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

Hao Peng, Roy Schwartz, Dianqi Li, Noah A. Smith

@ACL
July, 2020
A Mixture of $h-1$ Heads is Better Than $h$ Heads

Multihead attention
MAE: a mixture of attentive experts

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

Multihead attention
Attention

query

key

value

input $\mathbf{X}$
Attention

query

key

value

input $X$

softmax
Attention

query

key

value

input $X$

$h$
Multihead Attention

query

key

value

input $X$
Multihead Attention

\[
\begin{bmatrix}
    h_1^T \\
    h_2^T \\
    h_3^T \\
\end{bmatrix}^T
\]

output

\[\text{MultiHead}(x)\]
Multihead Attention: Overparameterization

Previous Attempt:
• Prune heads: Voita et al. (2019); Michel et al. (2019)
• Prune layers: Fan et al. (2020)
Previous Attempt:
• Prune heads: Voita et al. (2019); Michel et al. (2019)
• Prune layers: Fan et al. (2020)

Our approach:
• Activate different heads on different inputs
• View multihead attention as a mixture of experts (MoE)
• Applicable wherever multihead attention is used

This work: Mixture of Attentive Experts (MAE)
Mixture of Experts (MoE)

\[ \text{MoE}(x) = \sum_{i} g_i(x) \cdot f_i(x) \]
Multi-head Attention as MoE

MultiHead (x)

\[
\begin{bmatrix}
W_1 & W_2 & W_3
\end{bmatrix}
\begin{bmatrix}
h_1 \\
h_2 \\
h_3
\end{bmatrix}
\]
Multi-head Attention as MoE

MultiHead (x)

$$\begin{bmatrix} W_1 & W_2 & W_3 \\ h_1 \\ h_2 \\ h_3 \end{bmatrix} = \frac{1}{2} \begin{bmatrix} W_2 & W_3 \\ h_2 \\ h_3 \end{bmatrix} + \frac{1}{2} \begin{bmatrix} W_1 & W_3 \\ h_1 \\ h_3 \end{bmatrix} + \frac{1}{2} \begin{bmatrix} W_1 & W_2 \\ h_1 \\ h_2 \end{bmatrix}$$
Multi-head Attention as MoE

MultHead (x)

\[
\begin{bmatrix}
W_1 & W_2 & W_3 \\
\end{bmatrix}
\begin{bmatrix}
h_1 \\
h_2 \\
h_3 \\
\end{bmatrix}
= \frac{1}{2}
\begin{bmatrix}
W_2 & W_3 \\
\end{bmatrix}
\begin{bmatrix}
h_2 \\
h_3 \\
\end{bmatrix}
+ \frac{1}{2}
\begin{bmatrix}
W_1 & W_3 \\
\end{bmatrix}
\begin{bmatrix}
h_1 \\
h_3 \\
\end{bmatrix}
+ \frac{1}{2}
\begin{bmatrix}
W_1 & W_2 \\
\end{bmatrix}
\begin{bmatrix}
h_1 \\
h_2 \\
\end{bmatrix}
\]

gate \( g_1 \)
gate \( g_2 \)
gate \( g_3 \)

expert \( f_1(x) \)
expert \( f_2(x) \)
expert \( f_3(x) \)
Multi-head Attention as MoE:

- A mixture of \( \binom{h}{h-1} \) experts, each with \( h - 1 \) heads.
- Constant gates.

\[
\text{MultiHead}(x) = \frac{1}{2} \cdot f_1(x) + \frac{1}{2} \cdot f_2(x) + \frac{1}{2} \cdot f_3(x)
\]
Multi-head attention as MoE:
- A mixture of \( \binom{h}{h-1} \) experts, each with \( h-1 \) heads
- Constant gates

MAE:
- A mixture of \( \binom{h}{h-1} \) experts, each with \( h-1 \) heads.
- Learned gating function on the inputs
- Applicable wherever multihead attention is used
MAE: a Mixture of Attentive Experts

Multi-head attention as MoE:
- A mixture of $\binom{h}{h-1}$ experts, each with $h-1$ heads
- Constant gates

MAE:
- A mixture of $\binom{h}{h-1}$ experts, each with $h-1$ heads.
- Learned gating function on the inputs
- Applicable wherever multihead attention is used
MAE: Training

Undesired solutions:
• Equally weight the experts
• “The rich gets richer”
MAE: Training

Undesired solutions:
- Equally weight the experts
- “The rich gets richer”

Train with blockwise coordinate descent: alternate between
- Update the experts
- Update the gating function
Train with blockwise coordinate descent (BCD):

- **F step:**
  1. Randomly select expert $i$ according to

\[
g_i(x) / \sum g_j(x)
\]
  2. Fix gates, update expert $i$

- **G step:**
  1. Weight the experts by

\[
g_i(x) / \sum g_j(x)
\]
  2. Fix experts, update gates
MAE: Training

Train with blockwise coordinate descent (BCD):

• **F step:**
  1. Randomly select expert \( i \) according to
  \[ g_i (x) / \sum g_j (x) \]
  2. Fix gates, update expert \( i \)

• **G step:**
  1. Weight the experts by \( g_i (x) \)
  2. Fix experts, update gates

\[
g_1 (x) \cdot f_1 (x) + g_2 (x) \cdot f_2 (x) + g_3 (x) \cdot f_3 (x)
\]
MAE: Training

Train with blockwise coordinate descent (BCD):

- **F step:**
  1. Randomly select expert $i$ according to
     \[ g_i(x) / \sum g_j(x) \]
  2. Fix gates, update expert $i$

- **G step:**
  1. Weight the experts by $g_i(x)$
  2. Fix experts, update gates
MAE: Summary

- Applicable wherever multihead attention is used
- Alternate between updating experts and gates
- Gating functions implemented as MLPs over the inputs. Less than 5% additional parameters
- Connections to dropout in the paper
Experiments: Machine Translation

Dataset:

• WMT’14 EN->DE (Bojar et al., 2014). 4.5M training instances
• IWSLT’14 DE->EN (Cettolo et al., 2014). 160K training instances
Experiments: Machine Translation

Dataset:
- WMT’14 EN->DE (Bojar et al., 2014). 4.5M training instances
- IWSLT’14 DE->EN (Cettolo et al., 2014). 160K training instances

Implementation:
- Based on Vaswani et al. (2017)
- Comparable implementation and tuning
- More details in the paper
WMT'14 EN->DE test set BLEU

<table>
<thead>
<tr>
<th></th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base</td>
<td>27.6</td>
</tr>
<tr>
<td>MAE</td>
<td>28.4</td>
</tr>
</tbody>
</table>

Learned gate

BCD training
WMT'14 EN->DE test set BLEU

<table>
<thead>
<tr>
<th>Method</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base</td>
<td>27.6</td>
</tr>
<tr>
<td>MAE</td>
<td>28.4</td>
</tr>
<tr>
<td>Learned gate</td>
<td>27.5</td>
</tr>
<tr>
<td>- BCD</td>
<td>27.7</td>
</tr>
</tbody>
</table>

- Learned gate
- BCD training
IWSLT’14 DE->EN test set BLEU

<table>
<thead>
<tr>
<th></th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base</td>
<td>34.6</td>
</tr>
<tr>
<td>MAE</td>
<td>35.5</td>
</tr>
<tr>
<td>- learned gate</td>
<td>34.8</td>
</tr>
<tr>
<td>- BCD</td>
<td>34.9</td>
</tr>
</tbody>
</table>

Learned gate

BCD training
Experiments: Language Modeling

Dataset:
• WikiText-103 (Merity et al., 2016). 103M training data, 268K vocab size
Experiments: Language Modeling

Dataset:
• WikiText-103 (Merity et al., 2016). 103M training data, 268K vocab size

Implementation:
• Based on Baevski and Auli (2019)
• Comparable implementation and tuning
• More details in the paper
WikiText-103 test set perplexity

Lower is better

<table>
<thead>
<tr>
<th></th>
<th>Perplexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base</td>
<td>19.03</td>
</tr>
<tr>
<td>MAE</td>
<td>18.71</td>
</tr>
<tr>
<td>- learned gate</td>
<td>19.12</td>
</tr>
<tr>
<td>- BCD</td>
<td>19.26</td>
</tr>
</tbody>
</table>

Learned gate

BCD training
Other Results

- Mixture of $\binom{h}{h-2}$ experts, each with $h - 2$ heads
- Transfer learning by fine-tuning the gates
Other Results

- Mixture of $\binom{h}{h-2}$ experts, each with $h-2$ heads
- Transfer learning by fine-tuning the gates
- MAE learns to activate different experts on different inputs

Dev. instance clustered by experts they activate
MAE: a Mixture of Attentive Experts

- MoE perspective of multihead attention
- Learn to select different expert based on the inputs
- Strong performance on machine translation and language modeling
Thank You!
https://github.com/Noahs-ARK/MAE

• MoE perspective of multihead attention
• Learn to select different expert based on the inputs
• Strong performance on machine translation and language modeling