Show, Attend and Distill: Knowledge Distillation via Attention-based Feature Matching

Guangyao Zhou

November 15, 2022

1 Abstract

1.1 Motivation

Manual link selection does not consider the similarity between the teacher and student features, so there is a risk of forcing an incorrect intermediate process to the student. Furthermore, he link selection has a limitation on fully utilizing the whole knowledge of the teacher by choosing a few of all possible links.

1.2 Preworks

Most studies manually tie intermediate features of the teacher and student, and transfer knowledge through predefined links.

1.3 Approach

The authors proposed method utilizes an attention-based meta-network that learns relative similarities between features.

1.4 Contribution

The authors proposed method determines competent links more efficiently

2 Related Preworks

2.1 Learning to Transfer (L2T)

2.1.1 L2T Approach

To compensate for the limitation, Jang *et al.* [Learning What and Where to Transfer] apply a meta-networks, "learning to transfer (L2T)", automatically determining the links. In more details, the meta-network consists of individual

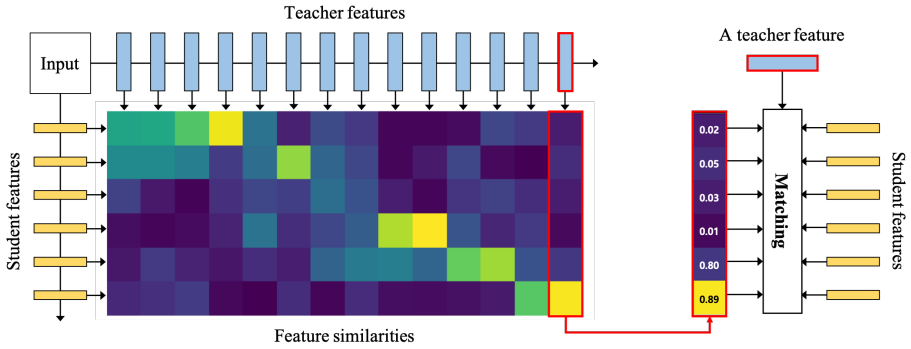

Figure 1: Overview of AFD. An attention-based model determines similarities between the teacher and student features. Knowledge from each teacher feature is transferred to the student with the identified similarities.

gates for all possible links, and each gate determines whether distillation through the link contributes to decreasing the classification loss of the student.

2.1.2 1st Con

The connections ae not aware that ther're affecting simultaneously
The individual gates are not aware of each other although the distillation through the gates simultaneously affect the student.

2.1.3 2nd Con

Computational Expensive
The meta-learning scheme requires expensive inner-loop procedures to learn their meta-networks, thus its application can be limited under practical scenarios.

2.2 Advantages Compare to L2T

i The authors proposed method considers the granularity of the teacher and student features to identify the importance of their links while L2T only uses information for a single pair in a narrow perspective.

ii AFD learns from feature similarities without any inner-loop procedure but L2T learns from the classification loss, which requires expensive Hessian computation.

3 Attention-based Feature Distillation

Let $\mathbf{h}^{\mathrm{T}} = \{h_1^{\mathrm{T}}, ..., h_T^{\mathrm{T}}\}$ be a set of the feature candidates from the teacher and $\mathbf{h}^{\mathrm{S}} = \{h_1^{\mathrm{S}}, ..., h_S^{\mathrm{S}}\}$ be a set of feature candidates from the student where T and

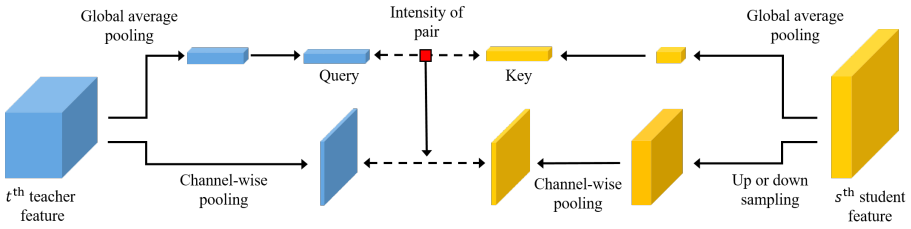

Figure 2: Overview of the proposed meta-network. The globally pooled features are utilized to estimate the similarities and the channel-wisely averaged features are used to calculate the distance between the features.

S indicate the numbers of the candidates from the teacher and student, respectively.
Specifically, each teacher feature generates a query, $\mathbf{q}_t$, and each student feature identifies a key, $\mathbf{k}_s$.

$$\begin{aligned} \mathbf{q}_t &= f_Q(W_t^{\mathrm{Q}} \cdot \phi^{HW}(h_t^{\mathrm{T}})), \\ \mathbf{k}_s &= f_K(W_s^{\mathrm{K}} \cdot \phi^{HW}(h_s^{\mathrm{S}})). \end{aligned} \tag{1}$$

$\phi^{HW}(\cdot)$ indicates a global average pooling. f_Q and f_K are activation function of the query and key. $W_t^{\mathrm{Q}} \in \mathbb{R}^{d \times d_t^{\mathrm{T}}}$ and $W_s^{\mathrm{K}} \in \mathbb{R}^{d \times d_s^{\mathrm{S}}}$ are linear transition parameters for the t-th query and the s-th key.

$$\begin{aligned} \alpha_t = \mathrm{softmax}([&(\mathbf{q}_t^\top W_1^{\text{Q-K}} \mathbf{k}_{t,1} + (\mathbf{p}_t^{\mathrm{T}})^\top \mathbf{p}_1^{\mathrm{S}})/\sqrt{d}, \\ &\cdots, (\mathbf{q}_t^\top W_S^{\text{Q-K}} \mathbf{k}_{t,S} + (\mathbf{p}_t^{\mathrm{T}})^\top \mathbf{p}_S^{\mathrm{S}})/\sqrt{d}]). \end{aligned} \tag{2}$$

Here, The authors introduce additional weight parameters; a bilinear weight, $W_t^{\text{Q-K}} \in \mathbb{R}^{d \times d}$, and positional encodings, $\mathbf{p}_t^{\mathrm{T}} \in \mathbb{R}^d$ and $\mathbf{p}_s^{\mathrm{S}} \in \mathbb{R}^d$. The bilinear weight is applied to generalize the attention value from different source ranks since the query and key are identified from different dimensional features [**?**, **?**]. The positional encodings are utilized to share common information over different instances [**?**]. α_t is the attention vector that capture relation between the t-th teacher feature and whole student features. By utilizing α_t, the teacher feature, h_t^{T}, enables to transfer its knowledge selectively to student features.
The final distillation term forms as

$$\mathcal{L}_{\mathrm{AFD}} = \Sigma_t \Sigma_s \alpha_{t,s} \left\| \tilde{\phi}^C(h_t^{\mathrm{T}}) - \tilde{\phi}^C(\hat{h}_s^{\mathrm{S}}) \right\|_2, \tag{3}$$

where $\tilde{\phi}^C$ indicates a combined function of a channel-wise average pooling layer with L2 normalization, $\mathbf{v}/\|\mathbf{v}\|_2$, by following [**?**]. In addition, $\hat{h}_s^{\mathrm{S}}$ is up-sampled or down-sampled from h_s^{S} to match the feature map size to those of the teacher features.
Finally, the regularization term is added to the total loss function as following;

$$\mathcal{L}_{\mathrm{Student}} = \mathcal{L}_{\mathrm{cls}} + \beta \mathcal{L}_{\mathrm{AFD}}, \tag{4}$$