

1 Attention Exploration

1.1

(a)

Let us note that due to the formula that defines each α_i the sum of all α_i is 1, and each value is positive. Hence, by looking at each value as category we easily get that this is a categorical distribution.

(b)

This would happen if

1. there exists some k_j where $k_j^T q$ is large and
2. for all other k_i where $i \neq j$ $k_i^T q$ is small.

Let us remember that $a^T b = |a||b| \cos(\theta)$ where θ is the angle between a and b . Hence for the conditions above to hold we need that k_j is parallel to q and all other k_i are in the opposite direction to q .

Plus, we need that either q is very large (i.e large norm) or that k_j is very large or that the other k_i vectors norm is very large.

(c)

Assuming that the conditions above are correct, we get that $c \approx v_j$

(d)

This means that the output c will depend almost only on a single v_j that corresponds to the k_j that is parallel to q . This is, obviously, not desirable since we want the output to be contextualized, i.e depend on the full sentence. Without this property the model will not be expressive enough.

1.2

(a)

Let us define the following matrix M :

$$M = c_1 \frac{a_1 a_1^T}{a_1^T a_1} + c_2 \frac{a_2 a_2^T}{a_2^T a_2} + \dots + c_m \frac{a_m a_m^T}{a_m^T a_m}$$

If we multiply M by a_i we get that

$$M a_i = c_i a_i$$

i.e, a_i is an eigenvector of M with eigenvalue c_i . That means that

$$Mv_a = M(a_1 + \dots + a_m) = Ma_1 + \dots + Ma_m = c_1 a_1 + \dots + c_m a_m = v_a$$

Plus, when multiplying M by v_b we get that

$$Mv_b = c_1 \frac{a_1 a_1^T v_b}{a_1^T a_1} + c_2 \frac{a_2 a_2^T v_b}{a_2^T a_2} + \dots + c_m \frac{a_m a_m^T v_b}{a_m^T a_m}$$

we know that $a_i^T v_b = 0$ for all i since

$$a_i^T v_b = a_i^T b_1 + \dots + a_i^T b_m = 0$$

which is true because the two basis are supposed to be orthogonal. Hence

$$Mv_b = 0$$

Hence we finally get that

$$Ms = Mv_a + Mv_b = v_a$$

(b)

we define q as follows:

$$q = (k_a + k_b) * H$$

where H is some Huge number. It holds that

$$k_a^T q = k_b^T q = H$$

Plus, it holds that

$$k_i^T q = 0$$

if $i \neq a, b$. We get that $\alpha_a = \alpha_b = \frac{e^H}{n-2+2e^H}$ and $\alpha_i = \frac{1}{n-2+2e^H}$ for $i \neq a, b$.

For large enough H we get that $\alpha_a \approx \alpha_b \approx \frac{1}{2}$ and $\alpha_i \approx 0$ for $i \neq a, b$. Hence $c \approx \frac{1}{2}(v_a + v_b)$

1.3

(a)

Let us note that for $k_i \sim N(\mu_i, \alpha_i I)$ we can rewrite k_i as $k_i = \mu_i + \sigma_i$ where $\sigma_i \sim N(0, \alpha_i I)$. Similarly to the clause before we define

$$q = (\mu_a + \mu_b)H$$

we get that for $k_a = \mu_a + \sigma_a$

$$(\mu_a + \sigma_a)(\mu_a + \mu_b)H = H(1 + \sigma_a(\mu_a + \mu_b)) \approx H$$

and similarly for k_b . For k_i where $i \neq a, b$ we get that

$$(\mu_i + \sigma_i)(\mu_a + \mu_b)H = \sigma_i(\mu_a + \mu_b)H \approx 0$$

Hence we get similarly to the previous clause that $c \approx \frac{1}{2}(v_a + v_b)$

(b)

We get that

$$k_a q = (\mu_a + \sigma_a)(\mu_a + \mu_b)H \approx H(1 \pm \|\sigma_a\|)$$

Let us note that $\|\sigma_a\|$ fluctuates from 0 to $\frac{1}{2}\|\mu_a\| = \frac{1}{2}$.

Hence, for different samples, we can see that the value of $k_a q$ fluctuates from $1.5H$ to $\frac{H}{2}$. Hence we get that α_a fluctuates from being equal to $\alpha_b * e^{0.5H}$ to being equal to $\alpha_b * e^{-0.5H}$.

Let us remind our selves that $\alpha_a, \alpha_b < 1$ and that H is a huge number. We that if $\alpha_a = \alpha_b * e^{0.5H}$ then $\alpha_a \approx 1$ and $\alpha_b \approx 0$. Similarly, if $\alpha_a = \alpha_b * e^{-0.5H}$ then $\alpha_a \approx 0$ and $\alpha_b \approx 1$.

Hence, the expectation is $E[c] \approx \frac{1}{2}(v_a + v_b)$ but c fluctuates from $c \approx v_a$ to $c \approx v_b$.

1.4

(a)

As in previous sections, we can design queries that copy a single mean vector. We define express the queries as:

$$q_1 = H\mu_a, \quad q_2 = H\mu_b, \quad \text{for some large } H$$

Since the means are orthogonal we have:

$$c_1 \approx v_a; \quad c_2 \approx v_b$$

And since multiheaded attention is just an average of the 2 values, we can see that:

$$c \approx \frac{1}{2}(v_a + v_b)$$

(b)

By the assumptions made in the question it holds that:

$$\begin{aligned} k_a^T q_1 &= \gamma \mu_a^T H \mu_a = \gamma H \quad \text{where } \gamma \sim \mathcal{N}(1, 0.5) \\ k_b^T q_2 &= \gamma \mu_b^T H \mu_b = H \end{aligned}$$

Again, since all other key-value products are insignificant with regards to H the softmax approaches the maximum:

$$\alpha_a^1 \approx 1, \quad \alpha_b^2 \approx 1$$

so by definition the final output approximates the average:

$$c \approx \frac{1}{2}(v_a + v_b)$$

2 Pretrained Transformers

2.d

The result of london baselinr is -

Correct: 25.0 out of 500.0: 5.0%

The result of the finetuned model is -

Correct: 10.0 out of 500.0: 2.0%

2.f

Correct: 115.0 out of 500.0: 23.0%

2.g

Let us note that the pretraining data contains the birthplace information about the individuals which are in the birth_dev dataset. It seems that the model is able to "remember" the birthplace knowledge that it acquires during pretraining, and it is able to infer that knowledge during evaluation. Let us also note that the finetuning process does not make the model forget the knowledge it acquired during pretraining.

3 In-Context Learning

3.1.b

Results for beam search were better for all prompt sizes. This is probably due to the trade-off between precision and creativity when considering "beam search" or "sampling decoding".

3.2.b

Validation question:

What are Jimmy Reed, Skip James, Elmore James and Big Bill Broonzy best known as?

Prompt:

Answer: Eleonore Roosevelt Question: On which Caribbean island did Princess Diana spend he first Christmas after her divorce was announced? Answer: Barbuda Dependency, Antigua and Barbuda Question: Who wrote the song Momma Told Me Not To Come? Answer: Randy Newman Question: What is the distance between bases on a little league baseball field? Answer: sixty distance Question: Which NASA space probe was launched to Venus in 1989? Answer: List of things named for Ferdinand Magellan Question: Which member of the Monkees came from Washington DC? Answer: Peter Tork Question: Who was the woman sentenced to six years in jail after the murder of Stompei Seipi? Answer:

Winnifred Mandela Question: Which Italian fashion designer was murdered on the orders of his ex-wife?

Answer: House of Gucci

Validation question:

For which movie did Katharine Hepburn win her second Oscar?

Prompt:

Question: Which African country is sandwiched between Ghana and Benin? Answer: Religion in Togo

Question: Which African country is bordered by Benin, Ghana, Ivory Coast, Niger, and Mali? Answer:

Burkina Faso Question: Which country mainly makes up the Horn of Africa? Answer: Somolia Question:

In The Banana Splits what sort of animal was Snorky? Answer: Elephantineness Question: What are

the two main arms of the River Nile called? Answer: Blue Nile and White Nile Question: Name the

East African country which lies on the equator. Answer: Prehistory of Kenya Question: What are

the international registration letters of a vehicle from Algeria? Answer: DZ (disambiguation) Question:

Banting and Best pioneered the use of what? Answer: Insullin

By selecting the closest vectors to the test question, we are effectively providing the model with contextual information that closely matches the specific question being asked. This approach helps the model focus on relevant training examples that are more likely to contain useful information for generating accurate answers.

3.3

In zero-shot learning, the model has not been explicitly trained on the specific task of question answering. Instead, it relies on its general language understanding and knowledge acquired during pre-training. Without task-specific training, the model may lack the fine-tuned knowledge and expertise necessary to generate accurate and relevant answers.

Some generation examples:

I'm not sure, but I'm pretty sure it's not me. I don't 'Run!|| †Run for your life! ‡†Run Gummo: "Chico" and "Harpo" are my last names. It depends on the type of plant. For example, if the plant is a shrub

4

4.a

The model's accuracy on the validation set was 0.1 (only correct on question 9).

4.b

The model's accuracy on the validation set remained 0.1 (again only correct on question 9). Note that the model failed to generate a consistent answer to question 7.

4.c

The model's accuracy on the validation set, evaluated using the provided in-context examples and CoT was 0.7.

We note that there was a one to one correspondence between correct final answers and correct explanations.

Out of all the answers, 70% of the explanations were correct (according to my opinion).

In none (0%) of the validation examples did the model produce a correct final answer with an unconvincing explanation.