Deep Attentional Structured Representation Learning for Visual Recognition

Krishna Kanth Nakka
krishna.nakka@epfl.ch

Mathieu Salzmann
mathieu.salzmann@epfl.ch

CVLab, EPFL
Switzerland

Abstract

Structured representations, such as Bags of Words, VLAD and Fisher Vectors, have proven highly effective to tackle complex visual recognition tasks. As such, they have recently been incorporated into deep architectures. However, while effective, the resulting deep structured representation learning strategies typically aggregate local features from the entire image, ignoring the fact that, in complex recognition tasks, some regions provide much more discriminative information than others.

In this paper, we introduce an attentional structured representation learning framework that incorporates an image-specific attention mechanism within the aggregation process. Our framework learns to predict jointly the image class label and an attention maps in an end-to-end fashion and without any other supervision than the target label. As evidenced by our experiments, this consistently outperforms attention-less structured representation learning and yields state-of-the-art results on standard scene recognition and fine-grained categorization benchmarks.

1 Introduction

In recent years, Convolutional Neural Networks (CNNs) have emerged as the de facto standard for visual recognition. Nevertheless, while they achieve tremendous success at classifying images containing iconic objects, their performance on more complex tasks, such as scene recognition and fine-grained categorization, remains comparatively underwhelming. This is partly due to their simple pooling schemes that fail to model the dependencies between local image regions. By contrast, in the realm of handcrafted features, structured representations, such as Bags of Words (BoW) [17, 53, 40], Vectors of Locally Aggregated Descriptors (VLAD) [1, 14] and Fisher Vectors (FV) [31, 35], have been shown to be highly discriminative thanks to their aggregation of local information. As a consequence, they have started to re-emerge in the deep networks realm, with architectures such as NetVLAD [2] and Deep FisherNet [41].

While effective for complex visual recognition tasks, these structured representations, whether based on handcrafted features or incorporated into deep networks, suffer from one drawback: They aggregate local information from the entire image, regardless of how relevant this information is to the recognition task. In practice, however, while certain image regions contain semantic information that contribute to the target label, others clearly don't.
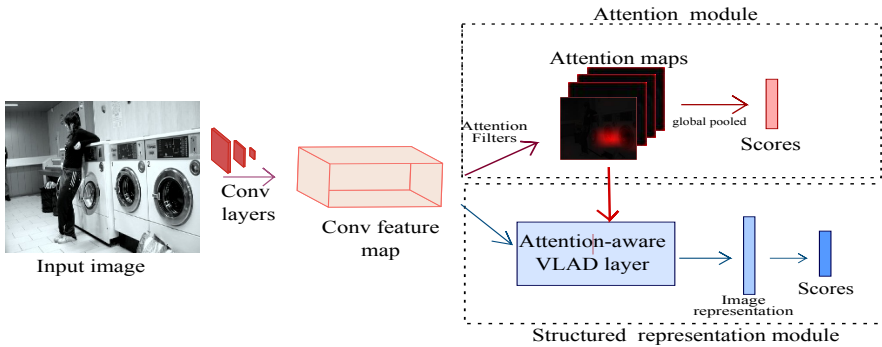
Figure 1: **Attentional structured representation network.** Our network consists of two branches with a shared base feature extraction CNN. The attention module produces class-specific attention maps, which are then incorporated into the VLAD module that outputs an attention-aware VLAD representation. Note that, while we focus on the VLAD case here, as evidenced by our experiments, our approach applies to any structured representation.

For example, in the image shown in Fig. 1, from the MIT-Indoor dataset [32], the region depicting washing machines gives us a much stronger cue of the class *laundry* than the regions containing the person and the background. Incorporating information from these latter two regions, which can appear in many other scene categories, will typically yield less discriminative image representations.

In this paper, we address this by introducing a novel deep attentional structured representation network for visual recognition. Our network incorporates an image-specific attention mechanism that encourages the learnt structured representation to focus on the discriminative regions of the image. We then learn to predict jointly the input image class label and the spatial attention map without requiring any annotations for the latter.

Our framework is depicted by Fig. 1 for the case of a VLAD aggregation strategy. It consists of two streams that share a base network extracting deep features: The attention module and the VLAD module. The attention module, based on the framework of [11], learns a set of filters that transform the deep features into C heatmaps encoding attention for the C classes of interest. The VLAD module then exploits these heatmaps to form an attention-aware VLAD vector from the deep features of the base network. We train our network with a combination of two losses that encourage both the attention maps and the final attention-aware VLAD representation to be discriminative. Note that, while Fig. 1 focuses on the VLAD case, as evidenced by our experiments, our approach generalizes to any local feature aggregation strategy.

In short, we contribute the first systematic integration of an attention mechanism within a structured image representation learning framework. We demonstrate the benefits of our approach on four challenging visual recognition tasks, including scene recognition on MIT-Indoor [32] and fine-grained categorization on the UCSD Birds [42], FGVC Aircrafts [27] and Stanford Cars [18] datasets. Our attentional structured representation learning strategy consistently outperforms its standard attention-less counterpart and yields state-of-the-art results on several of the above-mentioned datasets.

2 Related Work

Over the years, visual recognition has attracted a huge amount of attention in Computer Vision. Before the deep revolution in 2012, most methods adopted a two step pipeline consisting of extracting handcrafted features and training a classifier, such as Support Vector Machines [13] or Boosting [8]. In this pipeline, the core Computer Vision research was targeted towards extracting discriminative image features. In particular, Bags of Visual Words (BoW) [17, 33, 40], based on local features such as SIFT [26] or BRIEF [4], have proven effective for image recognition. Later, such histogram-based features were extended to VLAD [14] and Fisher Vectors [31, 35], which model higher-order statistics of the data w.r.t. the codewords. After the remarkable performance of AlexNet [20], much of the visual recognition research turned to deep learning strategies. While many new architectures do not explicitly focus on extracting structured representations, some work has nonetheless attempted to leverage the lessons learnt from handcrafted features. In particular, [12] performs multi-scale orderless pooling of deep CNN features, and [5, 7] compute Fisher encodings of similar deep features. In contrast with these approaches that still separate feature extraction from classifier learning, NetVLAD [2] and Deep FisherNet [41] constitute the first attempts at introducing learnable VLAD and Fisher Vector layers, respectively, within an end-to-end learning formalism. More recently, [22] proposed to make use of a mixture of factor analyzers to model an accurate Fisher Vector with full covariance matrix. While the previous methods all rely on histogram-based descriptors, in the context of fine-grained categorization and texture recognition, other structured representations, in the form of covariance matrices have been used [23, 24, 48]. In [37], a generalization of average and bilinear pooling was proposed to automatically learn an intermediate pooling strategy during training. In any event, all these methods, whether using hand-crafted features or relying on deep learning, aggregate local features from the entire image, without accounting for the fact that only parts of the image contain information that contributes to the target class label. This will typically reduce the discriminative power of the resulting representations.

For complex tasks, such as scene classification and fine-grained categorization, some research has nonetheless attempted to focus the feature extraction process on discriminative image regions. In the context of scene recognition, this was achieved by modeling the scene with mid- (or high-)level representations [21], such as detected semantic visual attributes [30], patch-based codewords obtained via discriminative clustering [39] and object-oriented representations learnt from a manually-created database of typical scene objects [46]. For fine-grained categorization, several works exploit bounding box annotations to learn part detectors [49, 51]. The use of such additional annotations was then removed in [47], which learns part templates by clustering deep features. More recently, [9, 52] introduced end-to-end learning strategies to automatically identify discriminative regions for fine-grained recognition.

The above-mentioned works typically reason about the notion of parts, or objects in a scene. In the rare cases that don't require part annotations during training [9, 52], the input image is first processed globally to identify regions with high attention, which are then cropped into multiple parts that are processed individually. By contrast, our network processes the input image in a single forward pass, without explicitly relying on the notion of parts. In essence, these methods are therefore tailored to the specific problem they tackle. By contrast, here, we exploit the more general notion of visual attention and produce heatmaps encompassing the discriminative regions in the image. This does therefore not require any prior knowledge about the data at hand. Our formalism builds upon the attention framework

of [11], but with the additional goal of leveraging structured representations. As a consequence, and as evidenced by our results, our approach yields higher accuracies than both attention-less methods and unstructured attentional pooling in all the tasks we tested it on.

3   Method

In this section, we introduce our novel attentional structured representation learning framework depicted by Fig. 1. We first present the structured representation and attention modules, and finally our approach to integrating them in an end-to-end learning formalism.

3.1   Structured Representation Module

Structured representations aggregate local descriptors into a global feature vector of fixed size using a visual codebook. In particular, here, we focus on VLAD, which has proven highly effective. As will be evidenced by our experiments, however, our framework generalizes to other aggregation strategies.

In contrast to BoW that only store information about which codeword each local descriptor is assigned to, VLAD also computes the residual distance of the descriptor to the codeword. To incorporate this into a deep learning framework, the hard codeword assignment of each descriptor is replaced by a soft one. More specifically, let $\mathbf{I}$ be an image input to a CNN, and $\mathbf{X} \in \mathbb{R}^{W \times H \times D}$ the feature map output by the last convolutional layer, with spatial resolution $W \times H$ and D channels. $\mathbf{X}$ can then be thought of as $N = W \times H$ local descriptors $\mathbf{x}_i$ of dimension D. Given a codebook $\mathbf{B}$ with K codewords, VLAD produces a DK-dimensional representation of the form

$$\mathbf{v} = [\mathbf{v}_0^T, \mathbf{v}_1^T, \cdots, \mathbf{v}_K^T]^T \ , \tag{1}$$

where $\mathbf{v}_k \in \mathbb{R}^D$ is given by

$$\mathbf{v}_k = \sum_{i=1}^{N} a_k(\mathbf{x}_i)(\mathbf{x}_i - \mathbf{b}_k) \ , \tag{2}$$

with $\mathbf{b}_k$ the k-th codeword of codebook $\mathbf{B}$. The values $a_k(\mathbf{x}_i)$ represent the assignment of descriptor $\mathbf{x}_i$ to codeword $\mathbf{b}_k$. In the standard VLAD formalism, these assignments are binary, with each descriptor being assigned to a single codeword. Within a deep learning context, for differentiability, these assignments can be relaxed and expressed as

$$a_k(\mathbf{x}_i) = \frac{e^{-\alpha \|\mathbf{x}_i - \mathbf{b}_k\|^2}}{\sum_{k'} e^{-\alpha \|\mathbf{x}_i - \mathbf{b}_{k'}\|^2}} \ , \tag{3}$$

with α a hyperparameter defining the softness of the assignments.

The resulting VLAD vector then acts as input to the classification layer of the deep network. While effective, as discussed above, the VLAD representation aggregates information from the entire image, regardless of whether the local descriptors correspond to discriminative regions or not. Below, we first discuss a general attention module, which is able to identify relevant image regions, and then introduce our approach to incorporating this information within our structured representation learning formalism.

3.2 Attention Module

It has been shown multiple times that CNNs were not only effective at predicting the class label of an image, but could also localize the image regions relevant to this label [15, 36, 53]. Most existing approaches to performing such a localization, however, work as a post-training step. By contrast, our attention module, based on the framework of [11], produces attention maps that are actively used during training. Furthermore, it combines top-down attention, modeling class-specific information, with bottom-up attention, modeling class-agnostic information, or, in other words, a form of image saliency. Such integration [29] of top-down cues [3, 50, 54] with bottom-up attention [54] modulates the image saliency map to ignore non-relevant background regions of the target.

Specifically, let $\mathbf{X}$ be the same final $W \times H \times D$ convolutional feature map as in Section 3.1. Our attention module consists of an additional 1×1 convolutional layer with one class-agnostic filter with parameters $\mathbf{w}_{ca} \in \mathbb{R}^{D \times 1}$ and C class-specific filters whose parameters can be grouped in a matrix $\mathbf{W}_{cs} \in \mathbb{R}^{D \times C}$, where C is the number of classes of the problem at hand. This convolutional layer produces a class-agnostic heatmap $\mathbf{H}_{ca}$ and class-specific heatmaps $(\mathbf{H}_{cs}^1, \cdots, \mathbf{H}_{cs}^C)$, each of spatial resolution $W \times H$. Each class-specific heatmap is then multiplied element-wise by the class-agnostic one, yielding C attention maps $(\mathbf{H}^1, \cdots, \mathbf{H}^C)$.

Training the attention module can be achieved by global average pooling of each of the attention maps, which produces a score for each class. These scores are then passed through a softmax layer, and the resulting probabilities $\{p_c\}$ used in a standard cross-entropy loss

$$L_{att} = -\frac{1}{S} \sum_{s=1}^{S} \log(p_{c^*}(\mathbf{I}_s)) \,, \tag{4}$$

where S is the number of samples in a mini-batch and $p_{c^*}(\mathbf{I}_s)$ is the probability of the ground-truth class for sample s. This was the procedure used in [11] to train an attentional deep network. Below, we propose to rather make use of the attention maps to further build a more discriminative structured representation. As evidenced by our results, this allows us to achieve consistently higher recognition accuracies.

3.3 Attention-aware Feature Aggregation

Our goal is to make use of the attention maps when aggregating the local descriptors into a structured representation. To this end, instead of global average pooling the maps, we generate a single attention map, which can be interpreted as a weight $w(\mathbf{x}_i)$ for every descriptor $\mathbf{x}_i$, and is defined as

$$w(\mathbf{x}_i) = \frac{\max_l \mathbf{H}_i^l}{\sum_{i'} \max_l \mathbf{H}_{i'}^l} \,, \tag{5}$$

where $\mathbf{H}_i^l$ indicates the attention-weight corresponding to feature $\mathbf{x}_i$ in the attention map of class l from Section 3.2. The resulting attention map has the same spatial resolution as the final deep feature map. We then use it to re-weight the aggregation scheme of Eq. 2. Specifically, we re-write Eq. 2 as

$$\mathbf{v}_k = \sum_{i=1}^{N} w(\mathbf{x}_i) a_k(\mathbf{x}_i) (\mathbf{x}_i - \mathbf{b}_k) \,. \tag{6}$$

Following common practice [2, 14], we perform L_2 normalization of each $\mathbf{v}_k$ to remove burstiness, followed by a final L_2 normalization of the entire vector $\mathbf{v}$. The resulting representation is then passed to a classification layer.

Ultimately, our network combines an attention module with a structured representation learning one. Both modules share the base network up to the final convolutional feature map. To train our network, we first pre-train the base network with the attention module only using L_{att} from Eq. 4. We then continue training the entire network in an end-to-end manner by minimizing a loss of the form

$$L = L_{cls} + \lambda L_{att} , \qquad (7)$$

where L_{cls} is a cross-entropy loss on the output of the classifier acting on the structured representation $\mathbf{v}$, and λ is a hyper-parameter setting the relative influence of both terms. At test time, we then take the prediction from the VLAD branch of the network.

Note that training is not only done w.r.t. the network parameters, but also w.r.t. to the codebook $\mathbf{B}$. As suggested in [2], and motivated by [1] to adapt VLAD descriptors to new datasets, we decouple the soft assignment $a_k(\mathbf{x}_i)$ from the codeword $\mathbf{b}_k$. That is, we re-write the assignment $a_k(\mathbf{x}_i)$ of Eq. (3) as

$$a_k(\mathbf{x}_i) = \frac{e^{\mathbf{s}_k^T \mathbf{x}_i + h_k}}{\sum_{k'} e^{\mathbf{s}_{k'}^T \mathbf{x}_i + h_{k'}}} , \qquad (8)$$

where $h_k = -\alpha \|\mathbf{b}_k\|^2$ and $s_k = 2\alpha \mathbf{b}_k$ are treated as independent parameters.

3.3.1 Geometric Interpretation of Attention

Consider the features of two images from same class with different backgrounds that are assigned to the same codeword, depicted as a Voronoi cell in Fig. 2. The features with high attention are shown in blue and those with low attention in red and orange, respectively. While ignoring attention would yield residual vectors pointing in almost opposite directions, our attention-aware aggregation produces vectors with high cosine similarity, shown as blue arrows. The inverse reasoning can be made for images from two different classes but containing common elements that are irrelevant to the class labels: By ignoring attention, these shared elements would yield components with high cosine similarity, thus decreasing the discriminative power of the complete VLAD vector. Attention allows us to discard these shared elements.
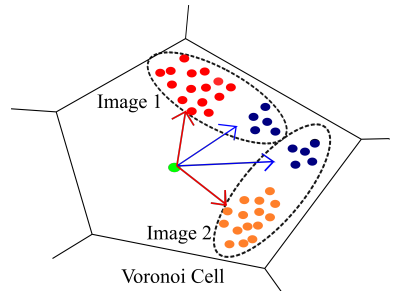

Figure 2: Geometric interpretation of attention

4   Experiments

We first present the datasets used in our experiments and implementation details for our model. Then, we demonstrate the benefits of our attention-aware structured representation learning framework over its attention-less counterpart and over unstructured attentional pooling, and finally compare our results to the state of the art on each dataset. We provide additional results and ablation studies in the supplementary material.

4.1 Datasets

We experiment on the MIT-Indoor scene recognition dataset and on three fine-grained categorization datasets, namely CUB-200, Stanford cars and aircraft. We discard the part annotations but conduct experiments with and without bounding box annotations on fine-grained datasets.

MIT-Indoor is a widely used benchmark dataset for scene classification with 67 classes. We use the train/test split of [32] consisting of roughly 80 training and 20 test images.

CUB-200 is a challenging dataset with $11,788$ images of 200 bird species, with an average of 60 images per class. The dataset has extremely large variations in pose, size and viewpoints. We use the standard train/test split of [42].

FGVC-Aircraft contains 100 different aircraft models with roughly 100 images for each model. We adopt the same train/test split as in [27].

Stanford Cars is a 196 class dataset [18] of 8144 training images and 8041 test images. Heavy background clutter makes this dataset challenging.

4.2 Implementation Details

We use the VGG-16 [38] model pre-trained on Imagenet [20] as our base model and that of the baselines, with the *conv5_3* features before *ReLU* activation as final convolutional features for aggregation. Following prior work [22, 24], we resize the images to 512×512 for MIT-Indoor, and 448×448 for the fine-grained datasets. Data augmentation is carried out on all datasets by performing random cropping and horizontal flipping. At test time, we flip the image and average the predictions for the original and flipped image. For structured representations, we fix the codebook size to $K = 64$ for VLAD and $K = 4096$ for the BoW experiments. We initialize the weights of the VLAD layer with K-means clustering of the *conv5_3* features. We set α in Eq. 3 to 100, and λ in Eq. 7 to 0.4.

Training: We use the ADAM optimizer [16] with parameter $\varepsilon = 10^{-4}$, batch size of 16 and a weight decay of 0.0005 for all experiments. We first pre-train the attention network with $\eta = 0.0001$ for 20 epochs. For scene recognition, we then train the classification layer with $\eta = 0.01$ for 5 epochs, and further train the layers above *conv5* with $\eta = 0.00001$ for 25 epochs. For the fine-grained datasets, we train with $\eta = 0.01$ for the classification layer and $\eta = 0.0001$ for the layers above *conv5* for 50 epochs, with a decay rate of 0.1 every 15 epochs.

4.3 Results

We first compare our approach to attention-less structured representation learning and to direct attentional pooling [11], and then to the state of the art on each dataset. To be consistent with prior work [22, 24], we report the average accuracy per class on MIT-Indoor and the average per image accuracy on the fine-grained datasets.

To evaluate the benefits of our attention-aware feature aggregation framework, we compare it with counterparts that do not rely on attention. In particular, we report results with VLAD pooling, as discussed in Section 3, but also with BoW representations, which can easily be obtained by using the soft assignments to form histograms. To further evidence the benefits of using structured representations, we compare our results with those of the direct attentional pooling strategy of [11], which relies on a global average pooling of the attention masks. The results of this comparison for all datasets are reported in Table 1,

Pooling	Anno.	Birds	Cars	Aircrafts	MIT-Indoor
VGG-16	BBox	79.9	88.4	86.9	-
Attention	BBox	77.2	90.3	85.0	-
NetBoW	BBox	74.4	89.1	85.6	-
Attentional-NetBoW	BBox	80.5	91.2	89.3	-
NetVLAD	BBox	82.4	89.8	88.0	-
Attentional-NetVLAD	BBox	**85.5**	**93.5**	**89.2**	-
VGG-16		76.0	82.8	82.3	76.6
Attention		77.0	87.4	81.4	77.2
NetBoW		68.9	85.2	79.9	76.1
Attentional-NetBoW		76.9	90.6	88.3	76.6
NetVLAD		80.6	89.4	86.4	79.2
Attentional-NetVLAD		**84.3**	**92.8**	**88.8**	**81.2**

Table 1: Comparison of our attentional structured pooling scheme with attention-less (VGG-16, NetBoW, NetVLAD) and structure-less (Attention) baselines. Our approach consistently outperforms these baselines, thus showing the benefits of pooling only the relevant local features into a structured representation.

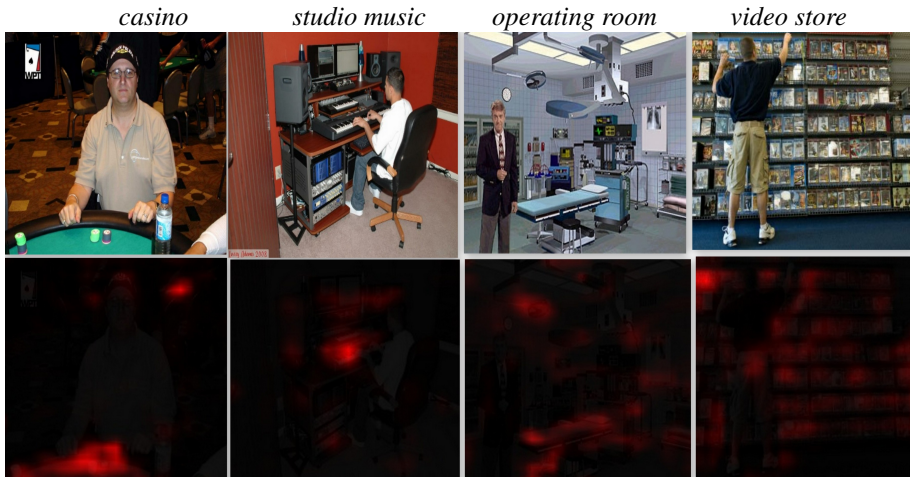| *casino* | *studio music* | *operating room* | *video store* |

Figure 3: **Attention maps for MIT-Indoor.** Each column shows an image from a different class (indicated above the image). Note that the maps focus on regions indicative of the label, ignoring the regions common to multiple classes, such as the people.

where we also show the accuracy of the standard VGG-16 model, with fully connected layers transformed into convolutional ones followed by global average pooling. Note that our Attentional-NetVLAD outperforms the baselines in all cases, both when using and not using bounding boxes for fine-grained recognition. Note also that using attention consistently helps improving the results, thus showing the importance of reasoning at the level of local features rather than combining information from the entire image in these challenging recognition tasks.

In Figs. 3 and 4, we provide some representative qualitative results of the attention maps obtained with our method for MIT-Indoor and the fine-grained datasets, respectively. For
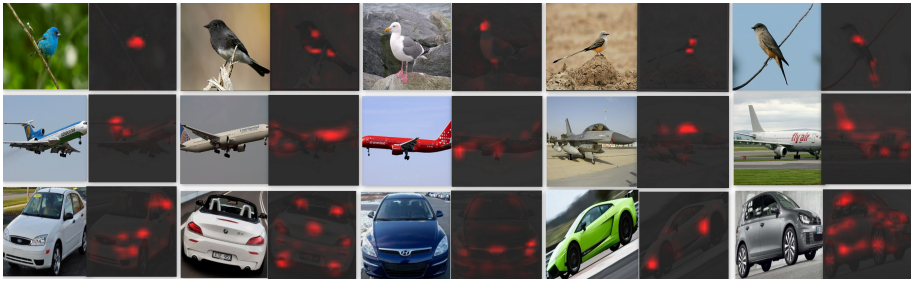
Figure 4: **Attention maps for fine-grained datasets**. Our method is able to localize discriminative parts of birds (tail, beak), aircrafts (engine, landing gear) and cars (lights, logo).

Method	Avg. Acc.
Deep FisherNet [41]	76.5
CBN [10]	77.6
NetVLAD [0]	79.1
H-Sparse [25]	79.5
B-CNN [24]	79.5
SMSO [43]	79.7
FV+FC [6]	81.0
MFAFVNet [22]	81.1
Ours	**81.2**

Table 2: Comparison with the state of the art on MIT-Indoor.

Method	Anno.	Birds	Cars	Aircraft
MG-CNN [43]	BBox	83.0	-	86.6
B-CNN [24]	BBox	85.1	-	-
PA-CNN [19]	BBox	82.8	92.8	-
Mask-CNN [45]	Parts	85.4	-	-
MDTP [44]	BBox	-	92.6	88.4
Ours	BBox	**85.5**	**93.5**	**89.2**
KP [6]		86.2	92.4	86.9
Boost-CNN [28]		86.2	92.1	88.5
B-CNN [24]		84.1	86.9	86.6
Imp. B-CNN [23]		85.8	92.0	88.5
α-pooling [37]		85.3	-	85.5
RA-CNN [9]		84.1	92.5	88.2
MA-CNN [52]		**86.5**	92.8	**89.9**
Ours		84.3	**92.8**	88.8

Table 3: Comparison with the state of the art on fine-grained datasets.

scene recognition, note that our network learnt to focus on the discriminative regions, such as the casino table and the piano, while ignoring regions shared by other classes, such as people. Similarly, for fine-grained categorization, the network is able to locate discriminative parts, such as the beak and the tail of birds, the brand logo and the head lights of cars, and the engine and landing gears of airplanes. This clearly evidences that our model can, in a single pass, find the regions of interest that define a class.

Finally, we compare our results with the state of the art on each individual dataset. These comparisons are provided in Table 2 for MIT-Indoor and Table 3 for the fine-grained datasets. In the case of scene recognition, we outperform all the baselines, including MFAFVNet [22], which relies on an accurate Fisher Vector encoding of 500K dimensions based on multi-scale image patches. For fine-grained recognition, we outperform all the baselines when relying on bounding boxes. Without bounding boxes, we achieve accuracies only slightly lower than the state-of-the-art methods, such as [9, 52], which were tailored to the fine-grained categorization problem, and rely on a multi-stage approach involving cropping parts and processing them separately. By contrast, our approach makes use of a single forward pass through a network and generalizes to any complex recognition scenario.

5   Conclusion

We have introduced an attention-aware structured representation network for complex visual recognition tasks. Our network jointly identifies the informative image regions and learns a structured representation. Our comprehensive experiments on scene recognition and fine-grained categorization have demonstrated the superiority of our approach over attention-less strategies. Our approach is general and can be extended to other feature aggregation techniques, or can make use of any generic attention module. This will be the focus of our future work.

Acnkowledgements

References

[1] Relja Arandjelovic and Andrew Zisserman. All about vlad. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 1578–1585. IEEE, 2013.

[2] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5297–5307, 2016.

[3] Farhan Baluch and Laurent Itti. Mechanisms of top-down attention. *Trends in neurosciences*, 34(4):210–224, 2011.

[4] Michael Calonder, Vincent Lepetit, Mustafa Ozuysal, Tomasz Trzcinski, Christoph Strecha, and Pascal Fua. Brief: Computing a local binary descriptor very fast. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(7):1281–1298, 2012.

[5] Mircea Cimpoi, Subhransu Maji, and Andrea Vedaldi. Deep filter banks for texture recognition and segmentation. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 3828–3836. IEEE, 2015.

[6] Yin Cui, Feng Zhou, Jiang Wang, Xiao Liu, Yuanqing Lin, and Serge Belongie. Kernel pooling for convolutional neural networks. In *Computer Vision and Pattern Recognition (CVPR)*, 2017.

[7] Mandar Dixit, Si Chen, Dashan Gao, Nikhil Rasiwasia, and Nuno Vasconcelos. Scene classification with semantic fisher vectors. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 2974–2983. IEEE, 2015.

[8] Yoav Freund, Robert E Schapire, et al. Experiments with a new boosting algorithm. In *Icml*, volume 96, pages 148–156. Bari, Italy, 1996.

[9] Jianlong Fu, Heliang Zheng, and Tao Mei. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In *Conf. on Computer Vision and Pattern Recognition*, 2017.

[10] Yang Gao, Oscar Beijbom, Ning Zhang, and Trevor Darrell. Compact bilinear pooling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 317–326, 2016.

[11] Rohit Girdhar and Deva Ramanan. Attentional pooling for action recognition. In *Advances in Neural Information Processing Systems*, pages 33–44, 2017.

[12] Yunchao Gong, Liwei Wang, Ruiqi Guo, and Svetlana Lazebnik. Multi-scale orderless pooling of deep convolutional activation features. In *European conference on computer vision*, pages 392–407. Springer, 2014.

[13] Marti A. Hearst, Susan T Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4):18–28, 1998.

[14] Hervé Jégou, Matthijs Douze, Cordelia Schmid, and Patrick Pérez. Aggregating local descriptors into a compact image representation. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3304–3311. IEEE, 2010.

[15] Albert Jimenez, Jose M Alvarez, and Xavier Giro-i Nieto. Class-weighted convolutional features for visual instance search. *In BMVC*, 2017.

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

[17] Jan J Koenderink and Andrea J Van Doorn. The structure of locally orderless images. *International Journal of Computer Vision*, 31(2-3):159–168, 1999.

[18] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Computer Vision Workshops (ICCVW), 2013 IEEE International Conference on*, pages 554–561. IEEE, 2013.

[19] Jonathan Krause, Hailin Jin, Jianchao Yang, and Li Fei-Fei. Fine-grained recognition without part annotations. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 5546–5555. IEEE, 2015.

[20] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[21] Li-Jia Li, Hao Su, Li Fei-Fei, and Eric P Xing. Object bank: A high-level image representation for scene classification & semantic feature sparsification. In *Advances in neural information processing systems*, pages 1378–1386, 2010.

[22] Yunsheng Li, Mandar Dixit, and Nuno Vasconcelos. Deep scene image classification with the mfafvnet. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5746–5754, 2017.

[23] Tsung-Yu Lin and Subhransu Maji. Improved bilinear pooling with cnns. *In BMVC*, 2017.

[24] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. Bilinear cnn models for fine-grained visual recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1449–1457, 2015.

[25] Lingqiao Liu, Peng Wang, Chunhua Shen, Lei Wang, Anton Van Den Hengel, Chao Wang, and Heng Tao Shen. Compositional model based fisher vector coding for image classification. *IEEE transactions on pattern analysis and machine intelligence*, 39(12): 2335–2348, 2017.

[26] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

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

[28] Mohammad Moghimi, Serge J Belongie, Mohammad J Saberian, Jian Yang, Nuno Vasconcelos, and Li-Jia Li. Boosted convolutional neural networks. In *BMVC*, 2016.

[29] Vidhya Navalpakkam and Laurent Itti. An integrated model of top-down and bottom-up attention for optimizing detection speed. In *null*, pages 2049–2056. ieee, 2006.

[30] Genevieve Patterson and James Hays. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2751–2758. IEEE, 2012.

[31] Florent Perronnin, Jorge Sánchez, and Thomas Mensink. Improving the fisher kernel for large-scale image classification. In *European conference on computer vision*, pages 143–156. Springer, 2010.

[32] Ariadna Quattoni and Antonio Torralba. Recognizing indoor scenes. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 413–420. IEEE, 2009.

[33] Pedro Quelhas, Florent Monay, J-M Odobez, Daniel Gatica-Perez, Tinne Tuytelaars, and Luc Van Gool. Modeling scenes with local descriptors and latent aspects. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 1, pages 883–890. IEEE, 2005.

[34] Ueli Rutishauser, Dirk Walther, Christof Koch, and Pietro Perona. Is bottom-up attention useful for object recognition? In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 2, pages II–II. IEEE, 2004.

[35] Jorge Sánchez, Florent Perronnin, Thomas Mensink, and Jakob Verbeek. Image classification with the fisher vector: Theory and practice. *International journal of computer vision*, 105(3):222–245, 2013.

[36] Ramprasaath R Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Why did you say that? *arXiv preprint arXiv:1611.07450*, 2016.

[37] Marcel Simon, Yang Gao, Trevor Darrell, Joachim Denzler, and Erik Rodner. Generalized orderless pooling performs implicit salient matching.

[38] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *In ICLR, pages 1409-1556*, 2015.

[39] Saurabh Singh, Abhinav Gupta, and Alexei A Efros. Unsupervised discovery of mid-level discriminative patches. In *Computer Vision–ECCV 2012*, pages 73–86. Springer, 2012.

[40] Josef Sivic and Andrew Zisserman. Video google: A text retrieval approach to object matching in videos. In *null*, page 1470. IEEE, 2003.

[41] Peng Tang, Xinggang Wang, Baoguang Shi, Xiang Bai, Wenyu Liu, and Zhuowen Tu. Deep fishernet for object classification. *arXiv preprint arXiv:1608.00182*, 2016.

[42] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[43] Dequan Wang, Zhiqiang Shen, Jie Shao, Wei Zhang, Xiangyang Xue, and Zheng Zhang. Multiple granularity descriptors for fine-grained categorization. In *Computer Vision (ICCV), 2015 IEEE International Conference on*, pages 2399–2406. IEEE, 2015.

[44] Yaming Wang, Jonghyun Choi, Vlad I Morariu, and Larry S Davis. Mining discriminative triplets of patches for fine-grained classification. *In CVPR, pages 1163Ű1172v*, 2016.

[45] Xiu-Shen Wei, Chen-Wei Xie, and Jianxin Wu. Mask-cnn: Localizing parts and selecting descriptors for fine-grained image recognition. *CoRR, abs/1605.06878*, 2016.

[46] Ruobing Wu, Baoyuan Wang, Wenping Wang, and Yizhou Yu. Harvesting discriminative meta objects with deep cnn features for scene classification. In *Computer Vision (ICCV), 2015 IEEE International Conference on*, pages 1287–1295. IEEE, 2015.

[47] Tianjun Xiao, Yichong Xu, Kuiyuan Yang, Jiaxing Zhang, Yuxin Peng, and Zheng Zhang. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 842–850. IEEE, 2015.

[48] Kaicheng Yu and Mathieu Salzmann. Statistically motivated second order pooling. *arXiv preprint arXiv:1801.07492*, 2018.

[49] Han Zhang, Tao Xu, Mohamed Elhoseiny, Xiaolei Huang, Shaoting Zhang, Ahmed Elgammal, and Dimitris Metaxas. Spda-cnn: Unifying semantic part detection and abstraction for fine-grained recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1143–1152, 2016.

[50] Jianming Zhang, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. In *European Conference on Computer Vision*, pages 543–559. Springer, 2016.

[51] Ning Zhang, Jeff Donahue, Ross Girshick, and Trevor Darrell. Part-based r-cnns for fine-grained category detection. In *European conference on computer vision*, pages 834–849. Springer, 2014.

[52] Heliang Zheng, Jianlong Fu, Tao Mei, and Jiebo Luo. Learning multi-attention convolutional neural network for fine-grained image recognition. In *Int. Conf. on Computer Vision*, 2017.

[53] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*, pages 2921–2929. IEEE, 2016.

[54] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2921–2929, 2016.