Paying More Attention to Attention: Improving the Performance of CNNs via Attention Transfer

Sergey Zagoruyko Nikos Komodakis

Universite Paris-Est, Ecole des Ponts ParisTech

ICLR, 2017

Presenter: Ritambhara Singh

Outline

Outline

Motivation

- Use attention to transfer knowledge from one network to another.

Motivation

- Use attention to transfer knowledge from one network to another.
- Teacher networks: deep networks, Student networks: shallow, wide networks (easier to parallelize).

Motivation

- Use attention to transfer knowledge from one network to another.
- Teacher networks: deep networks, Student networks: shallow, wide networks (easier to parallelize).
- **Basic idea:** Can a teacher network improve the performance of a student network...

Motivation

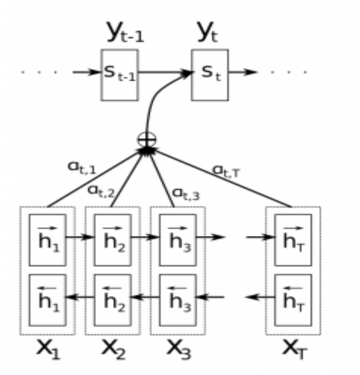
- Use attention to transfer knowledge from one network to another.
- Teacher networks: deep networks, Student networks: shallow, wide networks (easier to parallelize).
- **Basic idea:** Can a teacher network improve the performance of a student network...
- By providing to it information about where the teacher is looking?

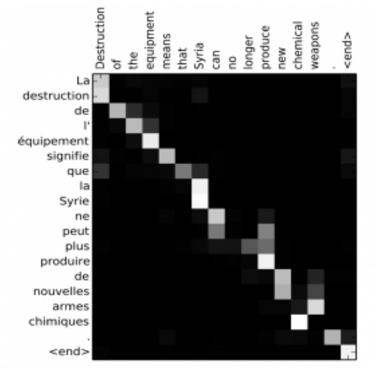
Motivation

- Use attention to transfer knowledge from one network to another.
- Teacher networks: deep networks, Student networks: shallow, wide networks (easier to parallelize).
- **Basic idea:** Can a teacher network improve the performance of a student network...
- By providing to it information about where the teacher is looking?

Outline

Attention Based Models (RNN)





Bahdanau et al. (2014)

Attention Based Models (CNN)

Simonyan et al. (2014)

Outline

Knowledge Transfer

- **Knowledge Distillation:** Training a student network by relying on knowledge borrowed from a powerful teacher network.

System	Test Frame Accuracy	WER
Baseline	58.9%	10.9%
10xEnsemble	61.1%	10.7%
Distilled Single model	60.8%	10.7%

Table 1: Frame classification accuracy and WER showing that the distilled single model performs about as well as the averaged predictions of 10 models that were used to create the soft targets.

Hinton et al. (2015)

Outline

1. Introduction
   - Motivation
   - Background
   - State-of-the-art
2. Proposed Approach
   - Attention Transfer
3. Evaluation
   - CIFAR Experiments
   - Imagenet Experiments

Activation-Based

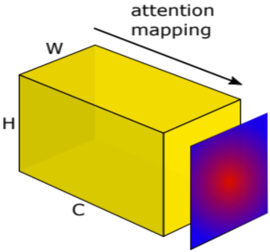


Figure 3: Attention mapping over feature dimension.

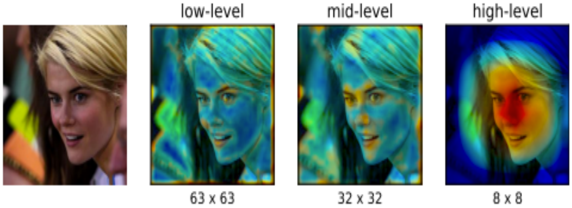
$$\mathcal{F} : R^{C \times H \times W} \rightarrow R^{H \times W} \quad (1)$$

Activation-Based: Attention Map

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

Activation-Based: Attention Map

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



Activation-Based: ResNet architectures

- Same depth: attention transfer after every residual block
- Different depth: attention transfer after groups of residual blocks

Activation-Based: ResNet architectures

- Same depth: attention transfer after every residual block
- Different depth: attention transfer after groups of residual blocks

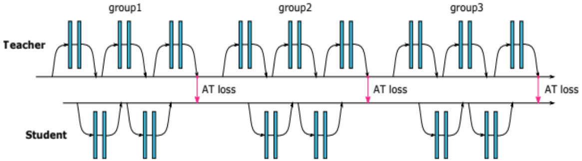


Figure 5: Schematics of teacher-student attention transfer for the case when both networks are residual, and the teacher is deeper.

Activation-Based: Attention Loss

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

Gradient Based

- Attention is defined as gradient w.r.t input (Saliency map in Simonyan et al. (2014))

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

Gradient Based

- Attention is defined as gradient w.r.t input (Saliency map in Simonyan et al. (2014))

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

- Minimize the distance between gradient attention maps of student and teacher

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

Gradient Based

- Attention is defined as gradient w.r.t input (Saliency map in Simonyan et al. (2014))

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

- Minimize the distance between gradient attention maps of student and teacher

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

- $\mathbf{W}_T$ and x are give, need to get derivative w.r.t $\mathbf{W}_S$

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

Gradient Based

- Attention is defined as gradient w.r.t input (Saliency map in Simonyan et al. (2014))

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

- Minimize the distance between gradient attention maps of student and teacher

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

- $\mathbf{W}_T$ and x are give, need to get derivative w.r.t $\mathbf{W}_S$

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

- Enforce horizontal flip invariance

$$\mathcal{L}_{sym}(\mathbf{W}, x) = \mathcal{L}(\mathbf{W}, x) + \frac{\beta}{2}||\frac{\partial}{\partial x}\mathcal{L}(\mathbf{W}, x) - \mathrm{flip}(\frac{\partial}{\partial x}\mathcal{L}(\mathbf{W}, \mathrm{flip}(x)))||_2 \;, \quad (6)$$

Outline

1. Introduction
   - Motivation
   - Background
   - State-of-the-art
2. Proposed Approach
   - Attention Transfer
3. Evaluation
   - CIFAR Experiments
   - Imagenet Experiments

Activation Based

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

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

Gradient Based

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**

Outline

Transfer learning

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

Summary

- Present different ways to transfer attention from one network to another.
- Demonstrate better performance for image recognition datasets.
- Future Direction
  - Understand how attention transfer works in cases where spatial information is important e.g. object detection