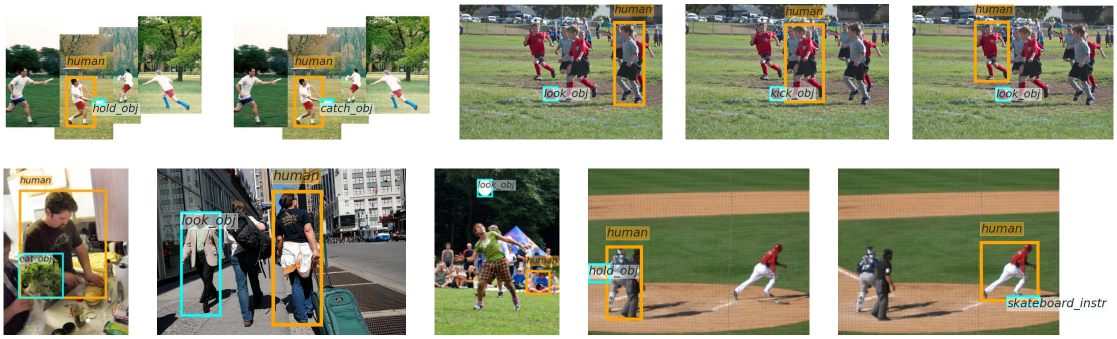

Figure 3. Example outputs of the top scoring detections by the proposed BAR-CNN model on *V-COCO*. The first row demonstrates correct predictions outputted by the model. The second row shows failures: image 1, 2, 3 — wrong target, images 4, 5 — hallucinated object.

are correct. A predicted box is correct if it has intersection-over-union with a ground-truth box of at least 50%. We use the publicly available code for computing this metric[2].

Qualitative results. Figure 3 show typical outputs of our model. By analyzing these qualitative results, we make a few observations. Most importantly, our model successfully learns to use the box attention map. Even for complex images with many objects it learns to correctly assign humans to their corresponding objects.

Interestingly, BAR-CNN can successfully predict that the same object can correspond to different humans through different actions (row 1, the right-most three images). Moreover, BAR-CNN can model long range interaction between objects, *e.g.* in row 1 the two football players looking at the ball are far from a ball and yet are predicted to be related to it through the action *look*.

In the V-COCO dataset most errors are caused by complex image semantics, which are hard to capture by a neural network trained on very limited amount of data (row 2).

Quantitative results. Results are presented in Table 1. The first two columns show quantitative comparison to the model from [8] and the approach from [5]. Our method BAR-CNN (third column) outperforms both of them.

The recently proposed ICAN model [4] achieves a 45.3 mean AP, which is slightly better than our model (43.6). However, we stress that (1) our model handles the generic visual relationship detection task, whereas ICAN focuses on human-object interaction; (2) the ICAN model is much more complex than ours, as it introduces numerous additional components on top of the object detection pipeline.

4.3. Results on Open Images VRD Challenge 2018

Data. The *Open Images* Dataset (OID) is a very large-scale dataset containing image-level labels, object bounding boxes, and visual relationships annotations. In total it contains 329 distinct relationship triplets and 374, 768 annotations on 100, 522 images in the training set.

Metric. We evaluate the model on the hidden Open Images Challenge 2018[3] test set using the official Kaggle server. The metric is the weighted average of the three metrics: mAP on phrase detection, mAP for relationship detection and Recall@50 for relationship detection. The task of phrase detection is to detect triplets of object with a single enclosing bounding box and three labels l^s, l^p, l^o. It was introduced in [17]. The two other metrics require detecting separately each object and their relationship label. For the mAP metrics, the mean is computed over relationship predicate, *i.e.* l^p.

	Score(public)	Score(private)
team MIL	21.8	19.7
team mission-pipeline	7.4	6.8
team toshif	25.6	22.8
BAR-CNN (proposed)	**26.6**	**25.0**

Table 2. Quantitative comparison of the proposed model with competing models on *Open Images*.

Quantitative results. In Table 2 we compare the results of our BAR-CNN to the results of *Open Images* VRD Challenge 2018, where comparable setting was used. We use values of the public leaderboard on the Kaggle server for validation and report the score both on the public and private leaderboard for all methods. We note, that among all submitted results, our model achieves the second best performance despite us not training a separate model for "is" relationship.

5. Conclusion

We presented a new model, BAR-CNN, for detecting visual relationships that relies on a box attention mechanism. Our model has several important benefits over previously proposed models. First, it is conceptually simple and theoretically sound: we tackle visual relationship detection by formulating it as a task of learning a probabilistic model and then decomposing this model into simpler sub-models using the chain rule. Second, our model does not introduce any new hyperparameters on top of those already required by the base detection model it builds on. Finally, BAR-CNN delivers strong performance on two challenging datasets.

[2] https://github.com/s-gupta/v-coco

[3] https://storage.googleapis.com/openimages/web/challenge.html

References

[1] B. Dai, Y. Zhang, and D. Lin. Detecting visual relationships with deep relational networks. In *CVPR*, 2017.

[2] C. Desai and D. Ramanan. Detecting actions, poses, and objects with relational phraselets. In *ECCV*, 2012.

[3] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *IJCV*, 2015.

[4] C. Gao, Y. Zou, and J.-B. Huang. ican: Instance-centric attention network for human-object interaction detection. In *British Machine Vision Conference*, 2018.

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

[6] A. Gupta and L. S. Davis. Beyond nouns: Exploiting prepositions and comparative adjectives for learning visual classifiers. In *ECCV*, 2008.

[7] A. Gupta, A. Kembhavi, and L. S. Davis. Observing human-object interactions: Using spatial and functional compatibility for recognition. *IEEE T-PAMI*, 2009.

[8] S. Gupta and J. Malik. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*, 2015.

[9] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[10] G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[11] 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 *CVPR*, 2017.

[12] R. Krishna, I. Chami, M. S. Bernstein, and L. Fei-Fei. Referring relationships. *CVPR*, 2018.

[13] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, T. Duerig, and V. Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *arXiv:1811.00982*, 2018.

[14] Y. Li, W. Ouyang, X. Wang, and X. Tang. ViP-CNN: Visual phrase guided convolutional neural network. In *CVPR*, 2017.

[15] X. Liang, L. Lee, and E. P. Xing. Deep variation-structured reinforcement learning for visual relationship and attribute detection. In *CVPR*, 2017.

[16] T. Lin, P. Goyal, R. B. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *ICCV*, 2017.

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

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

[19] J. Peyre, I. Laptev, C. Schmid, and J. Sivic. Weakly-supervised learning of visual relations. In *CVPR*, 2017.

[20] A. Prest, C. Schmid, and V. Ferrari. Weakly supervised learning of interactions between humans and objects. *IEEE T-PAMI*, 2012.

[21] M. A. Sadeghi and A. Farhadi. Recognition using visual phrases. In *CVPR*, 2011.

[22] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016.

[23] B. Yao and L. Fei-Fei. Modeling mutual context of object and human pose in human-object interaction activities. In *CVPR*, 2010.

[24] R. Yu, A. Li, V. I. Morariu, and L. S. Davis. Visual relationship detection with internal and external linguistic knowledge distillation. In *CVPR*, 2017.

[25] H. Zhang, Z. Kyaw, S.-F. Chang, and T.-S. Chua. Visual translation embedding network for visual relation detection. In *CVPR*, 2017.

[26] H. Zhang, Z. Kyaw, J. Yu, and S.-F. Chang. PPR-FCN: weakly supervised visual relation detection via parallel pairwise R-FCN. *ICCV*, 2017.