Are Sixteen Heads Really Better than One?

Summary

Alireza Moradi

This paper studies the importance of multiple attention heads in transformer models and how does it impact performance and speed of the model if we reduce the number of attention heads in each layer (even down to 1 head only).
The paper first starts by reviewing how vanilla attention layers operate and the math behind them and goes on to how multi-headed attention works in parallel and the math behind it.
Now in order to prune some attention heads, we introduce a new parameter ξ_h that is either 0 or 1. If its 0 it means that particular head is pruned.
In order to see if all the heads are important or not we used 2 models:

- **WMT**: which is the original large transformer (6 layers with 16 attention heads each) with encoder-decoder architecture with self-attention modules (Enc-Enc and Dec-Dec) and vanilla attention (Enc-Dec) modules in decoder network.
- **BERT**: a transformer (Encoder only) trained on MSM task (with 12 layers and 12 attention heads each). This model does not have vanilla attention modules.

In the first part we pruned heads of each layer (on pre-trained models) one by one (in the same layer) and it shows that the majority of attention heads can be removed without significant impact on the performance, even removing some of the heads can boost the performance of the model. This is showing that the majority of self-attention heads are redundant during test time.
In a new experiment we want to answer the question in the title of the paper, do we really need more than 1 head? So we pruned all the heads in each layer but 1. The result shows that for most of the layers, only 1 attention head is enough but for example the performance of the last layer of WMT drops significantly if we prune all the Enc-Dec attentions heads but one.
There's also the question of generalizability, as these tests have been held on a limited specific test set. But conducting these tests on a new out-of-domain test set shows that important heads tend to be the same across different datasets, therefore answering our earlier question.
To see the reaction of the model when we prune multiple heads from across the model, we first sort the heads (in ascending order) based on a proxy importance score I_h (which is a function of gradients of ℓ -loss on a specific sample x_i - and ξ_h) and then start pruning 10% of them at each step iteratively. The results show that only 20% of heads in WMT models and 40% of heads in BERT without significant performance drop, indicating we can't reduce these models to a pure single headed model, at least without retraining/fine-tuning.
We go further by measuring the speed increase and see that by pruning 50% of heads in we see a 17.5% speed increase in BERT model for large batch sizes. This is because each head in BERT contains 6.25% and 8.34% of all the parameters in the model, and by pruning heads we are also decreasing the trainable parameters and therefore we see the boost in speed.
Another experiment in a transformer-based translation model shows that Enc-Dec attention layers are more important and pruning them leads to a more rapid decrease in performance e.g.

pruning more than 60% of Enc-Dec attention heads results in significant performance degradation.

The above tests were all on trained models and showed the results on test time, but what about training time?

The experiments on a smaller version of WMT (6 layers with 8 heads per layer) trained to translate German to English show that in very early epochs (especially the first 2) all the heads are equally important (because of linear performance drop by pruning percentage and its independency from I_h). From the 10th epoch onwards, the results indicate that we can prune up to 40% of total heads while staying in 85-90% range of un-pruned model's BLEU score.

These results show that the important heads are determined early (but not immediately) during the training process.