CSE 5523: Lecture Notes 17
Transformers

Contents

The best neural net systems these days are ‘transformers’: GPT-2, BERT, GPT-3, . . .

Transformers associate ‘queries’ and ‘keys’ of K items to choose targets of attention.

These associations are modeled using ‘query’, ‘key’ and ‘value’ matrices $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{D \times D}$.

17.1 Attention Models [Vaswani et al., 2017]

Each item in a transformer is represented in a D-dimensional vector $\mathbf{H}_\ell \in \mathbb{R}^{D \times K}$ at each level ℓ.

At each level, each item may ‘attend’ to one other item per ‘head’ h.

This is done by comparing queries and keys, using inner products of these as a similarity measure.

Values, weighted by this similarity, are then passed to the next level:

$$\mathbf{H}_{\ell,h} = \overbrace{\mathbf{V}_{\ell,h}\,\mathbf{H}_{\ell-1}}^{\text{value for each target}} \operatorname{SoftMax}(\underbrace{(\overbrace{\mathbf{K}_{\ell,h}\,\mathbf{H}_{\ell-1}}^{\text{key for each target}})^\top \overbrace{\mathbf{Q}_{\ell,h}\,\mathbf{H}_{\ell-1}}^{\text{query for each source}}}_{\text{attention matrix}})$$

where SoftMax is our multinomial logistic function on $\mathbf{M} \in \mathbb{R}^{J \times N}$ with N instances of J values:

$$\operatorname{SoftMax}(\,\mathbf{M}\,) = \frac{\exp(\,\mathbf{M}\,)}{\mathbf{1}^\top \exp(\,\mathbf{M}\,)}$$

Again, we can stack the models for parallel multiplication: $\begin{bmatrix} \mathbf{Q}_{\ell,h} \\ \mathbf{K}_{\ell,h} \\ \mathbf{V}_{\ell,h} \end{bmatrix} \mathbf{H}_{\ell-1}$.

17.2 Multiple attention heads

The outputs $\mathbf{H}_{\ell,h}$ of the heads are then concatenated and fed into another (e.g. sigmoid) layer FF:

$$\mathbf{H}_\ell = \operatorname{FF}(\underbrace{\sum_h \delta_h \otimes \mathbf{H}_{\ell,h}}_{\text{concatenate}})$$

The backpropagation for each of these matrix operations is fairly straightforward.

The problem with these models for our purposes is that they take a lot of resources!

Usually, people use pre-trained models and train a feed-forward (e.g. sigmoid) layer on their task.

References

[Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *NIPS*, pages 5998–6008.