Tutorial on Transformers with Indices

Ref: Section I-0, “Crash Course on Transformers” (5 pages)
of [arXiv:2304.02034 (Emily Dinan, Sho Yaida, Susan Zhang)]

[cf. Andrej Karpathy’s YouTube tutorial]

Meta

Outline

0. Preliminary: Vocabulary and Task
1. Word Embedding and Positional Embedding
2. Layer Normalizations and Skip Connections
3. MLP Blocks
4. Multi-Head Self-Attention Blocks
5. Head

0. Preliminary: Vocabulary and Task

Example 1

Vocabulary=a set of characters

```
{a, b, ..., !, ?, [space]}
```

0. Preliminary: Vocabulary and Task

Example 1

Vocabulary=a set of characters

$$\underset{0}{\{\texttt{a}}, \underset{1}{\texttt{b}}, \ldots, \texttt{!}, \texttt{?}, \underset{n_{\text{vocab}}-1}{\texttt{[space]}}\}$$

0. Preliminary: Vocabulary and Task

Example 1

Vocabulary=a set of characters

$$\{\texttt{a},\ \texttt{b},\ \ldots,\ \texttt{!},\ \texttt{?},\ \texttt{[space]}\}$$

0 1 $n_{\text{vocab}} - 1$

Task=next-character prediction

Input: B|r|e|a|k|i|n|g|[space]|B|a
Output: B|r|e|a|k|i|n|g|[space]|B|a|d

0. Preliminary: Vocabulary and Task

Example 2

Vocabulary=a set of “tokens” (created by tokenization such as byte-pair encoding algorithm)

$$\{\texttt{the}, \texttt{of}, \ldots, \texttt{super}, \ldots, \texttt{symmetry}, \ldots, [\texttt{space}]\}$$

0. Preliminary: Vocabulary and Task

Example 2

Vocabulary=a set of “tokens” (created by tokenization such as byte-pair encoding algorithm)

$$\{\texttt{the}, \texttt{of}, \ldots, \texttt{super}, \ldots, \texttt{symmetry}, \ldots, [\texttt{space}]\}$$

0 1 $n_{\text{vocab}} - 1$

0. Preliminary: Vocabulary and Task

Example 2

Vocabulary=a set of “tokens” (created by tokenization such as byte-pair encoding algorithm)

$$\{\texttt{the}, \texttt{of}, \ldots, \texttt{super}, \ldots, \texttt{symmetry}, \ldots, [\texttt{space}]\}$$

0 1 $n_{\text{vocab}} - 1$

Task=next-token prediction

Input: Super|symmetry|[space]|is|[space]
Output: Super|symmetry|[space]|is|[space]|awesome

Input Notation

$$x_{\alpha;t;i}$$

sample 1: Super|symmetry|[space]|is|[space]|awesome|.
sample 2: Transformers|[space]|are|[space]|simple|.
sample 3: I|[space]|want|[space]|to|[space]|watch|[space]|Breaking|[space]|Bad|.
...

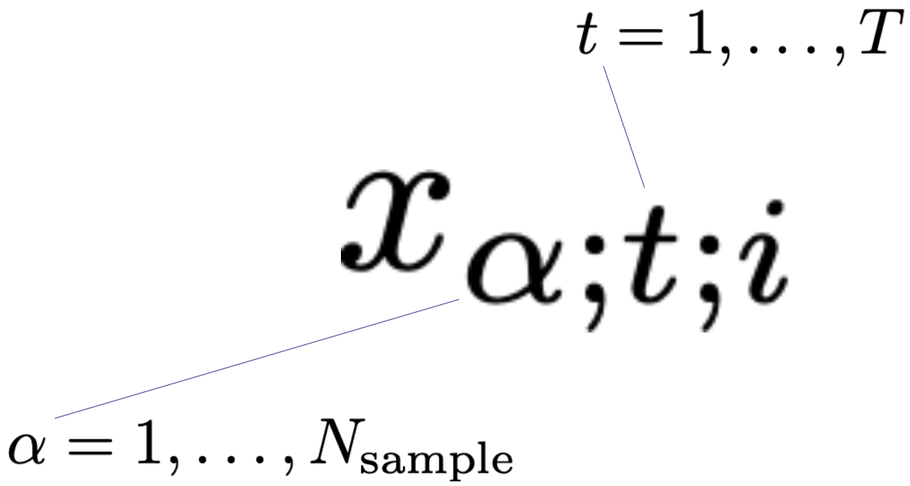
Input Notation

$$x_{\alpha;t;i}$$

$$\alpha = 1, \ldots, N_{\text{sample}}$$

sample 1: Super|symmetry|[space]|is|[space]|awesome|.
sample 2: Transformers|[space]|are|[space]|simple|.
sample 3: I|[space]|want|[space]|to|[space]|watch|[space]|Breaking|[space]|Bad|.
…

Input Notation



sample 1: Super|symmetry|[space]|is|[space]|awesome|.
sample 2: Transformers|[space]|are|[space]|simple|.
sample 3: I|[space]|want|[space]|to|[space]|watch|[space]|Breaking|[space]|Bad|.
...

Input Notation

$$x_{\alpha;t;i}$$

$t = 1, \ldots, T$

$\alpha = 1, \ldots, N_{\text{sample}}$

$i = 1, \ldots, n_{\text{vocab}}$

one-hot in i-direction: $(0, \ldots, 0, 1, 0, \ldots, 0)$

sample 1: Super|symmetry|[space]|is|[space]|awesome|.
sample 2: Transformers|[space]|are|[space]|simple|.
sample 3: I|[space]|want|[space]|to|[space]|watch|[space]|Breaking|[space]|Bad|.
...

1. Word Embedding and Positional Embedding

$$\tilde{z}_{\alpha;t;i} = \sum_{j=1}^{n_{\text{vocab}}} W_{ij}^{\text{WE}} x_{\alpha;t;j} = W_{ij_\star(\alpha;t)}^{\text{WE}} \qquad i = 1, \ldots, n$$

1. Word Embedding and Positional Embedding

$$\tilde{z}_{\alpha;t;i} = \sum_{j=1}^{n_{\text{vocab}}} W_{ij}^{\text{WE}} x_{\alpha;t;j} = W_{ij_\star(\alpha;t)}^{\text{WE}} \qquad i = 1, \ldots, n$$

$$z_{\alpha;t;i}^{(1)} = b_{t;i}^{\text{PE}} + \tilde{z}_{\alpha;t;i}$$

2. Layer Normalizations and Skip Connections

$$s_{\alpha;t;i}^{(\ell)} = \gamma_i^{(\ell)} \left[\frac{z_{\alpha;t;i}^{(\ell)} - \left(\frac{1}{n} \sum_{j=1}^{n} z_{\alpha;t;j}^{(\ell)} \right)}{\sqrt{\frac{1}{n} \sum_{j=1}^{n} \left(z_{\alpha;t;j}^{(\ell)} \right)^2 - \left(\frac{1}{n} \sum_{j=1}^{n} z_{\alpha;t;j}^{(\ell)} \right)^2 + \epsilon}} \right] + \beta_i^{(\ell)}$$

2. Layer Normalizations and Skip Connections

$$s_{\alpha;t;i}^{(\ell)} = \gamma_i^{(\ell)} \left[\frac{z_{\alpha;t;i}^{(\ell)} - \left(\frac{1}{n} \sum_{j=1}^{n} z_{\alpha;t;j}^{(\ell)} \right)}{\sqrt{\frac{1}{n} \sum_{j=1}^{n} \left(z_{\alpha;t;j}^{(\ell)} \right)^2 - \left(\frac{1}{n} \sum_{j=1}^{n} z_{\alpha;t;j}^{(\ell)} \right)^2 + \epsilon}} \right] + \beta_i^{(\ell)}$$

$$z_{\alpha;t;i}^{(\ell+1)} = \mathbf{R}_{t;i}^{(\ell+1)} \left(s_{\alpha}^{(\ell)}; \theta^{(\ell+1)} \right) + z_{\alpha;t;i}^{(\ell)}$$

3. MLP Blocks $\mathrm{R}_{t;i}^{\mathrm{MLP}}\left(s_{\alpha};\theta\right)$

$$w_{\alpha;t;i} = \sum_{j=1}^{n} W_{ij} s_{\alpha;t;j} \qquad \text{for} \quad i = 1, \ldots, Mn\,,$$

$$r_{\alpha;t;i} = \sum_{j=1}^{Mn} X_{ij} \sigma\left(w_{\alpha;t;j}\right) \qquad \text{for} \quad i = 1, \ldots, n\,,$$

4. Multi-Head Self-Attention Blocks $\mathrm{R}_{t;i}^{\mathrm{MHSA}}\left(s_\alpha;\theta\right)$

$$q_{\alpha;t;c}^h \equiv \sum_{i=1}^{n} Q_{ci}^h s_{\alpha;t;i}\,,$$

$$k_{\alpha;t;c}^h \equiv \sum_{i=1}^{n} K_{ci}^h s_{\alpha;t;i}\,,$$

$$v_{\alpha;t;c}^h \equiv \sum_{i=1}^{n} V_{ci}^h s_{\alpha;t;i}\,,$$

$$t = 1, \ldots, T$$

$$h = 1, \ldots, H$$

$$c = 1, \ldots, C \quad \text{where} \quad C = \frac{n}{H}$$

4. Multi-Head Self-Attention Blocks $\mathrm{R}^{\mathrm{MHSA}}_{t;i}(s_\alpha;\theta)$

$$q^h_{\alpha;t;c} \equiv \sum_{i=1}^{n} Q^h_{ci} s_{\alpha;t;i}\,,$$

$$k^h_{\alpha;t;c} \equiv \sum_{i=1}^{n} K^h_{ci} s_{\alpha;t;i}\,,$$

$$v^h_{\alpha;t;c} \equiv \sum_{i=1}^{n} V^h_{ci} s_{\alpha;t;i}\,,$$

$$\widetilde{\Omega}^h_{\alpha;tt'} \equiv \frac{1}{\sqrt{C}} \sum_{c=1}^{C} q^h_{\alpha;t;c} k^h_{\alpha;t';c}$$

$$t = 1, \ldots, T$$

$$h = 1, \ldots, H$$

$$c = 1, \ldots, C \quad \text{where} \quad C = \frac{n}{H}$$

4. Multi-Head Self-Attention Blocks $\mathrm{R}_{t;i}^{\mathrm{MHSA}}\left(s_{\alpha};\theta\right)$

$$q_{\alpha;t;c}^{h} \equiv \sum_{i=1}^{n} Q_{ci}^{h} s_{\alpha;t;i} \, ,$$

$$k_{\alpha;t;c}^{h} \equiv \sum_{i=1}^{n} K_{ci}^{h} s_{\alpha;t;i} \, ,$$

$$v_{\alpha;t;c}^{h} \equiv \sum_{i=1}^{n} V_{ci}^{h} s_{\alpha;t;i} \, ,$$

$$\widetilde{\Omega}_{\alpha;tt'}^{h} \equiv \frac{1}{\sqrt{C}} \sum_{c=1}^{C} q_{\alpha;t;c}^{h} k_{\alpha;t';c}^{h}$$

Encoder: $\Omega_{\alpha;tt'}^{h} \equiv \dfrac{\exp\left(\widetilde{\Omega}_{\alpha;tt'}^{h}\right)}{\sum_{t''=1}^{T} \exp\left(\widetilde{\Omega}_{\alpha;tt''}^{h}\right)}$

4. Multi-Head Self-Attention Blocks $\mathrm{R}^{\mathrm{MHSA}}_{t;i}\left(s_\alpha;\theta\right)$

$$q^h_{\alpha;t;c} \equiv \sum_{i=1}^{n} Q^h_{ci} s_{\alpha;t;i}\,,$$

$$k^h_{\alpha;t;c} \equiv \sum_{i=1}^{n} K^h_{ci} s_{\alpha;t;i}\,,$$

$$v^h_{\alpha;t;c} \equiv \sum_{i=1}^{n} V^h_{ci} s_{\alpha;t;i}\,,$$

$$\widetilde{\Omega}^h_{\alpha;tt'} \equiv \frac{1}{\sqrt{C}} \sum_{c=1}^{C} q^h_{\alpha;t;c} k^h_{\alpha;t';c}$$

Decoder:

$$\Omega^h_{\alpha;tt'} \equiv \begin{cases} \dfrac{\exp\left(\widetilde{\Omega}^h_{\alpha;tt'}\right)}{\sum_{t''=1}^{t}\exp\left(\widetilde{\Omega}^h_{\alpha;tt''}\right)} & \text{for} \quad t' \leq t\,, \\ 0 & \text{for} \quad t' > t\,, \end{cases}$$

4. Multi-Head Self-Attention Blocks $\mathrm{R}^{\mathrm{MHSA}}_{t;i}\left(s_\alpha;\theta\right)$

$$q^h_{\alpha;t;c} \equiv \sum_{i=1}^{n} Q^h_{ci} s_{\alpha;t;i}\,,$$

$$k^h_{\alpha;t;c} \equiv \sum_{i=1}^{n} K^h_{ci} s_{\alpha;t;i}\,,$$

$$v^h_{\alpha;t;c} \equiv \sum_{i=1}^{n} V^h_{ci} s_{\alpha;t;i}\,,$$

$$\widetilde{\Omega}^h_{\alpha;tt'} \equiv \frac{1}{\sqrt{C}} \sum_{c=1}^{C} q^h_{\alpha;t;c} k^h_{\alpha;t';c}$$

Decoder:

$$\Omega^h_{\alpha;tt'} \equiv \begin{cases} \dfrac{\exp\left(\widetilde{\Omega}^h_{\alpha;tt'}\right)}{\sum_{t''=1}^{t}\exp\left(\widetilde{\Omega}^h_{\alpha;tt''}\right)} & \text{for} \quad t' \le t\,, \\ 0 & \text{for} \quad t' > t\,, \end{cases}$$

$$r_{\alpha;t;i} \equiv \sum_{h=1}^{H}\sum_{c=1}^{C} U^h_{ic} \left(\sum_{t'=1}^{T} \Omega^h_{\alpha;tt'} v^h_{\alpha;t';c} \right)$$

5. Head

If we tie weights between word embedding and word unembedding, then

$$z^{(L)}_{\alpha;t;i} = \mathcal{N}_{\text{rescale}} \sum_{j=1}^{n} (W^{\text{WE}})^{\top}{}_{ij} s^{(L-1)}_{\alpha;t;j} = \mathcal{N}_{\text{rescale}} \sum_{j=1}^{n} W^{\text{WE}}_{ji} s^{(L-1)}_{\alpha;t;j} \quad \text{for} \quad i = 1, \ldots, n_{\text{out}} \,,$$

If we don’t, then

$$z^{(L)}_{\alpha;t;i} = W^{\text{WU}}_{ij} s^{(L-1)}_{\alpha;t;j}$$