Multi-Head Attention: Collaborate Instead of Concatenate

Author: Jean-Baptiste Cordonnier, Andreas Loukas and Martin Jaggi

Presented by: Shreya Modi, Ganesh Epili, Mukesh Jha and Arjun Vankani

Motivation

Widespread Use of Attention Layers.

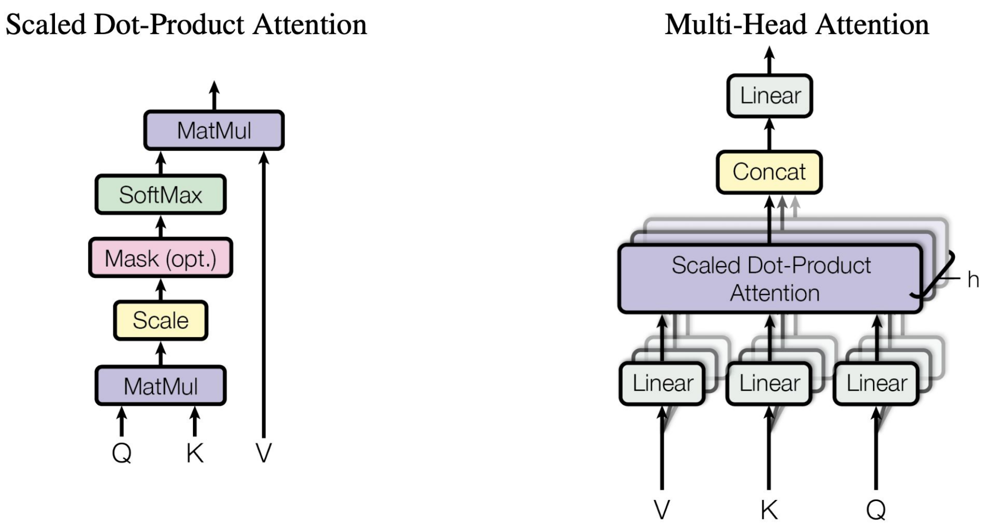


Figure (1): Left: Self-attention. Right: Multi-head self-attention. [1]

Multi-Head Attention: It significantly improves model performance by allowing diverse representation learning within the same architecture

Traditionally, the attention mechanism is replicated by concatenation to obtain multi-head attention defined for Nh heads as:

$$\text{MultiHead}(\boldsymbol{X}, \boldsymbol{Y}) = \underset{i \in [N_h]}{\text{concat}} \left[\boldsymbol{H}^{(i)} \right] \boldsymbol{W}_O$$
$$\boldsymbol{H}^{(i)} = \text{Attention}(\boldsymbol{X}\boldsymbol{W}_Q^{(i)}, \boldsymbol{Y}\boldsymbol{W}_K^{(i)}, \boldsymbol{Y}\boldsymbol{W}_V^{(i)}),$$

where distinct parameter matrices $\boldsymbol{W}_Q^{(i)}, \boldsymbol{W}_K^{(i)} \in \mathbb{R}^{D_{in} \times d_k}$ and $\boldsymbol{W}_V^{(i)} \in \mathbb{R}^{D_{in} \times d_{out}}$ are learned for

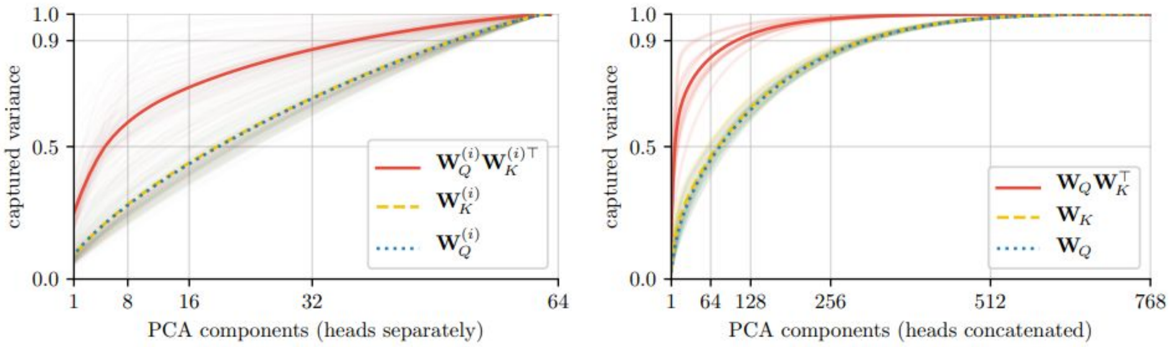
each head i ∈ $[N_h]$ and the extra parameter matrix $\boldsymbol{W}_O \in \mathbb{R}^{N_h d_{out} \times D_{out}}$ projects the concatenation of

the Nh head outputs (each in R dout) to the output space R Dout .In the multi-head setting, we call dk the dimension of each head and $D_k = N_h d_k$ the total dimension of the query/key space.

Motivation

Why the problem is considered important to be solved ?

- **Redundancy in Key/Query Projections:** It leads to over-parameterization and inefficiencies in the attention layer.



The observation from the figure is that individual heads are not low rank (indicating they have rich content), but when concatenated, they show a low rank (indicating redundancy), which supports the author’s claim that some projections in the attention mechanism are redundant.

Motivation

Why the problem is considered important to be solved ?

- **Inefficient Parameter Usage:** It results in a large number of parameters, which can lead to computational inefficiencies and increased training time.
- **Lack of Expressiveness:** It does not provide adaptive head expressiveness, as the dimensions of each head are fixed, potentially limiting the model's ability to capture complex attention patterns.
- **Limited Understanding of Interactions:** The interactions between heads in traditional MHA are not well understood, and it is unclear how independent heads learn overlapping or distinct concepts.

Problem formulation

- Replace concatenation-based multi-head attention with better alternative.
- The goal of mitigating the redundancy issue and enabling a decrease in the key/query dimension without sacrificing performance.
- How to re-parametrize pre-trained models and achieve compression while maintaining performance.
- The importance of fine-tuning the compressed models after re-parametrization to recover any performance loss.

Approach

- The collaborative approach aims to have all heads learn projections together, allowing each head to use a re-weighting of these projections rather than learning them independently.

$$\mathrm{CollabHead}(\boldsymbol{X}, \boldsymbol{Y}) = \operatorname*{concat}_{i \in [N_h]} \left[\boldsymbol{H}^{(i)}\right] \boldsymbol{W}_O$$

$$\boldsymbol{H}^{(i)} = \mathrm{Attention}(\boldsymbol{X}\tilde{\boldsymbol{W}}_Q \operatorname{diag}(\boldsymbol{m}_i), \boldsymbol{Y}\tilde{\boldsymbol{W}}_K, \boldsymbol{Y}\boldsymbol{W}_V^{(i)}).$$

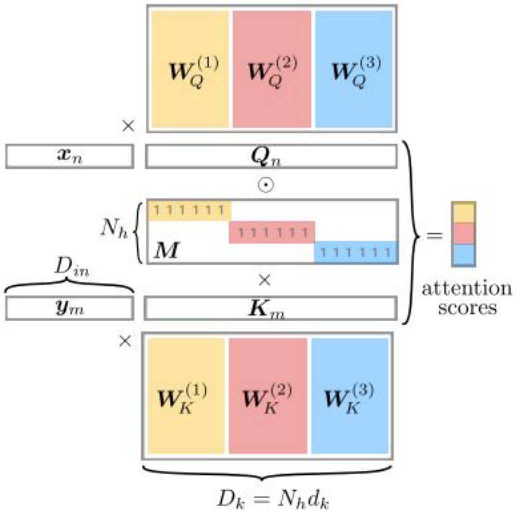
The key difference here is the introduction of a mixing vector m_i for each head, which allows for adaptive usage of the shared projection dimensions.

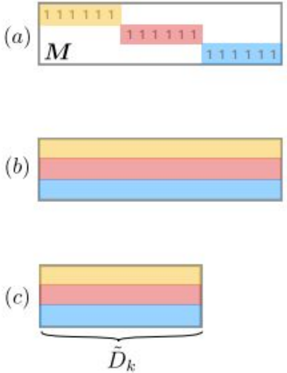
This approach results in two main benefits:

- **Adaptive head expressiveness**: Each head can use more or fewer dimensions from the shared projection space, allowing for more flexible and adaptive attention patterns.
- **Efficiency in parameters**: Since the projections are shared and only the re-weighting vectors are head-specific, the model is more parameter-efficient.

Standard Multi-Head Attention where each head computes attention scores independently with its dedicated W_Q and W_K matrices.

$$M := \underset{i \in [N_h]}{\text{concat}} [\boldsymbol{m}_i] \in \mathbb{R}^{N_h \times \tilde{D}_k},$$





Collaborative framework is visualized, suggesting alternative ways to mix shared projections, which can lead to more parameter efficiency and potentially better performance.

Tensor Decomposition:

- Canonical tensor decomposition is leveraged to reparametrize any pre-trained transformers to use collaborative attention. This allows for the application of collaborative attention to existing models without the need for retraining from scratch.

$$\mathbf{W}_{QK} := \underset{i \in [N_h]}{\text{stack}} \left[\boldsymbol{W}_Q^{(i)} \boldsymbol{W}_K^{(i)\top} \right] \in \mathbb{R}^{N_h \times D_{in} \times D_{in}}.$$

Following the notation[3] of Kolda & Bader (2009), the Tucker decomposition of a tensor $\mathbf{T} \in \mathbb{R}^{I \times J \times K}$ is written as

$$\mathbf{T} \approx \mathbf{G} \times_1 \boldsymbol{A} \times_2 \boldsymbol{B} \times_3 \boldsymbol{C} = \sum_{p=1}^{P} \sum_{q=1}^{Q} \sum_{r=1}^{R} g_{pqr}\, \boldsymbol{a}_p \circ \boldsymbol{b}_q \circ \boldsymbol{c}_r$$

with $\boldsymbol{A} \in \mathbb{R}^{I \times P}$, $\boldsymbol{B} \in \mathbb{R}^{J \times Q}$, and $\boldsymbol{C} \in \mathbb{R}^{K \times R}$ being factor matrices, whereas $\mathbf{G} \in \mathbb{R}^{P \times Q \times R}$ is the core tensor. Intuitively, the core entry $g_{pqr} = \mathbf{G}_{p,q,r}$ quantifies the level of interaction between the components $\boldsymbol{a}_p$, $\boldsymbol{b}_q$, and $\boldsymbol{c}_r$.

With this in place, the computation of the (unscaled) attention score for the i-th head is given by:

$$\left(\boldsymbol{X}\boldsymbol{W}_Q^{(i)} + \mathbf{1}_{T\times 1}\boldsymbol{b}_Q^\top\right)\left(\boldsymbol{Y}\boldsymbol{W}_K^{(i)} + \mathbf{1}_{T\times 1}\boldsymbol{b}_K^\top\right)^\top$$
$$\approx \boldsymbol{X}\tilde{\boldsymbol{W}}_Q \operatorname{diag}(\boldsymbol{m}_i)\tilde{\boldsymbol{W}}_K^\top \boldsymbol{Y}^\top + \mathbf{1}_{T\times 1}\boldsymbol{v}_i^\top \boldsymbol{Y}^\top,$$

Evaluation

1. Neural Machine Translation (NMT):

- Dataset: WMT14 English-to-German translation task.
- Metric: The evaluation metric used is compound split tokenized BLEU score.
- Results: Collaborative MHA matched the baseline BLEU 27.40 with 4x smaller shared key/query dimension, maintaining performance without sacrificing accuracy.

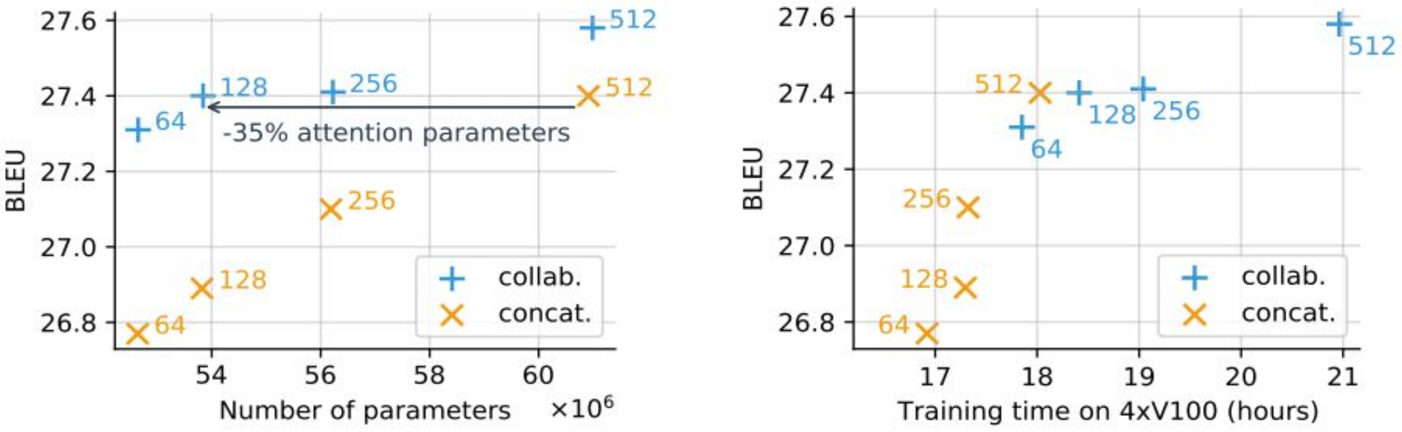


Figure : Comparison of BLEU score on WMT14 EN-DE translation task.

Evaluation

2. Natural Language Understanding (NLU):

- Dataset: GLUE benchmark, which consists of various NLU tasks.
- Metrics: Accuracy on individual tasks (MNLI, MRPC, STS-B, etc.).
- Results: Collaborative MHA applied to BERT-base, ALBERT, and DistilBERT. Compressed models showed comparable or slightly lower accuracy. E.g., BERT and DistilBERT key/query dimensions reduced by 2x and 3x with minimal performance impact.

Evaluation

3. Image Classification:

- Dataset: ImageNet

- Collaborative MHA for vision:

- Results:

It achieved an accuracy of 81.8%, which is comparable to the baseline model's accuracy 81.7%.

By reducing the shared key/query dimension, the model maintained its performance with only a minor 0.1% change in accuracy.

- Re-parametrization for vision:

- Results:

It’s pretrained DeiT-B model with collaborative attention for different shared key dimensions showed that compressing from Dk = 768 to 512 only altered the accuracy by 0.1%. A stronger compression to Dk = 256 resulted in a 1% change in accuracy on ImageNet.

Reference

1. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” 2017.
2. Cordonnier, J., Loukas, A., & Jaggi, M. (2020). Multi-Head Attention: Collaborate Instead of Concatenate. ArXiv. /abs/2006.16362