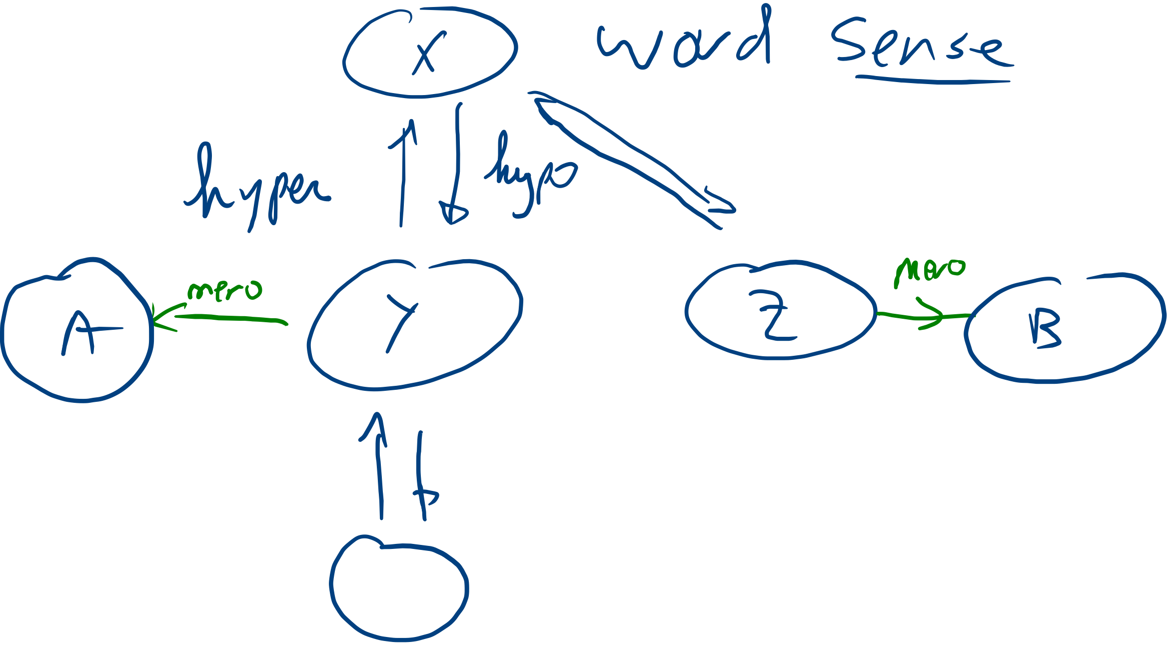
Semantics



RNN

$$\tilde{h}^d = \sum_i a_i h^e \qquad a = \text{softmax}(\text{attention}(h_i^e, h^d))$$

$$\sum_i a_i = 1$$

30% 40% 60-70%

$14 \cdot 10 + 5 \cdot 10 + 2.5 \times 6 \longrightarrow \alpha$

$14 \cdot 10 + 5 \cdot 10 + 2.6 \times 6 \longrightarrow \beta$

$$\frac{e^\alpha}{S} \quad \frac{e^\beta}{S}$$

HMM

$$\underset{t_1^n}{\mathrm{Argmax}}\ P(t_1 \ldots t_n \mid w_1 \ldots w_n)$$

$$P(t_1 \ldots t_n) = P(t_1)\, P(t_2 \mid t_1)\, P(t_3 \mid t_2) \ldots P(t_n \mid t_{n-1})$$

transformers

$$\tilde{H} = \underbrace{\text{softmax}\left(\frac{1}{\sqrt{d_k}}(QW^Q) \cdot (KW^K)\right)}_{\text{attention}} \cdot (VW^V)$$

self-attention: $Q = K = H$

CBOW

$$P(\text{word} \mid \text{context})$$

$$\text{softmax}\left(U \cdot \sum_{\text{context}} \text{embed}(w)\right)$$

④ **[5 pt]** Considering the probability of a **word** sequence $w_1...w_n$, what is the fundamental difference between a 2-gram language model and an order-1 HMM Part-of-Speech tagger?

Support your claim by providing the formula of $P(w_1, ..., w_n)$ in both cases.

⑤ **[12 pt]** Consider the following sentence:

the quick fox jumps over the lazy dog

and an order-1 HMM for Part-of-Speech tagging with the following parameters (not exhaustive, but no missing information to solve the question):

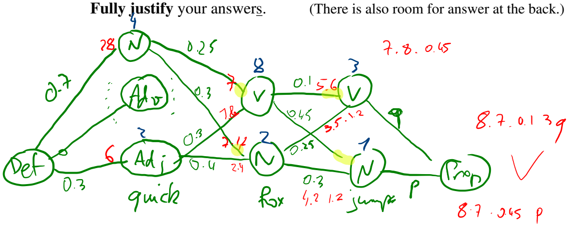
the: Det
quick: Adj: $2 \cdot 10^{-4}$, Adv: $9 \cdot 10^{-4}$, N: $4 \cdot 10^{-4}$
fox: N: $2 \cdot 10^{-4}$, V: $8 \cdot 10^{-4}$
jumps: N: 10^{-4}, V: $3 \cdot 10^{-4}$
over: Prep
lazy: Adj
dog: N: $6 \cdot 10^{-4}$, V: $7 \cdot 10^{-4}$

	Adj	Adv	Det	N	V	Prep
Adj	0.15	0.1	0.3	0.2	0.05	0.25
Adv	0.05	0.2	**0**	0.1	0.15	**0**
Det	0.02	0.1	**0**	0.04	0.05	0.3
N	0.4	0.1	0.7	0.3	0.45	r
V	0.3	0.4	**0**	0.25	0.1	s
Prep	0.02	0.1	**0**	p	q	**0**

(a) [8 pt] Provide the tightest possible condition(s) between p, q, r and s so that the tag of "*jumps*" in the most probable sequence of tags for the above sentence is V.

(b) [4 pt] If these conditions are fullfiled, what is the most probable sequence of tags for the above sentence?

Fully justify your answers. (There is also room for answer at the back.)



You decide to use the continuous bag of words algorithm to train your word embeddings. To test whether your training algorithm works correctly, you test it with a small vocabulary of five words and provide it the sequence of words "*what day is the exam*" with the following embeddings:

$$\begin{aligned} \text{what} &= [\ \ln 2, \quad \ln 0.5\ ] \\ \text{day} &= [\ \ln 0.5, \ \ln 2\ ] \\ \text{is} &= [\ \ln 0.5, \ \ln 0.5\ ] \\ \text{the} &= [\ \ln 1.5, \ \ln 0.5\ ] \\ \text{exam} &= [\ \ln 2, \quad \ln 2\ ] \end{aligned}$$

1st
ln 2 + ln 0.5 + ln 1.5 + ln 2
= ln 3
2nd
ln 0.5 + ln 2 + ln 0.5 + ln 2
= 0

(where $\ln$ is the natural logarithm function of base e);
and output vocabulary projection U:

$$U = \begin{pmatrix} 0 & 1 & 2 & 1 & 0 \\ 1 & 2 & 3 & 2 & 1 \end{pmatrix}$$

$\begin{pmatrix} \ln 3 \\ 0 \end{pmatrix} = \ln 3 \cdot U_1$

You can assume each column of U corresponds to the following vocabulary items:
what, day, is, the, exam.

④ **[6 pt]** Using a window size of 2, what is the probability of the word "*is*" according to the continuous bag of words network?
Justify your answer.

⑤ **[2 pt]** Using a window size of 1, what is the probability of the word "*the*" according to the continuous bag of words network?
Justify your answer.

Now that your embeddings are pretrained, you train your transformer language model. For the following questions, assume a single-headed attention function and use the following input embeddings as key vectors:

$$\begin{aligned} \text{what} &= [\,2,\quad 0.5\,] \\ \text{day} &= [\,0.5,\ 2\,] \\ \text{is} &= [\,0.5,\ 0.5\,] \\ \text{the} &= [\,2,\quad -2\,] \\ \text{exam} &= [\,1,\quad 1\,] \end{aligned}$$

⇒ K

⑥ **[6 pt]** Using scaled dot product attention, what is the attention distribution over key vectors for the word "*exam*" as the query in the first attention layer? You can ignore position embeddings. Assume that W^K, W^V are identity matrices and

$$W^Q = \begin{pmatrix} \sqrt{2}\,\ln(4) & 0 \\ 0 & \sqrt{2}\,\ln(4) \end{pmatrix}$$

$= \sqrt{2}\,\ln(4)\, I_2$

Justify your answer and provide all the steps of your computation.

$\frac{1}{\sqrt{2}} \sqrt{2} \ln 4 \; h \cdot K$

[1,1]

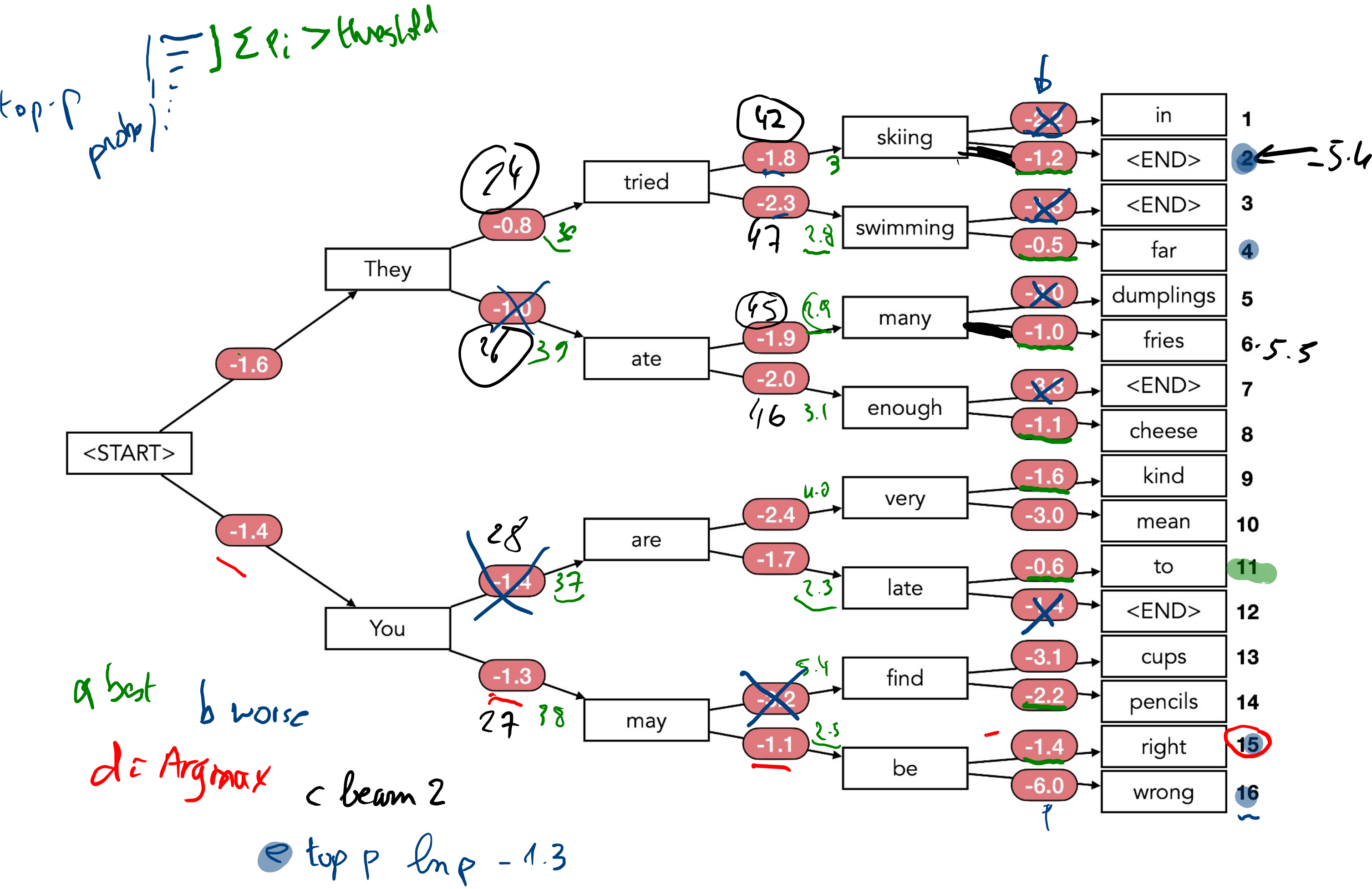
what	day	is	the	exam
2.5	2.5	1	0	2
32	32	4	1	16 → S

Softmax ↳ 4^i

S

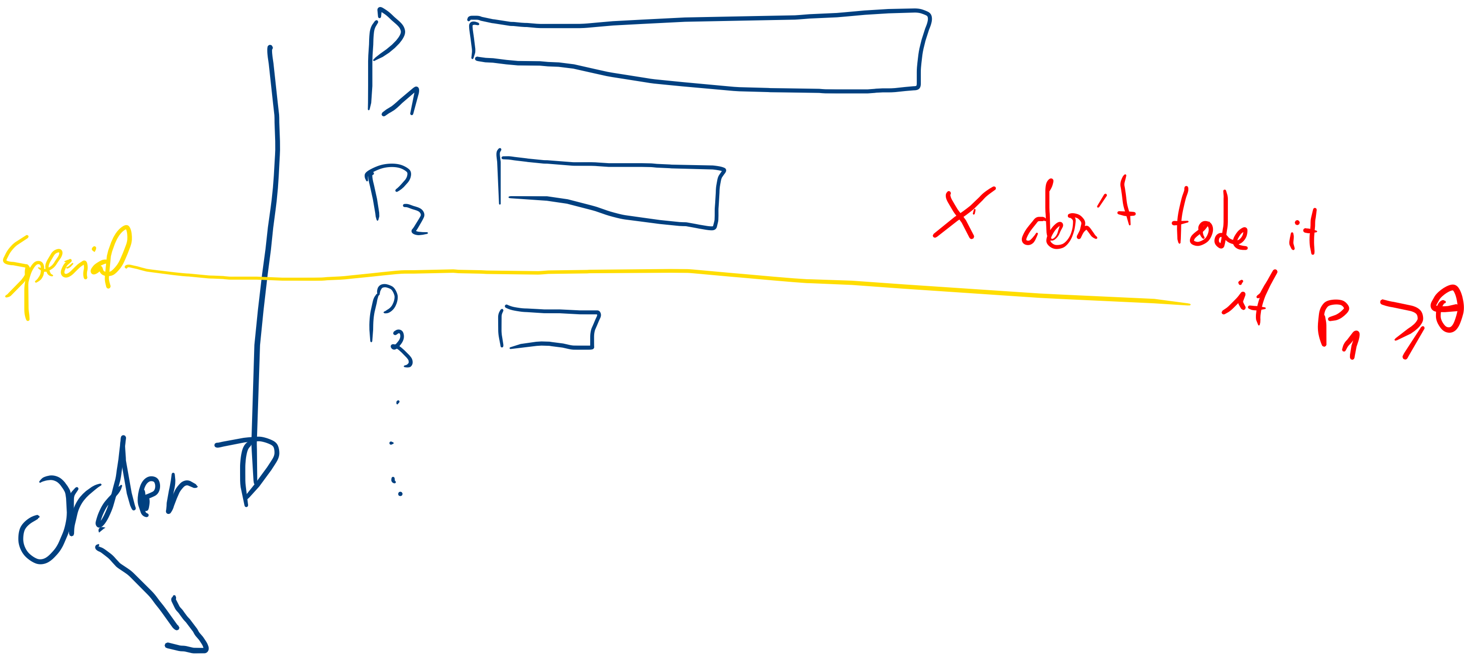
⑦ **[2 pt]** What is the attention distribution if the position embedding in the first position is $[-1,\ 0.5]$ and the others are $[0,\ 0]$?
Justify your answer.

continues on back ☞



<START>
-1.6
-1.4
They
You
-0.8
-1.0
-1.4
-1.3
tried
ate
are
may
-1.8
-2.3
-1.9
-2.0
-2.4
-1.7
-3.2
-1.1
skiing
swimming
many
enough
very
late
find
be
-2.2
-1.2
-3.3
-0.5
-3.0
-1.0
-3.3
-1.1
-1.6
-3.0
-0.6
-1.4
-3.1
-2.2
-1.4
-6.0
in
<END>
<END>
far
dumplings
fries
<END>
cheese
kind
mean
to
<END>
cups
pencils
right
wrong
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16

top-p



k best