Sequence Models

Jordan Boyd-Graber

University of Maryland

Backpack Models (Attention)

Slides adapted from Joshua Wagner, Matthias Assenmacher, Jay Alammar

Plan for Today

- Transition from linear NLM structures
- What made Transformer Language Models “a thing”
  - Attention
  - Multi-layer, multi-head
  - Transformers
- Hint at why they generalize so well (representations)

History of attention

NEURAL MACHINE TRANSLATION
BY JOINTLY LEARNING TO ALIGN AND TRANSLATE

Dzmitry Bahdanau
Jacobs University Bremen, Germany

KyungHyun Cho **Yoshua Bengio***
Université de Montréal

How to represent final state given previous words:

$$c = q(\{h_1, \ldots, h_T\}) \tag{1}$$

New context vector is weighted sum of the hidden states:

$$c_t = \sum_{i=1}^{T_x} \alpha_{t,i} h_i. \tag{2}$$

Where the coefficients are

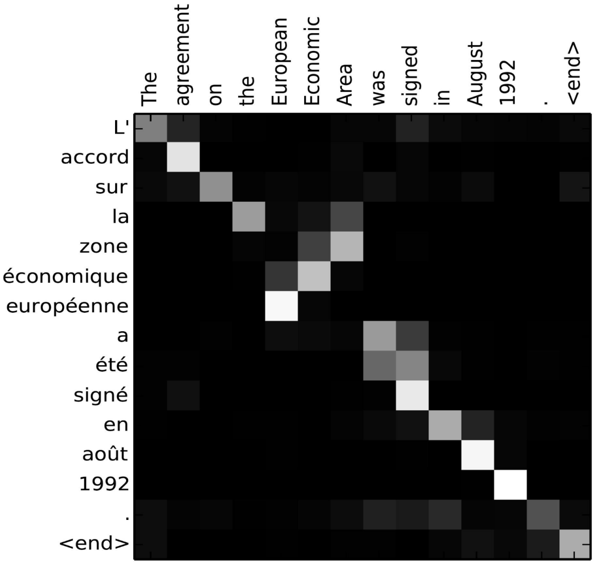
$$\alpha_{t,i} = \text{align}(y_t, x_i) = \frac{\exp(\text{score}(s_{t-1}, h_i))}{\sum_{i'=1}^{n} \exp(\text{score}(s_{t-1}, h_{i'}))} \tag{3}$$

History of attention

How to

New cc

Where



(1)

(2)

(3)

Key-Value Attention

Attention Is All You Need

Ashish Vaswani[*]
Google Brain
avaswani@google.com

Noam Shazeer[*]
Google Brain
noam@google.com

Niki Parmar[*]
Google Research
nikip@google.com

Jakob Uszkoreit[*]
Google Research
usz@google.com

Llion Jones[*]
Google Research
llion@google.com

Aidan N. Gomez[*, †]
University of Toronto
aidan@cs.toronto.edu

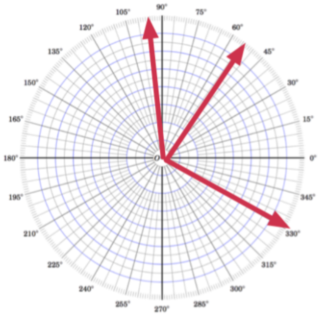
Łukasz Kaiser[*]
Google Brain
lukaszkaiser@google.com

Illia Polosukhin[*, ‡]
illia.polosukhin@gmail.com

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \tag{4}$$

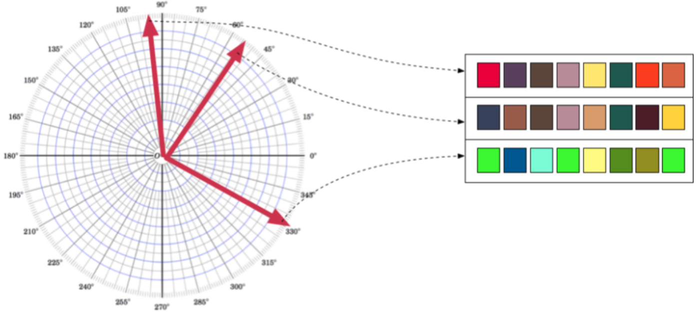
Key-Value Attention

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \tag{4}$$



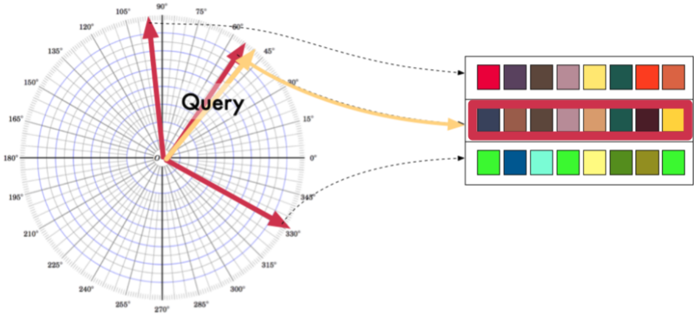
Key-Value Attention

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \tag{4}$$



Key-Value Attention

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \quad (4)$$



Returning to Backpack Language Model Example: "That trick was sick"

Sentence: "That trick was sick"

that	**trick**	**was**	**sick**
$C(\text{that})_0$	$C(\text{trick})_0$	$C(\text{was})_0$	$C(\text{sick})_0$
$\begin{pmatrix} 0 & 0 & 0 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 0 & 1 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 0 & 0 & 0 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 & 1 & 0 \end{pmatrix}$
$C(\text{that})_1$	$C(\text{trick})_1$	$C(\text{was})_1$	$C(\text{sick})_1$
$\begin{pmatrix} 0 & 0 & 0 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 0 & 0 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 0 & 0 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 1 & 0 & 1 \end{pmatrix}$

Explanation:

- Each word has two sense vectors: $C(x_i)_1$ and $C(x_i)_2$.
- "that" and "was" have zero vectors for both senses.
- "trick" has a non-zero vector for $C(\text{trick})_1$ at the third position (evokes scateboarding), while the second vector is zero.
- "sick" has two non-zero vectors: $C(\text{sick})_1$ and $C(\text{sick})_2$: positive for skateboard, negative for health.

Returning to Backpack Language Model Example: "That trick was sick"

Sentence: "That trick was sick"

that	**trick**	**was**	**sick**
$C(\text{that})_0$	$C(\text{trick})_0$	$C(\text{was})_0$	$C(\text{sick})_0$
$\begin{pmatrix} 0 & 0 & 0 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 0 & 1 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 0 & 0 & 0 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 & 1 & 0 \end{pmatrix}$
$C(\text{that})_1$	$C(\text{trick})_1$	$C(\text{was})_1$	$C(\text{sick})_1$
$\begin{pmatrix} 0 & 0 & 0 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 0 & 0 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 0 & 0 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 1 & 0 & 1 \end{pmatrix}$

Explanation:

- Each word has two sense vectors: $C(x_i)_1$ and $C(x_i)_2$.
- "that" and "was" have zero vectors for both senses.
- "trick" has a non-zero vector for $C(\text{trick})_1$ at the third position (evokes scateboarding), while the second vector is zero.
- "sick" has two non-zero vectors: $C(\text{sick})_1$ and $C(\text{sick})_2$: positive for skateboard, negative for health.

Returning to Backpack Language Model Example: "That trick was sick"

Sentence: "That trick was sick"

that	**trick**	**was**	**sick**
$C(\text{that})_0$	$C(\text{trick})_0$	$C(\text{was})_0$	$C(\text{sick})_0$
$\begin{pmatrix} 0 & 0 & 0 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 0 & 1 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 0 & 0 & 0 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 & 1 & 0 \end{pmatrix}$
$C(\text{that})_1$	$C(\text{trick})_1$	$C(\text{was})_1$	$C(\text{sick})_1$
$\begin{pmatrix} 0 & 0 & 0 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 0 & 0 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 0 & 0 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 1 & 0 & 1 \end{pmatrix}$

Explanation:

- Each word has two sense vectors: $C(x_i)_1$ and $C(x_i)_2$.
- "that" and "was" have zero vectors for both senses.
- "trick" has a non-zero vector for $C(\text{trick})_1$ at the third position (evokes scateboarding), while the second vector is zero.
- "sick" has two non-zero vectors: $C(\text{sick})_1$ and $C(\text{sick})_2$: positive for skateboard, negative for health.

Returning to Backpack Language Model Example: "That trick was sick"

Sentence: "That trick was sick"

that	**trick**	**was**	**sick**
$C(\text{that})_0$	$C(\text{trick})_0$	$C(\text{was})_0$	$C(\text{sick})_0$
$\begin{pmatrix} 0 & 0 & 0 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 0 & 1 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 0 & 0 & 0 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 & 1 & 0 \end{pmatrix}$
$C(\text{that})_1$	$C(\text{trick})_1$	$C(\text{was})_1$	$C(\text{sick})_1$
$\begin{pmatrix} 0 & 0 & 0 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 0 & 0 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 0 & 0 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 1 & 0 & 1 \end{pmatrix}$

Explanation:

- Each word has two sense vectors: $C(x_i)_1$ and $C(x_i)_2$.
- "that" and "was" have zero vectors for both senses.
- "trick" has a non-zero vector for $C(\text{trick})_1$ at the third position (evokes scateboarding), while the second vector is zero.
- "sick" has two non-zero vectors: $C(\text{sick})_1$ and $C(\text{sick})_2$: positive for skateboard, negative for health.

Returning to Backpack Language Model Example: "That trick was sick"

Sentence: "That trick was sick"

that	**trick**	**was**	**sick**
$C(\text{that})_0$	$C(\text{trick})_0$	$C(\text{was})_0$	$C(\text{sick})_0$
$\begin{pmatrix} 0 & 0 & 0 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 0 & 1 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 0 & 0 & 0 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 & 1 & 0 \end{pmatrix}$
$C(\text{that})_1$	$C(\text{trick})_1$	$C(\text{was})_1$	$C(\text{sick})_1$
$\begin{pmatrix} 0 & 0 & 0 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 0 & 0 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 0 & 0 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 1 & 0 & 1 \end{pmatrix}$

Explanation:

- Each word has two sense vectors: $C(x_i)_1$ and $C(x_i)_2$.
- "that" and "was" have zero vectors for both senses.
- "trick" has a non-zero vector for $C(\text{trick})_1$ at the third position (evokes scateboarding), while the second vector is zero.
- "sick" has two non-zero vectors: $C(\text{sick})_1$ and $C(\text{sick})_2$: positive for skateboard, negative for health.

Selecting the sense

Recap: Definition of $\alpha_{\ell,i,j}$:

$$\alpha_{\ell,i,j} = \begin{cases} 0 & \text{if } j \neq i \\ \sum_{r \neq i} \frac{C(\mathbf{x}_r)_{\ell+2}}{C(\mathbf{x}_r)_2 + C(\mathbf{x}_r)_3} & \text{if } j = i \end{cases} \tag{5}$$

Selecting the sense

Recap: Definition of $\alpha_{\ell,i,j}$:

$$\alpha_{\ell,i,j} = \begin{cases} 0 & \text{if } j \neq i \\ \sum_{r \neq i} \frac{C(\mathbf{x}_r)_{\ell+2}}{C(\mathbf{x}_r)_2 + C(\mathbf{x}_r)_3} & \text{if } j = i \end{cases} \tag{5}$$

How can you make this happen with attention?

Doing this with attention

What are the values?

What are the keys?

Doing this with attention

What are the values?

$$V = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

What are the keys?

Doing this with attention

What are the values?

What are the keys?

$$K = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

But what are the queries?

But what are the queries?

$$\alpha_{\ell,i,j} = \begin{cases} 0 & \text{if } j \neq i \\ \sum_{r \neq i} \frac{C(\mathbf{x}_r)_{\ell+2}}{C(\mathbf{x}_r)_2 + C(\mathbf{x}_r)_3} & \text{if } j = i \end{cases} \tag{6}$$

But what are the queries?

$$\alpha_{\ell,i,j} = \begin{cases} 0 & \text{if } j \neq i \\ \sum_{r \neq i} \frac{C(\mathbf{x}_r)_{\ell+2}}{C(\mathbf{x}_r)_2 + C(\mathbf{x}_r)_3} & \text{if } j = i \end{cases} \tag{6}$$

$$Q = \begin{pmatrix} 0 & 0 & \frac{2}{3} & \frac{1}{3} \\ 0 & 0 & \frac{2}{3} & \frac{1}{3} \\ 0 & 0 & \frac{2}{3} & \frac{1}{3} \\ 0 & 0 & \frac{2}{3} & \frac{1}{3} \\ 0 & 0 & \frac{2}{3} & \frac{1}{3} \\ 0 & 0 & \frac{2}{3} & \frac{1}{3} \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

Putting it all together

$$\text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right) = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \tag{7}$$

Putting it all together

$$\text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right) = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \tag{7}$$

So summing over all senses, we get the sentiment 1000 (Only contributing from skateboard sense of “sick”.