Attention!

Yifeng Liu

CS 495

UCLA Samueli Computer Science

Preliminaries

- Definition of token
- Vector representation of token
- Softmax function

Today's Learning Objectives

Students will be able to:

- Describe the **basic mechanism** of attention in neural network
- Understand the **function** of components in scaled dot-product attention

Attention in Life

Attention in Life

Attention in context

- Try to predict the next word:
  - Paris is the capital of _____
- What are the most important words for your attention to predict?
  - Paris is the capital of (France)
- Each word(token) can be represented as a vector (1×d)
  - Hence we need to predict the vector for "France" mainly based on attentions to "Paris" and "capital"
  - But how to measure the extent of the "attention"?

Attention in context

- Inner product as attention
  - z stands for the vector for "_____"
  - $X=[x_1, x_2, x_3, x_4, x_5]$ stands for the vectors for "Paris is the capital of"
  - Compute the inner product of z and $x_1 \sim x_5$ as $zX^T = [zx_1^T \dots zx_5^T]$
- Predict using the weighted average of given vectors (X)
  - $zX^T \cdot X$
  - Normalize the weight using Softmax: $\text{Softmax}(zX^T) \cdot X$
  - Suppose each element in z and $X \sim N(0, 1)$
    - $\text{Var}(zX^T)=d$
  - Therefore, divide by $\sqrt{d}$ to reduce the variance: $\text{Softmax}(\frac{zX^T}{\sqrt{d}}) \cdot X$

$$\text{Softmax}(\text{X})_i = \frac{e^{X_i}}{\sum_j e^{X_j}}$$

Attention in context

- How to learn with $\text{Softmax}(\frac{zX^T}{\sqrt{d}}) \cdot X$?
  - Weighted average of given vectors by weights of inner products of given vectors
  - ➢Weighted average of **representation** of given vectors by weights of inner products of **representation** of given vectors
  - $z \rightarrow W_q z = q$, $X^T \rightarrow W_k X^T = K^T$, $X \rightarrow W_v X = V$
- $\text{Softmax}(\frac{qK^T}{\sqrt{d}}) \cdot V$

Today's Learning Objectives

Students will be able to:

✓ Describe the **basic mechanism** of attention

- $\text{Softmax}(\frac{qK^T}{\sqrt{d}}) \cdot V$

✓ Understand the **function** of components in scaled dot-product attention

- To normalize the weight and control variance

Thank you very much!