Synthesizer: Rethinking Self-Attention in Transformer Models

Radek Bartyzal

GLAMI AI

26. 5. 2020

Motivation

Transformer uses self-attention:

- sentence = list of word (token) embeddings of dimension d
- sentence length = number of words = l
- each word is multiplied by Q, K, V $\in \mathbb{R}^{d \times d}$
- attention vector for word $x = A(x) \in \mathbb{R}^{1 \times l} = Q(x)K(x)^T$
- attention matrix with row for each word = $A \in \mathbb{R}^{l \times l}$
- Output = $Y = Softmax(A)G(x)$, G = Value matrix or other function.

Questions:

- Is the dot-product of $QK = O(d^3)$ necessary?
- replace it by calculating row vectors of A by FF net = don't look at other words?
- replace it by directly optimizing A?

Overview

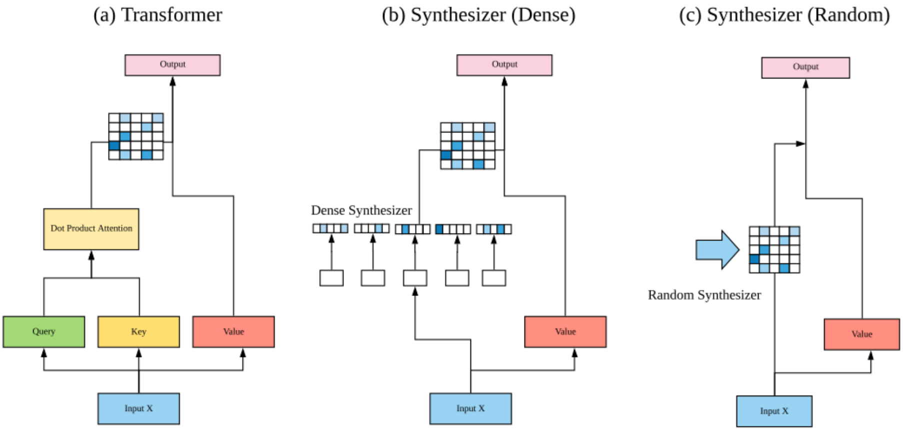

Figure: Sentence length = 5 words. a) Dot product of $Q, K \in \mathbb{R}^{d \times d}$.
b) 2-layer FF applied to each word, no dot product.
c) Weight matrix randomly inited $\implies$ optimized or fixed.

Overview

Dense variant adds $d \times l$ params:

- problem: if l is large that is more than d^2
- solution: factorize the weight matrix to $\mathbb{R}^{d \times a}$ and $\mathbb{R}^{d \times b}$ where $l = ab$

Model	$S(X)$	Condition On	Sample	Interact	$\|\theta\|$
Dot Product Attention	$F_Q(X)F_K(X_i)^\top$	$X_j \ \forall j$	Local	Yes	$2d^2$
Random	R	N/A	Global	No	. ℓ^2
Factorized Random	$R_1 R_2^\top$	N/A	Global	No	$2\ell k$
Dense	$F_1 \sigma(F_2(X_i))$	X_i	Local	No	$d^2 + d\ell$
Factorized Dense	$H_A(F_A(X_i))) * H_B(F_B(X_i)))$	X_i	Local	No	$d^2 + d(k_1 + k_2)$

Table 1: Overview of all Synthesizing Functions.

Figure: $S(X)$ = Synthesizing function returning A for the whole sentence. Dense variant is conditioned only on each word alone = no interaction with other words.

Experiments

Tasks:

- machine translation
- language modeling
- dialogue generation

Results:

- Fixed Random is significantly worse but not complete trash.
- Both Optimized Random and Dense have competitive results with classic Transformer.
- Mixing heads from diff. variants helps.

Sources

1. Tay, Yi, et al. "Synthesizer: Rethinking Self-Attention in Transformer Models." arXiv preprint arXiv:2005.00743 (2020).
https://arxiv.org/abs/2005.00743v1