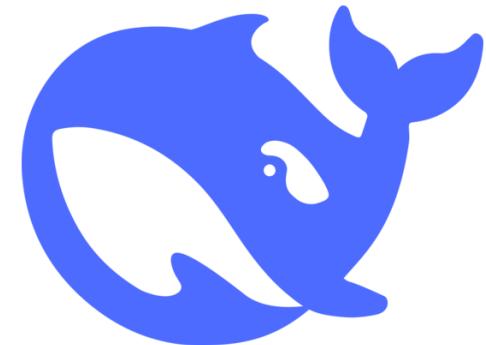


Gemini



VALDOM - NLP 2 LLM

LARGE LANGUAGE MODELS

JOSEBA DALMAU

LLM COMPARISON

Task: search online for a common set of criteria to compare LLMs against each other.

LLM COMPARISON

Homework: Pick a LLM from the list and research about it online in order to fill in the comparison table.

Gradient Updates

How has DeepSeek improved the Transformer architecture?

Published

Jan 17, 2025

Authors

Ege Erdil

Share

 Twitter

 LinkedIn

DeepSeek has recently released DeepSeek v3, which is currently state-of-the-art in benchmark performance among open-weight models, alongside [a technical report](#) describing in some detail the training of the model. Impressively, they've achieved this SOTA performance by only using 2.8 million H800 hours of training hardware time—equivalent to about 4e24 FLOP if we assume 40% MFU. This is about ten times less training compute than the similarly performing Llama 3.1 405B.

In this issue, I'll cover some of the important architectural improvements that DeepSeek highlight in their report and why we should expect them to result in better performance compared to a vanilla Transformer. The full technical report contains plenty of non-architectural details as well, and I strongly recommend reading it if you want to get a better idea of the

Face aux erreurs, Linagora suspend temporairement son assistant GenAI Lucie

Jacques Cheminat, publié le 27 Janvier 2025

Lancé en test la semaine dernière, l'assistant GenAI nommé Lucie a été suspendu au début du week-end. Plusieurs réponses erronées ont provoqué un déferlement de critiques sur l'initiative portée par Linagora. Ce dernier a reconnu une erreur de communication et plaide un travail de recherche en phase initiale.



LUCIE — The truly open source AI built on transparency, trust, and efficiency.

SUIVRE TOUTE L'ACTUALITÉ

✉ Newsletter

Recevez notre newsletter comme plus de 50 000 professionnels de l'IT!

JE M'ABONNE



A télécharger sur
guidescomparatifs.com



ERP - 454 critères

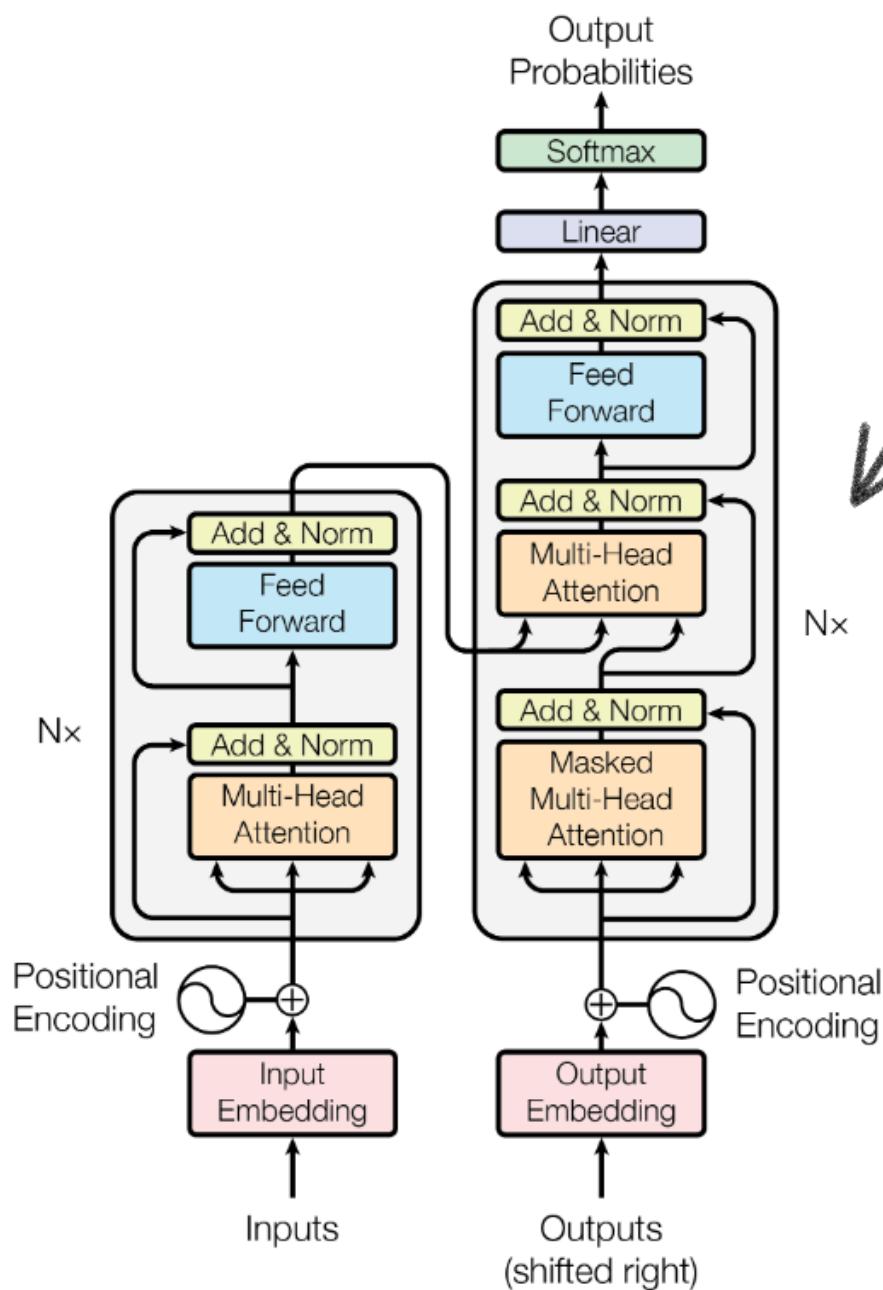
40 pages de **cahier des charges** pour préparer votre projet ERP

[PDF] [EXCEL]

OBJECTIVE

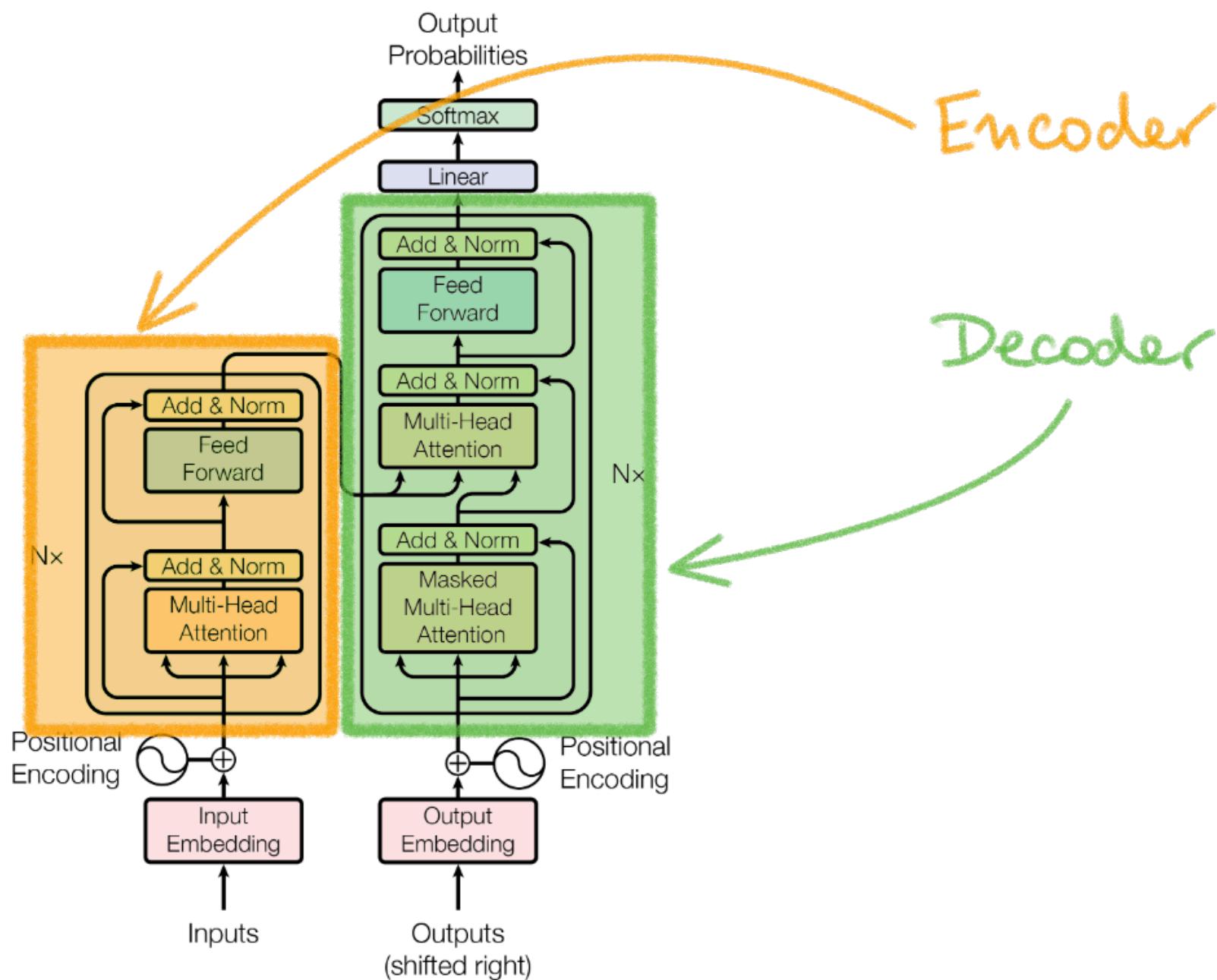
Master the Transformer
architecture for various
different NLP tasks

TRANSFORMERS



Main figure of
the transformer
architecture from
the original paper
"Attention is all
you need"
Vaswani et al. 2017

ENCODER - DECODER ARCHITECTURE



INFERENCE

Je suis enseignant

encoder

decoder

I am a teacher

INFERENCE

encoder

decoder

I am a teacher

"input sentence"

INFERENCE

encoder

decoder

I am a teacher

I, am, a, tea, cher ← "Input tokens"

↑ tokenize

INFERENCE

encoder

decoder

22 357 12 1012 3501

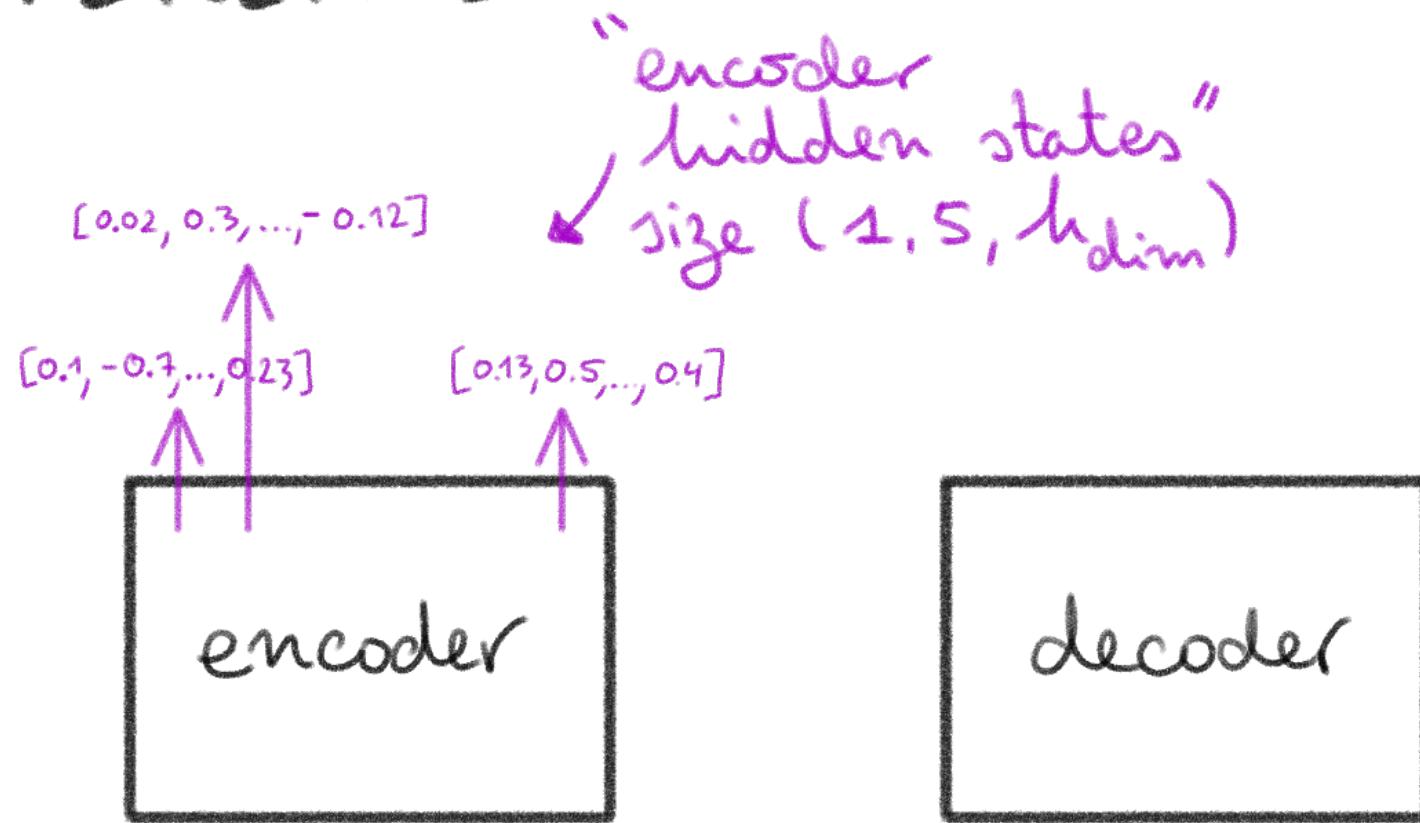
I, am, a, tea, cher

↑ tokenize

I am a teacher

← "input IDs"

INFERENCE



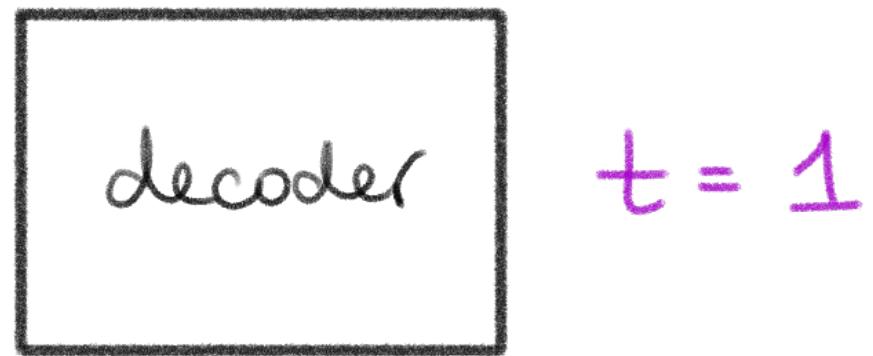
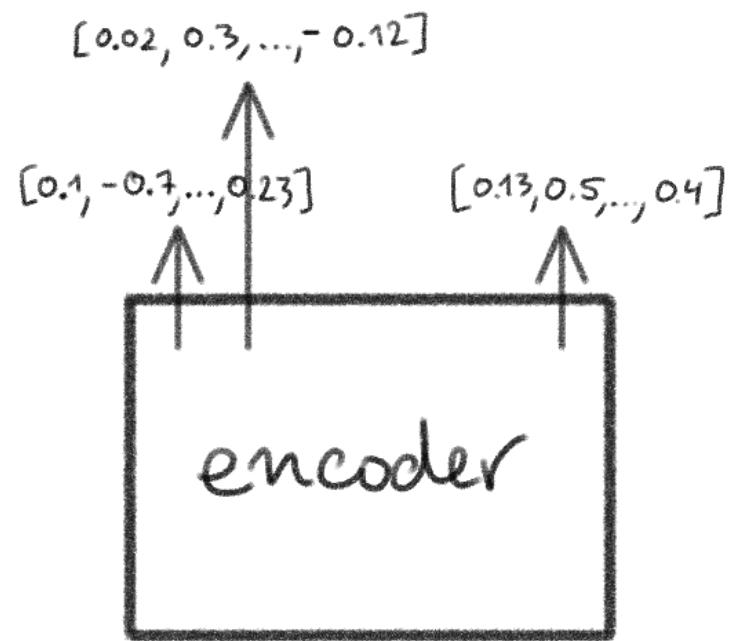
22 357 12 1012 3501

I, am, a, tea, cher

I am a teacher

↑ tokenize

INFERENCE

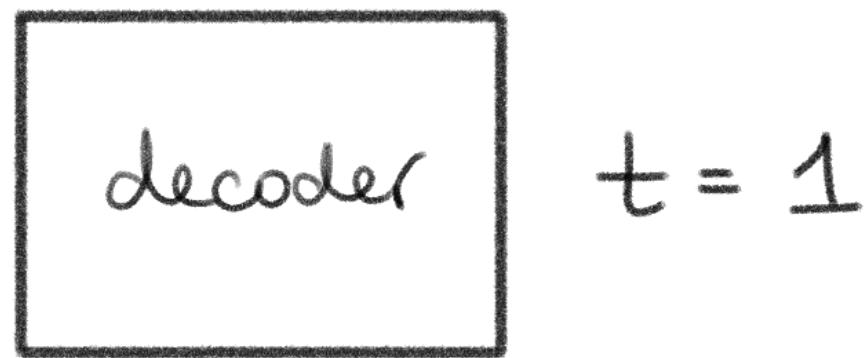
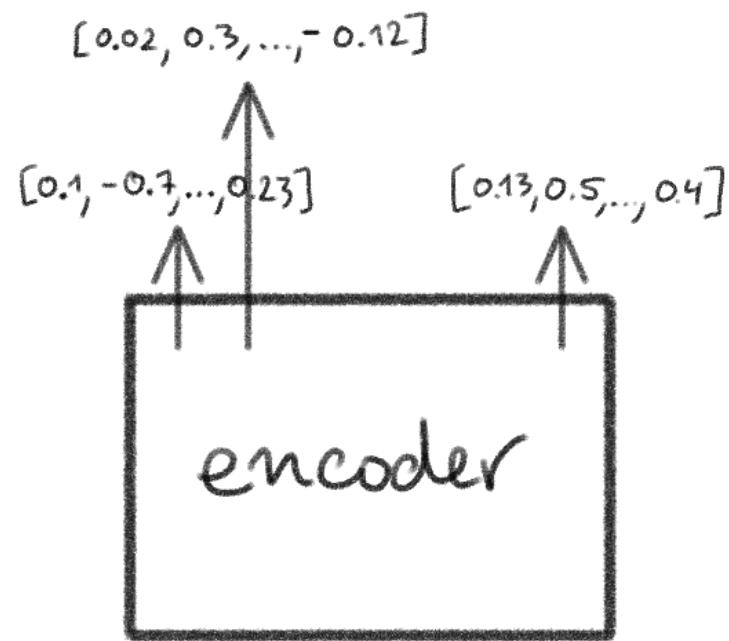


22 357 12 1012 3501

I, am, a, tea, cher
I am a teacher

$\langle s \rangle$
"decoder start token"

INFERENCE



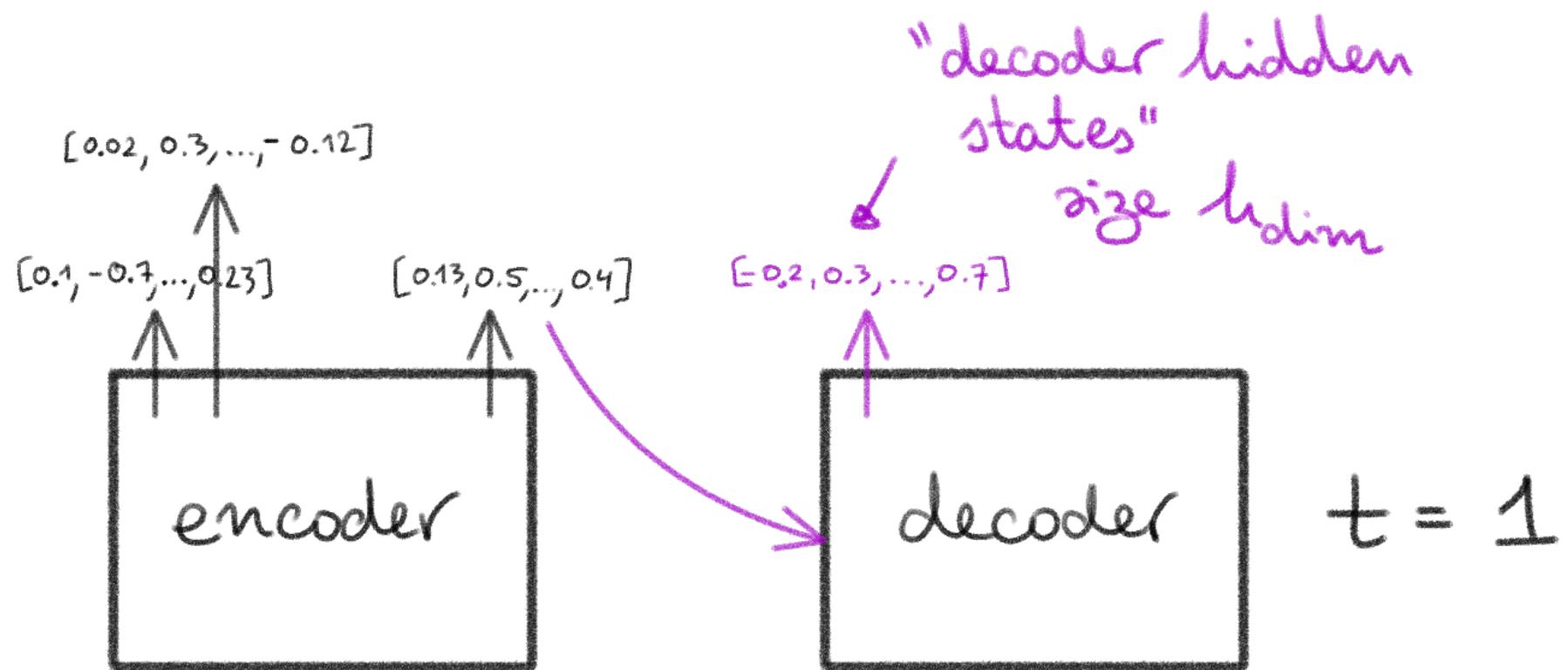
22 357 12 1012 3501

I, am, a, tea, cher

I am a teacher

212 ← "decoder
< s > input IDs"

INFERENCE

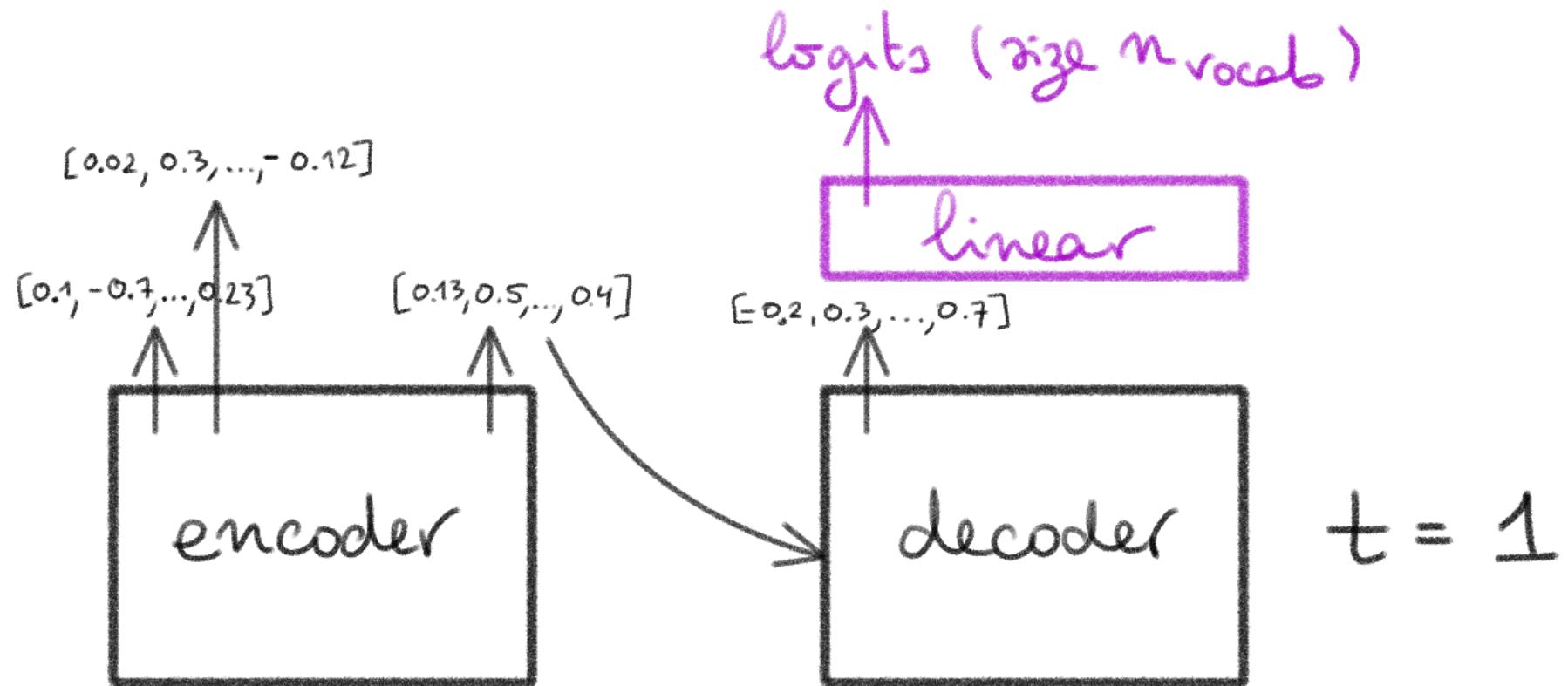


22 357 12 1012 3501 212

I, am, a, tea, cher < s >

I am a teacher
↑ tokenize

INFERENCE

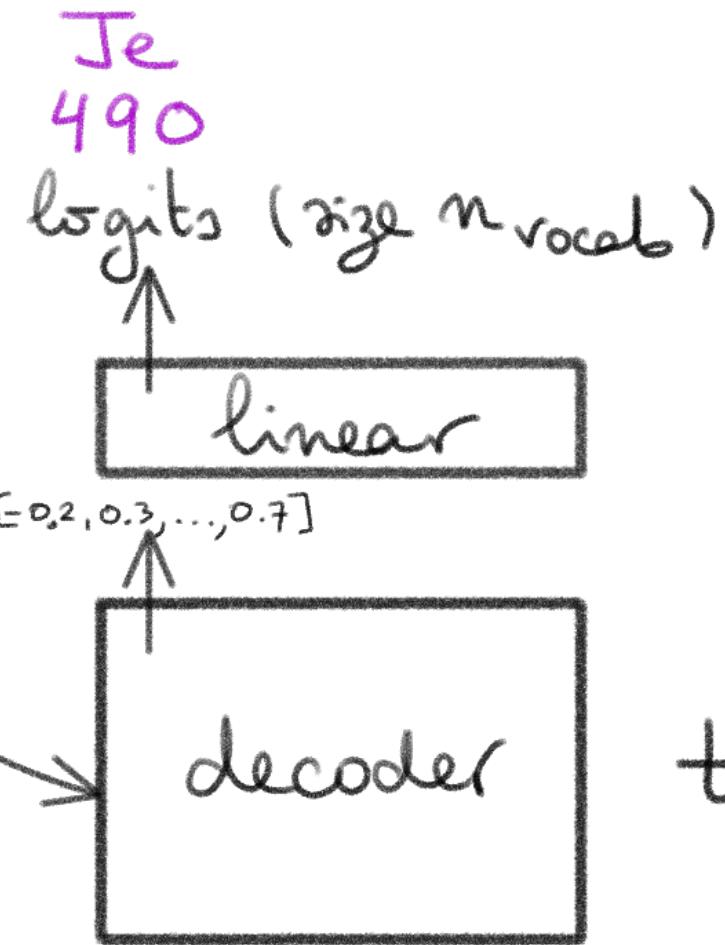
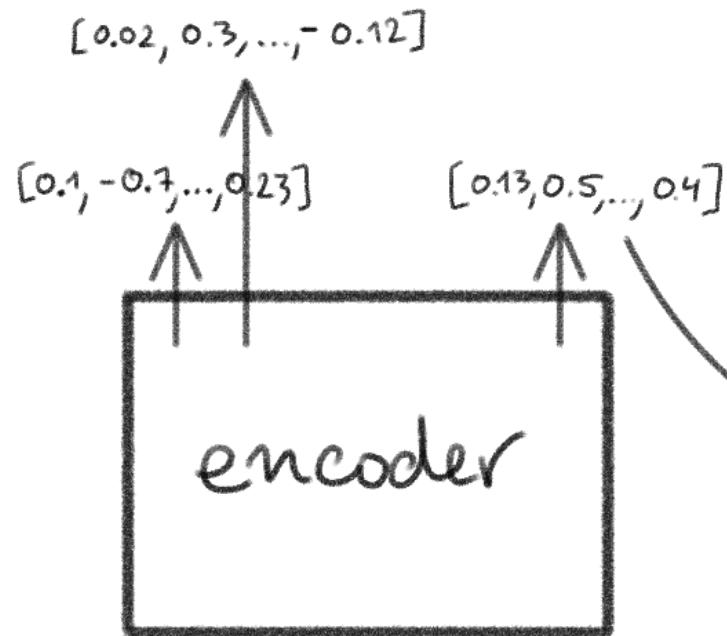


22 357 12 1012 3501 212

I, am, a, tea, cher < s >

I am a teacher
tokenize

INFERENCE



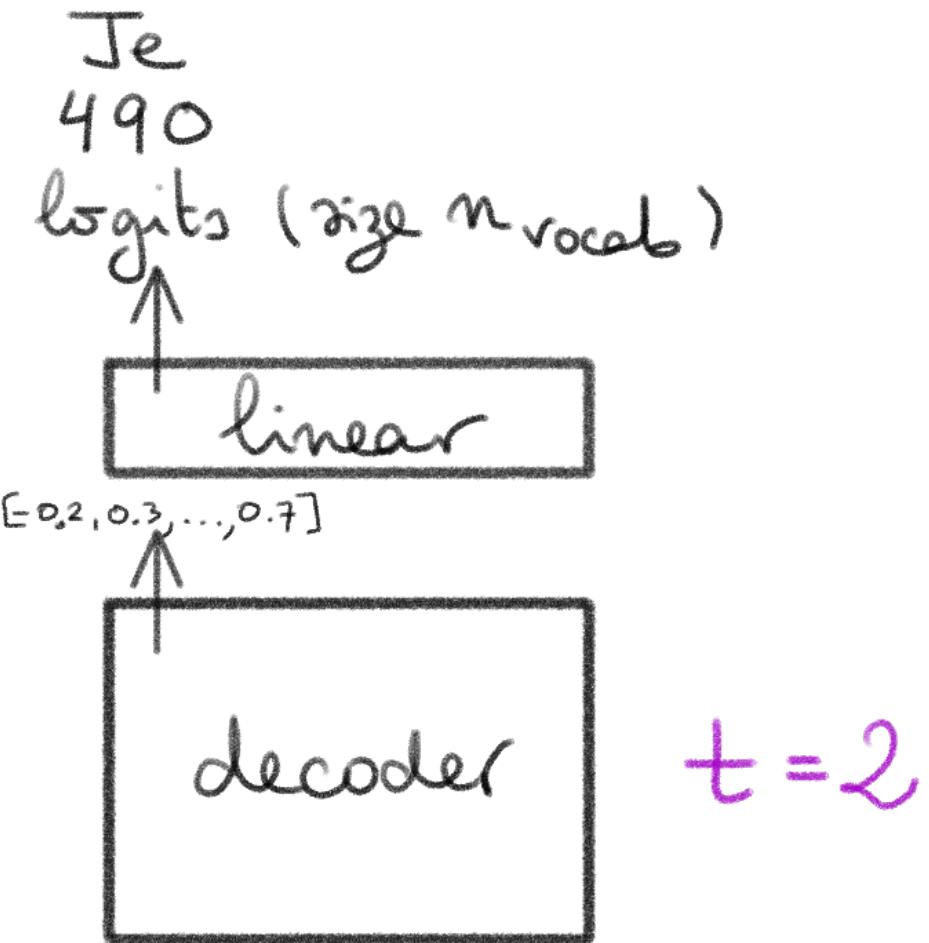
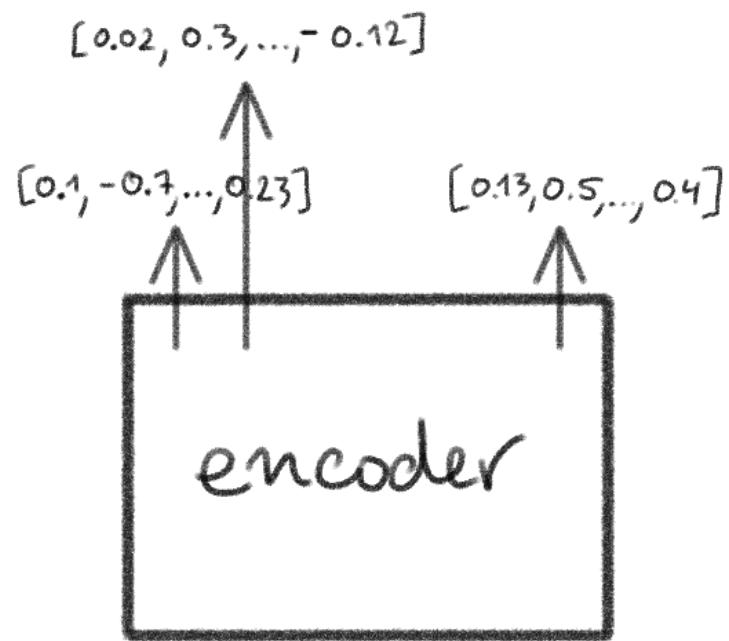
$t = 1$

22 357 12 1012 3501 212

I, am, a, tea, cher < s >

I am a teacher
tokenize

INFERENCE



22 357 12 1012 3501

I, am, a, tea, cher

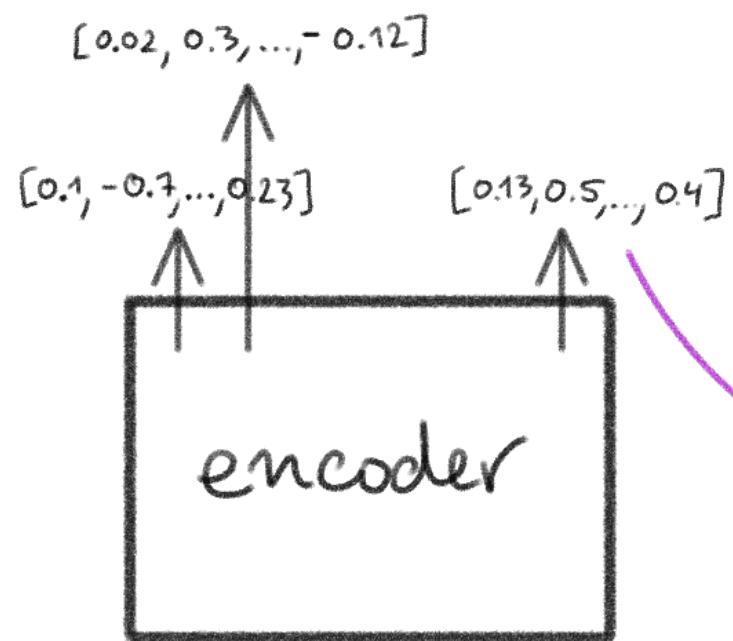
I am a teacher

tokenize

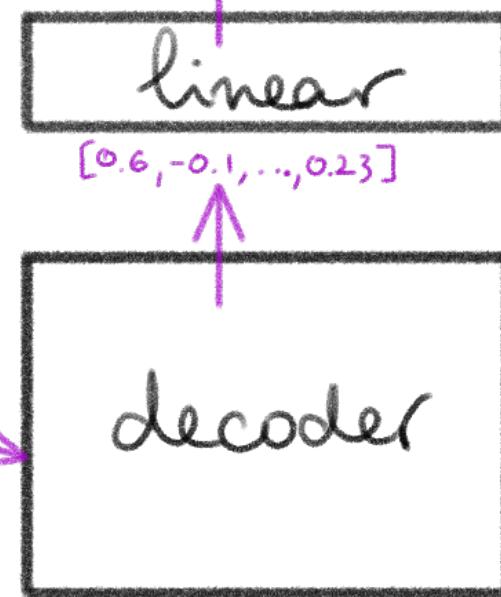
212 490

< s >

INFERENCE



Je suis
1131
logits (size n_vocab)



$t = 2$

22 357 12 1012 3501

I, am, a, tea, cher

I am a teacher

tokenize

212 490

< s >

INFERENCE

We iterate through the decoding process until we hit the "end of sequence" ($\langle \text{EOS} \rangle$) token!

INFERENCE

I am a teacher



Je suis enseignant

Je suis enseignante

⇒ We can randomize the output using softmax instead of argmax of logits !

TRAINING

labels

Je, suis, encei, gnant
490 1131 3212 2757

encoder

decoder

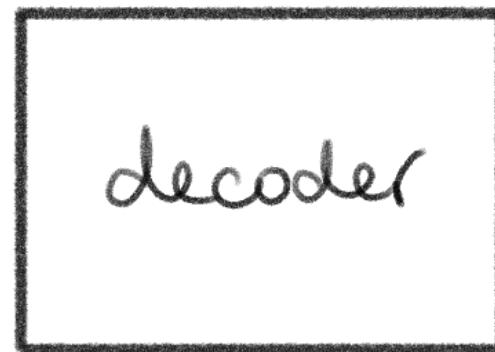
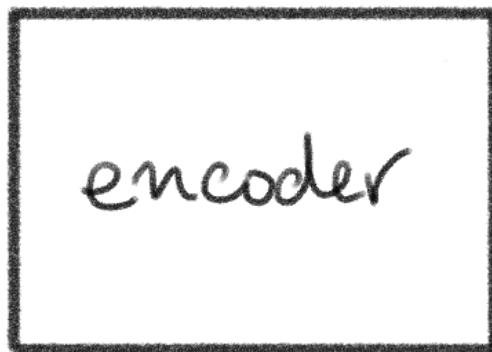
22 357 12 1012 3501

I, am, a, tea, cher

TRAINING

Je, suis, encei, gnant

490 1131 3212 2757



22 357 12 1012 3501

I, am, a, tea, cher

212 490 1131 3212 2757

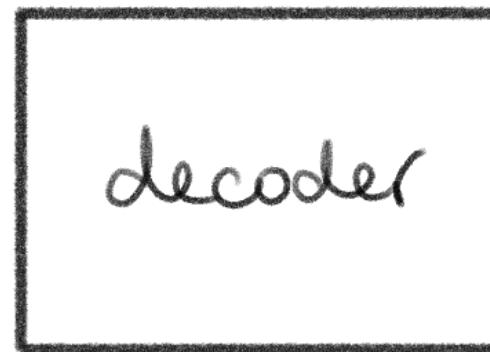
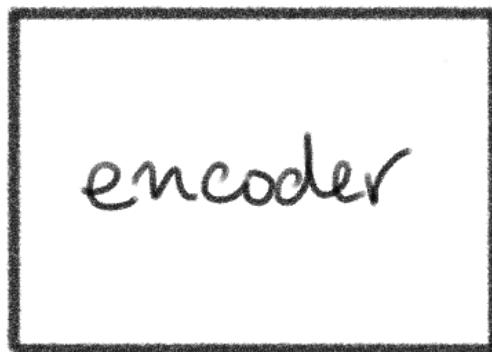
↑
<S>

TRAINING

Je, suis, encei, gnant $\langle \text{EOS} \rangle$

490 1131 3212 2757 5000

l_1 l_2 l_3 l_4 l_5



22 357 12 1012 3501

I, am, a, tea, cher

212 490 1131 3212 2757

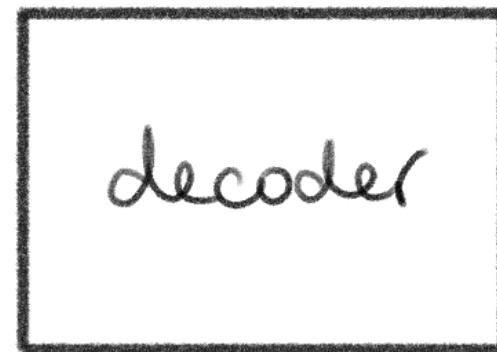
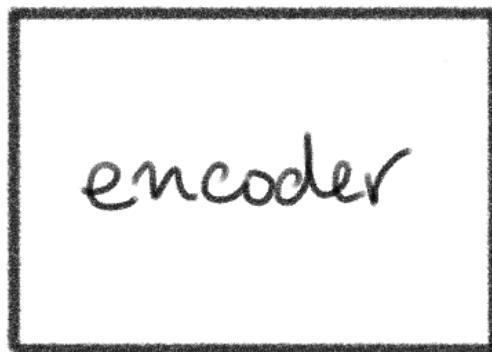
\uparrow
 $\langle S \rangle$

TRAINING

compute
cross-entropy loss
& back-propagation

Je, suis, encei, gnant (EOS)

{ 490 1131 3212 2757 5000
l₁ l₂ l₃ l₄ l₅



22 357 12 1012 3501

I, am, a, tea, cher

212 490 1131 3212 2757

↑
<S>

CROSS-ENTROPY LOSS

Used for multi-class classification

$$L(y, \hat{y}) := - \sum_{k=1}^c y_k \log(\hat{y}_k)$$

ground truth
prediction
probabilities
(one-hot)

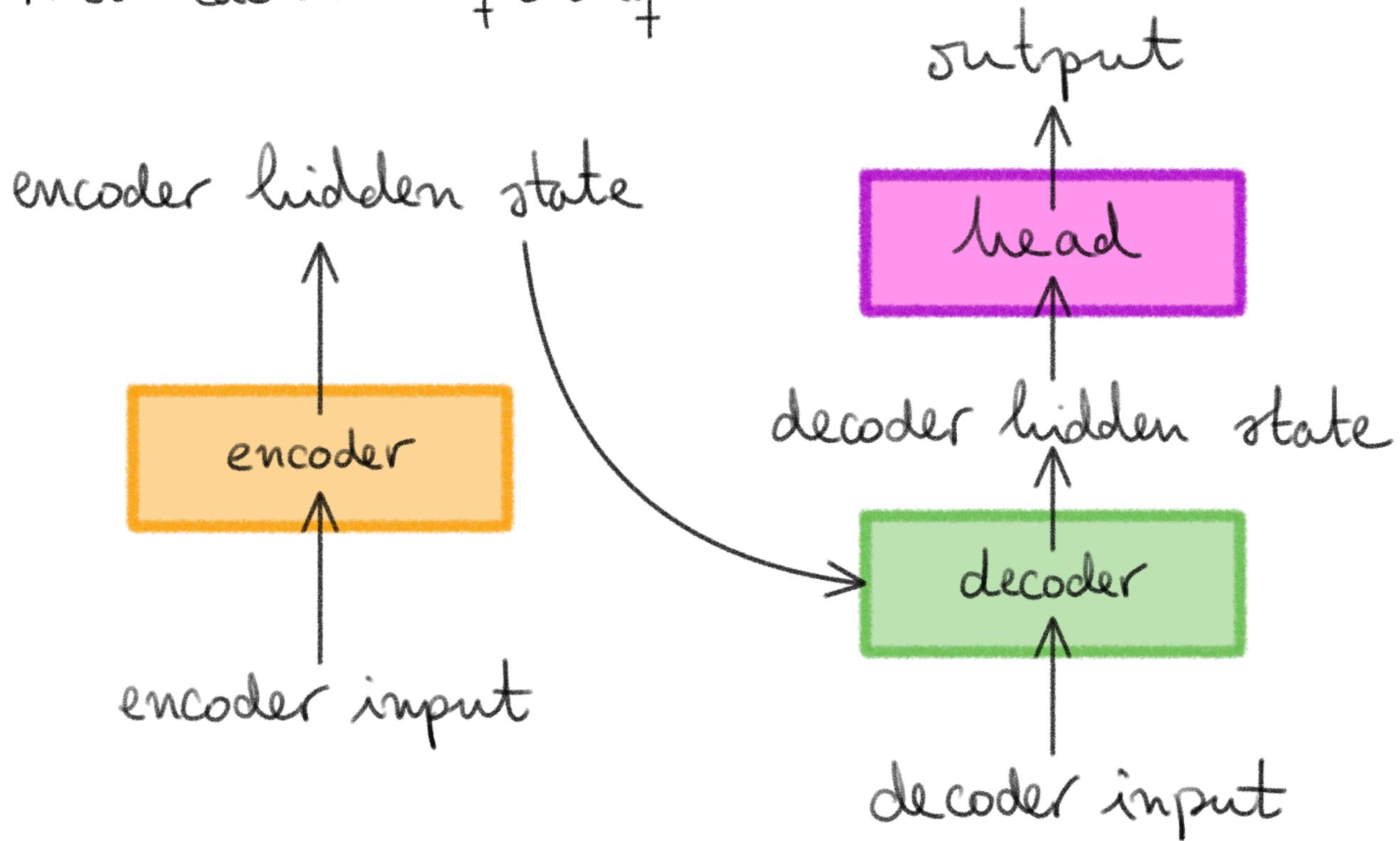
Transformers apply CE loss tokenwise!

EXERCISE

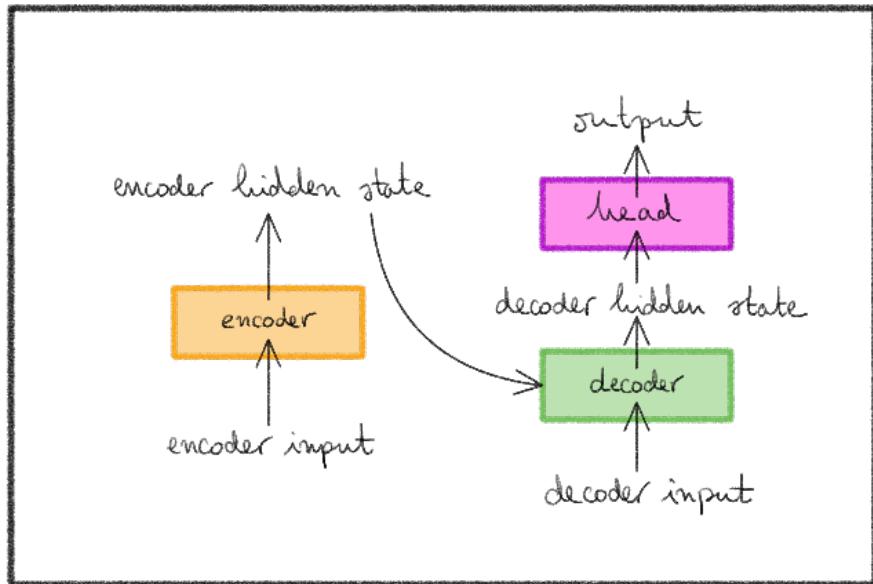
Given a 3-class classification problem,
and assuming a data point (x, y)
where the one-hot encoded label y
is $(1, 0, 0)$, $\hat{z} = \hat{f}(x) \in \mathbb{R}^3$ logits and
 $\hat{y} = \text{softmax}(\hat{z})$ i.e. $\hat{y}_k = \frac{e^{\hat{z}_k}}{e^{\hat{z}_1} + e^{\hat{z}_2} + e^{\hat{z}_3}}$.
Study the behavior (increasing/
decreasing) of the CE with respect to \hat{z} .

ENCODER- DECODER MODELS

Also called Seq2Seq



ENCODER- DECODER MODELS



Tasks:

- translation
- summarization
- question answering

Model examples:

- T5

ENCODER- DECODER MODELS

- BART

encoder hidden state

- MarianMT

encoder

encoder input

decoder

decoder input

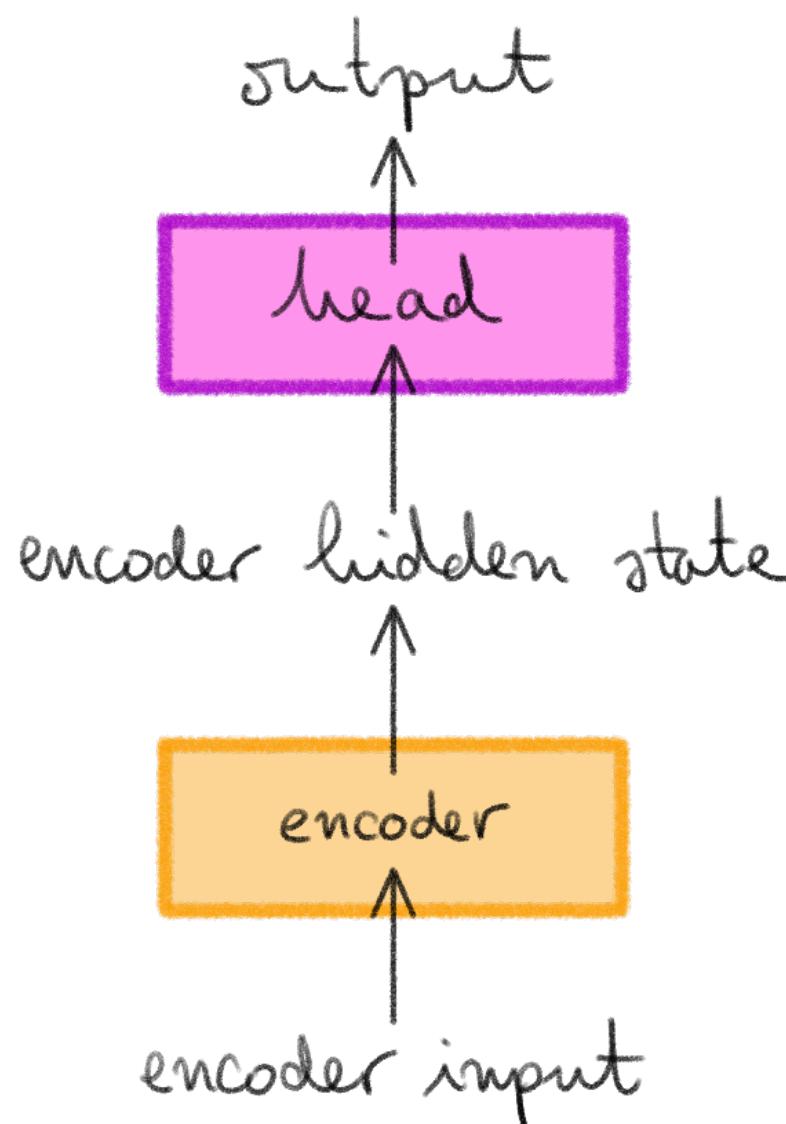


decoder hidden state

head

output

ENCODER-ONLY MODELS



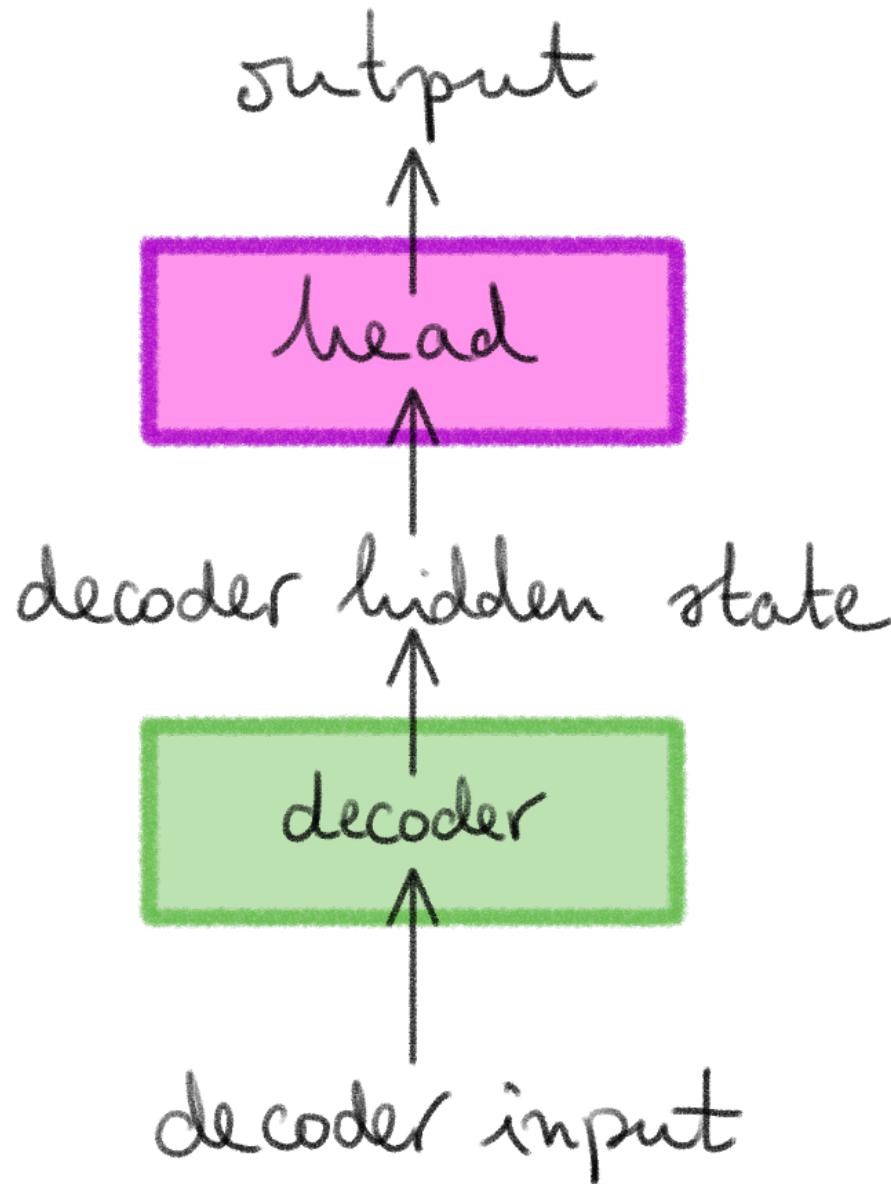
Tasks:

- classification
- named entity recognition

Models:

- BERT family

DECODER - ONLY MODELS



Tasks:

- Text generation
- Chatbots

Models:

- GPT family

Kahoot
time!

QUESTION TIME

- Write down something you understood well
- Write down something you did not fully understand
- Write something you would like to know more about

SUMMARY

- What are the three main types of transformer models ? What tasks is each of them used for ?
- In pairs : discuss the difference between the inference / training phases of a transformer model for the machine translation task.