Paying more attention to Attention:

Improving the performance of (student) Convolutional Neural Networks via Attention Transfer

Introduction:

Attention has played an important role in the context of applying artificial neural networks to a variety of tasks from fields such as computer vision (image captioning, visual question answering, etc) and NLP (neural machine translation). In this paper, authors show that the performance of student CNN can be improved significantly by forcing it to mimic the attention maps of a powerful teacher. They consider attention as a set of spatial maps that encode on which spatial areas of the input the network focuses most for taking its output decision. They use two types of spatial attention maps: activation-based and gradient-based.

Attention Transfer:

(i) Activation-based

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 $\mathcal{F} : R^{C \times H \times W} \to R^{H \times W}$.

To define such a spatial attention mapping function, we implicitly assume that the absolute value of a hidden neuron activation can be used as an indication about the importance of that neuron w.r.t the specific input.

Let S, T and W_S, W_T denote student, teacher and their weights correspondingly, and let $\mathcal{L}(W, x)$ denote a standard cross entropy loss. Let also $\mathcal{I}$ denote the indices of all teacher-student activation layer pairs for which we want to transfer attention maps. Then we can define the following total loss:

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

where $Q_S^j = vec(\mathcal{F}(A_S^j))$ and $Q_T^j = vec(\mathcal{F}(A_T^j))$ are respectively the j-th pair of student and teacher attention maps in vectorized form, and p refers to norm type. Normalization of attention maps is important for the success of the student training (here, l_2-normalized attention maps have been used).

(ii) Gradient-based

Here, attention is defined as gradient w.r.t input, which can be viewed as an input sensitivity map i.e., attention at an input spatial location encodes how sensitive the output prediction is w.r.t changes at that input location. Let's define the gradient of the loss w.r.t input for teacher and student as:

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

Then, if we want student gradient attention to be similar to teacher attention, we can minimise a distance between them (here, l_2 distance is used):

$$\mathcal{L}_{AT}(W_S, W_T, x) = \mathcal{L}(W_S, x) + \frac{\beta}{2} \sum_{j \in \mathcal{I}} \|J_S - J_T\|_2$$