A Mixture of $h - 1$ Heads is Better than h Heads

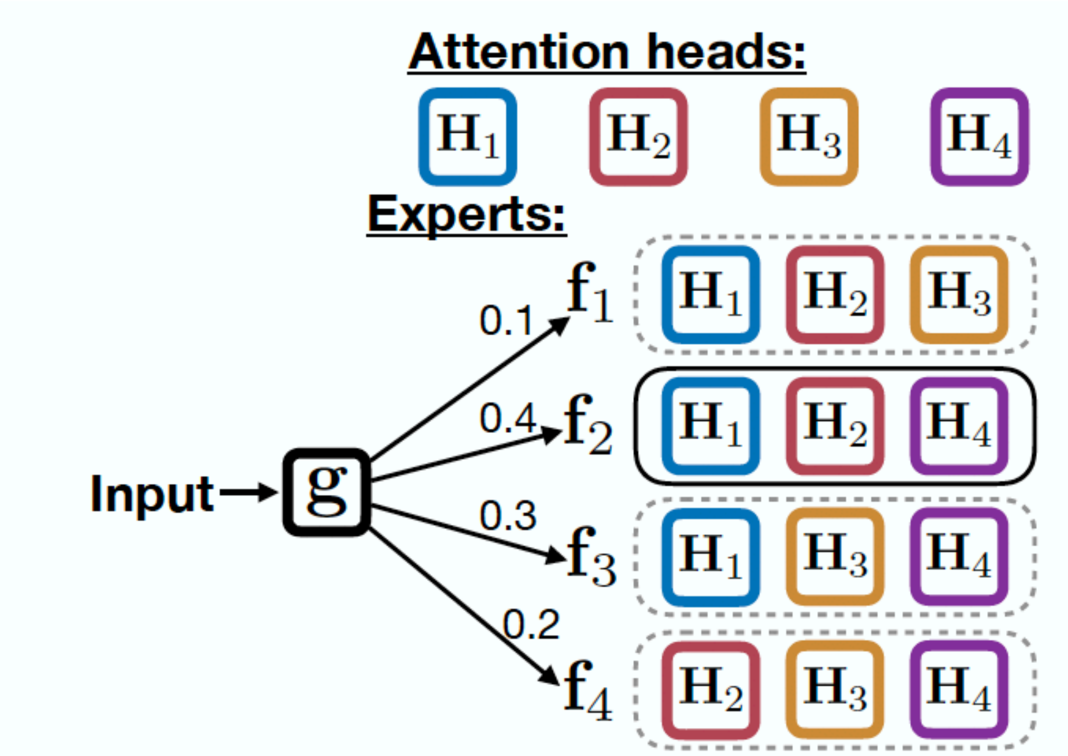
Hao Peng[♠] **Roy Schwartz**[♢♠] **Dianqi Li**[♣] **Noah A. Smith**[♢♠]

[♢]Allen Institute for Artificial Intelligence

[♠]Paul G. Allen School of Computer Science & Engineering, University of Washington

[♣]Department of Electrical & Computer Engineering, University of Washington

{hapeng,roysch,nasmith}@cs.washington.edu, dianqili@uw.edu


Attention heads:
H1 H2 H3 H4
Experts:
Input
g
0.1
0.4
0.3
0.2
f1 H1 H2 H3
f2 H1 H2 H4
f3 H1 H3 H4
f4 H2 H3 H4

$$\widetilde{\mathbf{H}}_i = \operatorname{softmax}\left(\mathbf{X}\mathbf{Q}_i\mathbf{K}_i^\top\mathbf{X}^\top\right)\mathbf{X}\mathbf{V}_i,$$

$$\mathbf{Z} \triangleq \operatorname{MultiHead}(\mathbf{X}) = \left[\widetilde{\mathbf{H}}_1; \ldots; \widetilde{\mathbf{H}}_h\right]\mathbf{W}$$

```
(0): BertLayer(
  (attention): BertAttention(
    (self): BertSelfAttention(
      (query): Linear(in_features=1024, out_features=1024, bias=True)
      (key): Linear(in_features=1024, out_features=1024, bias=True)
      (value): Linear(in_features=1024, out_features=1024, bias=True)
      (dropout): Dropout(p=0.1, inplace=False)
    )
    (output): BertSelfOutput(
      (dense): Linear(in_features=1024, out_features=1024, bias=True)
      (LayerNorm): LayerNorm((1024,), eps=1e-12, elementwise_affine=True)
      (dropout): Dropout(p=0.1, inplace=False)
    )
  )
  (intermediate): BertIntermediate(
    (dense): Linear(in_features=1024, out_features=4096, bias=True)
  )
  (output): BertOutput(
    (dense): Linear(in_features=4096, out_features=1024, bias=True)
    (LayerNorm): LayerNorm((1024,), eps=1e-12, elementwise_affine=True)
    (dropout): Dropout(p=0.1, inplace=False)
  )
)
```

tion into following sections. Let $\mathbf{H}_i = \widetilde{\mathbf{H}}_i \mathbf{W}_i$, where $\mathbf{W}_i$ is a block submatrix of $\mathbf{W}$, i.e., $\mathbf{W} = [\mathbf{W}_1^\top; \mathbf{W}_2^\top, \ldots; \mathbf{W}_h^\top]^\top$. Then

$$\mathbf{Z} = \left[\widetilde{\mathbf{H}}_1; \ldots; \widetilde{\mathbf{H}}_h\right] \mathbf{W} = \sum_{i=1}^{h} \mathbf{H}_i. \tag{4}$$

mixture-of-experts（MoE）

A mixture-of-experts perspective. Let us take a closer look at Eq. 4 and rewrite it:

$$\begin{aligned} \mathbf{Z} &= \frac{1}{h-1} \sum_{i=1}^{h} (-1+h)\, \mathbf{H}_i \\ &= \frac{1}{h-1} \left(-\sum_{i=1}^{h} \mathbf{H}_i + \sum_{i=1}^{h} \sum_{j=1}^{h} \mathbf{H}_j \right) \\ &= \sum_{i=1}^{h} \underbrace{\frac{1}{h}}_{\text{gate } g_i} \underbrace{\frac{h}{h-1} \left(-\mathbf{H}_i + \sum_{j=1}^{h} \mathbf{H}_j \right)}_{\text{expert } \mathbf{f}_i(\mathbf{X}; \boldsymbol{\theta}_i)} . \end{aligned} \tag{5}$$

A mixture-of-experts perspective. Let us take a closer look at Eq. 4 and rewrite it:

$$\begin{aligned}
\mathbf{Z} &= \frac{1}{h-1}\sum_{i=1}^{h}(-1+h)\ \mathbf{H}_i \\
&= \frac{1}{h-1}\left(-\sum_{i=1}^{h}\mathbf{H}_i + \sum_{i=1}^{h}\sum_{j=1}^{h}\mathbf{H}_j\right) \\
&= \sum_{i=1}^{h}\underbrace{\frac{1}{h}}_{\text{gate } g_i}\ \underbrace{\frac{h}{h-1}\left(-\mathbf{H}_i + \sum_{j=1}^{h}\mathbf{H}_j\right)}_{\text{expert } \mathbf{f}_i\left(\mathbf{X};\boldsymbol{\theta}_i\right)}.
\end{aligned} \tag{5}$$

Uniform => Gate (MLP+Softmax, condition on INPUT)

$$\sum_{i=1}^{h} g_i(\mathbf{X};\phi)\cdot\mathbf{f}_i(\mathbf{X};\boldsymbol{\theta}_i).$$

替换了Transformer Encoder/Decoder所有self-attention
所有block有自己用于控制Gate的MLP，参数量大概上升3%～5%
如果直接training，gate会逐渐收敛到uniform => Training in an interleaving way
Dropout?

Algorithm 1 A G step update for MAE, with step size η.

procedure MAEG($\mathbf{X}$)
  $\mathbf{Z} \leftarrow \sum_{i=1}^{h} g_i(\mathbf{X}; \boldsymbol{\phi}) \cdot \mathbf{f}_i(\mathbf{X}; \boldsymbol{\theta}_i)$
  Forwardprop with $\mathbf{Z}$ and calculate $\mathcal{L}$.
  Calculate $\nabla_{\boldsymbol{\phi}} \mathcal{L}$ with backprop.
  $\boldsymbol{\phi} \leftarrow \boldsymbol{\phi} - \eta \cdot \nabla_{\boldsymbol{\phi}} \mathcal{L}$.
end procedure

Algorithm 2 An F step update for MAE, with step size η.

procedure MAEF($\mathbf{X}$)
  Draw $i \sim \mathrm{Cat}(\mathbf{g}(\mathbf{X}; \boldsymbol{\phi}))$
  $\mathbf{Z} \leftarrow \mathbf{f}_i(\mathbf{X}; \boldsymbol{\theta}_i)$
  Forwardprop with $\mathbf{Z}$ and calculate $\mathcal{L}$.
  Calculate $\nabla_{\boldsymbol{\theta}_i} \mathcal{L}$ with backprop.
  $\boldsymbol{\theta}_i \leftarrow \boldsymbol{\theta}_i - \eta \cdot \nabla_{\boldsymbol{\theta}_i} \mathcal{L}$.
end procedure

Algorithm 3 Block coordinate descent (BCD) training for MAE, at epoch e. $\mathcal{D}$ denotes the training data.[8]

procedure BCD($\mathcal{D} = \{\mathbf{X}_i\}_i, e$)
  for $\mathbf{X}_i \in \mathcal{D}$ **do**
    ▷ Take G steps every 5 epochs.
    if $e \bmod 5 = 0$ **then**
      MAEG($\mathbf{X}_i$)
    end if
    ▷ Always do F step updates.
    MAEF($\mathbf{X}_i$)
  end for
end procedure

Experiments: NMT & LM

- MAE-7: paper提出的方法，8个expert
- MAE-6: h-2的版本，28个expert
- BASE：最普通Seq2Seq
- NOBCD：就是Joint，没有block下降
- UNI-MAE-7: gate是uniform，没有特别用神经网络得到
- UNI-MAE-6

Data	Train	Dev.	Test	Vocab.
WMT14	4.5M	3K	3K	32K
IWSLT14	160K	7K	7K	9K/7K

Table 1: Some statistics for WMT14 and IWSLT14 datasets. We use separate source and target vocabularies in IWSLT14 experiments.

Model	BLEU	# Params.
Base Transformer	27.3	65M
Large Transformer	28.4	213M
BASE	27.6	61M
‡NOBCD	27.5	63M
†UNI-MAE-7	27.7	61M
†UNI-MAE-6	27.6	61M
†‡MAE-7	**28.4**	63M
†‡MAE-6	28.1	63M

Table 2: WMT14 EN-DE translation test performance on newstest2014. † randomly select an expert to update for each training instance, and ‡ learns a gating function to weight the experts. Transformer performance in the first two rows are due to Vaswani et al. (2017).

Experiments: NMT & LM

Model	Perplexity	# Params.
*BASE (B&A, 2019)	18.70	247M
BASE (B&A, 2019)	19.03	247M
‡NOBCD	19.12	249M
†UNI-MAE-7	19.26	247M
†‡MAE-7	**18.71**	249M

Table 4: Language modeling performance on WikiText-103 test set (lower is better). ⋆Trains/evaluates with 3,072/2,048 context sizes and therefore not directly comparable to other models which use 512/480 sized ones. See Table 2 caption

为了证明这种ensemble看待问题的方式确实起作用

1. 计算gate的平均熵，发现MAE是最低的，说明它能够侧重选对应的某个expert来做。（MAE-7 => 1.91，NOBCD => 2.02, UNI-MAE-7 => 2.08)
2. 看了8个expert每一个gate的平均值，13/14/9/16/10/15/10/12%， 即不是全部uniform，也不是richer get richer
3. 如果每个layer都只使用gate最高的那个expert，而不是weighted sum

Model	**BLEU**	**Diff.**
UNI-MAE-7	26.6	-
One random expert	$25.8_{\pm 0.2}$	$\downarrow 0.8_{\pm 0.2}$
NOBCD	26.7	-
Most specialized expert	26.0	$\downarrow 0.7$
MAE-7	27.1	-
Most specialized expert	26.8	$\downarrow 0.3$

Table 5: Performance decrease for different models on WMT14 development set when only one expert is used for each multi-head attention layer (5.1).