Multi-Head Attention: Collaborate Instead of Concatenate

Jean-Baptiste Cordonnier
EPFL
Lausanne, Switzerland

Andreas Loukas
EPFL
Lausanne, Switzerland

Martin Jaggi
EPFL
Lausanne, Switzerland

Abstract

Attention layers are widely used in natural language processing (NLP) and are beginning to influence computer vision architectures. However, they suffer from over-parameterization. For instance, it was shown that the majority of attention heads could be pruned without impacting accuracy. This work aims to enhance current understanding on how multiple heads interact. Motivated by the observation that trained attention heads share common key/query projections, we propose a collaborative multi-head attention layer that enables heads to learn shared projections. Our scheme improves the computational cost and number of parameters in an attention layer and can be used as a drop-in replacement in any transformer architecture. For instance, by allowing heads to collaborate on a neural machine translation task, we can reduce the key dimension by a factor of eight without any loss in performance. We also show that it is possible to re-parametrize a pre-trained multi-head attention layer into our collaborative attention layer. Even without retraining, collaborative multi-head attention manages to reduce the size of the key and query projections by half without sacrificing accuracy. Our code is public.[1]

1 Introduction

Since the invention of attention (Bahdanau et al., 2014) and its popularization in the transformer architecture (Vaswani et al., 2017), multi-head attention (MHA) has become the de facto architecture for natural language understanding tasks (Devlin et al., 2019) and neural machine translation. Attention mechanisms have also gained traction in computer vision following the work of Ramachandran et al. (2019) and Bello et al. (2019). Nevertheless, despite their wide adoption, we currently lack solid theoretical understanding of how transformers operate. In fact, many of their modules and hyperparameters are derived from empirical evidences that are possibly circumstantial.

The uncertainty is amplified in multi-head attention, where both the roles and interactions between heads are still poorly understood. Empirically, it is well known that using multiple heads can improve model accuracy. However, not all heads are equally informative, and it has been shown that certain heads can be pruned without impacting model performance. For instance, Voita et al. (2019) present a method to quantify head utility and prune redundant members. Michel et al. (2019) go further to question the utility of multiple heads by testing the effect of heavy pruning in several settings. On the other hand, Cordonnier et al. (2020) prove that multiple heads are needed for self-attention to perform convolution, specifically requiring one head per pixel in the filter's receptive field.

This work aims to better detect and quantify head redundancy by asking whether independent heads learn overlapping or distinct concepts. We discover that some key/query projected dimensions are redundant, as trained concatenated heads tend to compute their attention patterns on common features. Our finding implies that MHA can be re-parametrized with better weight sharing for these common projections and a lower number of parameters.

Preprint. Under review.

[1] https://github.com/epfml/collaborative-attention

Contribution 1: Introducing the collaborative multi-head attention layer. Motivated by this observation, Section 3 describes a collaborative attention layer that allows heads to learn shared key and query features. The proposed re-parametrization significantly decreases the number of parameters of the attention layer without sacrificing performance. Our Neural Machine Translation experiments in Section 4 show that the number of FLOPS and parameters to compute the attention scores can be divided by 8 without affecting the BLEU score on the WMT17 English-to-German task.

Contribution 2: Re-parametrizing pre-trained models into a collaborative form renders them more efficient. Pre-training large language models has been central to the latest NLP developments. But pre-training transformers from scratch remains daunting for its computational cost even when using more efficient training tasks such as (Clark et al., 2020). Interestingly, our changes to the MHA layers can be applied post-hoc on pre-trained transformers, as a drop-in replacement of classic attention layers. To achieve this, we compute the weights of the re-parametrized layer using canonical tensor decomposition of the query and key matrices in the original layer. Our experiments in Section 4 show that the key/query dimensions can be divided by 3 without any degradation in performance.

As a side contribution, we identify a discrepancy between the theory and implementation of attention layers and show that by correctly modeling the biases of key and query layers, we can clearly differentiate between context and content-based attention. This finding could provide an explanation for the success of SYNTHESIZER (Tay et al., 2020): dot-product attention with biases already leverages content and its contextual part is often superfluous.

2 Multi-Head Attention

We first review standard multi-head attention introduced by Vaswani et al. (2017).

2.1 Attention

Let $\boldsymbol{X} \in \mathbb{R}^{T \times D_{in}}$ and $\boldsymbol{Y} \in \mathbb{R}^{T' \times D_{in}}$ be two input matrices consisting of respectively T and T' tokens of D_{in} dimensions each. An attention layer maps each of the T query token from D_{in} to D_{out} dimensions as follows:

$$\text{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^\top}{\sqrt{d_k}}\right)\boldsymbol{V}, \text{ with } \boldsymbol{Q} = \boldsymbol{X}\boldsymbol{W}_Q,\ \boldsymbol{K} = \boldsymbol{Y}\boldsymbol{W}_K,\ \boldsymbol{V} = \boldsymbol{Y}\boldsymbol{W}_V \quad (1)$$

The layer is parametrized by a query matrix $\boldsymbol{W}_Q \in \mathbb{R}^{D_{in} \times D_k}$, a key matrix $\boldsymbol{W}_K \in \mathbb{R}^{D_{in} \times D_k}$ and a value matrix $\boldsymbol{W}_V \in \mathbb{R}^{D_{in} \times D_{out}}$. Using attention on the same sequence (i.e. $\boldsymbol{X} = \boldsymbol{Y}$) is known as self-attention and is the basic building block of the transformer architecture.

2.2 Content vs. Context

The attention operator defined in eq. (1) differs from what is implemented in practice as it omits biases $\boldsymbol{b}_Q$ and $\boldsymbol{b}_K \in \mathbb{R}^{D_k}$. Key and query projections are computed as $\boldsymbol{K} = \boldsymbol{X}\boldsymbol{W}_K + \boldsymbol{1}_{T\times 1}\boldsymbol{b}_K$ and $\boldsymbol{Q} = \boldsymbol{Y}\boldsymbol{W}_Q + \boldsymbol{1}_{T\times 1}\boldsymbol{b}_Q$, respectively, where $\boldsymbol{1}_{a\times b}$ is an all one matrix of dimension $a \times b$. The exact computation of the (unscaled) attention scores can be decomposed as follows:

$$\begin{aligned}\boldsymbol{Q}\boldsymbol{K}^\top &= (\boldsymbol{X}\boldsymbol{W}_Q + \boldsymbol{1}_{T\times 1}\boldsymbol{b}_Q^\top)(\boldsymbol{Y}\boldsymbol{W}_K + \boldsymbol{1}_{T\times 1}\boldsymbol{b}_K^\top)^\top & (2)\\ &= \underbrace{\boldsymbol{X}\boldsymbol{W}_Q\boldsymbol{W}_K^\top\boldsymbol{Y}^\top}_{\text{context}} + \underbrace{\boldsymbol{1}_{T\times 1}\boldsymbol{b}_Q^\top\boldsymbol{W}_K^\top\boldsymbol{Y}^\top}_{\text{content}} + \boldsymbol{X}\boldsymbol{W}_Q\boldsymbol{b}_K\boldsymbol{1}_{1\times T} + \boldsymbol{1}_{T\times T}\boldsymbol{b}_Q^\top\boldsymbol{b}_K & (3)\end{aligned}$$

The reader can notice that the last two terms of eq. (3) have a constant contribution over all entries of the same row. As softmax is shift invariant,[2] these terms have no contribution to the computed attention probabilities. The first two terms (the only relevant ones) have a clear meaning: the first computes the attention from the context (between all key and query pairs), whereas the second computes attention solely based on key content. In particular, we show that $\boldsymbol{b}_K$, i.e. the bias of the key layer, is useless and can be disabled in all current implementations of attention layers.

[2] $\text{softmax}(\boldsymbol{x} + c) = \text{softmax}(\boldsymbol{x}),\ \forall c$

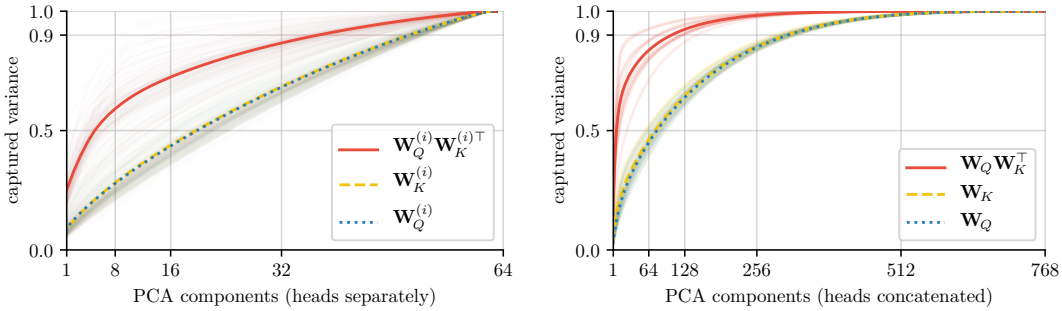


Figure 1: Cumulative captured variance of the key query matrices per head separately (*left*) and per layer with concatenated heads (*right*). Matrices are taken from a pre-trained BERT-base model with $N_h = 12$ heads of dimension $d_k = 64$. Bold lines show the means. Even though, by themselves, heads are not low rank (*left*), the product of their concatenation $\boldsymbol{W}_Q\boldsymbol{W}_K^\top$ is low rank (*right, in red*). Hence, the heads are sharing common projections in their column-space.

2.3 Multi-Head Attention

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

$$\text{MultiHead}(\boldsymbol{X}, \boldsymbol{Y}) = \underset{i\in[N_h]}{\text{concat}} \left[\boldsymbol{H}^{(i)}\right] \boldsymbol{W}^O \tag{4}$$

$$\boldsymbol{H}^{(i)} = \text{Attention}(\boldsymbol{X}\boldsymbol{W}_Q^{(i)}, \boldsymbol{Y}\boldsymbol{W}_K^{(i)}, \boldsymbol{Y}\boldsymbol{W}_V^{(i)}), \tag{5}$$

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 $i \in [N_h]$ and the extra parameter matrix $\boldsymbol{W}^O \in \mathbb{R}^{N_h d_{out}\times D_{out}}$ projects the concatenation of the N_h head outputs (each in $\mathbb{R}^{d_{out}}$) to the output space $\mathbb{R}^{D_{out}}$. In the multi-head setting, we call d_k the dimension of each head and $D_k = N_h d_k$ the total dimension of the query/key space.

3 Improving the Multi-Head Mechanism

Head concatenation is a simple and remarkably practical setup that gives empirical improvements. However, we show that another path could have been taken instead of concatenation. As the multiple heads are inherently solving similar tasks, they can collaborate instead of being independent.

3.1 How much do heads have in common?

We hypothesize that some heads might attend on similar features in the input space, for example computing high attention on the verb of a sentence or extracting some dimensions of the positional encoding. To verify this hypothesis, it does not suffice to look at the similarity between query (or key) matrices $\{\boldsymbol{W}_Q^{(i)}\}_{i\in[N_h]}$ of different heads. To illustrate this issue, consider the case where two heads are computing the same key/query representations up to a rotation matrix $\boldsymbol{R} \in \mathbb{R}^{d_k\times d_k}$ such that

$$\boldsymbol{W}_Q^{(2)} = \boldsymbol{W}_Q^{(1)}\boldsymbol{R} \quad \text{and} \quad \boldsymbol{W}_K^{(2)} = \boldsymbol{W}_K^{(1)}\boldsymbol{R}.$$

Even though the two heads are computing identical attention scores, i.e. $\boldsymbol{W}_Q^{(1)}\boldsymbol{R}\boldsymbol{R}^\top\boldsymbol{W}_K^{(1)\top} = \boldsymbol{W}_Q^{(1)}\boldsymbol{W}_K^{(1)\top}$, they can have orthogonal column-spaces and the concatenation $[\boldsymbol{W}_Q^{(1)}, \boldsymbol{W}_Q^{(2)}] \in \mathbb{R}^{D_{in}\times 2d_k}$ can be full rank.

To disregard artificial differences due to common rotations or scaling of the key/query spaces, we study the similarity of the product $\boldsymbol{W}_Q^{(i)}\boldsymbol{W}_K^{(i)\top} \in \mathbb{R}^{D_{in}\times D_{in}}$ across heads. Figure 1 shows the captured energy by the principal components of the key, query matrices and their product. It can be seen on the left that single head key/query matrices $\boldsymbol{W}_Q^{(i)}\boldsymbol{W}_K^{(i)\top}$ are not low rank on average. However, as seen on the right, even if parameter matrices taken separately are not low rank, their concatenation is

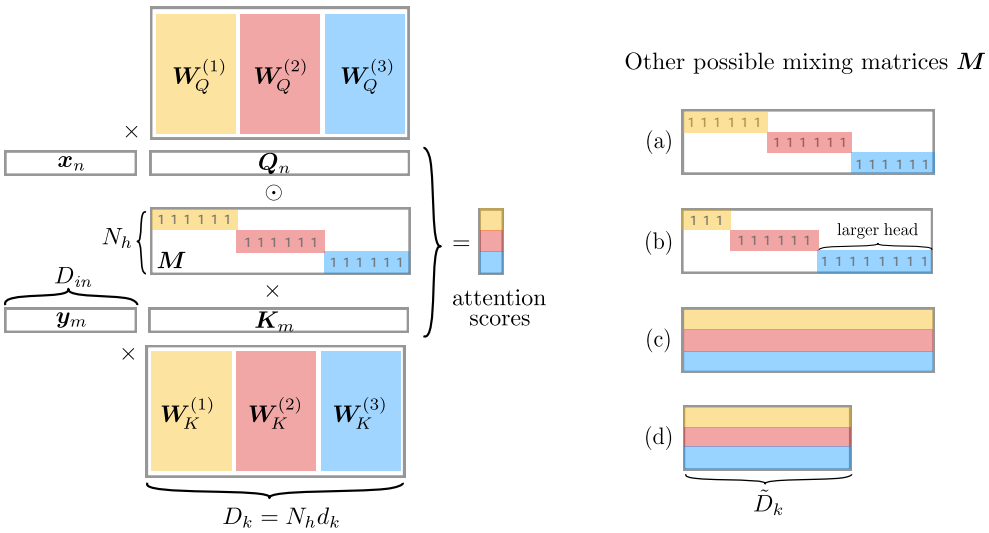


Figure 2: *Left*: computation of the attention scores between tokens $\boldsymbol{x}_n$ and $\boldsymbol{y}_m$ using a standard concatenated multi-head attention with $N_h = 3$ independent heads. The block structure of the mixing matrix $\boldsymbol{M}$ enforces that each head dot products non overlapping dimensions. *Right*: we propose to use more general mixing matrices than (a) heads concatenation, such as (b) allowing heads to have different sizes; (c) sharing heads projections; (d) compressing the number of projections from D_k to $\tilde{D}_k$ as heads can share redundant projections.

indeed low rank. This means that heads, though acting independently, learn to focus on the same subspaces. The phenomenon is quite pronounced: one third of the dimensions suffices to capture almost all the energy of $\boldsymbol{W}_Q \boldsymbol{W}_K^\top$, which suggests that there is inefficiency in the way multi-head attention currently operate.

3.2 Collaborative Multi-Head Attention

Following the observation that heads' key/query projections learn redundant projections, we propose to learn key/query projections for all heads at once and to let each head use a re-weighting of these projections. Our collaborative head attention is defined as follows:

$$\text{CollabHead}(\boldsymbol{X}, \boldsymbol{Y}) = \underset{i \in [N_h]}{\text{concat}} \left[\boldsymbol{H}^{(i)} \right] \boldsymbol{W}_O \tag{6}$$

$$\boldsymbol{H}^{(i)} = \text{Attention}(\boldsymbol{X} \tilde{\boldsymbol{W}}_Q \,\text{diag}(\boldsymbol{m}_i), \boldsymbol{Y} \tilde{\boldsymbol{W}}_K, \boldsymbol{Y} \boldsymbol{W}_V^{(i)}) \,. \tag{7}$$

The main difference with standard multi-head attention defined in eq. (5) is that we do not duplicate the key and query matrices for each head. Instead, each head learns a mixing vector $\boldsymbol{m}_i \in \mathbb{R}^{\tilde{D}_k}$ that defines a custom dot product over the $\tilde{D}_k$ projected dimensions of the shared matrices $\tilde{\boldsymbol{W}}_Q$ and $\tilde{\boldsymbol{W}}_K$ of dimension $D_{in} \times \tilde{D}_k$. This approach leads to:

(i) adaptive head expressiveness, with heads being able to use more or fewer dimensions according to attention pattern complexity;

(ii) parameter efficient representation, as learned projections are shared between heads, hence stored and learned only once.

It is instructive to observe how standard multi-head attention (where heads are simply concatenated) can be seen as a special case of our collaborative framework (with $\tilde{D}_k = N_h d_k$). The left of Figure 2 displays the standard attention computed between $\boldsymbol{x}_n$ and $\boldsymbol{y}_m$ input vectors with the mixing matrix

$$\boldsymbol{M} := \underset{i \in [N_h]}{\text{concat}} \left[\boldsymbol{m}_i \right] \in \mathbb{R}^{N_h \times \tilde{D}_k} \,, \tag{8}$$

laying out the mixing vectors $\boldsymbol{m}_i$ as rows. In the concatenated MHA, the mixing vector $\boldsymbol{m}_i$ for the i-th head is a vector with ones aligned with the d_k dimensions allocated to the i-th head among the $D_k = N_h d_k$ total dimensions.

Some alternative collaborative schema can be seen on the right side of Figure 2. By learning the mixing vectors $\{\boldsymbol{m}_i\}_{i\in[N_h]}$ instead of fixing them to this "blocks-of-1" structure, we increase the expressive power of each head for a negligible increase in the number of parameters. The size d_k of each head, arbitrarily set to 64 in most implementations, is now adaptive and the heads can attend to a smaller or bigger subspace if needed.

3.3 Head Collaboration as Tensor Decomposition

As we show next, there is a simple way to convert any standard attention layer to collaborative attention without retraining. To this end, we must extract the common dimensions between query/key matrices $\{\boldsymbol{W}_Q^{(i)}\boldsymbol{W}_K^{(i)\top} \in \mathbb{R}^{D_{in}\times D_{in}}\}_{i\in[N_h]}$ across the different heads. This can be solved using the Tucker tensor decomposition (Tucker, 1966) of the 3rd-order tensor

$$\mathbf{W}_{QK} := \underset{i\in[N_h]}{\text{stack}}\left[\boldsymbol{W}_Q^{(i)}\boldsymbol{W}_K^{(i)\top}\right] \in \mathbb{R}^{N_h\times D_{in}\times D_{in}}. \tag{9}$$

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

$$\mathbf{T} \approx \mathbf{G}\times_1 \boldsymbol{A}\times_2 \boldsymbol{B}\times_3 \boldsymbol{C} = \sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{r=1}^{R} g_{pqr}\,\boldsymbol{a}_p\circ\boldsymbol{b}_q\circ\boldsymbol{c}_r =: [\![\mathbf{G};\boldsymbol{A},\boldsymbol{B},\boldsymbol{C}]\!], \tag{10}$$

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 $\boldsymbol{c}_r$.

In the case of attention, it suffices to consider the dot product of the aligned key/query components of the $\boldsymbol{Q}$ and $\boldsymbol{K}$ matrices, which means that the core tensor is super-diagonal (i.e. $g_{pqr}\neq 0$ only if $q=r$). We further simplify the Tucker decomposition by setting the factors dimensions P, Q and R to $\tilde{D}_k$, a single interpretable hyperparameter equal to the dimension of the *shared* key/query space that controls the amount of compression of the decomposition into collaborative heads. These changes lead to a special case of Tucker decomposition called the canonical decomposition, also known as CP or PARAFAC (Harshman, 1970) in the literature (Kolda and Bader, 2009). Fix any positive rank R. The decomposition yields:

$$\mathbf{T} \approx \sum_{r=1}^{R}\boldsymbol{a}_r\circ\boldsymbol{b}_r\circ\boldsymbol{c}_r =: [\![\boldsymbol{A},\boldsymbol{B},\boldsymbol{C}]\!], \tag{11}$$

with $\boldsymbol{A}\in\mathbb{R}^{I\times R}$, $\boldsymbol{B}\in\mathbb{R}^{J\times R}$ and $\boldsymbol{C}\in\mathbb{R}^{K\times R}$.

What is remarkable is that the above can be used to express any (trained) attention layer parametrized by $\{\boldsymbol{W}_Q^{(i)}, \boldsymbol{b}_Q^{(i)}, \boldsymbol{W}_K^{(i)}, \boldsymbol{b}_K^{(i)}\}_{i\in[N_h]}$ as a collaborative layer. In particular, if we apply the decomposition to the stacked heads, i.e. $\mathbf{W}_{QK} \approx [\![\boldsymbol{M}, \tilde{\boldsymbol{W}}_Q, \tilde{\boldsymbol{W}}_K]\!]$, we obtain the three matrices that define a collaborative attention layer: the mixing matrix $\boldsymbol{M}\in\mathbb{R}^{N_h\times\tilde{D}_k}$, as well as the key and query projection matrices $\tilde{\boldsymbol{W}}_Q$, $\tilde{\boldsymbol{W}}_K \in \mathbb{R}^{D_{in}\times\tilde{D}_k}$.

On the other hand, biases can be easily dealt with based on the content/context decomposition of eq. (3), by storing for each head the vector

$$\boldsymbol{v}_i = \boldsymbol{W}_K^{(i)}\boldsymbol{b}_Q^{(i)} \in \mathbb{R}^{D_{in}}. \tag{12}$$

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

$$\left(\boldsymbol{X}\boldsymbol{W}_Q^{(i)} + \mathbf{1}_{T\times 1}\boldsymbol{b}_Q^\top\right)\left(\boldsymbol{Y}\boldsymbol{W}_K^{(i)} + \mathbf{1}_{T\times 1}\boldsymbol{b}_K^\top\right)^\top \approx \boldsymbol{X}\tilde{\boldsymbol{W}}_Q\,\mathrm{diag}(\boldsymbol{m}_i)\tilde{\boldsymbol{W}}_K^\top\boldsymbol{Y}^\top + \mathbf{1}_{T\times 1}\boldsymbol{v}_i^\top\boldsymbol{Y}^\top, \tag{13}$$

where $\boldsymbol{m}_i$ is the i-th row of $\boldsymbol{M}$. If $\tilde{D}_k \geq D_k$ the decomposition is exact (eq. (11) is an equality) and our collaborative heads layer can express any concatenation-based attention layer. We also note that the proposed re-parametrization can be applied to the attention layers of many transformer architectures, such as the ones proposed by Devlin et al. (2019); Sanh et al. (2019); Lan et al. (2020).

[3] $\circ$ represents the vector outer product

3.4 Parameter and Computation Efficiency

Table 1: Parameters and FLOPS gained in §4 with negligible performance loss.

	train		re-param.	
	concat.	collab.	concat.	collab.
$D_k \to \tilde{D}_k$	$1024 \to 128$		$768 \to 256$	
Params ($\times 10^6$)	2.1	0.26	1.2	0.40
FLOPS ($\times 10^6$)	2.1	0.27	1.2	0.40

Collaborative MHA introduces weight sharing across the key/query projections and decreases the number of parameters and FLOPS. While the size of the heads in the standard attention layer is set to $d_k = 64$ and the key/query layers project into a space of dimension $D_k = N_h d_k$, the shared key/query dimension $\tilde{D}_k$ of collaborative MHA can be set freely. According to our experiments in Section 4 (summarized in Table 1), a good rule of thumb when transforming a *trained* MHA layer to collaborative is to set $\tilde{D}_k$ to half or one third of D_k. When training from scratch, $\tilde{D}_k$ can even be set to $1/8$-th of D_k.

Parameters. Collaborative heads use $(2D_{in} + N_h)\tilde{D}_k$ parameters, as compared to $2D_{in}D_k$ in the standard case (ignoring biases). Hence, the compression ratio can be controlled by the shared key dimension $\tilde{D}_k$. The factorization introduces a new matrix $\boldsymbol{M}$ of dimension $N_h \times \tilde{D}_k$. Nevertheless, as the number of heads is small compared to the hidden dimension (in BERT-base $N_h = 12$ whereas $D_{in} = 768$), the extra parameter matrix yields a negligible increase as compared to the size of the query/key/values matrices of dimension $D_{in} \times D_k$.

Computational cost. Our layer decomposes two matrices into three, of modulable dimensions. To compute the attention scores between two tokens $\boldsymbol{x}_n$ and $\boldsymbol{y}_m$ for all the N_h heads, collaborative MHA requires $2(D_{in} + N_h)\tilde{D}_k$ FLOPS, while the concatenation-based uses $(2D_{in} + 1)D_k$ FLOPS. Assuming that $D_{in} \gg N_h = \mathcal{O}(1)$ (as is common in most implementations), we obtain a speedup of $\Theta(D_k/\tilde{D}_k)$.

4 Experiments

The goal of our experimental section is two fold. First, we show that concatenation-based MHA is a drop-in replacement for collaborative MHA in the transformer architecture. We obtain significant reduction in the number of parameters and number of FLOPS without sacrificing performance on a Neural Machine Translation (NMT) task with an encoder-decoder transformer. Secondly, we verify that our tensor decomposition allows one to reparametrize *pre-trained* transformers, such as BERT (Devlin et al., 2019) and its variants. To this end, we show that collaborative MHA performs on par with its concatenation-based counter-part on the GLUE benchmark (Wang et al., 2018) for Natural Language Understanding (NLU) tasks, even without retraining.

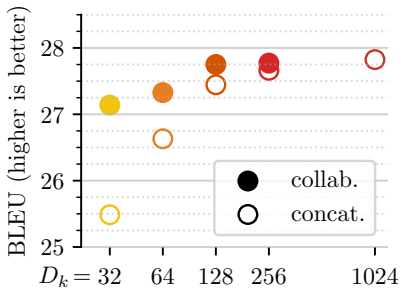
The NMT experiments use the MLPerf (Verma et al., 2019) encoder-decoder transformer code. For the NLU experiments, we implemented the collaborative MHA layer as an extension of the Transformers library (Wolf et al., 2019). The flexibility of our layer allows it to be applied to most of the existing transformer architectures, either at pre-training or after fine-tuning using tensor decomposition. We use the tensor decomposition library Tensorly (Kossaifi et al., 2019) with the PyTorch backend (Paszke et al., 2017) to reparameterize pre-trained attention layers. Our code and datasets are public[4] and all hyperparameters are specified in the Appendix.

4.1 Collaborative MHA for Neural Machine Translation

We replace the concatenation-based MHA layers of an encoder-decoder transformer by our collaborative MHA and evaluate it on the WMT 2017 English-to-German translation task. Results are shown in Figure 3. The original model uses $N_h = 16$ heads and $D_k = 1024$ key/query total dimensions and achieves a 27.8 BLUE score.

As observed in the original paper by Vaswani et al. (2017), decreasing the key/query head size d_k degrades the performance (*circles* in Figure 3). However, with collaborative heads (*disks* in Figure 3), the shared key/query dimension can be reduced by a factor 8 without decreasing the BLEU score. This translates to a linear decrease of the number of parameters and number of FLOPS for the computation of the attention scores at every layer. When we set an extreme total key/query dimension of $D_k = 32$, corresponding to $d_k = 2$ dimensions per head, the classic MHA model suffers a drop of

[4] `https://github.com/epfml/collaborative-attention`



D_k		concat.	collab.
1024	$\times 1$	**27.8**	-
256	$\times 4$	27.7	**27.8**
128	$\times 8$	27.4	**27.8**
64	$\times 16$	26.6	27.3
32	$\times 32$	25.5	27.1

Figure 3: Comparison of the BLEU score on WMT17 English German translation task for an encoder-decoder transformer (Vaswani et al., 2017) using collaborate vs. concatenate heads with key/query dimension D_k. With collaborative heads, D_k can be decreased by a factor of $\times 8$ without any drop in performance.

2.3 BLEU points, meanwhile the collaborative MHA performance only drops by less than one point. We conclude that sharing key/query projections across heads allows some attention features to be learned and stored only once. This weight sharing enables decreasing D_k without sacrificing heads' expressiveness.

4.2 Re-parametrize a Pre-trained MHA into Collaborative MHA

We turn to experiments on Natural Language Understanding (NLU) tasks, where transformers have been decisive in improving the state-of-the-art. As pre-training on large text corpora remains an expensive task, we leverage the post-hoc re-parametrization introduced in Section 3.3 to cast already pre-trained models into their collaborative form. We proceed in 3 steps for each GLUE task (Wang et al., 2018). First, we take a pre-trained transformer and fine-tune it on each task individually. Secondly, we replace all the attention layers by our collaborative MHA using tensor decomposition to compute $\tilde{\boldsymbol{W}}_Q$, $\tilde{\boldsymbol{W}}_K$ and $\boldsymbol{M}$ and re-parametrize the biases into $\boldsymbol{v}$. This step only takes a few minutes as shown in Figure 4. Finally, we fine-tune the compressed model again and evaluate its performance.

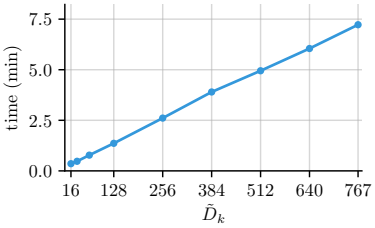


Figure 4: Time to decompose BERT-base from $D_k = 768$ to $\tilde{D}_k$.

We experiment with a pre-trained BERT-base model (Devlin et al., 2019). We also repurpose two variants of BERT designed to be more parameter efficient: ALBERT (Lan et al., 2020), an improved

Table 2: Performance of collaborative MHA on the GLUE benchmark (Wang et al., 2018). We report the median of 3 runs for BERT (Devlin et al., 2019), DistilBERT (Sanh et al., 2019) and ALBERT (Lan et al., 2020) with collaborative heads and different compression controlled by $\tilde{D}_k$. Comparing the original models with their compressed counter part shows that the number of parameters can be decreased with less than 1.5% performance drop (*gray rows*).

Model	$\tilde{D}_k$	$D_k/\tilde{D}_k$	CoLA	SST-2	MRPC	STS-B	QQP	MNLI	QNLI	RTE	**Avg.**
BERT-base	-		54.7	91.7	88.8/83.8	88.8/88.7	87.6/90.8	84.1	90.9	63.2	83.0
	768	$\times 1$	56.8	90.1	89.6/85.1	89.2/88.9	86.8/90.2	83.4	90.2	65.3	83.2
	384	$\times 2$	56.3	90.7	87.7/82.4	88.3/88.0	86.3/90.0	83.0	90.1	65.3	82.5
	256	$\times 3$	52.6	90.1	88.1/82.6	87.5/87.2	85.9/89.6	82.7	89.5	62.5	81.7
	128	$\times 6$	43.5	89.5	83.4/75.2	84.5/84.3	81.1/85.8	79.4	86.7	60.7	77.6
DistilBERT	-		46.6	89.8	87.0/82.1	84.0/83.7	86.2/89.8	81.9	88.1	60.3	80.0
	384	$\times 2$	45.6	89.2	86.6/80.9	81.7/81.9	86.1/89.6	81.1	87.0	60.7	79.1
ALBERT	-		58.3	90.7	90.8/87.5	91.2/90.8	87.5/90.7	85.2	91.7	73.7	85.3
	512	$\times 1.5$	51.1	86.0	91.4/88.0	88.6/88.2	87.2/90.4	84.2	90.2	69.0	83.1
	384	$\times 2$	40.7	89.6	82.3/71.1	86.0/85.6	87.2/90.5	84.4	90.0	49.5	77.9

transformer with a single layer unrolled, and DistilBERT (Sanh et al., 2019) a smaller version of BERT trained with distillation. We report in Table 2 the median performance of 3 independent runs of the models on the GLUE benchmark (Wang et al., 2018).

We first control that tensor decomposition without compression ($\tilde{D}_k = D_k = 768$) does not alter performance: both BERT-base and its decomposition performs similarly with an average score of 83.0% and 83.2% respectively. We then experiment with compressed decomposition using a smaller $\tilde{D}_k$. Comparing the original models with their well performing compressed counterpart (gray rows) shows that the key/query dimension of BERT and DistilBERT can be divided by 2 and 3 respectively without sacrificing more than 1.5% of performance. This is especially remarkable given that DistilBERT was designed to be a parameter efficient version of BERT. It seems that ALBERT suffers more from compression, but the dimension can be reduced by a factor 1.5 with minor performance degradation. We suspect that unrolling the same attention layer over the depth of the transformer forces the heads to use different projections and decreases their overlap, decreasing the opportunity for weight-sharing. Our hypothesis is that better performance may be obtained by pre-training the whole architecture from scratch.

Recovering from compression with fine-tuning. We further investigate the necessity of the second fine-tuning—step 3 of our experimental protocol—after the model compression. Figure 5 shows the performance of BERT-base on 3 GLUE tasks for different compression parameters $\tilde{D}_k$ with and without the second fine-tuning. We find that when the compression ratio is less than a third (from $D_k = 768$ to $\tilde{D}_k = 512$), the re-parametrization is accurate and performance is maintained without fine-tuning again. Further compressing the model starts to affect performance. However, up to two third of compression (to $\tilde{D}_k = 256$), this loss can be recovered by a second fine-tuning (*in orange*).

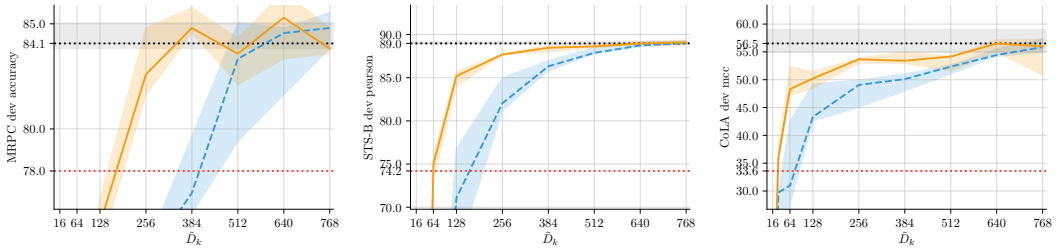


Figure 5: Performance on MRPC, STS-B and CoLA datasets of a fine-tuned BERT-base model, decomposed with collaborative heads of compressed dimension $\tilde{D}_k$ (*horizontal axis*). Repeating fine-tuning after compression can make the model recover the original performance when compression was drastic. The GLUE baseline gives a reference for catastrophic failure.

5 Conclusion

This work showed that trained concatenated heads in multi-head attention models can extract redundant query/key representations. To mitigate this issue, we propose to replace concatenation-based MHA by collaborative MHA. When our layer is used as a replacement for standard MHA in encoder/decoder transformers for Neural Machine Translation, it enables the decrease of effective individual head size from $d_k = 64$ to 8 without impacting performance. Further, without pre-training from scratch, switching a MHA layer to collaborative halves the number of FLOPS and parameters needed to compute the attentions score affecting the GLUE score by less than 1.5%.

Our model can impact every transformer architecture and our code (publicly available) provides post-hoc compression of already trained networks. We believe that using collaborative MHA in models pre-trained from scratch could force heads to extract meaningful shared query/key features. We are curious if this would translate to faster pre-training, better performance on downstream tasks and improved interpretability of the attention mechanism.

Acknowledgments

We acknowledge the support of Google Cloud for computational credits. Jean-Baptiste Cordonnier is thankful to the Swiss Data Science Center (SDSC) for funding this work. Andreas Loukas would like to thank the Swiss National Science Foundation for supporting him in the context of the project "Deep Learning for Graph-Structured Data" (grant number PZ00P2 179981).

Broader Impact

We study a well established deep learning model in natural language processing: the transformer architecture. This model has already demonstrated positive impact in machine translation by connecting people across cultures and question answering by extracting knowledge from large text database, to cite only two applications. Nevertheless, the automation of such language tasks allows cost savings and offers valuable services to all smart phone owners, it can also result in job losses.

Large models are expensive to train (months of GPU computation) and have negative environmental impact. Our work proposes a method to re-parametrize already trained models into our novel framework avoiding pre-training and making our method accessible to most academic researchers. The computational burden of pre-training raises the question of a monopoly: only a few companies/countries have the budget to exploit these powerful models.

We improve the understanding of transformers by questioning a specific part of the model: the concatenation of multiple heads. Better understanding of such networks is crucial to improve their performance but also to explain and avoid catastrophic failure at inference. Our work serves the scientific community as it looks back into design decisions made solely on empirical evidence and considers other paths motivated by theory.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate.

Irwan Bello, Barret Zoph, Ashish Vaswani, Jonathon Shlens, and Quoc V. Le. 2019. Attention augmented convolutional networks. In *The IEEE International Conference on Computer Vision (ICCV)*.

Lukas Biewald. 2020. Experiment tracking with weights and biases. Software available from wandb.com.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.

Jean-Baptiste Cordonnier, Andreas Loukas, and Martin Jaggi. 2020. On the relationship between self-attention and convolutional layers. In *International Conference on Learning Representations*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Richard A. Harshman. 1970. Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multi-modal factor analysis. *UCLA Working Papers in Phonetics*, 16:1–84.

Tamara G. Kolda and Brett W. Bader. 2009. Tensor decompositions and applications. *SIAM Review*, 51(3):455–500.

Jean Kossaifi, Yannis Panagakis, Anima Anandkumar, and Maja Pantic. 2019. Tensorly: Tensor learning in python. *Journal of Machine Learning Research*, 20(26):1–6.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

Paul Michel, Omer Levy, and Graham Neubig. 2019. Are sixteen heads really better than one? In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 14014–14024. Curran Associates, Inc.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch.

Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jon Shlens. 2019. Stand-alone self-attention in vision models. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 68–80.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Yi Tay, Dara Bahri, Donald Metzler, Da-Cheng Juan, Zhe Zhao, and Che Zheng. 2020. Synthesizer: Rethinking self-attention in transformer models.

Ledyard Tucker. 1966. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31(3):279–311.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 5998–6008.

Snehil Verma, Qinzhe Wu, Bagus Hanindhito, Gunjan Jha, Eugene B. John, Ramesh Radhakrishnan, and Lizy K. John. 2019. Demystifying the mlperf benchmark suite.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808, Florence, Italy. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Supplementary Material

A Hyperparameters for Neural Machine Translation Experiments

Our implementation is based on Fairseq implementation Ott et al. (2019) in MLPerf Verma et al. (2019). We report in the following tables the specification of the architecture and the training hyperparameters.

Transformer architecture parameters	
Max source positions	256
Max target positions	256
Dropout	0.3
Attention dropout	0.1
Relu dropout	0.1
Encoder embed dim	1024
Encoder ffn embed dim	4096
Encoder layers	6
Encoder attention heads	16
Encoder learned pos	False
Decoder embed dim	1024
Decoder ffn embed dim	4096
Decoder layers	6
Decoder attention heads	16
Decoder learned pos	False
Share decoder input output embed	False
Share all embeddings	False

Optimization hyperparameters	
Number of epochs	9
Learning rate	1e-9
β_1, β_2	0.9, 0.98
Warmup steps	1000
Math mode	fp16
Init scale	2^7
Scale factor	2
Scale window	2000

B Hyperparameters for Natural Language Understanding Experiments

We use standard models downloadable from HuggingFace repository along with their configuration.

Models		
BERT-base	Devlin et al. (2019)	`bert-base-cased`
DistilBERT	Sanh et al. (2019)	`distilbert-base-cased`
ALBERT	Lan et al. (2020)	`albert-base-v2`

We use HuggingFace default hyperparameters for GLUE fine-tuning in all our runs. We train with a learning rate of $2 \cdot 10^{-5}$ for 3 epochs for all datasets except SST-2 and RTE where we train for 10 epochs. In preliminary experiments, we tried to tune the tensor decomposition tolerance hyperparameter among $\{10^{-6}, 10^{-7}, 10^{-8}\}$ but did not see significant improvement and kept the default 10^{-6} for all our experiments.

GLUE fine-tuning hyperparameters	
Number of epochs	3 for all tasks but 10 for SST-2 and RTE
Batch size	32
Learning rate	2e-5
Adam ϵ	1e-8
Max gradient norm	1
Weight decay	0
Decomposition tolerance	1e-6