Paying More Attention to Attention: Improving the Performance of Convolutional Neural Networks via Attention Transfer

Sergey Zagoruyko, Nikos Komodakis
Université Paris-Est, École des Ponts ParisTech
Paris, France
{sergey.zagoruyko,nikos.komodakis}@enpc.fr

Abstract

Attention plays a critical role in human visual experience. Furthermore, it has recently been demonstrated that attention can also play an important role in the context of applying artificial neural networks to a variety of tasks from fields such as computer vision and NLP. In this work we show that, by properly defining attention for convolutional neural networks, we can actually use this type of information in order to significantly improve the performance of a student CNN network by forcing it to mimic the attention maps of a powerful teacher network. To that end, we propose several novel methods of transferring attention, showing consistent improvement across a variety of datasets and convolutional neural network architectures.

1 Introduction

As humans, we need to pay attention in order to be able to adequately perceive our surroundings. Attention is therefore a key aspect of our visual experience, and closely relates to perception - we need to keep attention to build a visual representation, possessing detail and coherence.

As artificial neural networks became more popular in fields such as computer vision and natural language processing in the recent years, artificial attention mechanisms started to be developed as well. Artificial attention lets a system "attend" to an object to examine it with greater detail. It has also become a research tool for understanding mechanisms behind neural networks, similar to attention used in psychology.

One of the popular hypothesis there is that there are non-attentional and attentional perception processes. Non-attentional processes help to observe a scene in general and gather high-level information, which, when associated with other thinking processes, helps us to control the attention processes and navigate to a certain part of the scene. This implies that different observers with different knowledge, different goals, and therefore different attentional strategies can literally see the same scene differently. This brings us to the main topic of this paper: how attention differs within artificial vision systems, and can we use attention information in order to improve the performance of convolutional neural networks ? More specifically, can a teacher network improve the performance of another student network by providing to it information about where it looks, i.e., about where it concentrates its attention into ?

To study these questions, one first needs to properly specify how attention is defined w.r.t. a given convolutional neural network. To that end, here we consider attention as a set of *spatial* maps that essentially try to encode on which spatial areas of the input the network focuses most for taking its output decision (e.g., for classifying an image), where, furthermore, these maps can be defined w.r.t. various layers of the network so that they are able to capture both low-, mid-, and high-level representation information. More specifically, in this work we define two types of attention maps: *activation-based* and *gradient-based*. We explore how both of these attention maps change over various datasets and architectures, and show that these actually contain valuable information that can be used for significantly improving the performance of convolutional neural network architectures (of various types and trained for various different tasks). To that end, we propose several novel ways

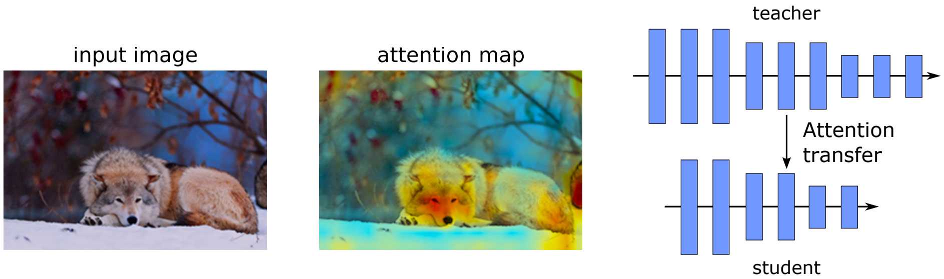

Figure 1: Schematic representation of attention transfer

of transferring attention from a powerful teacher network to a smaller student network with the goal of improving the performance of the latter (Fig. 1).

To summarize, the contributions of this work are as follows:

- We propose attention as a mechanism of transferring knowledge from one network to another
- We show experimental results showing benefits of activation-based and gradient-based attention transfer over residual and non-residual networks
- Our improvements are orthogonal to existing approaches, and can be combined to get more accurate/faster networks.

The rest of the paper is structured as follows: we first describe related work in section 2, we explain our approach for activation-based and gradient-based attention transfer in section 3, and then present experimental results for both methods in section 4.

2 Related work

Early work on attention based tracking Larochelle & Hinton (2010), Denil et al. (2012) was motivated by human attention mechanism theories Rensink (2000) and was done via Restricted Bolzmann Machines. It was recently adapted for neural machine translation with recurrent neural networks, e.g. Bahdanau et al. (2014) as well as in several other NLP-related tasks. It was also exploited in computer-vision-related tasks such as image captioning, visual question answering Yang et al. (2015), as well as in weakly-supervised object localization Oquab et al. (2015), to mention a few characteristic examples. In all these tasks attention proved to be useful.

Visualizing attention maps in deep convolutional neural networks is an open problem. The simplest gradient-based way of doing that is by computing a Jacobian of network output w.r.t. input (this leads to attention visualization that are not necessarily class-discriminative), as for example in Simonyan et al. (2014). Another approach was proposed by Zeiler & Fergus (2014) that consists of attaching a network called "deconvnet" that shares weights with the original network and is used to project certain features onto the image plane. A number of methods was proposed to improve gradient-based attention as well, for example guided backpropagation Springenberg et al. (2015), adding a change in *ReLU* layers during calculation of gradient w.r.t. previous layer output. Attention maps obtained with guided backpropagation are non-class-discriminative too. Among existing methods for visualizing attention, we should also mention class activation maps Zhou et al. (2016), which are based on removing top average-pooling layer and converting the linear classification layer into a convolutional layer, producing attention maps per each class. A method combining both guided backpropagation and CAM is Grad-CAM by Selvaraju et al. (2016), adding image-level details to class-discriminative attention maps.

Knowledge distillation with neural networks was pioneered by Bucila et al. (2006), which is a transfer learning method that aims to improve the training of a student network by relying on knowledge borrowed from a powerful teacher network. Although in certain special cases shallow networks had been shown to be able to approximate deeper ones without loss in accuracy Lei & Caruana (2014), later work related to knowledge distillation was mostly based on the assumption that deeper

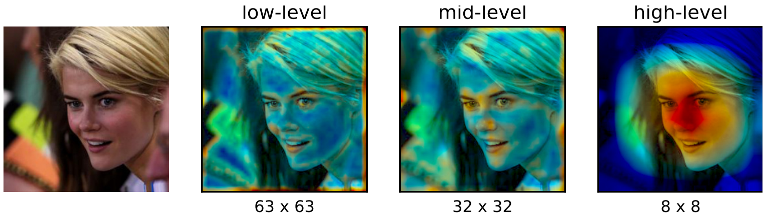

Figure 2: Sum of absolute values attention maps over different levels of a network trained for face recognition. Mid-level attention maps have higher activation level around eyes, nose and lips, high-level activations correspond to the whole face.

networks always learn better representations. For example, FitNets Romero et al. (2014) tried to learn a thin deep network using a shallow one with more parameters. The introduction of highway Srivastava et al. (2015) and later residual networks He et al. (2015) allowed training very deep architectures with higher accuracy, and generality of these networks was experimentally showed over a large variety of datasets. Although the main motivation for residual networks was increasing depth, it was later shown by Zagoruyko & Komodakis (2016) that, after a certain depth, the improvements came mostly from increased capacity of the networks, i.e. number of parameters (for instance, a wider deep residual network with only 16 layers was shown that it could learn as good or better representations as very thin 1000 layer one, provided that they were using comparable number of parameters).

Due to the above fact and due to that thin deep networks are less parallelizable than wider ones, we think that knowledge transfer needs to be revisited, and take an opposite to FitNets approach - we try to learn less deep student networks. Our attention maps used for transfer are similar to both gradient-based and activation-based maps mentioned above, which play a role similar to “hints” in FitNets, although we don’t introduce new weights.

3 Attention transfer

3.1 Activation-based attention transfer

Let us consider a CNN layer and its corresponding activation tensor $A \in R^{C\times H\times W}$, which consists of C feature planes with spatial dimensions $H \times W$. An activation-based mapping function $\mathcal{F}$ (w.r.t. that layer) takes as input the above 3D tensor A and outputs a spatial attention map, i.e., a flattened 2D tensor defined over the spatial dimensions, or

$$\mathcal{F} : R^{C\times H\times W} \rightarrow R^{H\times W} \quad . \tag{1}$$

To define such a spatial attention mapping function, the implicit assumption that we make is that the absolute value of a hidden neuron activation (that results when the network is evaluated on given input) can be used as an indication about the importance of that neuron w.r.t. the specific input. By considering, therefore, the absolute values of the elements of tensor A, we can construct a spatial attention map by computing statistics of these values across the channel dimension. More specifically, in this work we will consider the following activation-based spatial attention maps:

- sum of absolute values: $F_{\text{sum}}(A) = \sum_{i=1}^{C} |A_i|$
- sum of absolute values raised to the power of p (where $p > 1$): $F_{\text{sum}}^{p}(A) = \sum_{i=1}^{C} |A_i|^p$
- max of absolute values raised to the power of p (where $p > 1$): $F_{\text{max}}^{p}(A) = \max_{i=1,C} |A_i|^p$

where $A_i = A(i, :, :)$ (using Matlab notation).

We visualized activations of various networks on several datasets, including ImageNet classification and localization, COCO object detection, face recognition, and fine-grained recognition. We were mostly focused on modern architectures without top dense linear layers, such as Network-In-Network, ResNet and Inception, which have streamlined convolutional structure. We also examined networks of the same architecture, width and depth, but trained with different frameworks with significant difference in performance. We found that the above statistics of hidden activations not only have spatial correlation with predicted objects on image level, these correlations tend to be higher in networks with higher accuracy, and stronger networks have peaks in attention where weak networks don't.

Moreover, as expected, attention maps focus on different parts for different layers in the network. In the first layers neurons activation level is high for low-level gradient points, in the middle it is higher for the most discriminative regions such as eyes or wheels, and in the top layers it reflects full objects. For example, mid-level attention maps of a network trained for face recognition Parkhi et al. (2015) will have higher activations around eyes, nose and lips, and top level activation will correspond to full face (Fig. 2).

It is also apparent that different attention mapping functions have different properties. E.g.:

- Compared to $F_{\text{sum}}(A)$, the spatial map $F^p_{\text{sum}}(A)$ puts more weight to spatial locations that correspond to the neurons with the highest activations, i.e., puts weight to the most discriminative parts.
- Furthermore, among all neurons corresponding to a spatial location, $F^p_{\text{max}}(A)$ will consider only one of them to assign a weight to that spatial location (as opposed to $F_{\text{sum}}(A)$ that will favor spatial locations that carry multiple neurons with high activations).

where C is a number of feature planes, and $H \times W$ are spatial dimensions of the activation tensor (Fig. 3). In case of sum of absolute values the mapping will simply be $\sum_{i=1,C} |A_i|$, and will reflect activation level of neurons in a particular spatial location. In the first layers neurons activation level is high for low-level gradient points, in the middle it is higher for the most discriminative regions such as eyes or wheels, and in the top layers it reflects full objects. For example, mid-level attention maps of a network trained for face recognition Parkhi et al. (2015) will have higher activations around eyes, nose and lips, and top level activation will correspond to full face (Fig. 2).

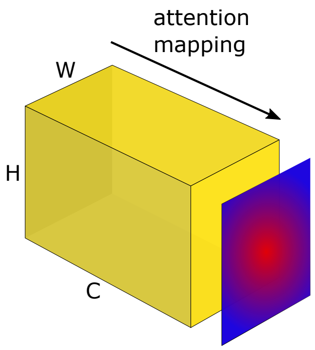

Figure 3: Attention mapping over feature dimension.

To illustrate the differences of these functions we visualized attention maps of 3 networks with sufficient difference in classification performance: Network-In-Network (62% top-1 val accuracy), ResNet-34 (73% top-1 val accuracy) and ResNet-101 (77.3% top-1 val accuracy). In each network we took last pre-downsampling activation maps, on the left for mid-level and on the right for top pre-average pooling activations on fig. 4. Top-level maps are blurry because their original spatial resolution is 7×7. It is clear that most discriminative regions have higher activation levels, e.g. face of the wolf, and how shape details disappear as power decreases.

In case of ResNet, there are several possible ways to place attention transfer layers, depending on depth of teacher and student:

- Same depth: possible to have attention transfer layer after every residual block;
- Different depth: have attention transfer on output activations of each group of residual blocks;

Here we require student and teacher have the same resolution, so attention transfer losses placing does not depend on width. In general it attention maps can be interpolated to match their shapes. We illustrate the different depth case in fig. 5. We also call "group" a set of residual block in ResNet, or a block of 3×3, 1×1, 1×1 convolutions in NIN.

We should note importance of normalization of attention maps before transfer, and perform l_1 or l_2 normalization in experimental section. Let S, T and $\mathbf{W}_S$, $\mathbf{W}_T$ denote student, teacher and

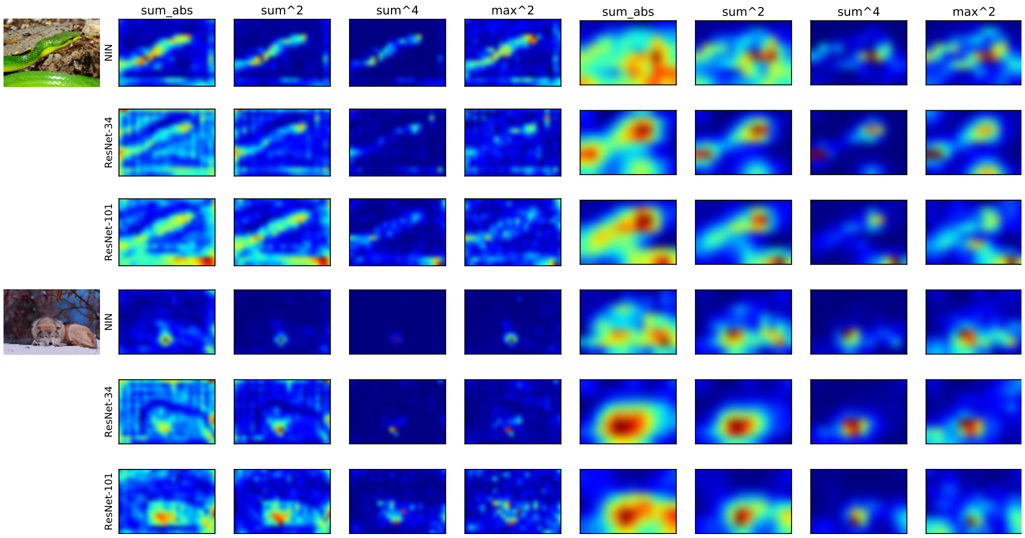

Figure 4: Activation attention maps for various ImageNet networks: Network-In-Network (62% top-1 val accuracy), ResNet-34 (73% top-1 val accuracy), ResNet-101 (77.3% top-1 val accuracy). Left part: mid-level activations, right part: top-level pre-softmax acivations

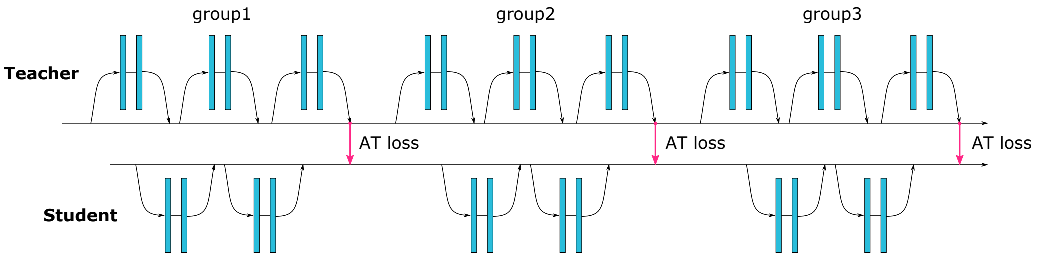

Figure 5: Schematics of teacher-student attention transfer for the case when both networks are residual, and the teacher is deeper.

their weights correspondingly, and let $\mathcal{L}(\mathbf{W}, x)$ be a network with cross entropy loss. Then for all teacher-student activation layers pairs we can define the following total loss:

$$\mathcal{L}_{AT} = \mathcal{L}(\mathbf{W}_S, x) + \sum_j \frac{1}{2} \| \frac{Q_S^j}{\|Q_S^j\|_2} - \frac{Q_T^j}{\|Q_T^j\|_2} \|_p \tag{2}$$

where Q_S and Q_T and student and teacher attention maps correspondingly, and p refers to norm type. Attention transfer can be combined with knowledge distillation, the the total loss will have an additional term. When combined, attention transfer adds very little computational cost, as attention maps for teacher can be easily computed during forward propagation, needed for distillation.

3.2 Gradient-based attention transfer

In this section we define attention as gradient w.r.t. input, which can be viewed as a pixel sensitivity map, i.e. how much we should change a pixel to have an effect on output prediction, e.g. Simonyan et al. (2014). To transfer gradient-based attention we use double backpropagation technique developed by Drucker & LeCun (1992). It was originally used to constrain weights to be smaller making them spend more time in linear regions, when *tanh* activations were popular, by minimizing a norm on Jacobian w.r.t input. Let's define Jacobian w.r.t input for teacher and student as:

$$J_S = \frac{\partial}{\partial x}\mathcal{L}(\mathbf{W_S}, x), J_T = \frac{\partial}{\partial x}\mathcal{L}(\mathbf{W_T}, x) \tag{3}$$

Then if we want student gradient attention to be similar to teacher attention, we can minimize a distance between them (let's assume l_2 for simplicity):

$$\mathcal{L}_{AT}(\mathbf{W_S}, \mathbf{W_T}, x) = \mathcal{L}(\mathbf{W_S}, x) + \beta\frac{1}{2}||J_S - J_T||_2 \tag{4}$$

As $\mathcal{W}_T$ and x are given, to get the needed derivative w.r.t. $\mathcal{W}_S$:

$$\frac{\partial}{\partial \mathbf{W_S}}\mathcal{L}_{AT} = \frac{\partial}{\partial \mathbf{W_S}}\mathcal{L}(\mathbf{W_S}, x) + \beta(J_S - J_T)\frac{\partial^2}{\partial \mathbf{W_S}\partial x}\mathcal{L}(\mathbf{W_S}, x) \tag{5}$$

So to do an update we first need to do forward and back propagation to get J_S and J_T, compute the second error $\frac{1}{2}||J_S - J_T||_2$ and propagate it second time. The second propagation is similar to forward propagation in this case, and involves second order mixed partial derivative calculation $\frac{\partial^2}{\partial W_S \partial x}$. The above computation can be done efficiently in a framework with automatic differentiation support, even for modern architectures with sophisticated graphs. The second backpropagation has approximately the same cost with first backpropagation, excluding forward propagation.

We also propose to enforce horizontal flip invariance on gradient attention maps. To do that we propagate horizontally flipped images as well originals, backpropagate and flip gradient attention maps back. We then add l_2 losses on the obtained attentions and outputs, and do second backpropagation. This is similar to Group Equivariant CNN approach by Cohen & Welling (2016), however is not a hard constraint. We experimentally find that this has regularization effect on training.

It might also be possible to have the proposed attention transfer and symmetry constraints in higher layers of the network.

4 Experimental section

In the following section we explore attention transfer on various datasets, including CIFAR, Scenes, CUB and ImageNet. For activation-based attention transfer we used Network-In-Network and ResNet architectures, as they are most performant and set strong baselines in terms of number of parameters compared to AlexNet or VGG, and have been explored in various papers across small and large datasets. On Scenes, CUB and ImageNet we experimented with ResNet-18 and ResNet-34. As for gradient-based attention, we constrained ourselves to Network-In-Network without batch normalization and CIFAR dataset, due to the need of complex automatic differentiation.

4.1 Activation-based attention transfer

We start with CIFAR dataset which has small 32×32 images, and after downsampling top activations have even smaller resolution, so there is not much space for attention transfer. Interestingly we find that even in this case attention transfer seems to give reasonable benefits. We use horizontal flips and random crops data augmentations, and all networks have batch normalization. Interestingly, we find that ZCA whitening has negative effect on validation accuracy, and omit it in favor of simpler meanstd normalization. We raise KD temperature for ResNet transfers to 4.

Results of attention transfer for various networks on CIFAR-10 can be found in table 1. It was reported that KD struggles to work if teacher has different width, that's also the case for residual networks. Attention transfer results are also slightly worse for such pairs, but overall provide stable improvement over various architectures. By combining attention transfer with knowledge distillation we're able to achieve 7.5% accuracy with only 175K parameters.

student	teacher	student	AT	AT+KD	teacher
NIN-thin, 0.2M	NIN-wide, 1M	9.38	8.93	8.70	7.28
WRN-16-1, 0.2M	WRN-16-2, 0.7M	8.77	8.43	7.47	6.31
WRN-16-1, 0.2M	WRN-40-1, 0.6M	8.77	8.59	8.27	6.58
WRN-16-2, 0.7M	WRN-40-2, 2.2M	6.31	5.91	5.80	5.23

Table 1: Activation attention transfer with various architectures on CIFAR-10. Error is computed as median of 5 runs with different seed.

function	error
$\sum_{i=1}^{C} \lvert A_i \rvert$	5.58
$\sum_{i=1}^{C} \lvert A_i \rvert^2$	8.60
$\sum_{i=1}^{C} \lvert A_i \rvert^4$	8.58
$\max_{i=1,C} \lvert A_i \rvert^2$	8.69
KD	8.46
KD & $\sum_{i=1}^{C} \lvert A_i \rvert^2$	8.27

Table 2: Test error of WRN-16-1-WRN-40-1 student-teacher pair for various attention mapping functions. Median of 5 runs test errors are reported.

To find that having at least one activation attention transfer loss per group in WRN transfer is important, we trained three networks with just one additional loss per network in `group1`, `group2` and `group3` separately, and compared to a network trained with all three losses. Each loss provides some degree of attention transfer, and combined improvement is proportional to the result when all group losses are used.

We also explore which attention mapping functions tend to work best using WRN-16-1-WRN-40-1 student-teacher pair (table 2). Interestingly, sum functions work very similar, and better than max. On this combination KD also works slightly better than attention transfer, but they stack nicely. We further use sum of squared value function for simplicity.

4.1.1 Gradient-based attention transfer

For simplicity we use small Network-In-Network model in these experiments, and don't apply random crop data augmentation with batch normalization, just horizontal flips augmentation. We also only use deterministic algorithms and sampling with fixed seed, so reported numbers are for single run experiments. We find that in this setting network struggles to fit into training data already, and turn off weight decay even for baseline experiments.

We explore the following methods:

- Minimizing Jacobian w.r.t. input, i.e. simple double backpropagation;
- Symmetry norm on gradient attention maps;
- Student-teacher attention transfer;

Results for various methods are shown in table 3. Interestingly, just minimizing l_2 norm of Jacobian already works pretty well. Also, symmetry norm is one the best performing attention norms, which we plan to investigate in future on other datasets. Similar to hidden attention based transfer gradient based attention transfer works similar to knowledge distillation, and stacking both leads to better results.

4.2 Large networks

4.2.1 Attention transfer learning

To see how attention transfer works in finetuning we choose too datasets: Caltech-UCSD Birds-200-2011 fine-grained classification (CUB) by Wah et al. (2011), and MIT indoor scene classification (Scenes) by Quattoni & Torralba (2009), both containing around 5K images training images. We took ResNet-18 and ResNet-34 pretrained on ImageNet and finetuned on both datasets. On CUB

norm type	error
baseline	13.5
min	12.5
symmetry norm	11.8
AT-NIN-wide	12.1
KD	12.1

Table 3: Performance of various gradient-based attention methods on CIFAR-10. Baseline is a thin NIN network with 0.2M parameters trained on hflip augmenated data, min refers to minimizing l_2 norm of Jacobian w.r.t. input, symmetry norm - to minimizing l_2 norm on difference between original and flipped Jacobian, and AT-NIN-wide - gradient attention transfer

type	model	ImageNet→CUB	ImageNet→Scenes
student	ResNet-18	28.5	28.2
AT	ResNet-18	27 (-1.5)	27.1 (-1.1)
teacher	ResNet-34	26.5	26

Table 4: Finetuning with attention transfer error on Scenes and CUB datasets

we crop bounding boxes, rescale to 256 in one dimension and then take a random crop. Batch normalization layers are fixed for finetuning, and first group of residual blocks is frozen. We then took finetuned ResNet-34 networks and used them as teachers for ResNet-18 pretrained on ImageNet, with sum squared attention losses on 2 last groups. In both cases attention transfer provides significant improvements, closing the gap between ResNet-18 and ResNet-34 in accuracy. Moreover, after finetuning student's attention maps indeed look more similar to teacher's (Fig. 6).

4.2.2 IMAGENET

To showcase activation attention transfer on ImageNet we took ResNet-18 as a student, and ResNet-34 as a teacher, and tried to improve ResNet-18 accuracy. Due to lack of time and computational resources, we plugged attention transfer losses in the middle of training, on epoch 60 out of 100, right after second learning rate drop, and trained until the end. We added only two losses in the 2 last groups of residual blocks, used l_2 normalization and squared sum attention. We also didn't have time to tune any hyperparameters and kept them from finetuning experiments. Nevertheless, ResNet-18 with attention transfer achieved 0.64% top-1 and 0.43% top-5 better validation accuracy. We plan to train ResNet-18 from scratch using several different teacher networks and update the paper.

4.3 IMPLEMENTATION DETAILS

The experiments were conducted in Torch machine learning framework. Double propagation can be implemented in a modern framework with grad of grad automatic differentiation support, e.g. Torch, Theano, Tensorflow. For ImageNet experiments we used fb.resnet.torch code, and used 2 Titan X cards with data parallelizm in both teacher and student to speed up training. We plan to release code and models for all our experiments.

Model	top1, top5
ResNet-18	30.4, 10.8
AT	29.7, 10.4
ResNet-34	26.1, 8.3

Table 5: Attention transfer validation error (single crop) on ImageNet. Transfer losses are added on epoch 60/100.

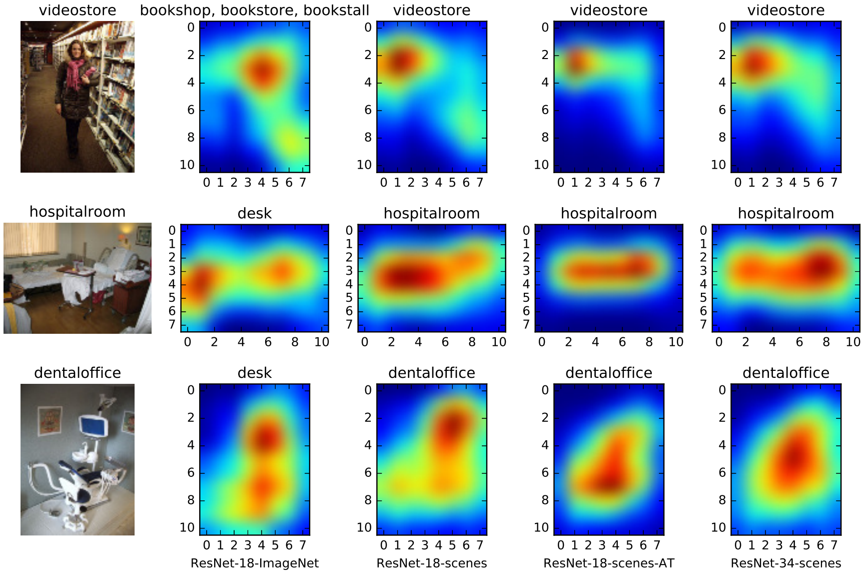

Figure 6: Top activation attention maps for different Scenes networks: original pretrained ResNet-18 (ResNet-18-ImageNet), ResNet-18 trained on Scenes (ResNet-18-scenes), ResNet-18 trained with attention transfer (ResNet-18-scenes-AT) with ResNet-34 as a teacher, ResNet-34 trained on Scenes (ResNet-34-scenes). Predicted classes for each task are shown on top. Attention maps look more similar after transfer (images taken from test set).

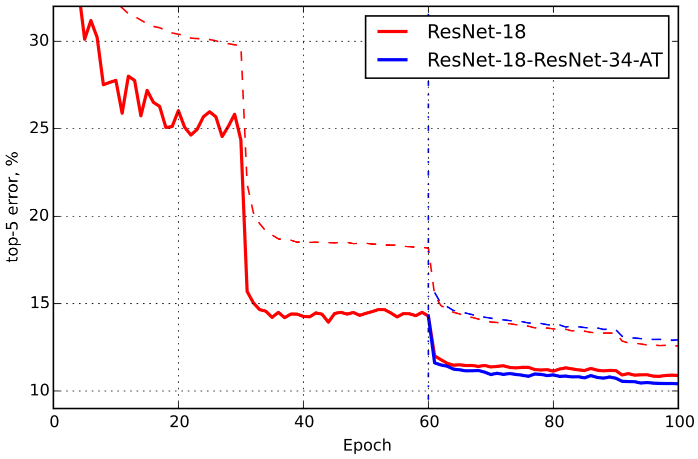

Figure 7: Attention transfer on ImageNet between ResNet-18 and ResNet-34. Solid lines represent top-5 validation error, dashed - top-5 training error. Transfer losses are added on epoch 60/100 (dashed vertical line), no KD losses used.

5 CONCLUSIONS

We presented several ways of transferring attention from one network to another, with experimental results over several image recognition datasets. It would be interesting to see how attention transfer works in cases where spatial information is more important, e.g. object detection or weakly-supervised localization, which we plan to do in future.

Relative improvements in accuracy with ImageNet-sized networks turned out to be higher than for small CIFAR-sized networks, suggesting that attention maps are more important for larger networks.

Overall, we think that our interesting findings will help further advance knowledge distillation, and understanding neural networks in general.

REFERENCES

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473, 2014. URL `http://arxiv.org/abs/1409.0473`.

Cristian Bucila, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *KDD*, pp. 535–541, 2006.

Taco S. Cohen and Max Welling. Group equivariant convolutional networks. *CoRR*, abs/1602.07576, 2016. URL `http://arxiv.org/abs/1602.07576`.

Misha Denil, Loris Bazzani, Hugo Larochelle, and Nando de Freitas. Learning where to attend with deep architectures for image tracking. *Neural Computation*, 2012.

H. Drucker and Y LeCun. Improving generalization performance using double backpropagation. *IEEE Transaction on Neural Networks*, 3(6):991–997, 1992.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

Hugo Larochelle and Geoffrey E. Hinton. Learning to combine foveal glimpses with a third-order boltzmann machine. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta (eds.), *Advances in Neural Information Processing Systems 23*, pp. 1243–1251. Curran Associates, Inc., 2010.

Jimmy Ba Lei and Rich Caruana. Do deep nets really need to be deep? In *NIPS*, 2014.

M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Is object localization for free? weakly-supervised learning with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *British Machine Vision Conference*, 2015.

A. Quattoni and A. Torralba. Recognizing indoor scenes. In *CVPR*, 2009.

Ronald A. Rensink. The dynamic representation of scenes. In *Visual Cognition*, pp. 17–42, 2000.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. FitNets: Hints for thin deep nets. Technical Report Arxiv report 1412.6550, arXiv, 2014.

Ramprasaath R. Selvaraju, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization. 2016.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *ICLR Workshop*, 2014.

J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. Striving for simplicity: The all convolutional net. In *arXiv:1412.6806, also appeared at ICLR 2015 Workshop Track*, 2015. URL `http://arxiv.org/abs/1412.6806`.

Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. Highway networks. *CoRR*, abs/1505.00387, 2015.

C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alexander J. Smola. Stacked attention networks for image question answering. *CoRR*, abs/1511.02274, 2015. URL `http://arxiv.org/abs/1511.02274`.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *BMVC*, 2016.

Matthew Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *ECCV*, 2014.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Computer Vision and Pattern Recognition*, 2016.