DS-GA 1011 NLP

Fall 2023

Recitation 4

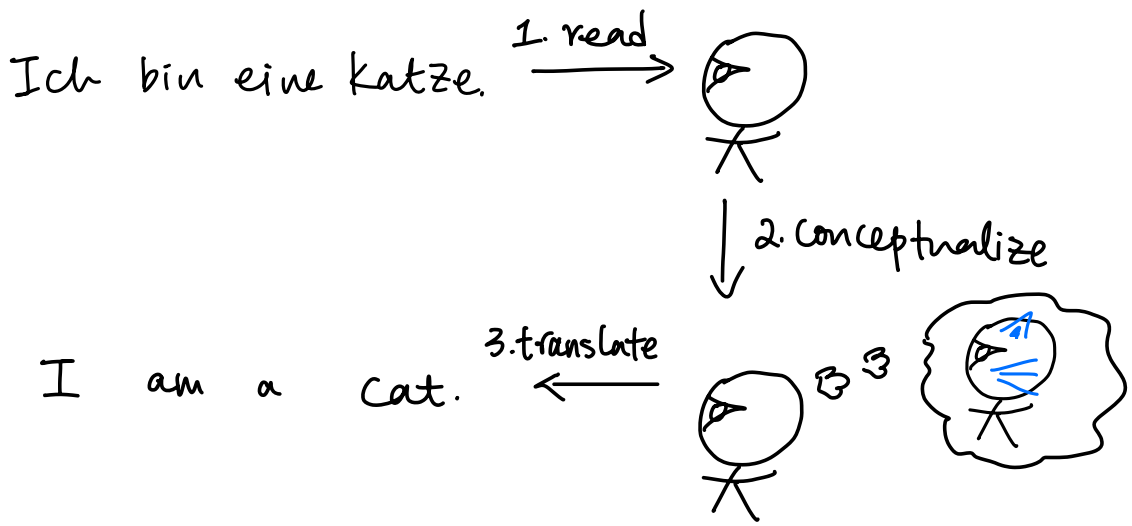
Encoder-Decoder Model for Machine Translation

Lavender Jiang

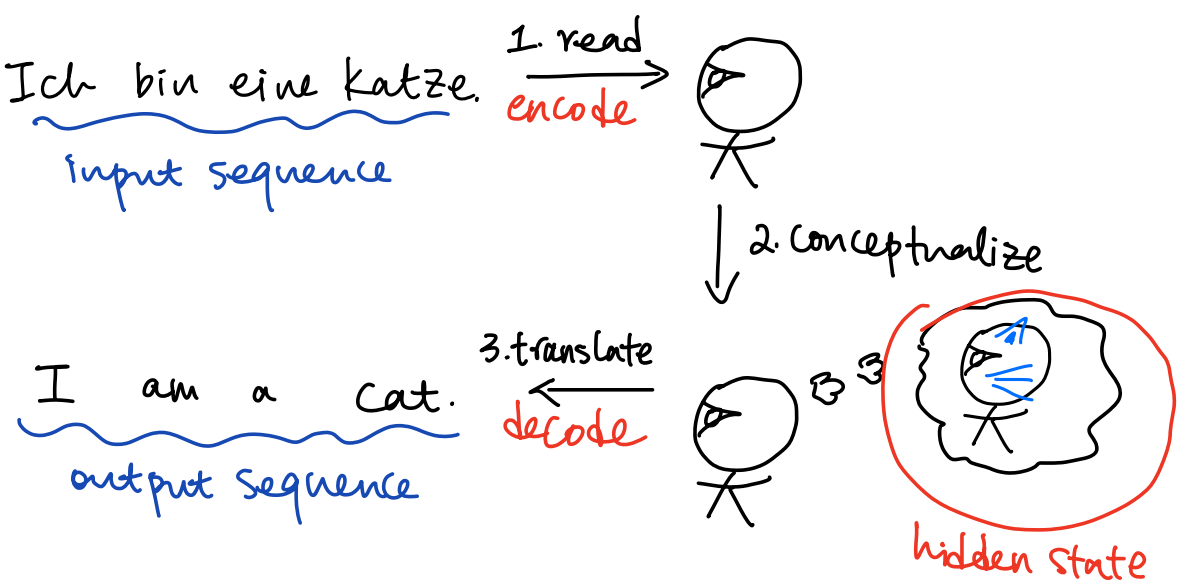
Sep 29, 2023

German	English
Ich bin eine katze.	I am a cat.
Ich bin eine frau.	I am a woman.
Ich bin Bürger.	I am a citizen.

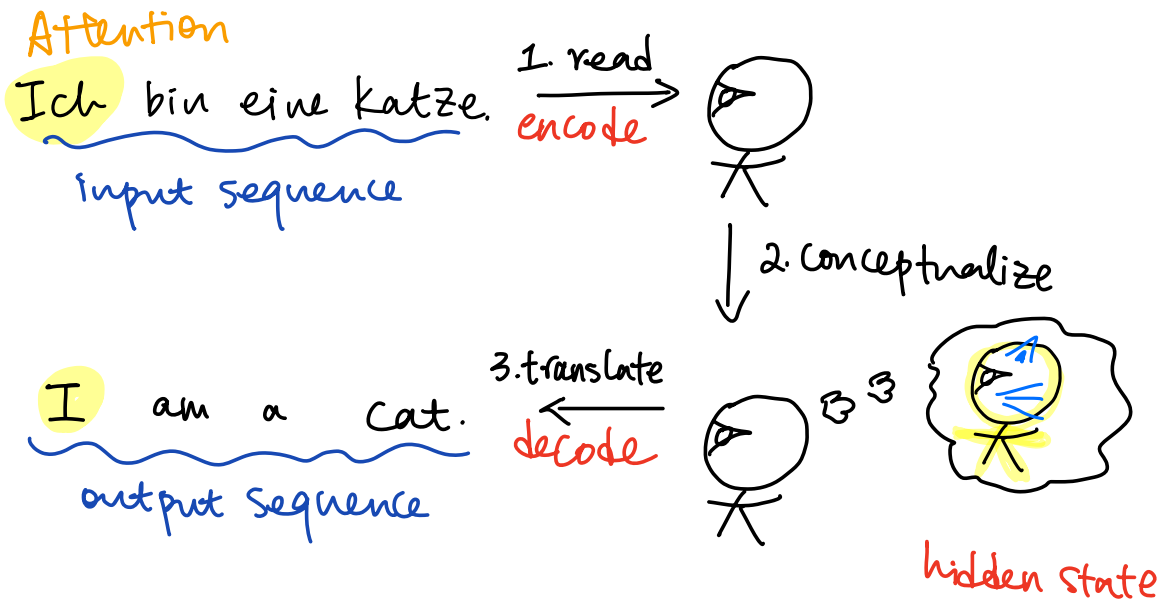
1



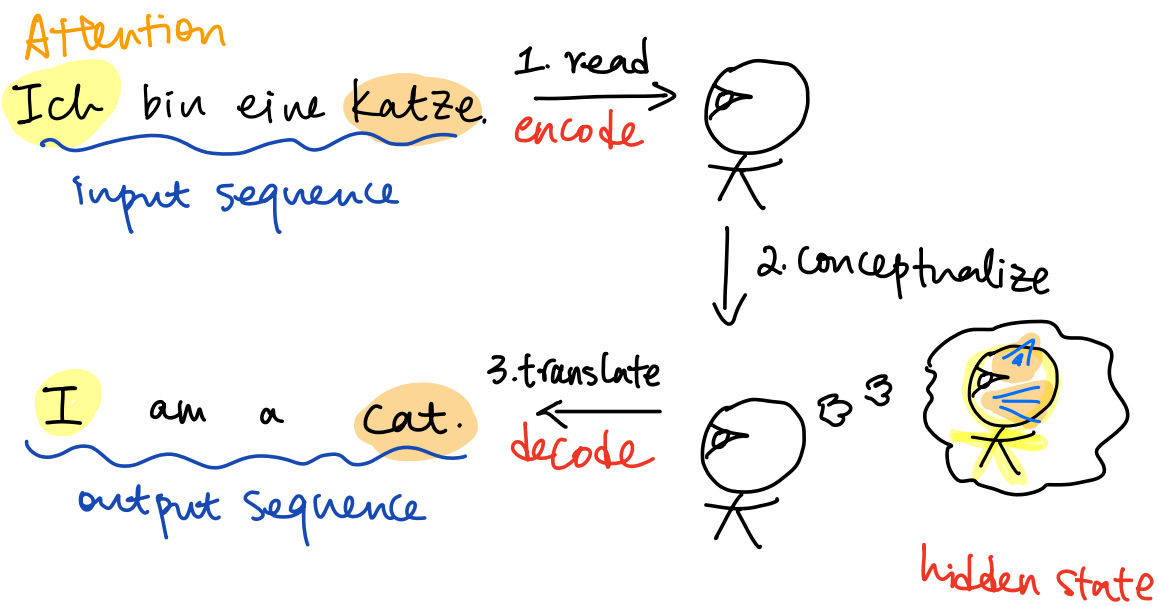
2



3

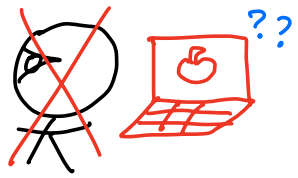

Attention
Ich bin eine katze.
input sequence
1. read
encode
2. conceptualize
I am a cat.
output sequence
3. translate
decode
hidden state

4


Attention
Ich bin eine katze.
input sequence
1. read
encode
2. conceptualize
I am a cat.
output sequence
3. translate
decode
hidden state

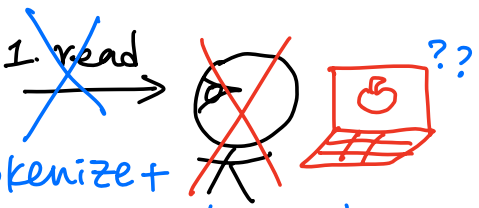
5

Ich bin eine katze. 1. read ??



6

Ich bin eine katze. ~~1. read~~ ??



tokenize + parameterized encoder

7

Ich bin eine katze

↓ tokenize

[1, 3, 4, 5]

vocab	id
Ich	1
du	2
bin	3
eine	4
katze	5
frau	6

8

<bos> Ich bin eine katze <eos>

↓ tokenize

[7, 1, 3, 4, 5, 8]

vocab	id
Ich	1
du	2
bin	3
eine	4
katze	5
frau	6
special tokens: <bos>	7
special tokens: <eos>	8

9

<bos> Ich bin eine katze <eos>

↓ tokenize

[7, 1, 3, 4, 5, 8]

<bos> Ich bin eine studentin <eos>

↓ tokenize

[7, 1, 3, 4, 9, 10, 8]

vocab	id
Ich	1
du	2
bin	3
eine	4
katze	5
frau	6
<bos>	7
<eos>	8
student	9
##in	10

Suppose encoder is linear. i.e. $Enc := A$.

vocab size = 10. hidden size = 20. $A \in \mathbb{R}^{10 \times 20}$

Then hidden for input s is $h = Enc(\overbrace{tokenize(s)}^{x}) = Enc(x) = Ax$

<bos> Ich bin eine katze <eos> s

↓ tokenize

[7, 1, 3, 4, 5, 8] $x \in \mathbb{R}^{1\times 6}$

each token has a learned embedding

↓ index for embedding

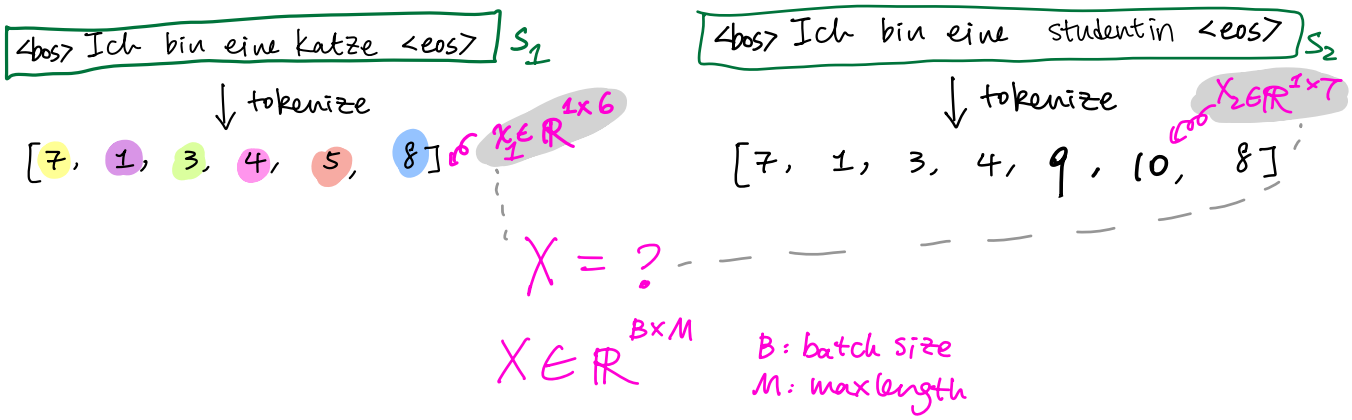
$x' \in \mathbb{R}^{1\times 6\times 10}$

$x' = [e_7, e_1, e_3, e_4, e_5, e_8]$

$A =$ rows $v_1, v_2, v_3, v_4, v_5, v_6, v_7, v_8$ (columns $h_1, h_2, \ldots, h_{20}$)

$h \in \mathbb{R}^{1\times 6\times 20}$ $(1\times 6\times 10) \times (10\times 20)$ (x × A)

$\rightarrow [v_7, v_1, v_3, v_4, v_5, v_8]$



12

$M = 8, B = 2$

[<bos> Ich bin eine katze <eos> <pad>]
[<bos> Ich bin eine student ##in <eos>]

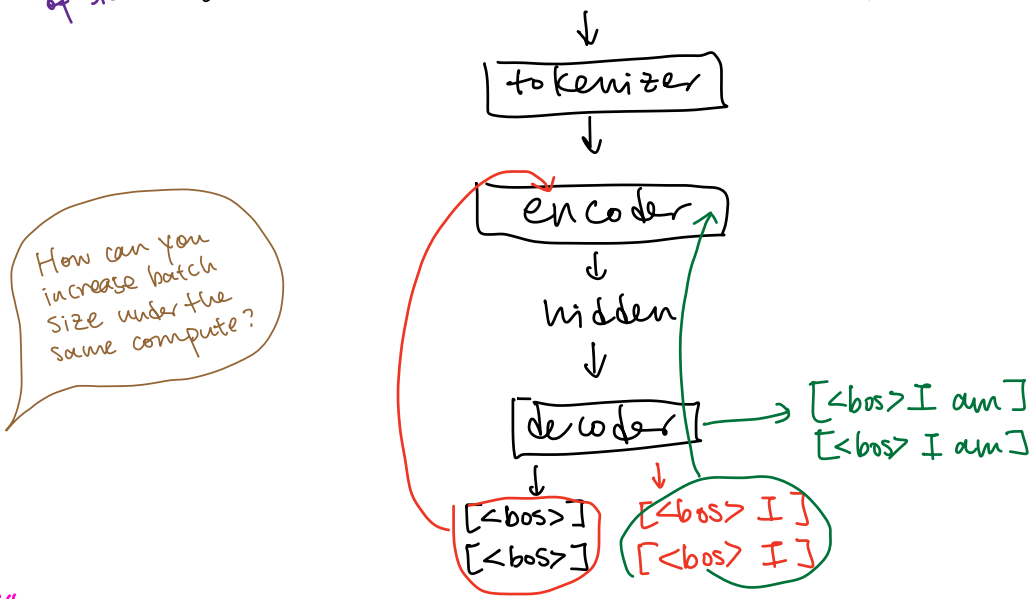
$M = 9, B = 2$

[<bos> Ich bin eine katze <eos> <pad> <pad>]
[<bos> Ich bin eine student ##in <eos> <pad>]

13

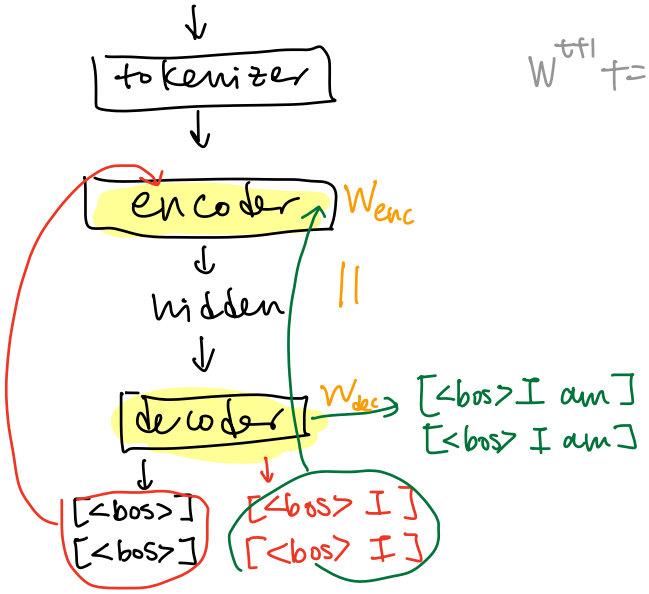
---

batch of size 2 {
[<bos> Ich bin eine katze <eos> <pad> <pad>]
[<bos> Ich bin eine student ##in <eos> <pad>]



14

batch of size 2 { [<bos> Ich bin eine katze <eos> <pad> <pad>]
[<bos> Ich bin eine student ##in <eos> <pad>]



$W^{tfl} += \eta(\nabla W_{enc} + \nabla W_{dec})$

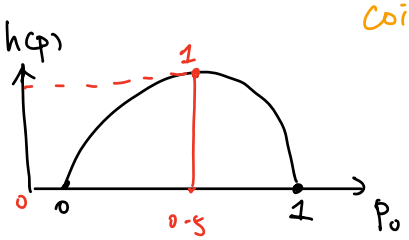
weight sharing!
also helps w
- overfitting
- robustness
- semantic matching

14

---

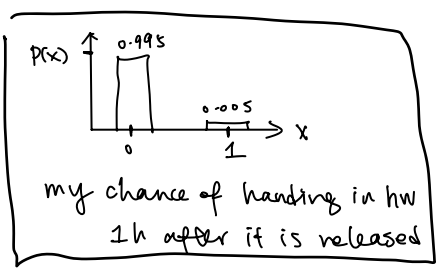
entropy measures the "uncertainty" / "chaoticness" of a dist.

coin flip $1

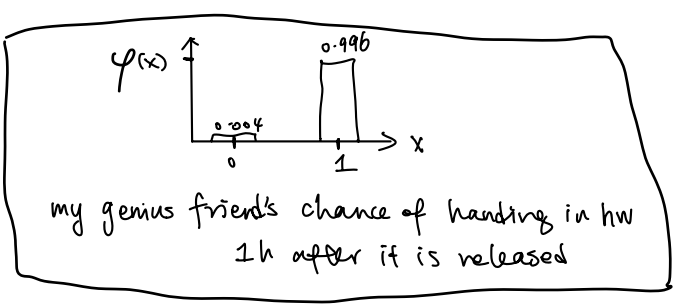


$$h(p) = -\sum_i p_i \log p_i = \mathbb{E}_p[-\log p] \in [0,1]$$

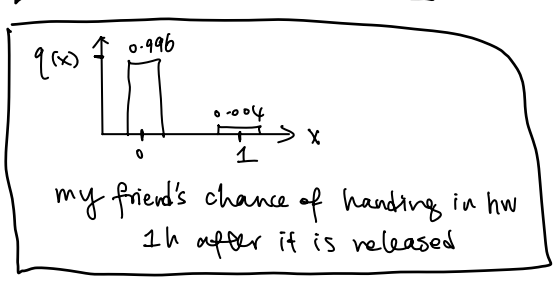
cross entropy measures how "close" two distributions are



my chance of handing in hw 1h after it is released



my genius friend's chance of handing in hw 1h after it is released



my friend's chance of handing in hw 1h after it is released

$$H(p,q) = -\sum_i p_i \log q_i = \mathbb{E}_p[-\log q_i]$$

$$KL(p,q) = -\sum_i p_i \log\left(\frac{q_i}{p_i}\right) = \mathbb{E}_p\left[-\log \frac{q_i}{p_i}\right]$$

$$H(p,q) = KL(p,q) - H(p)$$

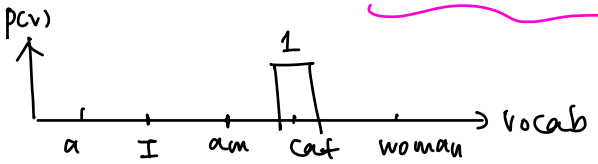
minimize $H(p,q) \Rightarrow$ minimize $KL(p,q)$

15

In our case, $p = \mathbb{P}(y_{t+1} \mid \underbrace{\text{context}}_{y_{\leq t},\ \vec{x}})$,

$$q = \mathbb{P}(\hat{y}_{t+1} \mid \text{context})$$

Concrete example: $p = \mathbb{P}(y_{t+1} \mid [\text{I}, \text{Ich}, \text{bin}, \text{eine}, \text{Katz}])$



$q = \mathbb{P}(\hat{y}_{t+1} \mid [\text{I}, \text{Ich}, \text{bin}, \text{eine}, \text{Katz}])$