AI 539 Assignment 4

ONID: sawantam

1 Scaled Dot-Product Attention

TASK 1.1

We get $\mathbf{a} \approx v_j$ when the query $\mathbf{q}$ aligns with the corresponding key, say k_j where $j \in (1, m)$. Since the values of that corresponding key will be only value considered after softmax function. Since dot product of $\mathbf{q}$ with any other element of $\mathbf{k}$ will be ≈ 0 after softmax function, except for the k_j, $\mathbf{q} = ck_j$ will give us desired output, where c is just some constant.

TASK 1.2

Referring to the previous answer, $a \approx \frac{1}{2}(v_a + v_b)$ if we consider $\mathbf{q} = C(k_a + k_b)$, where C is some scalar constant
given that,

$$\mathbf{a} = \sum_{j=1}^{m} \alpha_j v_j$$

and

$$\because \mathbf{a} \approx \frac{1}{2}v_a + \frac{1}{2}v_b$$

$$\Rightarrow \frac{1}{2} = \frac{\exp\left(\frac{\mathbf{q}\cdot k_a}{\sqrt{d}}\right)}{\sum_{j=1}^{m} \exp\left(\frac{\mathbf{q}\cdot k_j}{\sqrt{d}}\right)} \text{ and } \frac{1}{2} = \frac{\exp\left(\frac{\mathbf{q}\cdot k_b}{\sqrt{d}}\right)}{\sum_{j=1}^{m} \exp\left(\frac{\mathbf{q}\cdot k_j}{\sqrt{d}}\right)}$$

Substituting value of $\mathbf{q}$, we get

$$\frac{1}{2} = \frac{\exp\left(\frac{C(k_a+k_b)\cdot k_a}{\sqrt{d}}\right)}{\sum_{j=1}^{m} \exp\left(\frac{\mathbf{q}\cdot k_j}{\sqrt{d}}\right)}$$

Since all the $\mathbf{k}$ vectors are orthonormal, $k_i \cdot k_j = 0$ where $i \neq j$. Hence,

$$\frac{\exp\left(\frac{C(k_a+k_b)\cdot k_a}{\sqrt{d}}\right)}{\sum_{j=1}^{m} \exp\left(\frac{C(k_a+k_b)\cdot k_j}{\sqrt{d}}\right)} \approx 0.5 \Rightarrow \frac{\exp\left(\frac{C}{\sqrt{d}}\right)}{\sum_{j=1}^{m} \exp\left(\frac{C(k_a+k_b)\cdot k_j}{\sqrt{d}}\right)} \approx 0.5$$

$$\frac{\exp\left(\frac{C(k_a+k_b)\cdot k_b}{\sqrt{d}}\right)}{\sum_{j=1}^{m} \exp\left(\frac{C(k_a+k_b)\cdot k_j}{\sqrt{d}}\right)} \approx 0.5 \Rightarrow \frac{\exp\left(\frac{C}{\sqrt{d}}\right)}{\sum_{j=1}^{m} \exp\left(\frac{C(k_a+k_b)\cdot k_j}{\sqrt{d}}\right)} \approx 0.5$$

To make the output of softmax function closer to 0.5, we can make the constant C bigger. Thus,

$$\mathbf{q} = C(k_a + k_b)$$

TASK 1.3

Considering the strat that we used in TASK 1.2, we continue to consider $\mathbf{q} = C(k_a + k_b)$.

$$\mathbf{q} \cdot k_a = C(k_a + k_b) \cdot k_a$$

given that,

$$k_i = \mu_i * \lambda_i$$

we get,

$$\mathbf{q} \cdot k_a = C(\mu_a * \lambda_a + \mu_b * \lambda_b) \cdot \mu_a * \lambda_a$$
$$\mathbf{q} \cdot k_a = C(\mu_a * \lambda_a^2)$$

Similarly,

$$\mathbf{q} \cdot k_b = C(\mu_b * \lambda_b^2)$$

For multiple resamples of $\lambda_1 ... \lambda_m$ from the normal distribution, $\mathbf{a}$ will be closer to the average of the v_a and v_b

TASK 1.4

Given that,

$$\mathbf{a} = \frac{1}{2}(\mathbf{a}_1 + \mathbf{a}_2)$$
$$\mathbf{a} \approx \frac{1}{2}(v_a + v_b)$$

This implies,

$$\frac{1}{2}(\mathbf{a}_1 + \mathbf{a}_2) \approx \frac{1}{2}(v_a + v_b)$$
$$(\mathbf{a}_1 + \mathbf{a}_2) \approx (v_a + v_b)$$

From TASK 1.2 and TASK 1.3, we get

$$\mathbf{a}_1 \approx \frac{1}{2}(v_a + v_b)$$

$$\mathbf{a}_2 \approx \frac{1}{2}(v_a + v_b)$$

This implies,

$$\mathbf{q}_1 = C_1(k_a + k_b)$$
$$\mathbf{q}_2 = C_2(k_a + k_b)$$

Since, the keys are in the format described in TASK 1.3, we get

$$\mathbf{q}_1 = C_1(\mu_a * \lambda_a + \mu_b * \lambda_b)$$

$$\mathbf{q}_2 = C_2(\mu_a * \lambda_a + \mu_b * \lambda_b)$$

2 Attention in German-to-English Machine Translation

TASK 2.1

BLEU on test set = 33.82

TASK 2.2

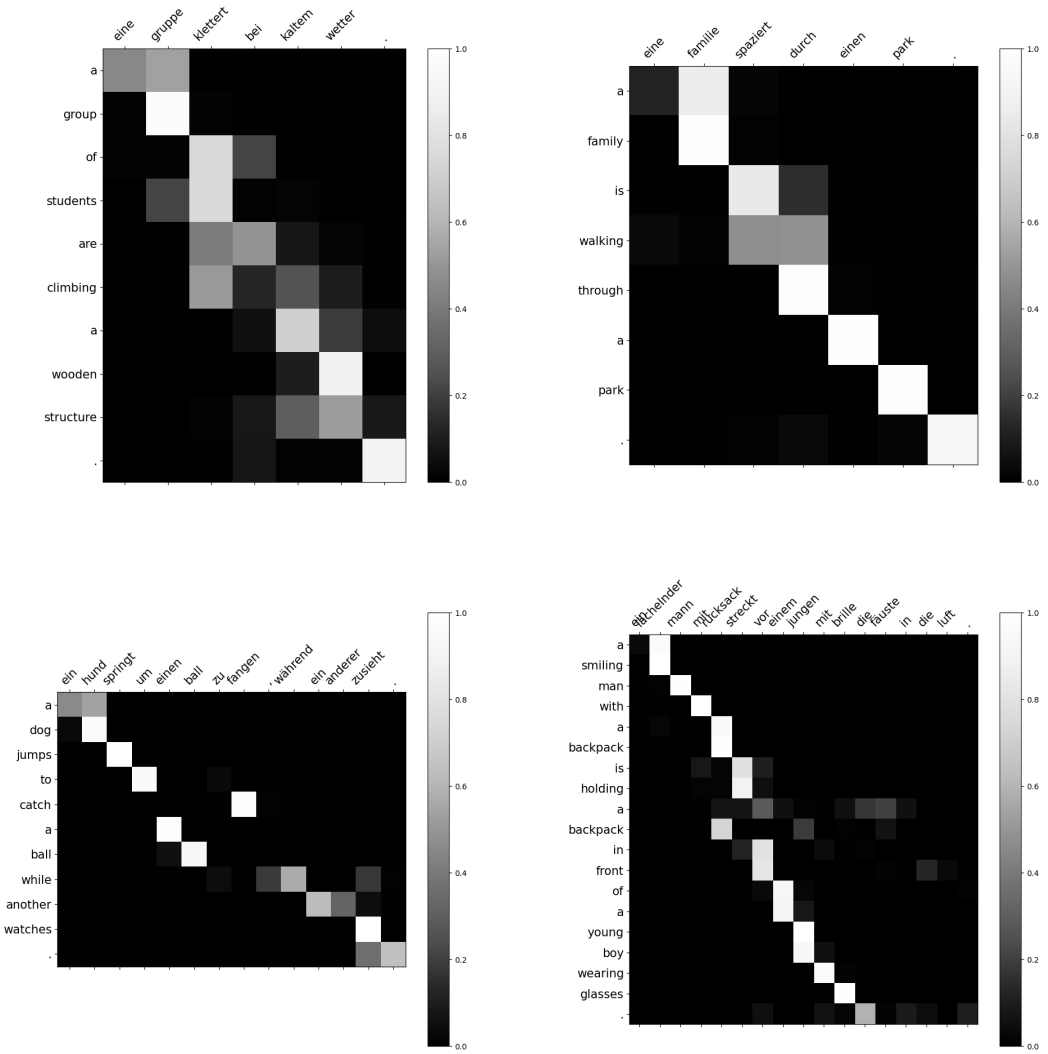

Figure 1: Attention Diagram for German-to-English MT

Common pattern that I observe is that the attention mechanism focuses on a single German word and produces many different English words. In other words, German sentences are more compact while English sentences are more verbose. Another thing that I noticed

is that in some attention diagram, the attention mechanism focuses on next words in the sentences from German sentences, while writing in English. This is because German is Subject-Object-Verb language.

TASK 2.3

Attn.	**PPL**		**BLEU**	
	Mean	**Variance**	**Mean**	**Var**
none	18.0224	0.0204	17.6303	0.1787
sdp	10.9917	0.0011	33.9706	0.0004
mean	15.4491	0.0017	19.5746	0.3239

Table 1: Mean and Variance over three runs

We can see that the scaled-dot product attention is way better than when we don't use attend to any other words at all. Scaled-dot product also performs better when we attend equally to all the words. This is because when we use SDP attention, we only attend to the important words when translating.