FAN: FOCUSED ATTENTION NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Attention networks show promise for both vision and language tasks, by emphasizing relationships between constituent elements through weighting functions. Such elements could be regions in an image output by a region proposal network, or words in a sentence, represented by word embedding. Thus far the learning of attention weights has been driven solely by the minimization of task specific loss functions. We introduce a method for learning attention weights to better emphasize informative pair-wise relations between entities. The key component is a novel center-mass cross entropy loss, which can be applied in conjunction with the task specific ones. We further introduce a focused attention backbone to learn these attention weights for general tasks. We demonstrate that the focused supervision leads to improved attention distribution across meaningful entities, and that it enhances the representation by aggregating features from them. Our focused attention module leads to state-of-the-art recovery of relations in a relationship proposal task and boosts performance for various vision and language tasks.

1 INTRODUCTION

Complex tasks involving visual perception or language interpretation are inherently contextual. In an image of an office scene, for example, a computer mouse may be too small to detect but the recognition of a computer keyboard might hint at its presence and its possible locations. The study of objects in their context is a cornerstone of much past computer vision work (Rabinovich et al., 2007). Scene categories are often determined by the relationships between objects or environments commonly found in them (Zhou et al., 2017) while in natural language processing words must be interpreted in relation to other words or phrases in sentences. Machine learning algorithms that learn object to object or word to word relationships have thus been sought. Among them, attention networks have shown great promise for the task of learning relationship attention weights between entities (Veličković et al., 2017; Vaswani et al., 2017). As a recent example, the scaled dot product attention module from Vaswani et al. (2017) achieves state of the art performance in language translation tasks.

We here propose to explicitly supervise the learning of attention weights between elements of a data source using a novel center-mass cross entropy loss. The minimization of this loss increases relation weights between entity pairs which are more commonly observed in the data, but without the need for handcrafted frequency measurements. We design a focused attention network that is end-to-end trainable and which explicitly learns pairwise element affinities without the need for relationship annotations in the data. Multiple experiments demonstrate that such focused attention improves upon the baseline as well as attention without focus, for both computer vision and natural language processing tasks. In a relationship proposal task the use of this backbone achieves results comparable to the present state-of-the-art (Zhang et al., 2017), even without the use of ground truth relationship labels. The use of ground truth labels for focused attention learning leads to a further 25% relative improvement, as measured by a relationship recall metric.

2 MOTIVATION

Attention Networks – The Present State The modeling of relations between objects as well as objects in their common contexts has a rich history in computer vision (Rabinovich et al., 2007; Torralba, 2003; Galleguillos & Belongie, 2010). Deep learning based object detection systems leverage attention models to this end, to achieve impressive performance in recognition tasks. The scaled dot product attention module of Vaswani et al. (2017), for example, uses learned pairwise

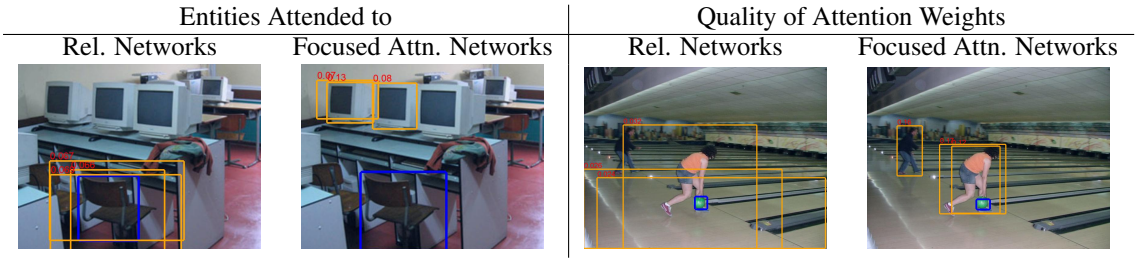


Figure 1: A comparison of recovered relationships on the MIT67 dataset, with training *only* on the minicoco dataset. The reference object is surrounded by a blue box and regions with which it learns relationships are shown with orange boxes with the relationship weights visible in red text (zoom-in on the PDF). Left: Relation Networks (Hu et al., 2018) often learn weights between a reference object and its immediate surrounding context, while Focused Attention Networks better emphasize relationships between distinct and spatially separated objects. Right: Relation Networks can suffer from a poor selection of regions to pair, or low between object relationship weights in comparison to Focused Attention Networks.

attention weights between region proposal network (RPN) generated bounding boxes in images of natural scenes (Hu et al., 2018) to boost object detection. Pixel level attention models have also been explored to aid semantic segmentation (Zhao et al., 2018) and video classification (Wang et al., 2018).

Current approaches to learn the attention weights often do not reflect relations *between* entities in a typical visual scene. In fact, for a given reference object (region), relation networks (Hu et al., 2018) tend to predict high attention weights with scaled or shifted bounding boxes surrounding the same object instance. This is likely because including surrounding context, or simply restoring missing parts of the reference object, boosts object detection. The learned relationship weights between distinct objects (regions) are also often small in magnitude. Typical qualitative examples comparing Relation Networks with our Focused Attention Network are shown in Figure 1, with a quantitative comparison reported in Section 5. Similar situations can occur in applications of attention networks to natural language processing tasks. In document classification, for example, attention weights learned using Hierarchical Attention Networks (Yang et al., 2016) tend to concentrate on a few words in a sentence, while our Focused Attention Network leads to more distributed attention, allowing for more comprehensive sentence level features, as illustrated in Figure 2.

Figure 2: Left: Visualization of the word importance factor, which models the contribution of a given word to its sentence level context (see Section 4.5). The sentence is in a clockwise direction starting from 12 o'clock. "ait": weights learned using Hierarchical Attention Networks (Yang et al., 2016), "unsup": weights from the unsupervised case of our Focused Attention Network module, and "sup": weights from the supervised case. Right: semantic word-to-word relationship labels used in the supervision of our network (see Section 3.2).

Attention Networks – Limitations A present limitation of attention networks in various applications is the use of only task specific losses as their training objectives. There is little work thus far on explicitly supervising the learning of weights, so as to be more distributed across meaningful entities. For example, Relation Networks (Hu et al., 2018) and those applied to segmentation problems, such as PSANet (Zhao et al., 2018), learn attention weights *solely* by minimizing categorical cross entropy for classification, L1 loss for bounding box localization or pixel-wise cross entropy loss for semantic segmentation (Zhao et al., 2018). In language tasks including machine translation (Vaswani et al., 2017) and document classification (Yang et al., 2016), the attention weights are also solely learned by

minimizing the categorical cross entropy loss. In this article we refer to such attention networks as being *unsupervised*.

Whereas attention aggregation with learned weights boosts performance for these specific tasks, our earlier examples provide evidence that relationships between distinct entities may not be adequately captured. We shall address this limitation by focusing the attention weight learning using a novel center-mass cross entropy loss, as discussed in the following section.

3 FOCUSING THE ATTENTION

Given our goal of better reflecting learned attention weights between distinct entities, we propose to explicitly supervise the learning of attention relationship weights. We accomplish this by introducing a novel center-mass cross entropy loss.

3.1 PROBLEM STATEMENT

Given a set of N entities that are generated by a feature embedding framework, which can be a region proposal network (RPN) (Ren et al., 2015) or a word embedding layer with a bidirectional LSTM (Yang et al., 2016), for the i-th entity we define $\mathbf{f}^i$ as the embedding feature. To compute the relatedness or affinity between entity m and entity n, we define an attention function $\mathcal{F}$ which computes the pairwise attention weight as

$$\omega^{mn} = \mathcal{F}(\mathbf{f}^m, \mathbf{f}^n). \tag{1}$$

A specific form of this attention function applied in this paper is reviewed in Section 4.1, and it originates from the scaled dot product attention module of Vaswani et al. (2017).

We can now build an attention graph G whose vertices m represent entities in a data source with features $F = \{\mathbf{f}^m\}$ and whose edge weights $\{\omega^{mn}\}$ represent pairwise affinities between the vertices. We define the graph adjacency matrix for this attention graph as $\mathcal{W}$. We propose to supervise the learning of $\mathcal{W}$ so that the matrix entries ω^{mn} corresponding to entity pairs with high co-occurrence in the training data gain higher attention weights.

3.2 SUPERVISION TARGET

We now discuss how to construct ground truth supervision labels in matrix form to supervise the learning of the entries of $\mathcal{W}$. For **visual recognition** tasks we want our attention weights to focus on relationships between objects from different categories, so for each entry t^{mn} of the ground truth relationship label matrix $\mathcal{T}$, we assign $t^{mn} = 1$ only when: 1) entities m and n overlap with two different ground truth objects' bounding boxes with intersection over union (IOU) > 0.5 and 2) their category labels c_m and c_n are different. We provide a visualization of such ground truth relationships in Figure 4 of the appendix. For **language** tasks we want the attention weights to reveal meaningful word pairs. For example, semantic relationships between nouns and nouns, verbs and nouns, nouns and adjectives, adverbs and verbs, and adverbs and adjectives should be encouraged. To this end, we build a simple word category pair dictionary of valid pairings (see Figure 2) and assign label $t^{mn} = 1$ when the word category pair c_m and c_n is found in this semantic pair dictionary.

Center-Mass Intuitively, we would like $\mathcal{W}$ to have high affinity weights at those entries where $t^{mn} = 1$, and low affinity weights elsewhere, i.e., the attention weights should concentrate on the 1's in the ground truth relationship label matrix $\mathcal{T}$. We capture this via a notion of *center-mass* $\mathcal{M}$ of ground truth relation weights, defined as

$$\mathcal{M} = \sum \tilde{\mathcal{W}} \odot \mathcal{T}, \tag{2}$$

where $\tilde{\mathcal{W}} = softmax(\mathcal{W})$ is a matrix-wise softmax operation.

3.3 CENTER-MASS CROSS ENTROPY LOSS

The key to our approach is the introduction of a center-mass cross entropy loss, which aims to focus attention weight learning so that ω^{mn} is high for pairs of commonly occurring distinct entities. The

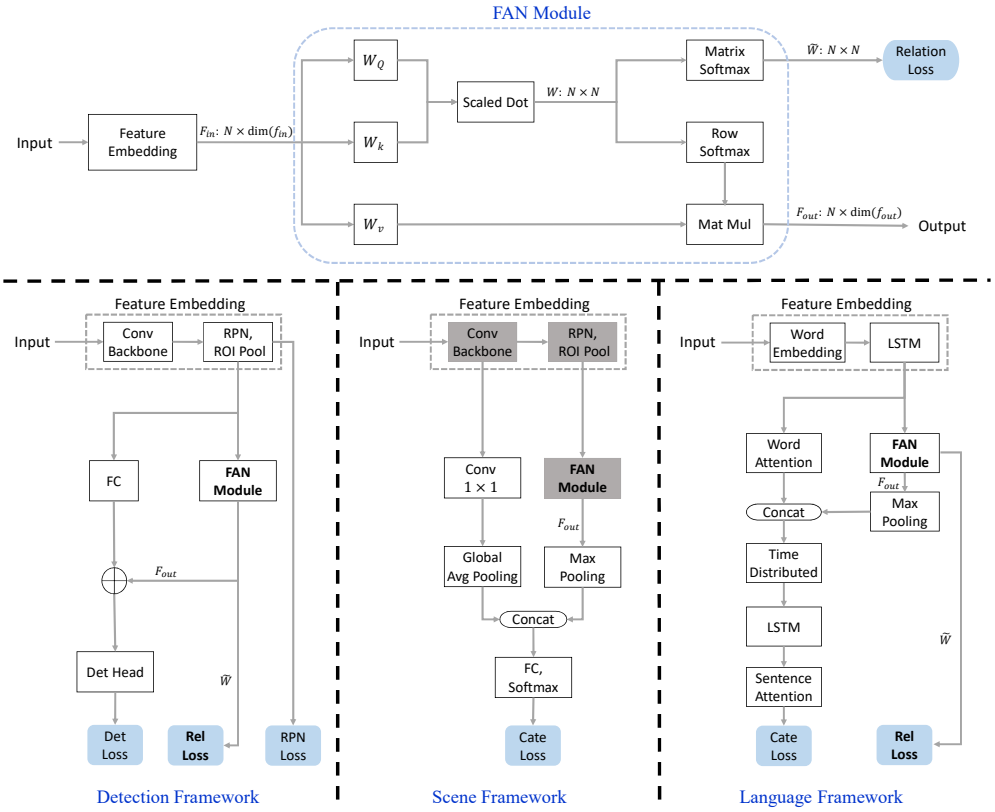


Figure 3: Top: The Focused Attention Network backbone. Bottom left: we add a detection branch to the backbone, similar to that of Hu et al. (2018). Bottom middle: we add a scene recognition branch to the backbone. Bottom right: we insert the Focused Attention Module into a Hierarchical Attention Network (Yang et al., 2016). Here “FC" stands for fully connected layer and “Conv" stands for convolutional layer.

loss is computed as

$$\mathcal{L} = -(1 - \mathcal{M})^r \log(\mathcal{M}). \tag{3}$$

When minimizing this center-mass loss more frequently occurring 1-labeled pairs in the matrix will cumulatively receive stronger emphasis, for example, human-horse pairs versus horse-chair pairs in natural images. Furthermore, when supervising the attention learning in conjunction with another task specific loss, the matrix entries that reduce the task loss will also be optimized. The resultant dominant ω^{mn} entries will not only reflect entity pairs with high co-occurrence, but will also help improve the main objective. The focal term $(1 - \mathcal{M})^r$ (Lin et al., 2017) helps narrow the gap between well converged center-masses and those that are far from convergence. For example, with a higher center-mass value the gradient on the log loss will be scaled down, whereas for a lower center-mass the gradient will be scaled up. The focal term prevents the network from committing only to the most dominant ω^{mn} entries, and thus promotes diversity. We choose $r = 2$ in our experiments, motivated by the model ablation study conducted in Section 5.3.

4 Network Architecture

Our focused attention module originates from the scaled dot product attention module in Vaswani et al. (2017). We discuss our network structures for the focused attention weight learning backbone and various specific tasks, as shown in Figure 3, with implementation details provided in the appendix.

4.1 Scaled Dot Product Attention Network

We briefly review the computation of attention weights in Vaswani et al. (2017), given a pair of nodes from the attention graph defined in Section 3.1. Let an entity node consist of its feature embedding,

defined as $\mathbf{f}$. Given a reference entity node m, such as one of the blue boxes in Figure 1, the attention weight $\tilde{\omega}^{mn}$ indicates its affinity to a surrounding entity node n. It is computed using a softmax activation over the scaled dot products ω^{mn} defined as:

$$\tilde{\omega}^{mn} = \frac{\exp(\omega^{mn})}{\sum_k \exp(\omega^{kn})}, \quad \omega^{mn} = \frac{dot(W_K \mathbf{f}^m, W_Q \mathbf{f}^n)}{\sqrt{d_k}}. \tag{4}$$

Both W_K and W_Q are matrices and so this linear transformation projects the embedding features $\mathbf{f}^m$ and $\mathbf{f}^n$ into metric spaces to measure how well they match. The feature dimension after projection is d_k. With the above formulation, the attention graph affinity matrix is defined as $\mathcal{W} = \{\omega^{mn}\}$.

4.2 Focused Attention Network (FAN) Backbone

In Figure 3 (top) we illustrate the base Focused Attention Network architecture. Here the dot product attention weights $\mathcal{W}$ go through a matrix-wise softmax operation to generate the attention matrix output $\tilde{\mathcal{W}}$, that is used for the focused supervision with the center-mass cross entropy loss defined in Section 3.3. We shall refer to this loss term as *relation loss*. In parallel, a row-wise softmax is applied to $\mathcal{W}$ to output the coefficients $\mathcal{W}_{agg}$, which are then used for attention weighted aggregation: $\mathbf{f}^m_{out} = \sum_n \omega^{mn}_{agg} f^n$. The aggregated output from the FAN module is sent to a task specific loss function. The entire backbone is end-to-end trainable, with both the task loss and the relation loss.

4.3 Object Detection and Relationship Proposals

In Figure 3 (bottom left) we demonstrate how to generalize the FAN module for object detection and relationship proposal generation. The network is end-to-end trainable with detection loss, RPN loss and our relation loss. In addition to the ROI pooling features $\mathbf{F} \in \mathcal{R}^{N_{obj} \times 1024}$ from the Faster R-CNN backbone of Ren et al. (2015), contextual features $\mathbf{F}_c$ from attention aggregation are applied to boost detection performance: $\mathbf{F}_c = \mathcal{W}_{agg}\mathbf{F}$. The final feature descriptor for the detection head is $\mathbf{F}_d = \mathbf{F} + \mathbf{F}_c$, following Hu et al. (2018). In parallel, the attention matrix output $\tilde{\mathcal{W}} \in \mathcal{R}^{N \times N}$ is used to generate relationship proposals by finding the top K weighted pairs in the matrix.

4.4 Scene Categorization Task

In Figure 3 (bottom middle) we demonstrate how to apply the FAN module to scene categorization. Since there are no bounding box annotations in most scene recognition datasets, we adopt a pre-trained FAN detection module, described in Section 4.3, in conjunction with a newly added convolution branch, to perform scene recognition. From the convolution backbone, we apply an additional convolution layer followed by a global average pooling to acquire the scene level feature descriptor $\mathbf{F}_s$. The FAN module takes as input the object proposals' visual features $\mathbf{F}$, and outputs the aggregation result as the scene contextual feature $\mathbf{F}_c$. The input to the scene classification head thus becomes $\mathbf{F}_{meta} = concat(\mathbf{F}_s, \mathbf{F}_c)$, and the class scores are output.

4.5 Document Categorization Task

In Figure 3 (bottom right) we demonstrate how to apply the FAN module to a document classification task, using Hierarchical Attention Networks (Yang et al., 2016). We insert the FAN module into the word level attention layer, but making it parallel to the original word-to-sentence attention module, and denote it as "FAN-hatt". FAN module learns word-to-word attention and outputs a sentence level descriptor. The FAN descriptor is concatenated with the output from the word-to-sentence attention to result in enhanced sentence level embedding. The rest of the computation follows the original paper (Yang et al., 2016), where a sentence level attention model is applied and a final document level descriptor is abstracted.

5 Experiments

5.1 Datasets and Network Training

Datasets. We evaluate our FAN architecture using the following datasets: VOC07, MSCOCO, Visual Genome, MIT67, 20 Newsgroups, Yahoo Answers. Details are provided in the appendix.

Vision Network Training Following Section 4.3, we first train the detection-and-relation joint framework end-to-end with a detection task loss and a relation loss on the minicoco dataset, dubbing this network "FAN-minicoco". We report detection results as well as relation learning quality. We then fine tune the scene task structure on the MIT67 dataset, using the pre-trained FAN-minicoco network (see Section 4.4), and report scene categorization performance. Details on the network architecture, the input/output dimensions and the hyper-parameters are in the appendix.
Language Network Training We train the FAN models with Hatt (Yang et al., 2016) as the backbone. Details on the architecture and hyper-parameters are in the appendix.

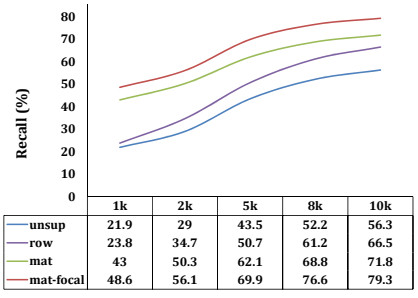
5.2 Evaluation Metrics

Object Detection Task. For the VOC07 and MSCOCO datasets the evaluation metric is mAP (mean average precision), as defined in (Everingham et al., 2010; Lin et al., 2014).
Relationship Proposal Task. We evaluate the learned relationships using a recall metric which measures the percentage of ground truth relations that are covered in the predicted top K relationship list, which is consistent with (Zhang et al., 2017; Zellers et al., 2018; Xu et al., 2017). A detailed definition is in the appendix.
Classification Tasks. For MIT67, 20 Newsgroups and Yahoo Answers, where the task is to classify scenes or documents, we use classification accuracy as the evaluation metric.

5.3 Model Ablation Study

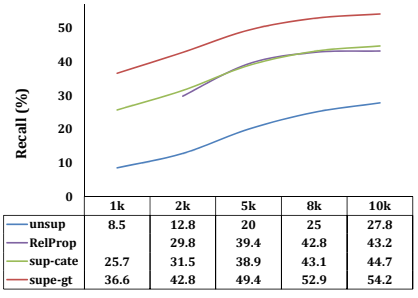
Prior to evaluating the FAN model we carry out ablation studies to examine ways of supervising the center mass and different choices of loss functions.

Focused Supervision Strategies We consider different approaches to training the focused attention, using the detection-and-relation joint framework from Section 4.3 on the VOC07 dataset for each case. First, we apply a row-wise softmax over the pre-activation matrix $\mathcal{W}$ and calculate the center-mass in a row-wise manner and apply the center-mass cross entropy loss accordingly. We refer to this as "row". Second, we apply the supervision explained in Section 3.3 but without the use of the focal term, and refer to this as "mat". Third, we add the focal term to the matrix supervision, referring to this as "mat-focal". The results in Table 1 show that the focused attention weights, when supervised using the center-mass cross entropy loss with a focal term (Section 3.3), are better concentrated on inter-object relationships, as reflected by the recall metric (Section 5.2) when compared with the unsupervised case. In all further experiments, unless stated otherwise, we apply the matrix supervision with the focal term, since it provides the best performance.



	1k	2k	5k	8k	10k
unsup	21.9	29	43.5	52.2	56.3
row	23.8	34.7	50.7	61.2	66.5
mat	43	50.3	62.1	68.8	71.8
mat-focal	48.6	56.1	69.9	76.6	79.3

Table 1: Evaluating different supervision strategies with varying top K, using the VOC07 test-set, with ground truth relation labels as described in Section 3.2.



	1k	2k	5k	8k	10k
unsup	8.5	12.8	20	25	27.8
RelProp		29.8	39.4	42.8	43.2
sup-cate	25.7	31.5	38.9	43.1	44.7
supe-gt	36.6	42.8	49.4	52.9	54.2

Table 2: Recall comparison for the Visual Genome dataset with varying top K, where the ground truth relation labels are human annotated. See text in Section 5.4 for a discussion.

Design of Loss Functions. We conduct a second ablation study to examine additional loss functions for optimizing the center mass as well as varying focal terms r, as introduced in Section 3.3. Defining

VOC07	smooth L1	L2	$r = 0$	$r = 1$	$r = 2$	$r = 3$	$r = 4$
mAP@0.5	79.6 ± 0.2	79.7 ± 0.2	79.4 ± 0.1	79.5 ± 0.2	$\mathbf{79.9 \pm 0.2}$	79.8 ± 0.1	79.7 ± 0.2
recall@5k	60.3 ± 0.3	64.6 ± 0.5	62.1 ± 0.3	66.5 ± 0.2	$\mathbf{69.9 \pm 0.3}$	68.7 ± 0.4	67.1 ± 0.3

Table 3: An evaluation of smooth L_1 and L_2 loss functions and variations of the focal loss factor r, on the VOC07 dataset. The results are reported as percentages (%) averaged over 3 runs.

$x = 1 - \mathcal{M} \in [0, 1]$, we consider loss function variants $L_2 = x^2$ and

$$\text{smooth}_{L_1}(x) = \begin{cases} x^2 & \text{if} \quad |x| < 0.5 \\ |x| - 0.25 & \text{otherwise,} \end{cases} \tag{5}$$

in addition to the focal loss. The results in Table 3 show that focal loss is in general better than smooth L1 and L2 when supervising the center mass. In subsequent experiments we apply focal loss with $r = 2$, which empirically gives the best performance.

5.4 Relationship Proposal Task

Table 2 shows the evaluation of relationships learned on the Visual Genome dataset, by the unsupervised Focused Attention Network model "unsup" (similar to (Hu et al., 2018)), the Focused Attention Network supervised with weak relation labels described in Section 3.2 "sup-cate", and supervision with human annotated ground truth relation labels "sup-gt". We also include the reported recall metric from Relationship Proposal Networks (Zhang et al., 2017), which is a state-of-the-art level relationship learning network with strong supervision, using ground truth relationships. Our center-mass cross entropy loss does not require potentially costly human annotated relationship labels for learning, yet it achieves the same level of performance as the present state-of-the-art (Zhang et al., 2017) (the green curve in Table 2). When supervised with the ground truth relation labels instead of the weak labels (Section 3.2), we significantly outperform relation proposal networks (by about 25% in relative terms for all K thresholds) with this recall metric (the red curve in Table 2).

5.5 Object Detection Task

In Table 4 we provide object detection results on the VOC07 and MSCOCO datasets. In both cases we improve upon the baseline and slightly outperform the unsupervised case (similar to Relation Networks (Hu et al., 2018)). This suggests that relation weights learned using our focused attention network are at least as good as those from (Hu et al., 2018), in terms of object detection performance.

VOC07	base F-RCNN	FAN + $\mathcal{L}_{det}$ (Hu et al., 2018)	FAN + $\mathcal{L}_{det}$ + $\mathcal{L}_{rel}$
avg mAP (%)	47.0	47.7 ± 0.1	48.2 ± 0.2
mAP@0.5 (%)	78.2	79.3 ± 0.2	79.9 ± 0.2
mini COCO	base F-RCNN	FAN + $\mathcal{L}_{det}$ (Hu et al., 2018)	FAN + $\mathcal{L}_{det}$ + $\mathcal{L}_{rel}$
avg mAP (%)	26.8	27.5	27.9
mAP@0.5 (%)	46.6	47.4	47.8

Table 4: Object Detection Results. mAP@0.5: average precision over a bounding box overlap threshold as $IOU = 0.5$. avg mAP: averaged mAP over multiple bounding box overlap thresholds. VOC07 experiments are reported over 3 runs, demonstrating stability. $\mathcal{L}_{det}$ stands for detection task loss as defined in Ren et al. (2015) and $\mathcal{L}_{rel}$ for the center mass relation loss defined in section 3.3.

5.6 Scene Categorization Task

We adopt the FAN-minicoco network (Section 5.1), and add an additional scene task branch to fine tune it on MIT67, as discussed in Section 4.4. Table 5 shows the results of applying this model to the MIT67 dataset. We refer to the backbone as "CNN" (first column), which is the left branch in Figure 3 (bottom middle). In the second column we apply the same network further fine-tuned on minicoco before training on MIT67. In the third column we include the detection branch, which is the right branch in Figure 3 (bottom middle) but remove its FAN module. In the fourth and fifth

	CNN	CNN	CNN + ROIs	CNN + FAN-unsup	CNN + FAN-sup
Pretraining	Imgnet	Imgnet+COCO	Imgnet+COCO	Imgnet+COCO	Imgnet+COCO
Features	$\mathbf{F}_S$	$\mathbf{F}_S$	$\mathbf{F}_S, max(\mathbf{F})$	$\mathbf{F}_S, \mathbf{F}_C$	$\mathbf{F}_S, \mathbf{F}_C$
Accuracy (%)	75.1	76.8	78.0 ± 0.3	77.1 ± 0.2	80.2 ± 0.3

Table 5: MIT67 Scene Categorization Results. The important entries are averages over 3 runs. See the text in Section 5.6 for a discussion. For details regarding F_s, F_c and F, see Section 4.4.

columns we apply the full scene architecture in Figure 3 (bottom middle), using FAN-minicoco network pretrained without (unsupervised) and with (supervised) relation loss, respectively. The FAN supervised case (fifth column) demonstrates a non-trivial improvement over the baseline (third column) and also significantly outperforms the unsupervised case (fourth column). This suggests that the relation weights learned solely by minimizing detection loss do not generalize well to a scene task, whereas those learned by our Focused Attention Network supervised by weak relations labels can. We hypothesize that recovering informative relations between distinct objects, which is what our Focused Attention Network is designed to do, is particularly beneficial for scene categorization.

5.7 Document Categorization Task

Datasets	Train Vol. Per Cate	Hatt (Yang et al., 2016)	FAN-hatt-unsup	FAN-hatt-sup
20 News	550	64.0 ± 0.3	64.6 ± 0.2	65.6 ± 0.2
Yahoo-mini	5,000	64.9 ± 0.2	65.2 ± 0.2	66.0 ± 0.1
Yahoo-half	70,000	72.2 ± 0.1	72.1 ± 0.1	72.4 ± 0.1

Table 6: Document categorization results for the 20 Newsgroups and Yahoo Answers datasets, with the results averaged over 5 trials. For the Yahoo dataset, we train on sub-sampled training sets and report results on the full test set.

We present document classification results in Table 6. We provide a comparison between the base Hierarchical Attention Networks (Hatt), and FAN-hatt, explained in Section 4.5, with and without relation loss supervision. The supervised (semantic) focused attention supervision results in an improvement over both the unsupervised case and the baseline, particularly when the training data per category is small (top two rows). This advantage diminishes, however, when the training volume is dramatically increased, suggesting that in the latter case a baseline network is able to perform equally well. In addition, the qualitative distributions of word importance factors (see appendix for an exact definition) in Figure 2 suggest that focused semantic attention encourages more diversity in attention weight learning, which in turn leads to better document classification performance.

6 Conclusion

Our Focused Attention Network is versatile, and allows the user to direct the learning of attention weights in the manner they choose. The application of the FAN module allows multiple informative entities in the data source to contribute to the learned feature representation, abstracting diverse aspects of the data. In multiple experiments we have demonstrated the benefit of learning relations between distinct objects for computer vision tasks, and between lexical categories (words) for document classification tasks. It not only boosts performance in object detection, scene categorization and document classification, but also leads to state-of-the-art performance in a relationship proposal task. In the future we envision its use as a component for deep learning architectures where supervised control of relationship weights is desired, since it is adaptable, modular, and end-to-end trainable in conjunction with a task specific loss.

References

M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 2010.

Carolina Galleguillos and Serge Belongie. Context based object categorization: A critical survey. *Computer vision and image understanding*, 2010.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. 2016.

Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. Relation networks for object detection. *CVPR*, 2018.

Thorsten Joachims. A probabilistic analysis of the rocchio algorithm with tfidf for text categorization. *Carnegie-Mellon University Technical Report*, 1996.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *ICLR*, 2015.

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. *IJCV*, 2017.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. *ECCV*, 2014.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *CVPR*, 2017.

Ariadna Quattoni and Antonio Torralba. Recognizing indoor scenes. *CVPR*, 2009.

Andrew Rabinovich, Andrea Vedaldi, Carolina Galleguillos, Eric Wiewiora, and Serge J. Belongie. Objects in context. *ICCV*, 2007.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *NIPS*, 2015.

Antonio Torralba. Contextual priming for object detection. *IJCV*, 2003.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NIPS*, 2017.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. *CVPR*, 2018.

Danfei Xu, Yuke Zhu, Christopher Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. *CVPR*, 2017.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical attention networks for document classification. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016.

Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. *CVPR*, 2018.

Ji Zhang, Mohamed Elhoseiny, Scott Cohen, Walter Chang, and Ahmed Elgammal. Relationship proposal networks. *CVPR*, 2017.

Hengshuang Zhao, Yi Zhang, Shu Liu, Jianping Shi, Chen Change Loy, Dahua Lin, and Jiaya Jia. Psanet: Point-wise spatial attention network for scene parsing. *ECCV*, 2018.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

Appendix

A Further Experimental Details

A.1 Datasets

VOC07: is part of the PASCAL VOC detection dataset (Everingham et al., 2010). It consists of 5k images for training and 5k for testing. We used the full training/test split in our experiments.
MSCOCO: consists of 80 object categories (Lin et al., 2014). Within the 35k validation images of the COCO2014 detection benchmark, a selected 5k subset named "minival" is commonly used when reporting test time performance, for ablation studies (Hu et al., 2018). We used the 30k validation images for training and the 5k "minival" images for testing. We define this split as "minicoco".
Visual Genome: is a large scale relationship understanding benchmark (Krishna et al., 2017), consisting of 150 object categories and human annotated relationship labels between objects. We used 70k images for training and 30K for testing, as in the scene graph literature (Zellers et al., 2018; Xu et al., 2017).
MIT67: is a scene categorization benchmark which consists of 67 scene categories, with each category having 80 training images and 20 test images (Quattoni & Torralba, 2009). We used the full training/testing split in our experiments.
20 Newsgroups: is a document classification dataset consisting of text documents from 20 categories (Joachims, 1996), with 11314 training ones and 7532 test ones. We used the full training/testing split in our experiments.
Yahoo Answers: is a topic classification dataset consisting of 10 categories and is organized from the Yahoo! Answers Comprehensive Questions and Answers version 1.0 dataset. Each class contains 140,000 training samples and 6,000 testing samples. In our experiments, we evaluate on the full test set but sub-sample the training set during training. First, we randomly sample 5,000 training samples per category within the 140,000 and refer to this configuration as "Yahoo-mini". Next we randomly sample 70,000 training samples per category within the 140,000 and refer to this configuration as "Yahoo-half". We report full test set results using the above different training sets.

A.2 Relationship Recall Metric

We evaluate the learned relationships using a recall metric defined as $R_{rel} = \frac{C(rel|topK)}{C(rel)}$. Here $C(rel)$ stands for the number of unique ground truth relations in a given image and $C(rel|topK)$ stands for the number of unique matched ground truth relations in the top-K ranked relation weight list. In the calculation of $C(rel|topK)$, we only consider a match when both bounding boxes in a given relationship pair have overlaps of more than 0.5, with the corresponding ground truth boxes in a ground truth relationship pair. Therefore, R_{rel} measures how well the top-K ranked relation weights capture the ground truth labeled relationships.

A.3 Word Importance Factor

The word importance factor models the contribution of a given word to its sentence level representation. In Hierarchical Attention Networks (Yang et al., 2016), the word-to-sentence attention module directly models the word importance given a learned sentence representation. Whereas it is different from the FAN module, which models word-to-word attention, the resultant attention weights can be comparable at the sentence level. To this end, we define the word importance factor as $\beta_i = \sum_j \tilde{\omega}^{ji}$, because it represents the contribution of the i-th word in the final aggregation result $\mathcal{W}_{agg}\mathbf{F}$. A visualization of the distribution of learned word importance factors, for Hierarchical Attention Networks and Focused Attention Networks, is provided in Figure 2 in the main text and Figure 6 in this appendix.

A.4 Supervision Targets $\mathcal{T}$

We consider different possibilities for supervision targets $\mathcal{T}$, as defined in Section 3.2.

Vision tasks. In our paper we focused attention between objects from different categories, and we refer to this as **different category focused supervision**. We now consider the case that attention

between different object instances is focused. That is, as long as object proposal a and object proposal b are different object instances in an image, we consider a possible relationship between them and assign $t^{ab} = \mathcal{T}[a, b] = 1$. We refer to this as **different instance focused supervision**. We provide a visual example of the above mentioned supervision targets in Figure 4.

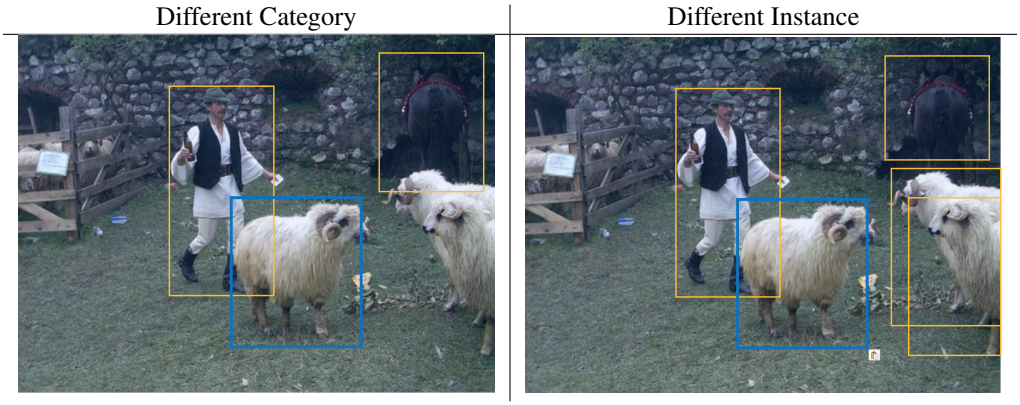


Figure 4: The visualization of supervision targets for vision tasks. The blue box indicates a fixed reference object a and the orange boxes indicate the objects b that have ground truth relationship with a, for which we assign $\mathcal{T}[a, b] = 1$. Left: different category supervision. Note that the sheep in the blue box is *not* related to the other sheep in the image. Right: different instance supervision. The sheep in the blue box now has a relationship to other sheep (in yellow boxes).

In Table 7, we provide object detection results on VOC07 when training the FAN model using different supervision targets.

VOC07 varying $\mathcal{T}$	Diff Instance	Diff Category
avg mAP (%)	47.6 ± 0.1	48.2 ± 0.2
mAP@0.5 (%)	79.5 ± 0.2	79.9 ± 0.2

Table 7: Detection results on the VOC07 dataset when varying supervision targets, where we show maen accuracies over 3 runs.

Language Tasks. In our paper we focused attention between word categories according to English grammar. Here we consider additional ways of supervising the word-to-word relationships. In **different category supervision**, we consider a word pair (a, b) to have a valid ground truth relationship when a and b belongs to different lexical categories. In **same category supervision**, we consider a word pair (a, b) to have a valid ground truth relationship when a and b belong to the same lexical categories. In **different word supervision**, we consider a word pair (a, b) to have a valid ground truth relationship when a and b are different words. We provide a comparison of using the aforementioned supervision targets on the 20News dataset, in Table 8, with the supervision constraints becoming stricter from the leftmost column to the rightmost column.

20news varying $\mathcal{T}$	Diff Word	Same Category	Diff Category	Semantic
Accuracy (%)	64.7 ± 0.2	64.9 ± 0.2	65.1 ± 0.1	65.6 ± 0.2

Table 8: Document categorization results for the 20 Newsgroups dataset when varying the supervision target, where we show mean accuracies over 5 runs.

A.5 Convergence of Center-Mass

We provide additional results illustrating the convergence of center-mass $\mathcal{M}$ training in Table 9. The center mass is a element wise multiplication between the post softmax attention weight matrix $\tilde{\mathcal{W}}$ and the ground truth label matrix $\mathcal{T}$, defined as $\mathcal{M} = \tilde{\mathcal{W}} \odot \mathcal{T} \in [0, 1]$. More details are in Section

COCO $\mathcal{M}$	Training	Testing
un-sup Hu et al. (2018)	0.020	0.013
sup-obj	0.747	0.459

Table 9: We compare center-mass values for the FAN-minicoco network between training and testing. The values reported are evaluated on the minicoco train/test set.

3.2 of the main article. Applying the FAN-minicoco network described in Section 5 of the main article, "sup-obj" stands for focusing the attention using the ground truth label constructed following Section 3.2 in the main article, and "un-sup" stands for the unfocused case of removing the relation loss during training, which is similar to Hu et al. (2018). The converged center-mass value for the supervised case is much higher than that for the unsupervised case. Empirically this suggests that our relation loss, which is designed to increase the center-mass during learning, is effective. Furthermore, the gap between the training center-mass and the testing one is reasonable for the supervised case, i.e., we do not appear to be suffering from over-fitting. We have observed the same general trends for the other tasks as well.

B Network Training Details

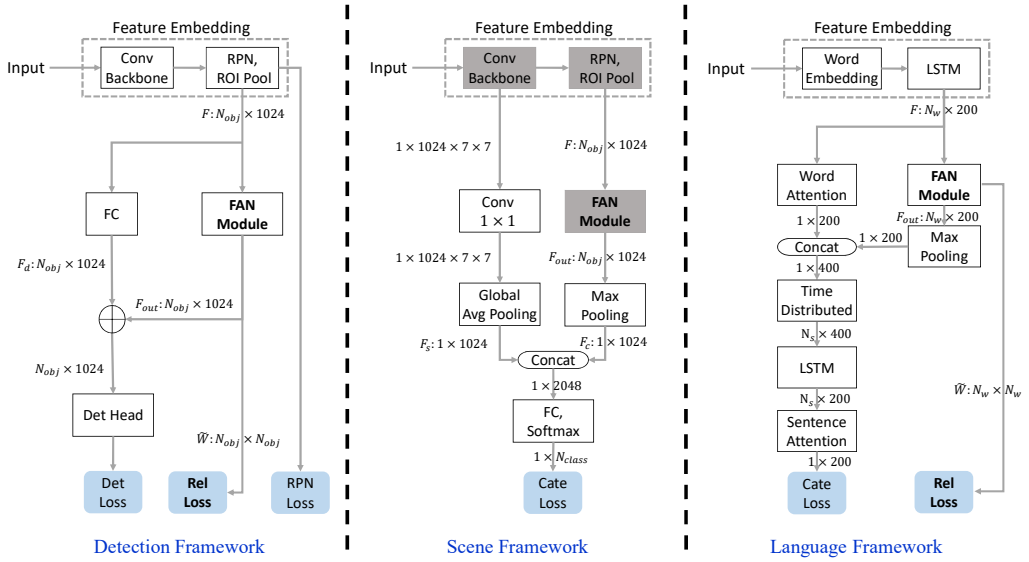


Figure 5: The input/output dimension details related to Figure 3 of the main article. The dimensions shown are for the case of a batch size of 1. Left: We add a detection branch to the backbone. Middle: We add a scene recognition branch to the backbone. Right: We insert the Focused Attention Module into a Hierarchical Attention Network.

Vision Tasks Unless stated otherwise, all the vision task networks are based on a ResNet101 (He et al., 2016) structure trained with a batch size of 2 (images), using a learning rate of $5e-4$ which is decreased to $5e-5$ after 5 epochs, with 8 epochs in total for the each training session. SGD with a momentum optimizer is applied with the momentum set as 0.9. The number of RPN proposals is fixed at $N_{obj} = 300$. Thus the attention weight matrix $\mathcal{W}$ has a dimension of 300×300 for a single image. Further details regarding input/output dimensions of the intermediate layers can be found in Figure 5 (left and middle).

Language Task For the document classification task, the network structure is based on a Hierarchical Attention Network (Yang et al., 2016). For all experiments, the batch size is set to be 256

(documents), and the word embedding dimension is set to 100. The maximum number of words in a sentence is set to be $N_w = 30$, and the maximum number of sentences in a document is set to be $N_s = 15$. Therefore, the word level Focused Attention Network's attention weight matrix $\mathcal{W}$ has a dimension of 30×30, for a single sentence. The output dimension for Bi-LSTMs is set to be 100, and the attention dimension in attention models is also set to be 100. The Adam optimizer (Kingma & Ba, 2015) is applied with an initial learning rate of $1e-3$. The network is trained end-to-end with categorization loss and relation loss for 15 epochs. Further details regarding input/output dimensions of intermediate layers can be found in Figure 5 (right).

C SCALING OF LOSS TERMS

We applied a balancing factor λ to properly weight the relation loss when training in conjunction with the main objective loss.

Detection and relation proposal We applied

$$\mathcal{L} = \mathcal{L}_{det} + \lambda \mathcal{L}_{rel}, \lambda = 0.01 \tag{6}$$

where $\mathcal{L}_{det}$ is defined in (Ren et al., 2015), which is a combination of RPN losses and detection head losses.

Scene categorization The benefit of focused attention comes from a pretrained detection model and the scene task itself does not optimize the relation loss.

Document Classification For language tasks, we applied

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda \mathcal{L}_{rel}, \lambda = 0.1 \tag{7}$$

where $\mathcal{L}_{cls}$ is a standard cross entropy loss for the classification task.

D IMPLEMENTATION DETAILS

We ran multiple trials (3-5) of our experiments and reported error bars in our results, and observed that the numbers are relatively stable and are reproducible. Furthermore, we plan to release our code upon acceptance of this article. Given that all our datasets are publicly available this will allow other researchers to both reproduce our experiments and use our Focused Attention Network module for their own research.

D.1 VISION TASKS

We implemented the center-mass cross entropy loss as well as the Focused Attention Module using MxNet. For the Faster R-CNN backbone, we adopted the source code from Relation Networks (Hu et al., 2018).

Scene Categorization. In order to maintain the learned relationship weights from the pre-trained module, which helps encode object co-occurrence context in the aggregation result, we fix the parameters in the convolution backbone, RPN layer and Focused Attention Network module, but make all other layers trainable. Fixed layers are shaded in grey in Figure 3 (bottom middle).

D.2 LANGUAGE TASKS

We implemented the Hierarchical Attention Networks according to Yang et al. (2016) in Keras with a TensorFlow backend. The word-to-word Focused Attention Network module as well as the center-mass cross entropy loss, are also implemented in the same Keras based framework.

D.3 RUNTIME AND MACHINE CONFIGURATION

All our experiments are carried out on a linux machine with 2 Titan XP GPUs, an Intel Core i9 CPU and 64GBs of RAM. The Figures and Tables referred to in the following text are those in the main article.

- **Figure 4, Relationship Proposal.** For a typical run of Visual Genome Focused Attention Network training, it takes 55 hours for 8 epochs using the above machine configuration.
- **Table 1 Object, Detection.** For a typical run of VOC07 Focused Attention Network training, it takes 4 hours when training for 8 epochs. For a typical run on minicoco , it takes 26 hours using the same setup.
- **Table 2, Scene Categorization.** For a typical run of the MIT67 dataset, it takes 2 hours when training for 8 epochs.
- **Table 3, Document Classification.** For a typical run of the 20 Newsgroup dataset, it takes 30 minutes for 15 epochs.

We also determined that when compared with unsupervised cases of the above experiments, the use of the Focused Attention Network module does not add any noticeable run time overhead.

E ADDITIONAL VISUALIZATIONS

Word Importance in a Sentence In Figure 6, we provide additional visualizations of the word importance factor in a sentence (defined in Section 4.5 of the main article), using the same format as that used in Figure 2 in the main article.

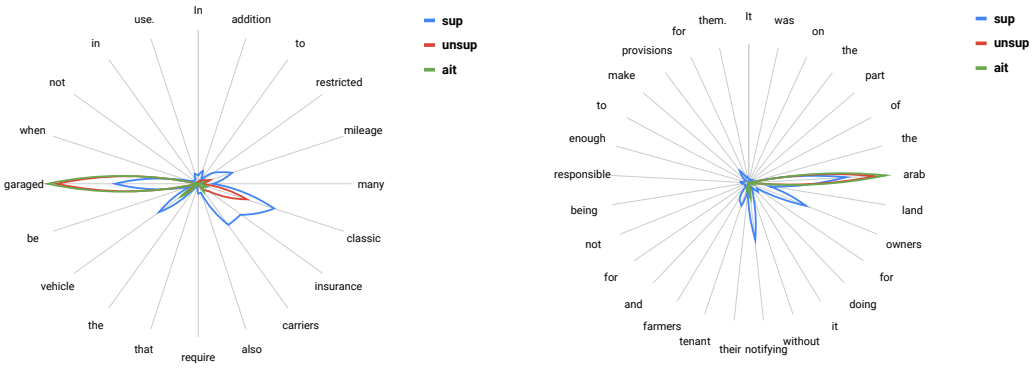


Figure 6: Additional visualization of the word importance factor in a sentence. See Section 3.2 and the caption of Figure 2 of the main article for an explanation.

Visual Relationships We now provide additional qualitative visualizations showing typical relationship weights learned by our method. In Figure 7, we visualize the predicted relationship on images from the MIT67 dataset, using a pre-trained Focused Attention Network on the minicoco dataset, referred to as FAN-minicoco, as discussed in Section 5.1 of the main article. We compare this with the corresponding unsupervised case, which is similar to Relation Networks (Hu et al., 2018).

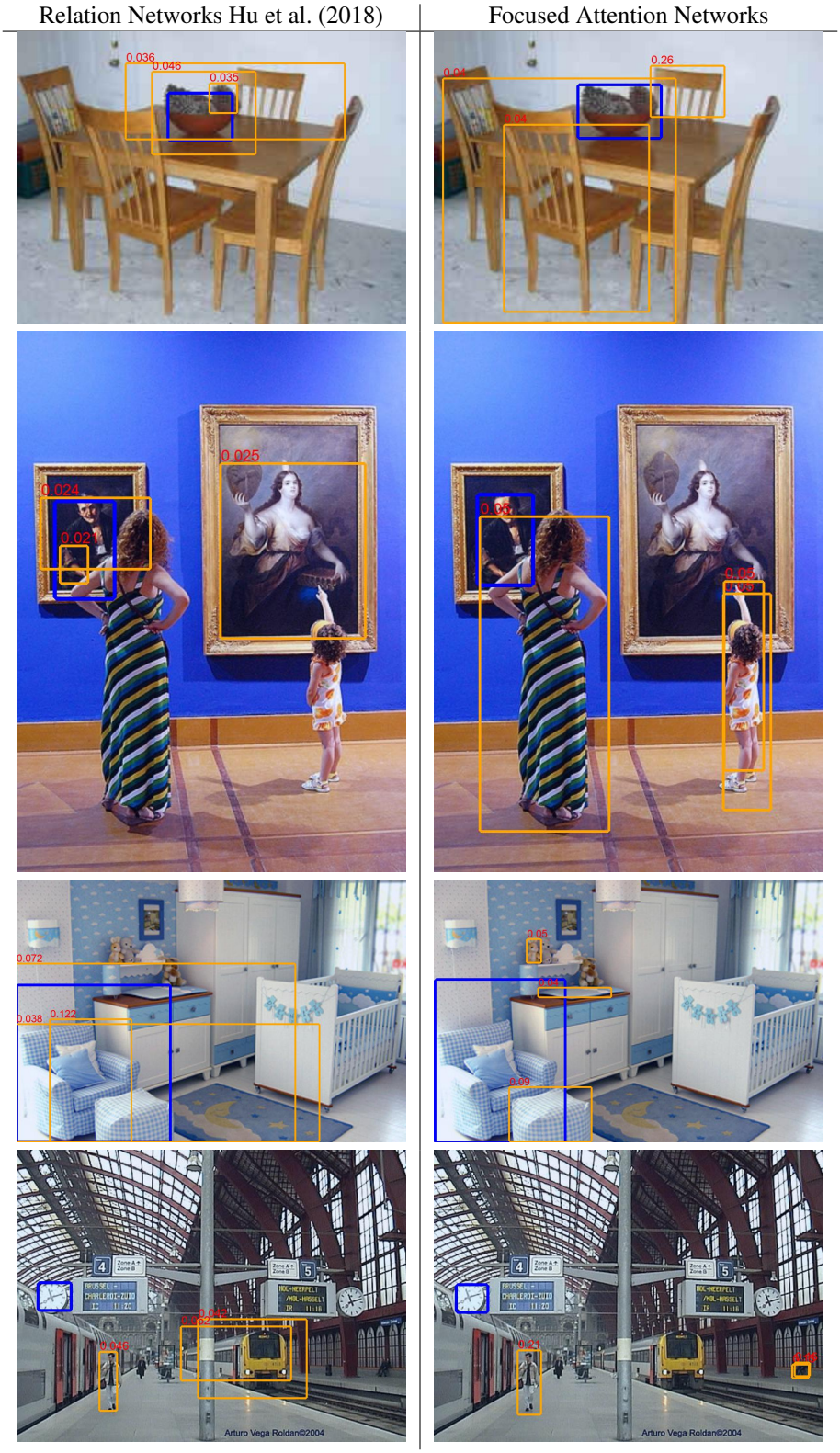


Figure 7: The visualization of relationships recovered on additional images of the MIT67 dataset. See the caption of Figure 1 of the main article for an explanation.