A. Notation

Embedding size	d
Number of layers	l
Number of heads	h
Sequence length	n
Vocab size	v
Head size	d_p

B. Proofs

Proof of Theorem 2. For simplicity of notation, we drop the dependence on parameters d, h and d_p for functions $f_{\mathbf{W}}(\cdot)$ and $g_{\mathbf{V}}(\cdot)$ in the proof.

First let us rewrite the MultiHead and FixedMultiHead layers as follows. The MultiHead layer can be rewritten as

$$f_{\mathbf{W}}(\mathbf{X}) = \mathbf{W}_o \cdot \text{MultiHead}(\mathbf{X}) = \sum_{i=1}^{h} \mathbf{W}_o^i \mathbf{W}_v^i \mathbf{X} \cdot \text{Softmax}\left[(\mathbf{W}_k^i \mathbf{X})^T (\mathbf{W}_q^i \mathbf{X}) / \sqrt{\tfrac{d}{h}} \right],$$

where $\mathbf{W}_o^i$ are $d \times d/h$ matrices and $\mathbf{W}_v^i$, $\mathbf{W}_k^i$, and $\mathbf{W}_q^i$ are $d/h \times d$ matrices. We denote the collection of all parameter matrices as $\mathbf{W}$.

Similarly, rewrite the fixed head size attention layer as

$$g_{\mathbf{V}}(\mathbf{X}) = \mathbf{V}_o \cdot \text{FixedMultiHead}(\mathbf{X}) = \sum_{i=1}^{h} \mathbf{V}_o^i \mathbf{V}_v^i \mathbf{X} \cdot \text{Softmax}\left[(\mathbf{V}_k^i \mathbf{X})^T (\mathbf{V}_q^i \mathbf{X}) / \sqrt{d_p} \right],$$

where $\mathbf{V}_o^i \in \mathbb{R}^{d \times d_p}$, and $\mathbf{V}_v^i, \mathbf{V}_k^i, \mathbf{V}_q^i \in \mathbb{R}^{d_p \times d}$. Let $\mathbf{V}$ be the collection of all these matrices.

The outline of the proof is basically a case analysis: we divide possible values of $\mathbf{W}$ into three categories, and show in each case that there exists a $\mathbf{X}$ such that $f_{\mathbf{W}}(\mathbf{X}) \neq g_{\mathbf{V}}(\mathbf{X})$. Here are the three cases:

- **Case 1**: $\sum_{i=1}^h \mathbf{W}_o^i \mathbf{W}_v^i \neq \sum_{i=1}^h \mathbf{V}_o^i \mathbf{V}_v^i$.
- **Case 2**: $\sum_{i=1}^h \mathbf{W}_o^i \mathbf{W}_v^i = \sum_{i=1}^h \mathbf{V}_o^i \mathbf{V}_v^i$, and there exists $i \in \{1, \dots, h\}$ such that $\mathbf{U}/\sqrt{d_p} - (\mathbf{W}_k^i)^T (\mathbf{W}_q^i)/\sqrt{d/h}$ is not skew-symmetric.
- **Case 3**: $\sum_{i=1}^h \mathbf{W}_o^i \mathbf{W}_v^i = \sum_{i=1}^h \mathbf{V}_o^i \mathbf{V}_v^i$, and all $\mathbf{U}/\sqrt{d_p} - (\mathbf{W}_k^i)^T (\mathbf{W}_q^i)/\sqrt{d/h}$ are skew-symmetric.

Case 1. In the first case, we can choose any $\mathbf{v}$ such that $(\sum_{i=1}^h \mathbf{W}_o^i \mathbf{W}_v^i - \sum_{i=1}^h \mathbf{V}_o^i \mathbf{V}_v^i)\mathbf{v} \neq \mathbf{0}$. Choose $\mathbf{X} = \mathbf{v}\mathbf{1}^T = \begin{bmatrix} \mathbf{v} & \mathbf{v} & \dots & \mathbf{v} \end{bmatrix}$. Then, note that for any column stochastic matrix $\mathbf{P}$, we have $\mathbf{X}\mathbf{P} = \mathbf{X}$. Therefore,

$$\begin{aligned} &\sum_{i=1}^{h} \mathbf{W}_o^i \mathbf{W}_v^i \mathbf{X} \cdot \text{Softmax}\left[(\mathbf{W}_k^i \mathbf{X})^T (\mathbf{W}_q^i \mathbf{X}) / \sqrt{d/h} \right] - \sum_{i=1}^{h} \mathbf{V}_o^i \mathbf{V}_v^i \mathbf{X} \cdot \text{Softmax}\left[(\mathbf{V}_k^i \mathbf{X})^T (\mathbf{V}_q^i \mathbf{X}) / \sqrt{d_p} \right] \\ =&\sum_{i=1}^{h} \mathbf{W}_o^i \mathbf{W}_v^i \mathbf{X} - \sum_{i=1}^{h} \mathbf{V}_o^i \mathbf{V}_v^i \mathbf{X} = (\sum_{i=1}^{h} \mathbf{W}_o^i \mathbf{W}_v^i - \sum_{i=1}^{h} \mathbf{V}_o^i \mathbf{V}_v^i) \mathbf{v}\mathbf{1}^T \neq \mathbf{0}. \end{aligned}$$

Case 2. In cases where $\sum_{i=1}^h \mathbf{W}_o^i \mathbf{W}_v^i = \sum_{i=1}^h \mathbf{V}_o^i \mathbf{V}_v^i$, since $\sum_{i=1}^h \mathbf{V}_o^i \mathbf{V}_v^i$ is full rank by assumption and each $\mathbf{W}_o^i \mathbf{W}_v^i$ is at most rank d/h, it follows that all columns in $\mathbf{W}_o^i \in \mathbb{R}^{d \times d/h}$ must be linearly independent. Therefore, for any $\mathbf{v} \neq \mathbf{0}$, $\{\mathbf{W}_o^i \mathbf{W}_v^i \mathbf{v}, i = 1, \dots, h\}$ is a set of linearly independent vectors, because each $\mathbf{W}_o^i \mathbf{W}_v^i \mathbf{v}$ is a linear combination of d/h column vectors of $\mathbf{W}_o^i$ that are linearly independent of other column vectors in $\mathbf{W}_o^j$, $j \neq i$.

Now consider any $\mathbf{v} \in \mathbb{R}^d$, and $\mathbf{X} = \mathbf{v}\mathbf{e}_1^T$, where $\mathbf{e}_1 = (1, 0, \dots, 0) \in \mathbb{R}^n$. Define $\phi(t) = \exp(t)/(\exp(t) + n - 1)$. Then, we have

$$g_{\mathbf{V}}(\mathbf{X}) = \sum_{i=1}^{h} \mathbf{V}_o^i \mathbf{V}_v^i \mathbf{X} \cdot \text{Softmax}\left[{\mathbf{X}^T \mathbf{U} \mathbf{X}}/{\sqrt{d_p}}\right] = \sum_{i=1}^{h} \mathbf{V}_o^i \mathbf{V}_v^i \mathbf{X} \cdot \text{Softmax}\begin{bmatrix} \frac{\mathbf{v}^T \mathbf{U} \mathbf{v}}{\sqrt{d_p}} & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 \end{bmatrix}$$

$$= \left(\sum_{i=1}^{h} \mathbf{V}_o^i \mathbf{V}_v^i\right) \begin{bmatrix} \phi\left(\frac{\mathbf{v}^T \mathbf{U} \mathbf{v}}{\sqrt{d_p}}\right) \mathbf{v} & \frac{\mathbf{v}}{n} & \dots & \frac{\mathbf{v}}{n} \end{bmatrix} = \left(\sum_{i=1}^{h} \mathbf{W}_o^i \mathbf{W}_v^i\right) \begin{bmatrix} \phi\left(\frac{\mathbf{v}^T \mathbf{U} \mathbf{v}}{\sqrt{d_p}}\right) \mathbf{v} & \frac{\mathbf{v}}{n} & \dots & \frac{\mathbf{v}}{n} \end{bmatrix}.$$

Similarly, we can calculate

$$\begin{aligned} f_{\mathbf{W}}(\mathbf{X}) &= \sum_{i=1}^{h} \mathbf{W}_o^i \mathbf{W}_v^i \mathbf{X} \cdot \text{Softmax}\left[{(\mathbf{W}_k^i \mathbf{X})^T (\mathbf{W}_q^i \mathbf{X})}/{\sqrt{d/h}}\right] \\ &= \sum_{i=1}^{h} \mathbf{W}_o^i \mathbf{W}_v^i \begin{bmatrix} \phi\left(\frac{\mathbf{v}^T (\mathbf{W}_k^i)^T \mathbf{W}_q^i \mathbf{v}}{\sqrt{d/h}}\right) \mathbf{v} & \frac{\mathbf{v}}{n} & \dots & \frac{\mathbf{v}}{n} \end{bmatrix}. \end{aligned}$$

Notice that all the columns of $f_{\mathbf{W}}(\mathbf{X})$ and $g_{\mathbf{V}}(\mathbf{X})$, from the second columns to the last ones, are the same. We now compare the first columns:

$$f_{\mathbf{W}}(\mathbf{X})_{:,1} - g_{\mathbf{V}}(\mathbf{X})_{:,1} = \sum_{i=1}^{h} \left(\phi\left(\frac{\mathbf{v}^T (\mathbf{W}_k^i)^T \mathbf{W}_q^i \mathbf{v}}{\sqrt{d/h}}\right) - \phi\left(\frac{\mathbf{v}^T \mathbf{U} \mathbf{v}}{\sqrt{d_p}}\right) \right) \mathbf{W}_o^i \mathbf{W}_v^i \mathbf{v}.$$

Recall that for any $\mathbf{v} \neq \mathbf{0}$, $\mathbf{W}_o^i \mathbf{W}_v^i \mathbf{v}$ are linearly independent, so $f_{\mathbf{W}}(\mathbf{X})_{:,1} - g_{\mathbf{V}}(\mathbf{X})_{:,1} = \mathbf{0}$ if and only if all $\phi\left(\frac{\mathbf{v}^T (\mathbf{W}_k^i)^T \mathbf{W}_q^i \mathbf{v}}{\sqrt{d/h}}\right) - \phi\left(\frac{\mathbf{v}^T \mathbf{U} \mathbf{v}}{\sqrt{d_p}}\right)$ are zero. However, since there exists $i \in \{1, \dots, h\}$ such that $\mathbf{U}/\sqrt{d_p} - (\mathbf{W}_k^i)^T (\mathbf{W}_q^i)/\sqrt{d/h}$ is not skew-symmetric, we can choose $\mathbf{v}$ to be one that satisfies $\frac{\mathbf{v}^T (\mathbf{W}_k^i)^T \mathbf{W}_q^i \mathbf{v}}{\sqrt{d/h}} \neq \frac{\mathbf{v}^T \mathbf{U} \mathbf{v}}{\sqrt{d_p}}$, hence making $\phi\left(\frac{\mathbf{v}^T (\mathbf{W}_k^i)^T \mathbf{W}_q^i \mathbf{v}}{\sqrt{d/h}}\right) - \phi\left(\frac{\mathbf{v}^T \mathbf{U} \mathbf{v}}{\sqrt{d_p}}\right) \neq 0$, therefore $f_{\mathbf{W}}(\mathbf{X})_{:,1} - g_{\mathbf{V}}(\mathbf{X})_{:,1} \neq \mathbf{0}$.

Case 3. Now consider any $\mathbf{X} = \begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \mathbf{0} & \dots & \mathbf{0} \end{bmatrix}$, where $\mathbf{v}_1$ and $\mathbf{v}_2$ will be chosen later. Define $\phi_1(t_1, t_2) = \exp(t_1)/(\exp(t_1) + \exp(t_2) + n - 2)$, $\phi_2(t_1, t_2) = \exp(t_2)/(\exp(t_1) + \exp(t_2) + n - 2)$. Then, we have

$$g_{\mathbf{V}}(\mathbf{X}) = \sum_{i=1}^{h} \mathbf{V}_o^i \mathbf{V}_v^i \mathbf{X} \cdot \text{Softmax}\begin{bmatrix} \frac{\mathbf{v}_1^T \mathbf{U} \mathbf{v}_1}{\sqrt{d_p}} & \frac{\mathbf{v}_1^T \mathbf{U} \mathbf{v}_2}{\sqrt{d_p}} & 0 & \dots & 0 \\ \frac{\mathbf{v}_2^T \mathbf{U} \mathbf{v}_1}{\sqrt{d_p}} & \frac{\mathbf{v}_2^T \mathbf{U} \mathbf{v}_2}{\sqrt{d_p}} & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 0 \end{bmatrix}.$$

Therefore, the first column of $g_{\mathbf{V}}(\mathbf{X})$ can be written as

$$g_{\mathbf{V}}(\mathbf{X})_{:,1} = \left(\sum_{i=1}^{h} \mathbf{W}_o^i \mathbf{W}_v^i\right) \left[\phi_1\left(\frac{\mathbf{v}_1^T \mathbf{U} \mathbf{v}_1}{\sqrt{d_p}}, \frac{\mathbf{v}_2^T \mathbf{U} \mathbf{v}_1}{\sqrt{d_p}}\right) \mathbf{v}_1 + \phi_2\left(\frac{\mathbf{v}_1^T \mathbf{U} \mathbf{v}_1}{\sqrt{d_p}}, \frac{\mathbf{v}_2^T \mathbf{U} \mathbf{v}_1}{\sqrt{d_p}}\right) \mathbf{v}_2 \right].$$

Similarly, the first column of $f_{\mathbf{W}}(\mathbf{X})$ is

$$\begin{aligned} f_{\mathbf{W}}(\mathbf{X})_{:,1} = \sum_{i=1}^{h} \mathbf{W}_o^i \mathbf{W}_v^i \Bigg[&\phi_1\left(\frac{\mathbf{v}_1^T (\mathbf{W}_k^i)^T \mathbf{W}_q^i \mathbf{v}_1}{\sqrt{d/h}}, \frac{\mathbf{v}_2^T (\mathbf{W}_k^i)^T \mathbf{W}_q^i \mathbf{v}_1}{\sqrt{d/h}}\right) \mathbf{v}_1 + \\ &\phi_2\left(\frac{\mathbf{v}_1^T (\mathbf{W}_k^i)^T \mathbf{W}_q^i \mathbf{v}_1}{\sqrt{d/h}}, \frac{\mathbf{v}_2^T (\mathbf{W}_k^i)^T \mathbf{W}_q^i \mathbf{v}_1}{\sqrt{d/h}}\right) \mathbf{v}_2 \Bigg]. \end{aligned}$$

Since $\mathbf{U}/\sqrt{d_p} - (\mathbf{W}_k^1)^T(\mathbf{W}_q^1)/\sqrt{d/h}$ is skew-symmetric by assumption, we have $\mathbf{v}_1^T\left(\frac{\mathbf{U}}{\sqrt{d_p}} - \frac{(\mathbf{W}_k^1)^T(\mathbf{W}_q^1)}{\sqrt{d/h}}\right)\mathbf{v}_1 = 0$ for all $\mathbf{v}_1$. Recall that $\mathbf{U}$ is rank-d_p by assumption, so $\mathbf{U}/\sqrt{d_p} - (\mathbf{W}_k^1)^T(\mathbf{W}_q^1)/\sqrt{d/h}$ is at least rank $d_p - d/h \geq 1$, so we can choose any $\mathbf{v}_1$ such that $\left(\frac{\mathbf{U}}{\sqrt{d_p}} - \frac{(\mathbf{W}_k^1)^T(\mathbf{W}_q^1)}{\sqrt{d/h}}\right)\mathbf{v}_1 \neq \mathbf{0}$.

If both $\frac{\mathbf{U}}{\sqrt{d_p}}\mathbf{v}_1$ and $\frac{(\mathbf{W}_k^1)^T(\mathbf{W}_q^1)}{\sqrt{d/h}}\mathbf{v}_1$ are nonzero, We can always choose $\tilde{\mathbf{v}}_2$ such that $\tilde{\mathbf{v}}_2^T\left(\frac{\mathbf{U}}{\sqrt{d_p}}\right)\mathbf{v}_1 > 0$ and $\tilde{\mathbf{v}}_2^T\left(\frac{(\mathbf{W}_k^1)^T(\mathbf{W}_q^1)}{\sqrt{d/h}}\right)\mathbf{v}_1 < 0$. This means that if we choose $\mathbf{v}_2 = \alpha\tilde{\mathbf{v}}_2$ and scale $\alpha \to \infty$,

$$\phi_1\left(\frac{\mathbf{v}_1^T\mathbf{U}\mathbf{v}_1}{\sqrt{d_p}}, \frac{\mathbf{v}_2^T\mathbf{U}\mathbf{v}_1}{\sqrt{d_p}}\right) \to 0, \;\; \phi_2\left(\frac{\mathbf{v}_1^T\mathbf{U}\mathbf{v}_1}{\sqrt{d_p}}, \frac{\mathbf{v}_2^T\mathbf{U}\mathbf{v}_1}{\sqrt{d_p}}\right) \to 1,$$

$$\phi_1\left(\frac{\mathbf{v}_1^T(\mathbf{W}_k^1)^T\mathbf{W}_q^1\mathbf{v}_1}{\sqrt{d/h}}, \frac{\mathbf{v}_2^T(\mathbf{W}_k^1)^T\mathbf{W}_q^1\mathbf{v}_1}{\sqrt{d/h}}\right) \to \frac{\exp(\mathbf{v}_1^T(\mathbf{W}_k^1)^T\mathbf{W}_q^1\mathbf{v}_1/\sqrt{d/h})}{\exp(\mathbf{v}_1^T(\mathbf{W}_k^1)^T\mathbf{W}_q^1\mathbf{v}_1/\sqrt{d/h}) + n - 2},$$

$$\phi_2\left(\frac{\mathbf{v}_1^T(\mathbf{W}_k^1)^T\mathbf{W}_q^1\mathbf{v}_1}{\sqrt{d/h}}, \frac{\mathbf{v}_2^T(\mathbf{W}_k^1)^T\mathbf{W}_q^1\mathbf{v}_1}{\sqrt{d/h}}\right) \to 0.$$

Then, consider the difference $f_{\mathbf{W}}(\mathbf{X})_{:,1} - g_{\mathbf{V}}(\mathbf{X})_{:,1}$. Recall that for any $\mathbf{v}$, $\mathbf{W}_o^1\mathbf{W}_v^1\mathbf{v}$ is independent of $\{\mathbf{W}_o^i\mathbf{W}_v^i\mathbf{v}, i \neq 1\}$. This means that, to show $f_{\mathbf{W}}(\mathbf{X})_{:,1} - g_{\mathbf{V}}(\mathbf{X})_{:,1} \neq \mathbf{0}$, it suffices to show that

$$\left[\phi_1\left(\frac{\mathbf{v}_1^T(\mathbf{W}_k^1)^T\mathbf{W}_q^1\mathbf{v}_1}{\sqrt{d/h}}, \frac{\mathbf{v}_2^T(\mathbf{W}_k^1)^T\mathbf{W}_q^1\mathbf{v}_1}{\sqrt{d/h}}\right) - \phi_1\left(\frac{\mathbf{v}_1^T\mathbf{U}\mathbf{v}_1}{\sqrt{d_p}}, \frac{\mathbf{v}_2^T\mathbf{U}\mathbf{v}_1}{\sqrt{d_p}}\right)\right]\mathbf{W}_o^1\mathbf{W}_v^1\mathbf{v}_1 +$$
$$\left[\phi_2\left(\frac{\mathbf{v}_1^T(\mathbf{W}_k^1)^T\mathbf{W}_q^1\mathbf{v}_1}{\sqrt{d/h}}, \frac{\mathbf{v}_2^T(\mathbf{W}_k^1)^T\mathbf{W}_q^1\mathbf{v}_1}{\sqrt{d/h}}\right) - \phi_2\left(\frac{\mathbf{v}_1^T\mathbf{U}\mathbf{v}_1}{\sqrt{d_p}}, \frac{\mathbf{v}_2^T\mathbf{U}\mathbf{v}_1}{\sqrt{d_p}}\right)\right]\mathbf{W}_o^1\mathbf{W}_v^1\mathbf{v}_2 \neq \mathbf{0}.$$

If we scale $\mathbf{v}_2 = \alpha\tilde{\mathbf{v}}_2$ with large enough α, the second term will dominate the first term and the first term will never be able to cancel the second one. Thus, by choosing large enough $\alpha > 0$, we can make sure that the sum is nonzero.

Even in case where one of $\frac{\mathbf{U}}{\sqrt{d_p}}\mathbf{v}_1$ and $\frac{(\mathbf{W}_k^1)^T(\mathbf{W}_q^1)}{\sqrt{d/h}}\mathbf{v}_1$ is zero (say $\frac{(\mathbf{W}_k^1)^T(\mathbf{W}_q^1)}{\sqrt{d/h}}\mathbf{v}_1 = \mathbf{0}$), we can choose $\tilde{\mathbf{v}}_2 = \frac{\mathbf{U}}{\sqrt{d_p}}\mathbf{v}_1$ and use a similar scaling argument. By choosing large enough $\alpha > 0$ and $\mathbf{v}_2 = \alpha\tilde{\mathbf{v}}_2$, one can show that the difference $f_{\mathbf{W}}(\mathbf{X})_{:,1} - g_{\mathbf{V}}(\mathbf{X})_{:,1}$ is nonzero. □

C. Experimental settings

For our experiments with the language modeling (LM1B dataset), we train 6 layer Transformer models. We use a batch size of 4096 and train for 250k steps. We use a learning rate of 0.1 with a linear warm up for the first 10k steps. We decay the learning rate with the square root of the number of steps. We train the baseline models, with the prevalent head size heuristic, with the embedding dimension varying from 256 to 512. We fix the width of the feed forward layer in the Transformer to be 1024. In addition, we use weight decay of 0.01 and dropout with probability of 0.1 on all the layers.

For our experiments with BERT, we follow the same experimental settings as in (Devlin et al., 2018). We present the key details here and refer the reader to (Devlin et al., 2018). We train with a batch size of 1024 for 450k steps with inputs of sequence length n = 128 followed by 50k steps with inputs of sequence length 512. In contrast the BERT paper uses a batch size of 512, and does the pre-training for 900K steps with 128 sequence length inputs and 100k steps with 512 sequence length inputs. We train using ADAM with a learning rate of 1e-4, and a linear warmup and decay schedule as in BERT. We use 5k warmup steps for the first stage, and a re-warmup of 3k steps for the second stage (You et al., 2019). Again, we use weight decay of 0.01 and dropout with probability of 0.1 on all the layers.

For the language modeling task, training is performed on 4 TPUv2 chips for a couple of hours. For BERT models training is performed on 16 TPUv3 chips in the first stage and 64 TPUv3 chips for the second stage. Pre-training with this configuration takes between 2 to 3 days. We did not attempt to find the optimal hyper-parameters for the fixed head size architecture, and use the same hyper-parameters as used for training the BERT models.

# heads	8	12	16	20
# params	214M	252M	290M	327M
SQuAD - F1	90.35±0.14	90.48±0.09	90.92±0.14	90.89±0.08
SQuAD - EM	83.37±0.12	83.67±0.03	84.16±0.35	84.29±0.16
MNLI	84.4±0.2	84.4±0.2	84.7±0.1	85.1±0.4

(A) Increasing number of heads

Table 3: (A): 24 layer Transformer trained with a fixed head size of 128 and an embedding size of 768 shows an improvement in the accuracy with the increasing number of heads.

D. Additional experimental results

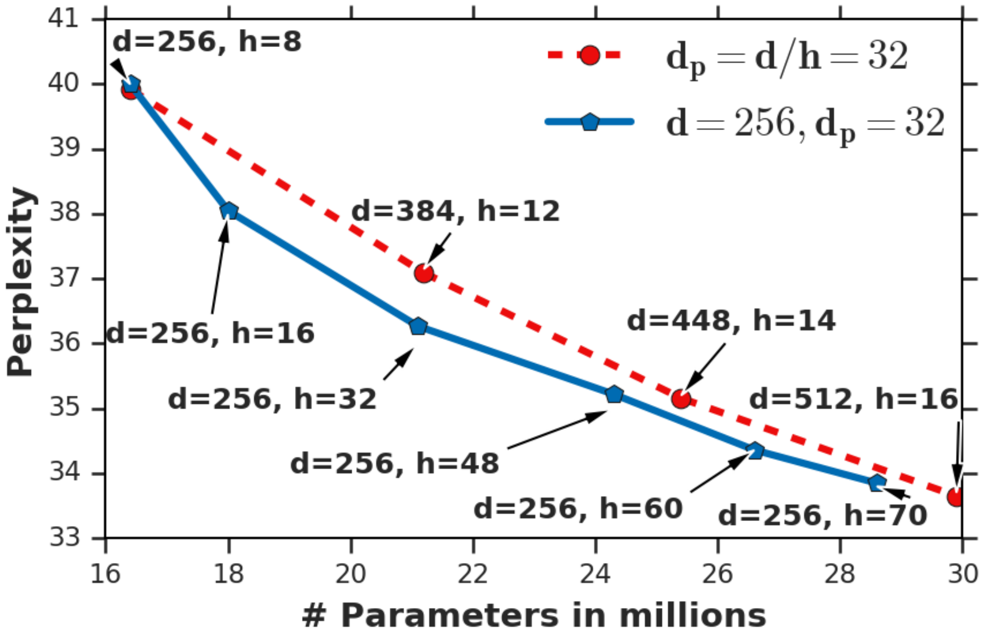

Figure 4: Performance of the Transformers trained with the prevalent head size heuristic (baseline) compared with the fixed head size (d_p) models for a language modeling task (LM1B) on the test set. Unlike Fig. 1, we vary both the embedding size and the number of heads of the baseline models to keep their head size fixed to 32. We train the fixed head size models with a fixed embedding size of 256 and a head size of 32, and vary the number of heads from 4 to 70, while matching the number of parameters. The plot again clearly indicates the advantage of the fixed head size models. The main issue with the baseline models is that fixing the head size to 32 forces the number of heads to be small when the embedding size is small. Reducing the number of heads below certain threshold hurts the performance of the Transformer.