CS539 - NLP - Homework 3

Nuttaree Busarapongpanich

1 Scaled Dot-Product Attention

TASK 1.1 Copying

From equation 2, we can see that if we want $\mathbf{a} \approx \mathbf{v}_j$, the α_j should be very close to 1 ($\alpha_j \approx 1$) and the other α values should be close to 0 ($\alpha_i \approx 0$, where $\forall i \neq j$). From equation 1, $\mathbf{q}\mathbf{k}_i^T$ should be large to make α_i to be large. In this case, the $\mathbf{q}\mathbf{k}_j^T$ should be a lot larger than the other $\mathbf{q}\mathbf{k}_i^T$, where $\forall i \neq j$.

TASK 1.2 Average of Two

Regarding equation 2, if we want to make $\mathbf{a} \approx \frac{1}{2}(\mathbf{v}_a + \mathbf{v}_b)$, we have to make α_a and α_b a lot larger than the other α_i, where $\forall i \neq a$ and $\forall i \neq b$. Therefore, we should have α_i ($\forall i \neq a$ and $\forall i \neq b$) be very low or even close to 0. Both α_a and α_b should be close to 0.5.

The query should be $\mathbf{q} = c(\mathbf{k}_a + \mathbf{k}_b)$, where c is a scaling constant.

We will have:

$$\mathbf{q}\mathbf{k}_a^T = c(\mathbf{k}_a + \mathbf{k}_b)\mathbf{k}_a^T = c(\mathbf{k}_a\mathbf{k}_a^T + \mathbf{k}_b\mathbf{k}_a^T)$$

As we have: $\mathbf{k}_b\mathbf{k}_a^T = 0$, then:

$$\mathbf{q}\mathbf{k}_a^T = c(\mathbf{k}_a\mathbf{k}_a^T)$$

From the definition that $||\mathbf{k}_i|| = 1$, we know that $\mathbf{k}_a\mathbf{k}_a^T = 1$. Therefore, we have:

$$\mathbf{q}\mathbf{k}_a^T = c$$

The same thing happens to $\mathbf{q}\mathbf{k}_b^T$. We will have:

$$\mathbf{q}\mathbf{k}_b^T = c(\mathbf{k}_b\mathbf{k}_b^T) = c$$

For the other $\mathbf{q}\mathbf{k}_i^T$, where $\forall i \neq a$ and $\forall i \neq b$, we will have:

$$\mathbf{q}\mathbf{k}_i^T = c(\mathbf{k}_a + \mathbf{k}_b)\mathbf{k}_i^T = c(\mathbf{k}_a\mathbf{k}_i^T + \mathbf{k}_b\mathbf{k}_i^T) = 0$$

From equation 1,

$$\alpha_i = \frac{exp(0)}{\sum_{j=1}^{m} \exp(\mathbf{q}\mathbf{k}_j^T/\sqrt{d})} = \frac{1}{\sum_{j=1}^{m} \exp(\mathbf{q}\mathbf{k}_j^T/\sqrt{d})}, \text{ where } \forall i \neq a \text{ and } \forall i \neq b$$

We know that the $\sum_{j=1}^{m} \exp(\mathbf{q}\mathbf{k}_j^T/\sqrt{d})$ will always be the same value for every α_i, including α_a and α_b. Also, the scaling constant will affect only on $\mathbf{q}\mathbf{k}_a^T$ and $\mathbf{q}\mathbf{k}_b^T$, so we can make $\mathbf{q}\mathbf{k}_a^T$ and $\mathbf{q}\mathbf{k}_b^T$ arbitrarily larger than the others, which will only affect α_a and α_b. We have $\mathbf{q}\mathbf{k}_a^T = \mathbf{q}\mathbf{k}_b^T = c$, so $\alpha_i = \alpha_j$ and each of them is close to 0.5. Hence we will get $\mathbf{a} \approx \frac{1}{2}(\mathbf{v}_a + \mathbf{v}_b)$.

TASK 1.3 Noisy Average

We have $\mathbf{q} = c(\mathbf{k}_a + \mathbf{k}_b)$, where c is a scaling constant, from **TASK 1.2**. Let $\mathbf{k}_i = \mu_i * \lambda_i$. We will have:

$$\mathbf{q} = c(\mu_a\lambda_a + \mu_b\lambda_b)$$

We, now, will derive $\mathbf{q}\mathbf{k}_a^T$.

$$\begin{aligned}\mathbf{q}\mathbf{k}_a^T &= c(\mu_a\lambda_a + \mu_b\lambda_b)\mathbf{k}_a^T \\ &= c(\mu_a\lambda_a + \mu_b\lambda_b)\mu_a^T\lambda_a \\ &= c(\mu_a\mu_a^T\lambda_a^2 + \mu_b\mu_a^T\lambda_b\lambda_a)\end{aligned}$$

We know that vectors $\mu_1, ..., \mu_m$ are orthogonal unit vectors, so $\mu_b\mu_a^T = 0$ and $\mu_a\mu_a^T = 1$. We will have:

$$\begin{aligned}\mathbf{q}\mathbf{k}_a^T &= c((1)\lambda_a^2 + 0) \\ &= c(\lambda_a^2)\end{aligned}$$

We can get $\mathbf{q}\mathbf{k}_b^T$ in the same way we just did, so we will have:

$$\mathbf{q}\mathbf{k}_b^T = c(\lambda_b^2)$$

For the other $\mathbf{q}\mathbf{k}_i^T$, where $\forall i \neq a$ and $\forall i \neq b$, we will have:

$$\begin{aligned}\mathbf{q}\mathbf{k}_i^T &= c(\mu_a\lambda_a + \mu_b\lambda_b)\mathbf{k}_i^T \\ &= c(\mu_a\lambda_a + \mu_b\lambda_b)\mu_i^T\lambda_i \\ &= c(\mu_a\mu_i^T\lambda_a\lambda_i + \mu_b\mu_i^T\lambda_b\lambda_i) \\ &= c(0(\lambda_a\lambda_i) + 0(\lambda_b\lambda_i)) \\ &= 0\end{aligned}$$

We can see that α_i ($\forall i \neq a$ and $\forall i \neq b$) still be very low or even close to 0.

From **TASK 1.2**, we have $\mathbf{q}\mathbf{k}_a^T = \mathbf{q}\mathbf{k}_b^T = c$, but for this task, we have $\mathbf{q}\mathbf{k}_a^T = c(\lambda_a^2)$ and $\mathbf{q}\mathbf{k}_b^T = c(\lambda_b^2)$. If $\lambda_a = \lambda_b$, we will still keep the same behavior as in **TASK 1.2**, which is $\mathbf{a} \approx \frac{1}{2}(\mathbf{v}_a + \mathbf{v}_b)$. However, $\lambda_1, ..., \lambda_m$ in this task are sampled from standard distribution, so it is more likely that $\lambda_a \neq \lambda_b$ resulting $\mathbf{a} \not\approx \frac{1}{2}(\mathbf{v}_a + \mathbf{v}_b)$. As $\mathbf{a}$ get the influence mostly from the value of α_a and α_b which is from $\mathbf{q}\mathbf{k}_a^T$ and $\mathbf{q}\mathbf{k}_b^T$, when we resample $\lambda_1, ..., \lambda_m$ (including λ_a and λ_b) many times, we possibly get the different value of $\mathbf{a}$ each time.

TASK 1.4 Noisy Average with Multi-head Attention

The output of a simple version of multi-head attention computation is $\mathbf{a} = \frac{1}{2}(\mathbf{a}_1 + \mathbf{a}_2)$. The goal is to design query $\mathbf{q}_1$ and $\mathbf{q}_2$ to have $\mathbf{a} \approx \frac{1}{2}(\mathbf{v}_a + \mathbf{v}_b)$. We, then, have $\frac{1}{2}(\mathbf{a}_1 + \mathbf{a}_2) \approx \frac{1}{2}(\mathbf{v}_a + \mathbf{v}_b)$; $\mathbf{a}_1 + \mathbf{a}_2 \approx \mathbf{v}_a + \mathbf{v}_b$. We, now, can construct $\mathbf{a}_1 \approx \mathbf{v}_a$ and $\mathbf{a}_2 \approx \mathbf{v}_b$, which is very similar on the idea in **TASK 1.1**. Now, if we want to only have a scaling constant for both $\mathbf{q}_1\mathbf{k}_a^T$ and $\mathbf{q}_2\mathbf{k}_b^T$, we could have $\mathbf{q}_1 = c\mathbf{k}_a$ and $\mathbf{q}_2 = c\mathbf{k}_b$. Calculate $\mathbf{q}_1\mathbf{k}_a^T$:

$$\begin{aligned}\mathbf{q}_1 &= c\mu_a\lambda_a \\ \mathbf{q}_1\mathbf{k}_a^T &= c\mu_a\lambda_a\mathbf{k}_a^T \\ &= c\mu_a\lambda_a\mu_a^T\lambda_a \\ &= c\lambda_a^2\end{aligned}$$

Calculate $\mathbf{q}_1\mathbf{k}_i^T$, where $\forall i \neq a$:

$$\begin{aligned}
\mathbf{q}_1\mathbf{k}_i^T &= c\mu_a\lambda_a\mathbf{k}_i^T \\
&= c\mu_a\lambda_a\mu_i^T\lambda_i \\
&= c\mu_a\mu_i^T\lambda_a\lambda_i \\
&= c(0)\lambda_a\lambda_i \\
&= 0
\end{aligned}$$

The calculations of $\mathbf{q}_2\mathbf{k}_b^T$ and $\mathbf{q}_2\mathbf{k}_i^T$, where $\forall i \neq b$, are the same. We will have:

$$\begin{aligned}
\mathbf{q}_2\mathbf{k}_b^T &= c\lambda_b^2 \\
\mathbf{q}_2\mathbf{k}_i^T &= 0
\end{aligned}$$

α_a depends only on $\mathbf{q}_1\mathbf{k}_a^T = c\lambda_a^2$, so we can set an arbitrarily large number to c to make the α_a have a lot larger than the other α. The same thing happens to α_b as well. Hence, we will get $\mathbf{a}_1 \approx \mathbf{v}_a$ and $\mathbf{a}_2 \approx \mathbf{v}_b$, which yields the result of $\mathbf{a} \approx \frac{1}{2}(\mathbf{v}_a + \mathbf{v}_b)$.

2 Attention in German-to-English Machine Translation

TASK 2.1 Scaled-Dot Product Attention

```
2021-03-03 15:51:40 INFO     | Test Loss: 1.815 | Test PPL:   6.139 | Test BLEU 34.07
```

Figure 1: The perplexity and BLUE score on the test set for Scaled-Dot Product Attention

Figure 1 shows the perplexity and BLUE score on the test set for Scaled-Dot Product Attention.

TASK 2.2 Attention Diagrams

I have got the same result as the assignment description for the Subject-Object-Verb language pattern of German as shown in Figure 2.

The next observation is German doesn't have a present continuous tense. We can see in Figure 3 and 4 that the words "is cleaning" and "are riding" are "putzt" and "fahren" in German respectively.

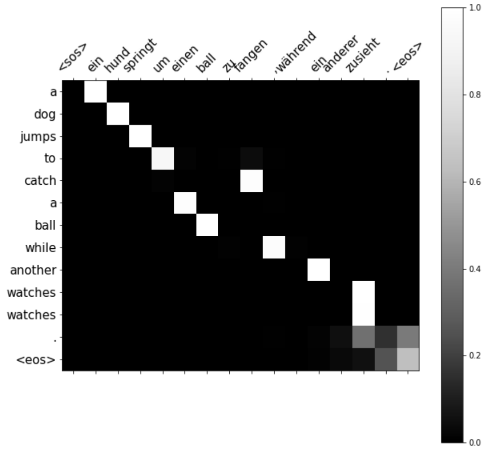

Figure 2: The Subject-Object-Verb language pattern in German

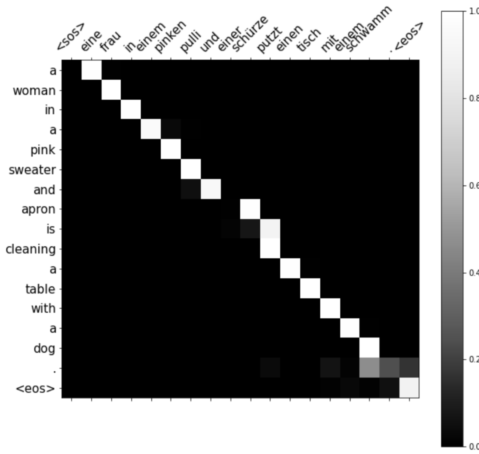

Figure 3: No present continuous tense in German 1

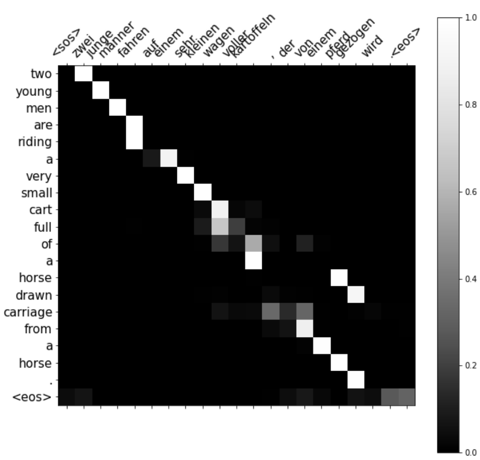

Figure 4: No present continuous tense in German 2

TASK 2.3 Comparison

```
2021-03-06 01:49:36 INFO     | Test Loss: 2.399 | Test PPL:  11.018 | Test BLEU 18.16
2021-03-06 03:09:26 INFO     | Test Loss: 2.390 | Test PPL:  10.918 | Test BLEU 18.61
2021-03-06 03:23:47 INFO     | Test Loss: 2.395 | Test PPL:  10.967 | Test BLEU 18.35
```

Figure 5: The perplexity and BLUE score on the test set for Dummy Attention over 3 runs

```
2021-03-06 02:49:48 INFO     | Test Loss: 2.202 | Test PPL:   9.043 | Test BLEU 22.85
2021-03-06 03:16:44 INFO     | Test Loss: 2.211 | Test PPL:   9.127 | Test BLEU 21.97
2021-03-06 03:31:10 INFO     | Test Loss: 2.201 | Test PPL:   9.035 | Test BLEU 22.61
```

Figure 6: The perplexity and BLUE score on the test set for MeanPool Attention over 3 runs

```
2021-03-07 21:54:20 INFO     | Test Loss: 1.840 | Test PPL:   6.294 | Test BLEU 35.16
2021-03-07 22:16:00 INFO     | Test Loss: 1.819 | Test PPL:   6.165 | Test BLEU 33.70
2021-03-08 07:10:43 INFO     | Test Loss: 1.834 | Test PPL:   6.261 | Test BLEU 34.39
```

Figure 7: The perplexity and BLUE score on the test set for Scaled-Dot Product Attention over 3 runs

Table 1: Mean and variance.

	Dummy		**MeanPool**		**Scaled-Dot Product**	
	PPL	BLEU	PPL	BLEU	PPL	BLEU
Mean	10.97	18.37	9.07	22.48	6.240	34.42
Variance	0.002	0.034	0.002	0.138	0.003	0.356

From the above results, we can see that the perplexity and BLEU scores seem stable for each type of attention. With the Dummy attention, the blue score is around 18 with the perplexity score around 10.9-11. The MeanPool attention have higher BLEU score than Dummy but have lower perplexity scores. For the Scaled-Dot Product attention, this type of attention has the highest BLEU score comparing to Dummy and MeanPool and have the lowest perplexity score among these mechanisms. Hence, Scaled-Dot Product attention seems to have the best performance. The MeanPool attention is in the second rank, and Dummy attention mechanism is the last.

TASK 2.EC Beam Search and BLEU

```
2021-03-08 16:50:11 INFO     | Test Loss: 1.807 | Test PPL:   6.095 | Test BLEU 36.33
```

Figure 8: BLEU scores on the test set for the scaled dot-product attention model with B=5

```
2021-03-09 18:52:44 INFO     | Test Loss: 1.833 | Test PPL:   6.252 | Test BLEU 35.90
```

Figure 9: BLEU scores on the test set for the scaled dot-product attention model with B=10

```
2021-03-09 19:07:06 INFO     | Test Loss: 1.834 | Test PPL:   6.260 | Test BLEU 37.36
```

Figure 10: BLEU scores on the test set for the scaled dot-product attention model with B=20

```
2021-03-09 19:28:28 INFO     | Test Loss: 1.824 | Test PPL:   6.198 | Test BLEU 35.99
```

Figure 11: BLEU scores on the test set for the scaled dot-product attention model with B=50

Figure 8-11 show the perplexity and BLEU scores of different beam widths.