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

(2016)

Sergey Zagoruyko, Nikos Komodakis

Notes

Contributions

Instead of using the activation to transfer knowledge from a powerful teacher model to a smaller student model. The authors propose to use attention maps that can be extracted from convnets as an alternative mechanism of transferring knowledge, so that the generated attention maps of the teacher and student match at various levels.

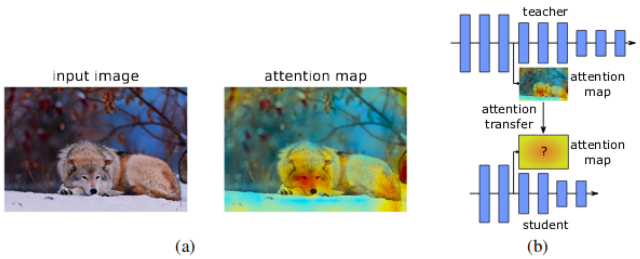

Figure 1: **(a)** An input image and a corresponding spatial attention map of a convolutional network that shows where the network focuses in order to classify the given image. Surely, this type of map must contain valuable information about the network. The question that we pose in this paper is the following: can we use knowledge of this type to improve the training of CNN models ? **(b)** Schematic representation of attention transfer: a student CNN is trained so as, not only to make good predictions, but to also have similar spatial attention maps to those of an already trained teacher CNN.

These attention maps can either be activation-based and gradient-based.

Method

Attention: Activation-Based

Given an activation tensor at a given CNN layer of size $A \in R^{C \times H \times W}$, the goal is to map this tensor into a 2D attention maps, where each spatial location represents the important of the neurons at that location with respect to the specific input. We can obtain these attention maps in three ways (where $A_i = A(i, :, :)$):

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

Compared to $F_{\text{sum}}(A)$, $F^p_{\text{sum}}(A)$ puts more weight to spatial locations that correspond to the neurons with the highest activations (most discriminative parts). On the other hand, $F^p_{\text{max}}(A)$ will only consider one neuron's activations and assign it to the spatial location even if the rest neurons

are not as active, on the other $F_{\text{sum}}(A)$ favors multiple neurons with high activations at the same spatial location.

Using these function, we can then compute the attention maps of a teacher network and train a student network to have similar attention maps to those of the teacher. The attention transfer can be done at each level of the two network if they have the same depth, or at at the different but corresponding levels (groups, such as resnet blocks).

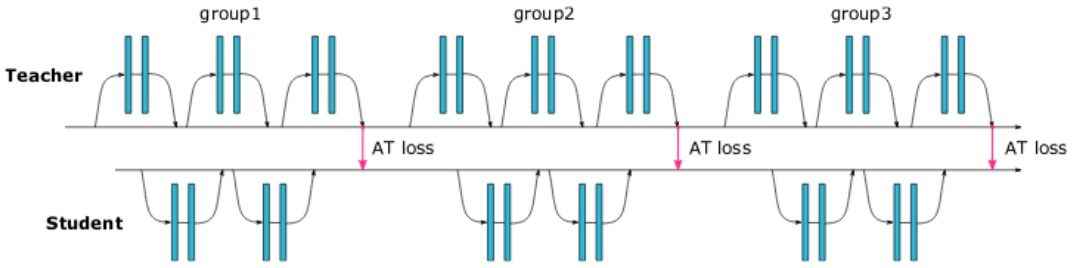

Figure 5: Schematics of teacher-student attention transfer for the case when both networks are residual, and the teacher is deeper.

The loss can then be defined as the MSE between normalized activation maps at the corresponding levels (plus the classification loss).

$$\mathcal{L}_{AT} = \mathcal{L}\left(\mathbf{W}_S, x\right) + \frac{\beta}{2} \sum_{j \in \mathcal{I}} \left\| \frac{Q_S^j}{\left\| Q_S^j \right\|_2} - \frac{Q_T^j}{\left\| Q_T^j \right\|_2} \right\|_p$$

where $Q_S^j = \text{vec}(F(A_S^j))$ and $Q_T^j = \text{vec}(F(A_T^j))$

An additional term can be added to the attention loss, corresponding to the cross entropy between softened distributions over labels of teacher and student (knowledge distillation)

Attention: Gradient-Based

In this case we define attention as gradient w.r.t. input, in other words, attention at an input spatial location encodes how sensitive the output prediction is w.r.t. changes at that input location.

If we want student gradient attention to be similar to teacher gradient attention, we first compute the gradient of the loss w.r.t the input for the teacher and the student:

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

And then we can minimize the $l2$ distance between the two (plus the classification loss):

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

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 $\frac{\beta}{2} \left\| J_S - J_T \right\|_2$ and propagate it second time.

An additional regularization constraint can be added, such as enforcing horizontal flip invariance on gradient attention map (the image is flipped, passed through the net, we then get the gradients and then flip them and add it to the loss):

$$\mathcal{L}_{sym}(\mathbf{W}, x) = \mathcal{L}(\mathbf{W}, x) + \frac{\beta}{2} \left\| \frac{\partial}{\partial x}\mathcal{L}(\mathbf{W}, x) - \text{flip}\left(\frac{\partial}{\partial x}\mathcal{L}(\mathbf{W}, \text{flip}(x))\right) \right\|_2$$

Results

student	teacher	student	AT	F-ActT	KD	AT+KD	teacher
NIN-thin, 0.2M	NIN-wide, 1M	9.38	8.93	9.05	8.55	8.33	7.28
WRN-16-1, 0.2M	WRN-16-2, 0.7M	8.77	7.93	8.51	7.41	7.51	6.31
WRN-16-1, 0.2M	WRN-40-1, 0.6M	8.77	8.25	8.62	8.39	8.01	6.58
WRN-16-2, 0.7M	WRN-40-2, 2.2M	6.31	5.85	6.24	6.08	5.71	5.23

Table 1: Activation-based attention transfer (AT) with various architectures on CIFAR-10. Error is computed as median of 5 runs with different seed. F-ActT means full-activation transfer (see §4.1.2).

attention mapping function	error
no attention transfer	8.77
F_{sum}	7.99
F^2_{sum}	7.93
F^4_{sum}	8.09
F^1_{max}	8.08

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

norm type	error
baseline (no attention transfer)	13.5
min-l_2 Drucker & LeCun (1992)	12.5
grad-based AT	12.1
KD	12.1
symmetry norm	11.8
activation-based AT	**11.2**

Table 3: Performance of various gradient-based attention methods on CIFAR-10. Baseline is a thin NIN network with 0.2M parameters (trained only on horizontally flipped augmented data and without batch normalization), min-l_2 refers to using l_2 norm of gradient w.r.t. input as regularizer, symmetry norm - to using flip invariance on gradient attention maps (see eq. 6), AT - to attention transfer, and KD - to Knowledge Distillation (both AT and KD use a wide NIN of 1M parameters as teacher).

type	model	ImageNet→CUB	ImageNet→Scenes
student	ResNet-18	28.5	28.2
KD	ResNet-18	27 (-1.5)	28.1 (-0.1)
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