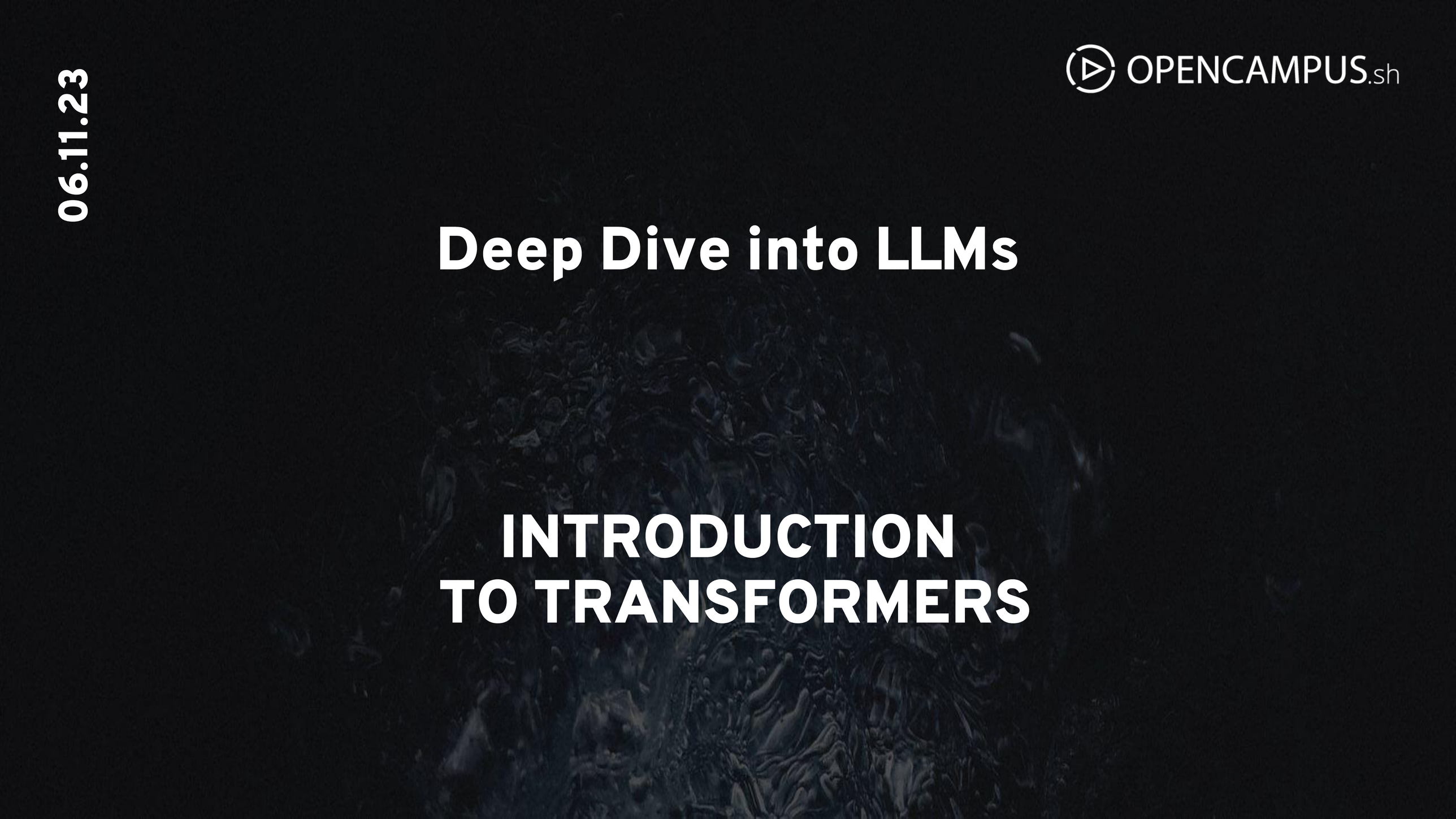
06.11.23

OPENCAMPUS.sh

Deep Dive into LLMs



INTRODUCTION TO TRANSFORMERS

QUIZ

https://forms.office.com/e/PAnsTPRr4y

PROJECT MILESTONES

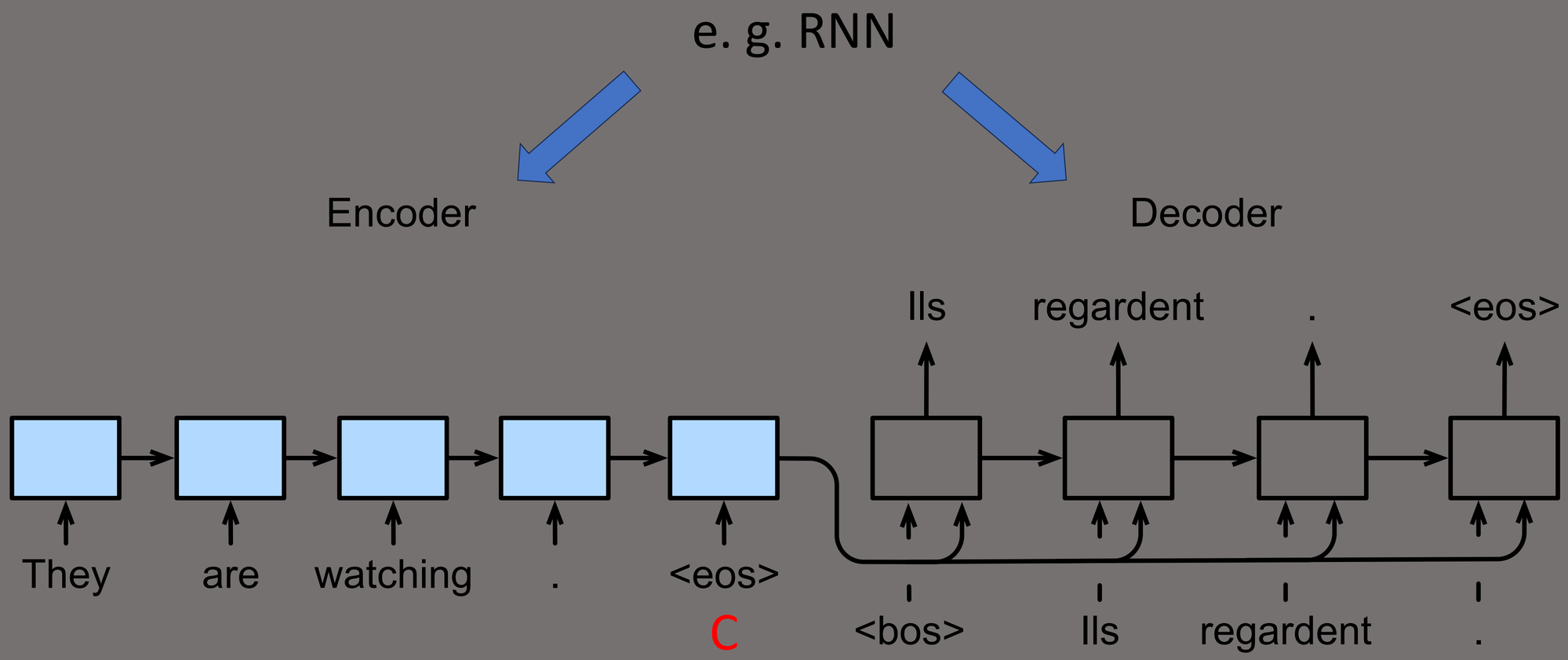
- 30.10 Form Groups
- 06.11 Literature Review I
- 13.11 Literature Review II
- 20.11 TBD
- 27.11 TBD
- 04.12 TBD
- 11.12 TBD
- 18.12 TBD
- 08.01 Project presentations

LEARNING GOALS

1. **Understand the Evolution of Language Models**: Gain a broad understanding of how language models have evolved, from early methods like n-grams to advanced transformer architectures.
2. **Grasp Core Concepts of Word Embeddings and RNNs**: Understand the significance and limitations of word embeddings and recurrent neural networks, including LSTMs.
3. **Master the Basics of Attention and Transformers**: Comprehend the fundamentals of attention mechanisms and the architecture of Transformer models, including key components like Multi-Head Attention and Positional Encoding.
4. **Familiarize with Advanced Models**: Be aware of advancements beyond basic Transformers, including models like BERT and GPT.
5. **Synthesize and Reflect**: Summarize key takeaways from the session and formulate thoughts or questions for further exploration.

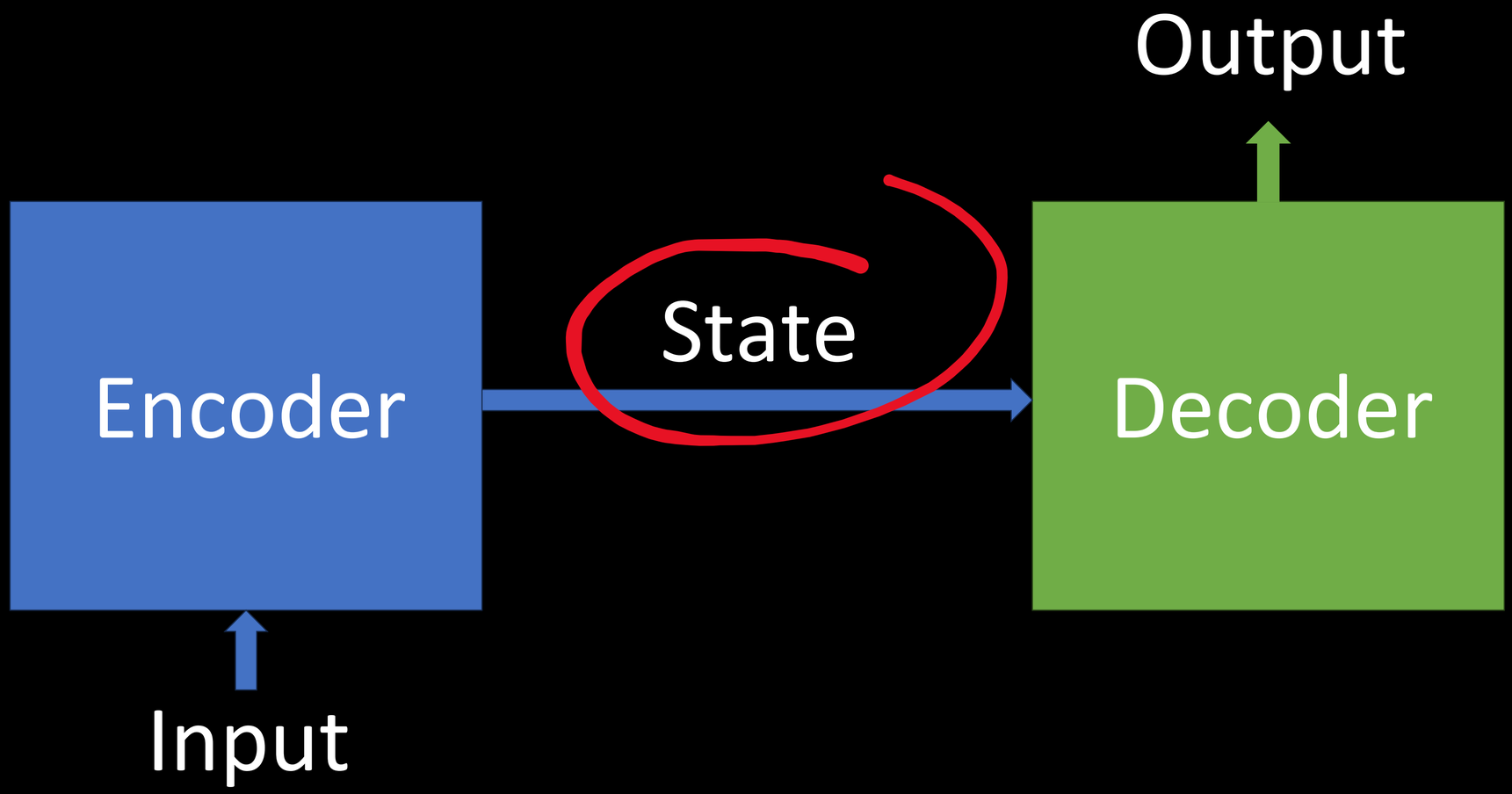
RECAP

SEQ2SEQ



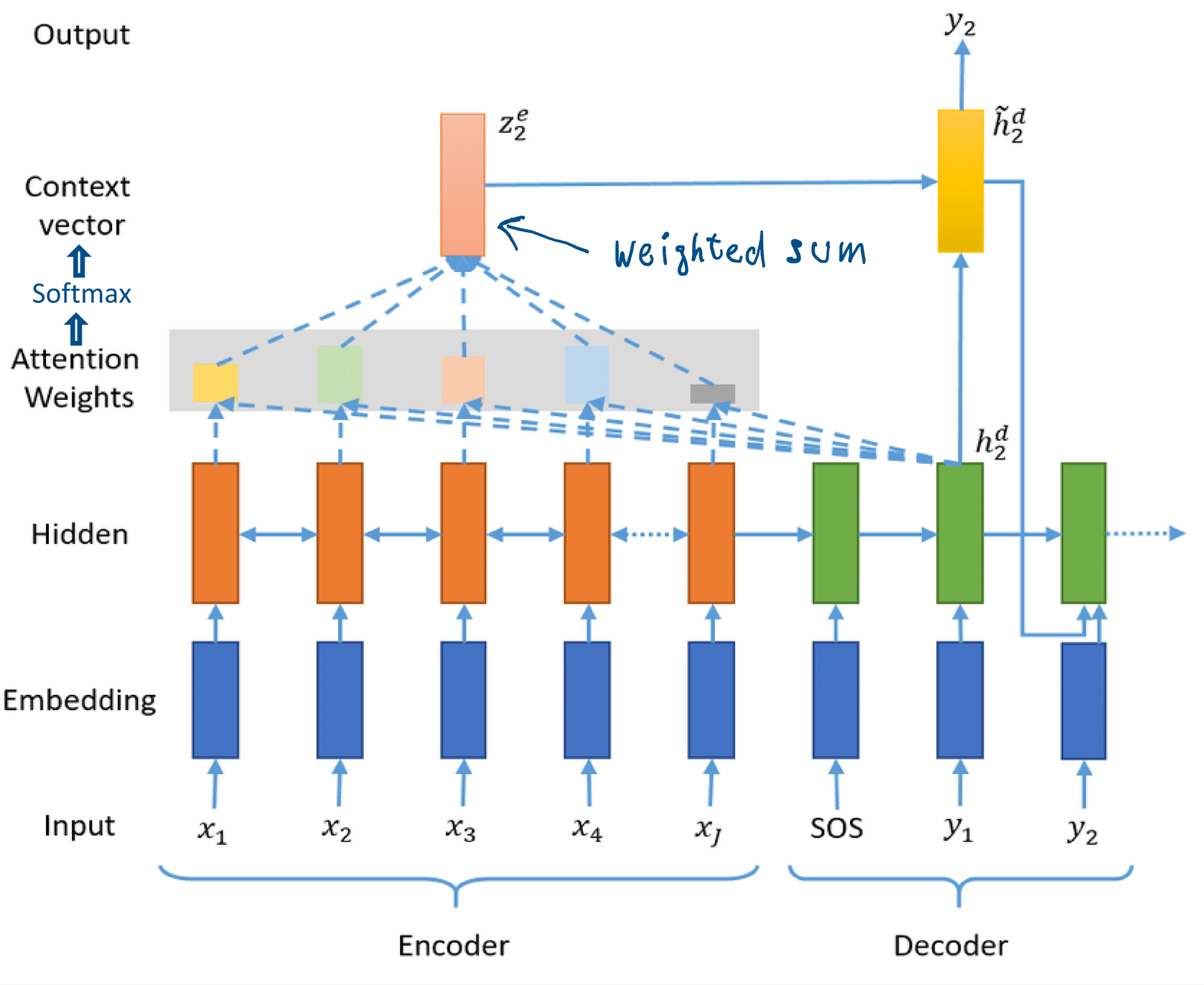
https://d2l.ai/chapter_recurrent-modern/seq2seq.html

PROBLEM



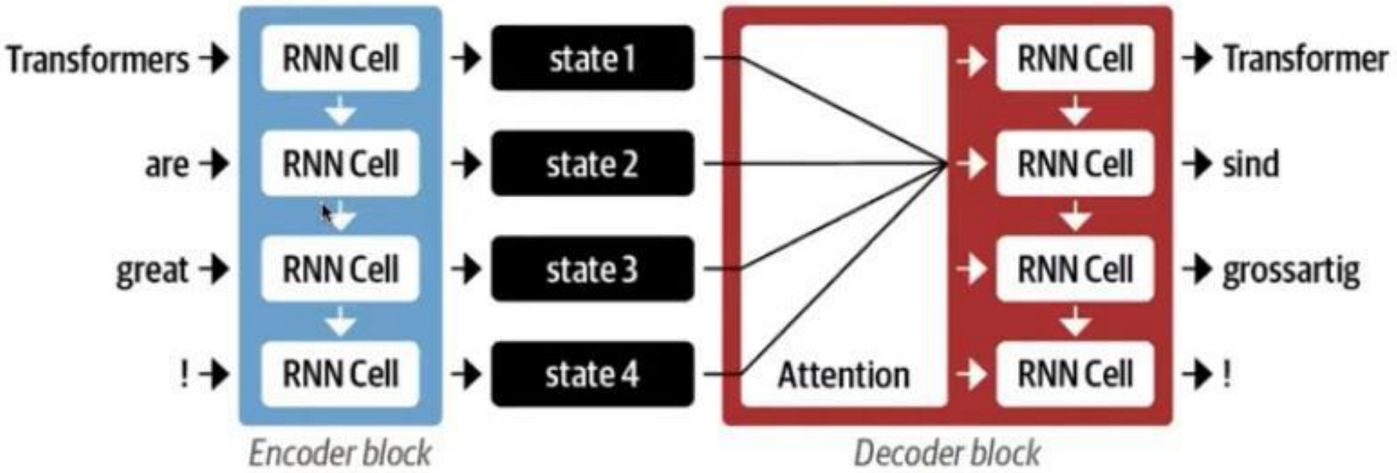
ATTENTION

ATTENTION



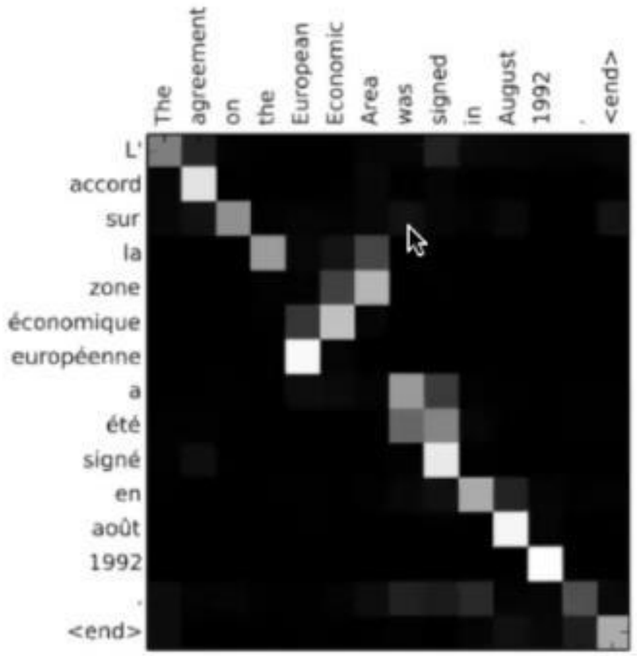
Neural Abstractive Text Summarization with Sequence-to-Sequence Models - Scientific Figure on ResearchGate.
Available from: https://www.researchgate.net/figure/An-attention-based-seq2seq-model_fig3_329464533 [accessed 1 Nov, 2023]

ATTENTION MECHANISM



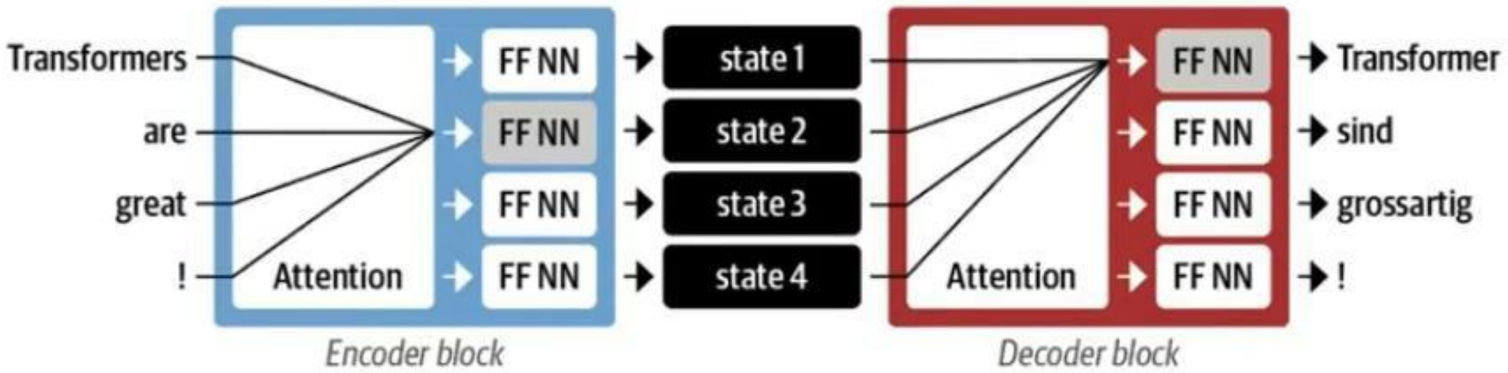
Assign a weight or "pay attention" to specific states

INTERPRETATION

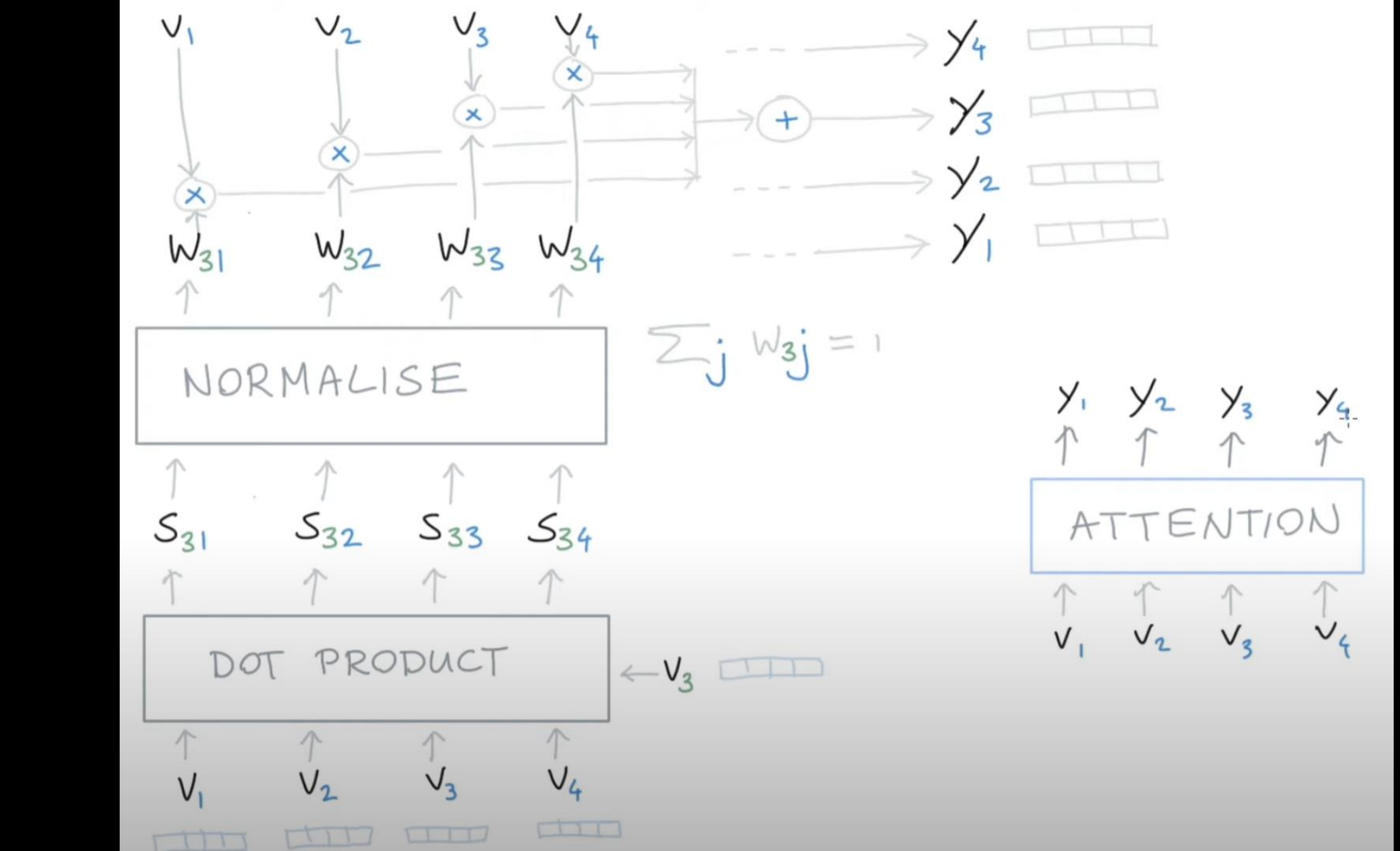


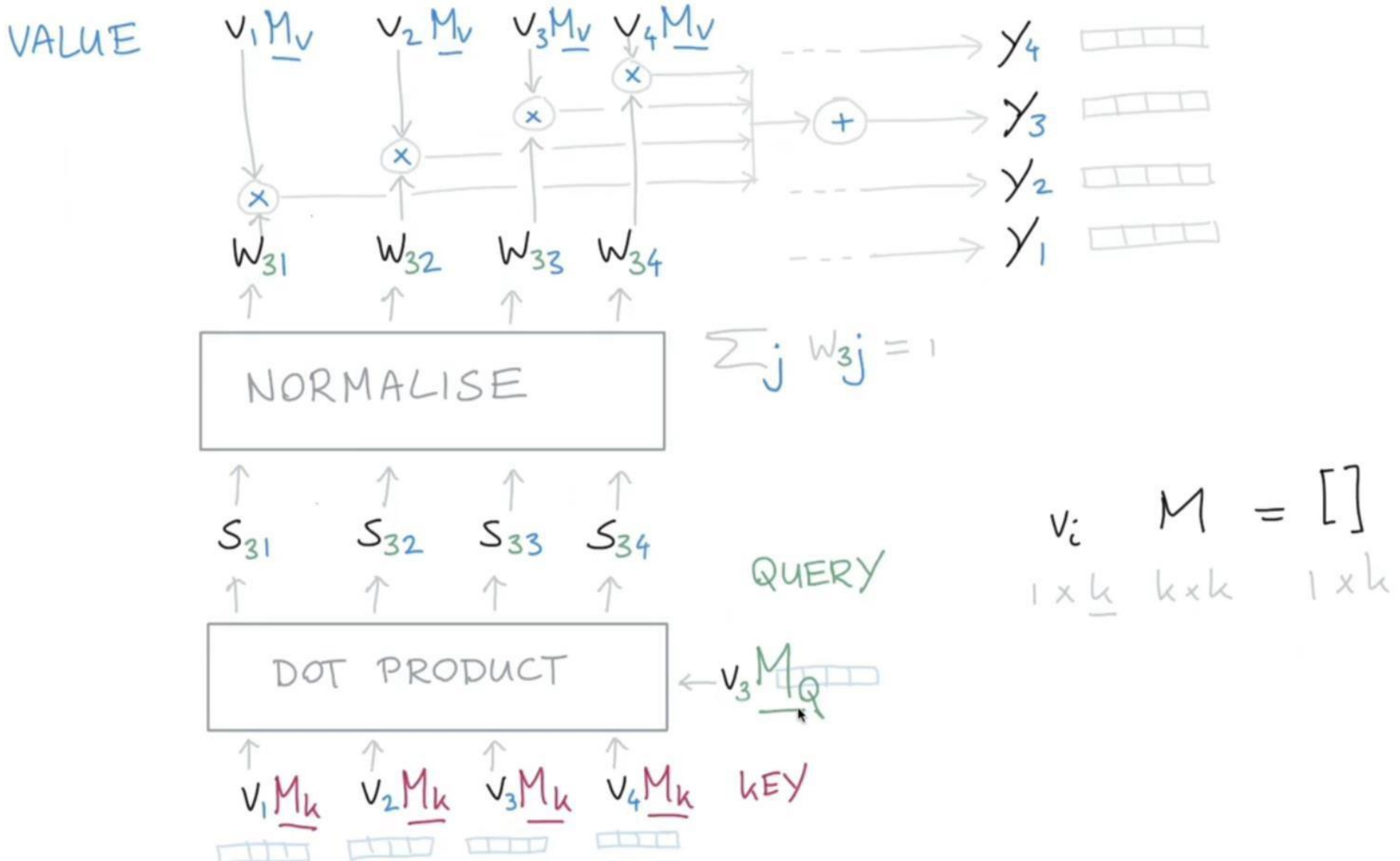
Attention gives better modelling of word order

ATTENTION IS ALL YOU NEED



Transformers much easier to scale with compute & data

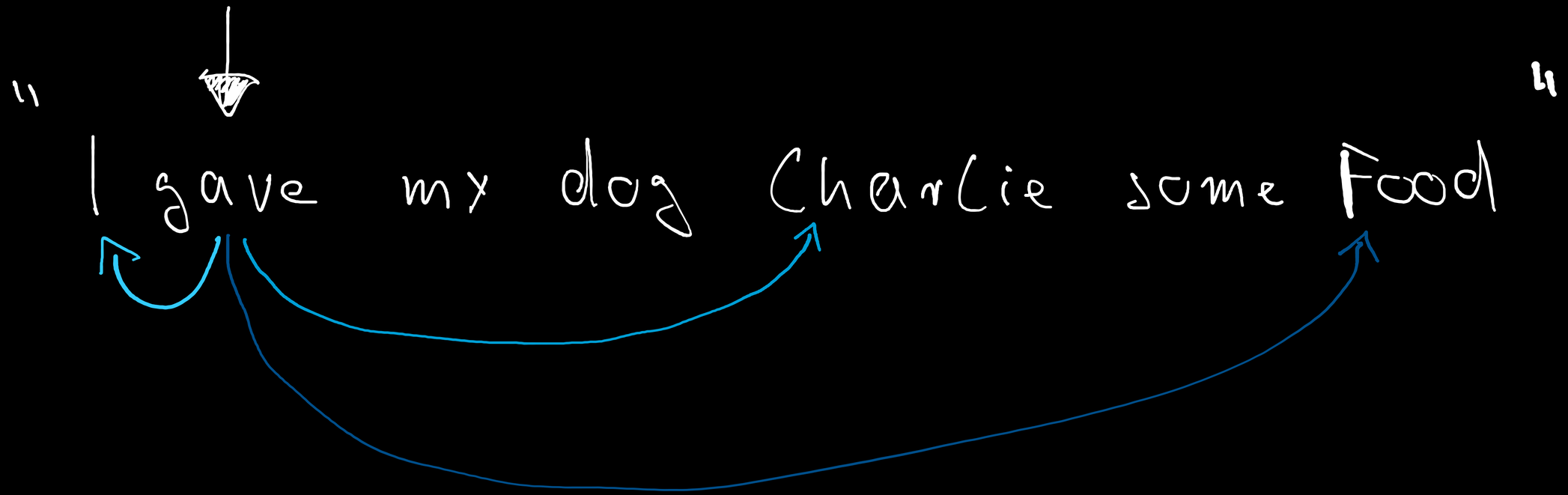

v_1 v_2 v_3 v_4
y_4 y_3 y_2 y_1
w_{31} w_{32} w_{33} w_{34}
NORMALISE
$\sum_j w_{3j} = 1$
s_{31} s_{32} s_{33} s_{34}
DOT PRODUCT
$\leftarrow v_3$
v_1 v_2 v_3 v_4
y_1 y_2 y_3 y_4
ATTENTION
v_1 v_2 v_3 v_4


VALUE
$v_1 M_V$
$v_2 M_V$
$v_3 M_V$
$v_4 M_V$
y_4
y_3
y_2
y_1
w_{31}
w_{32}
w_{33}
w_{34}
NORMALISE
$\sum_j w_{3j} = 1$
s_{31}
s_{32}
s_{33}
s_{34}
QUERY
DOT PRODUCT
$v_3 M_Q$
$v_1 M_k$
$v_2 M_k$
$v_3 M_k$
$v_4 M_k$
KEY
v_i $M = []$
$1 \times k$ $k \times k$ $1 \times k$

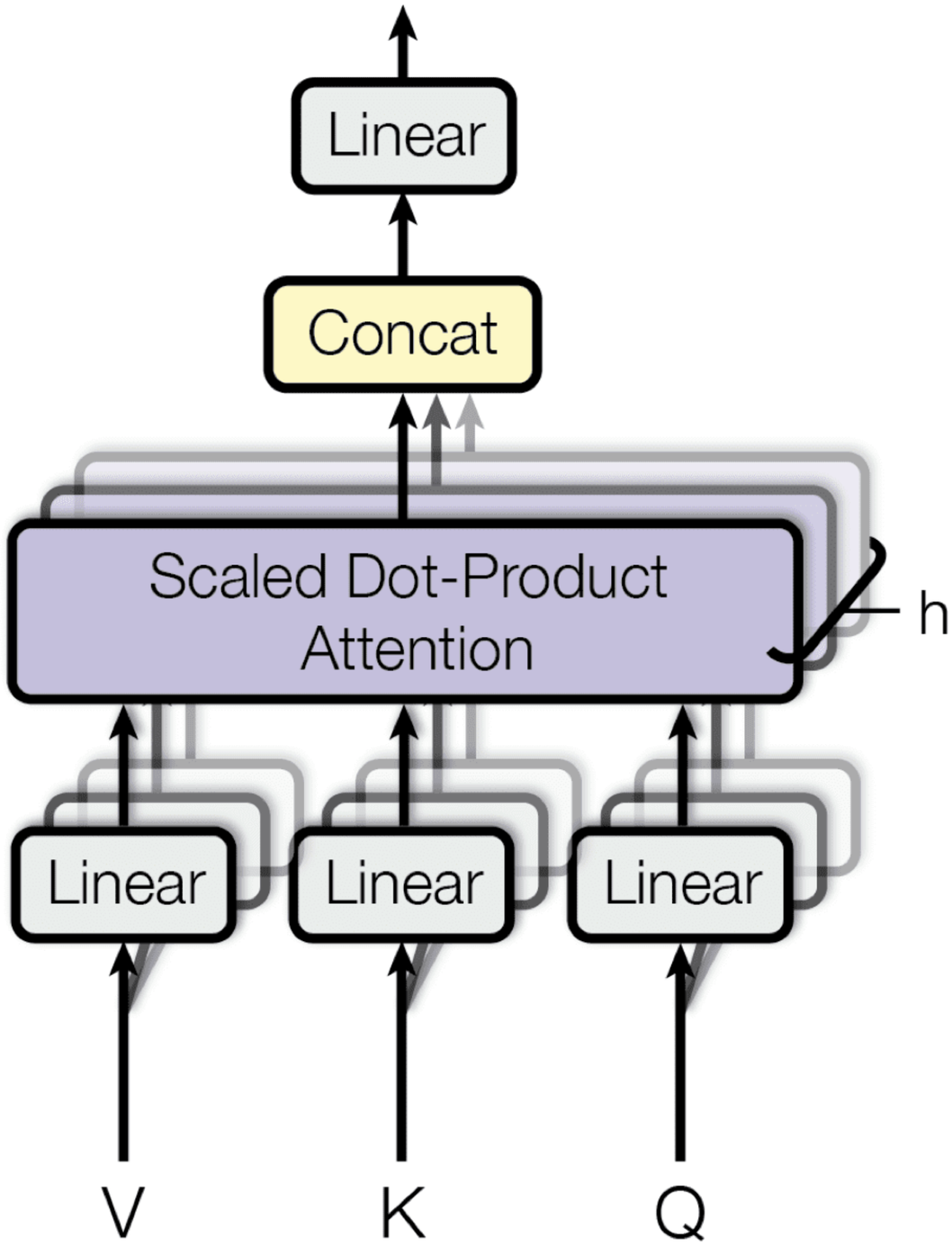
BREAKOUT DISCUSSION

- What problem do you see with the basic self attention in the context of neural networks (Hint: Learning)?
- What new method is introduced in the drawing, compared to basic self attention?
- Why is it useful?

DO WE HAVE ENOUGH ATTENTION



MULTI HEAD ATTENTIO



Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). Attention Is All You Need. *ArXiv:1706.03762 [Cs]*. Retrieved from http://arxiv.org/abs/1706.03762

ATTENTION HANDS ON

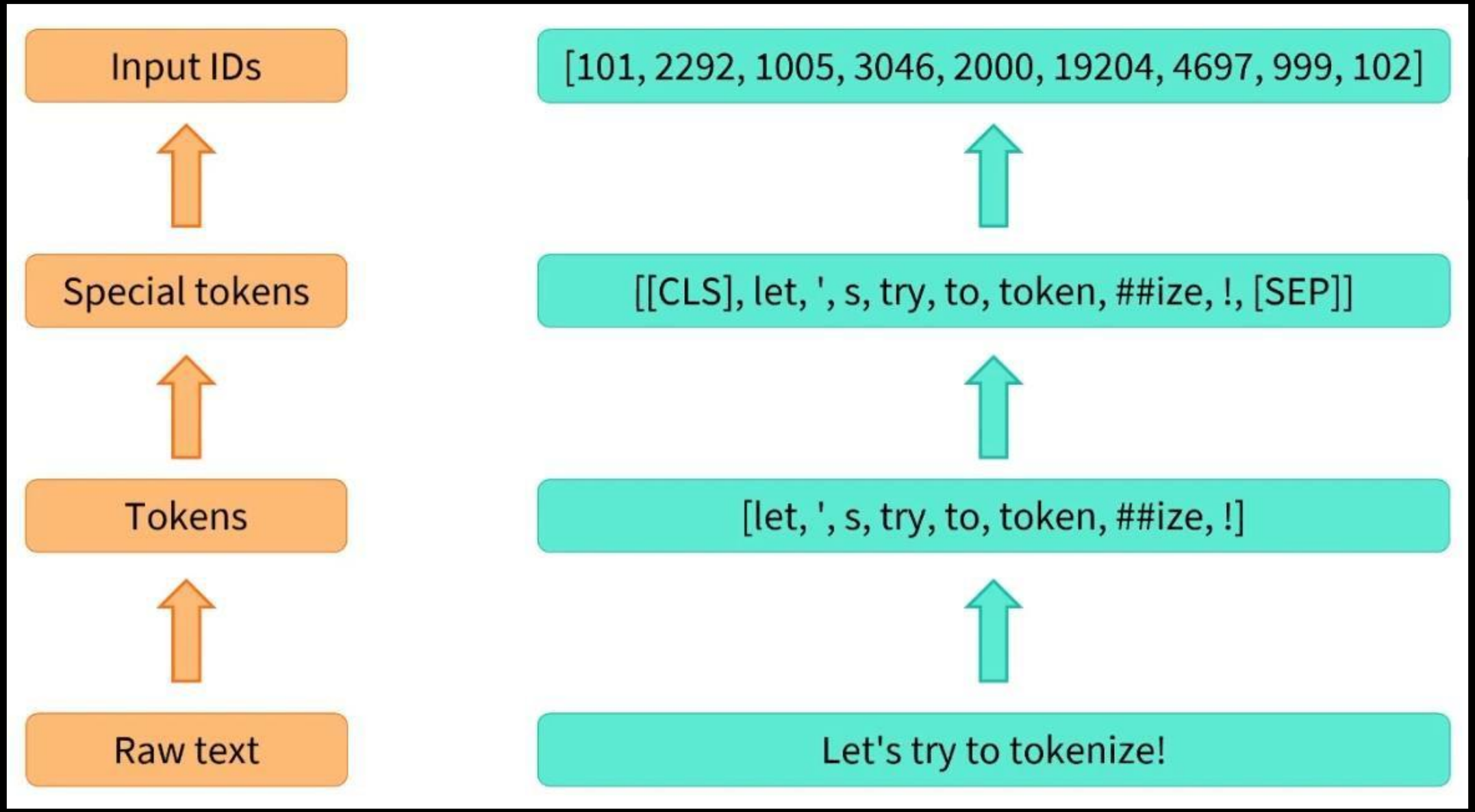
https://colab.research.google.com/drive/1hXIQ77A4TYS4y3UthWF-Ci7V7vVUoxmQ?usp=sharing#scrollTo=twSVFOM9SopW

WHY IS ATTENTION GREAT?

- Significantly improves NMT
- Solves the bottleneck problem
- Helps with vanishing gradient problem
- Provides some interpretability
- Can be run in parallel

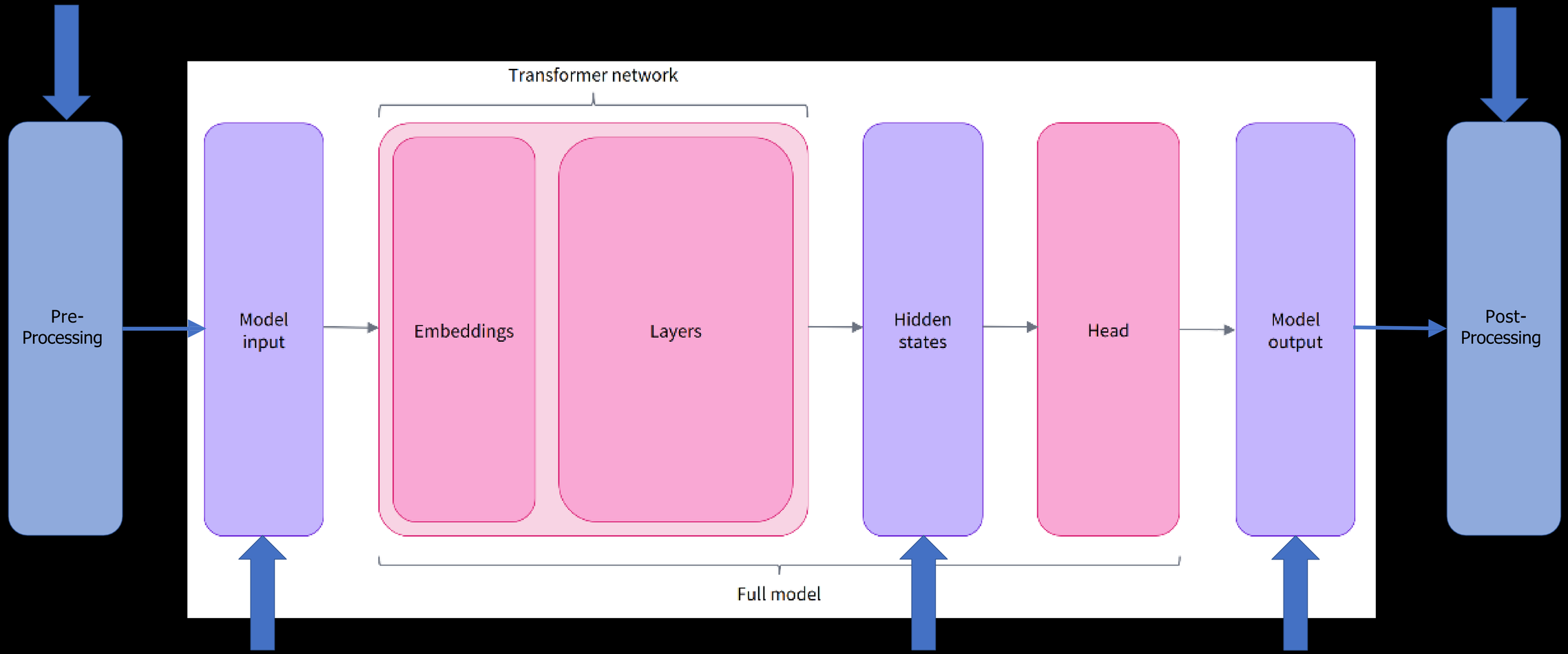
TRANSFORMERS

REMINDER TOKENIZING



- Splitting
- Mapping to integers
- Adding model dependent tokens/integers

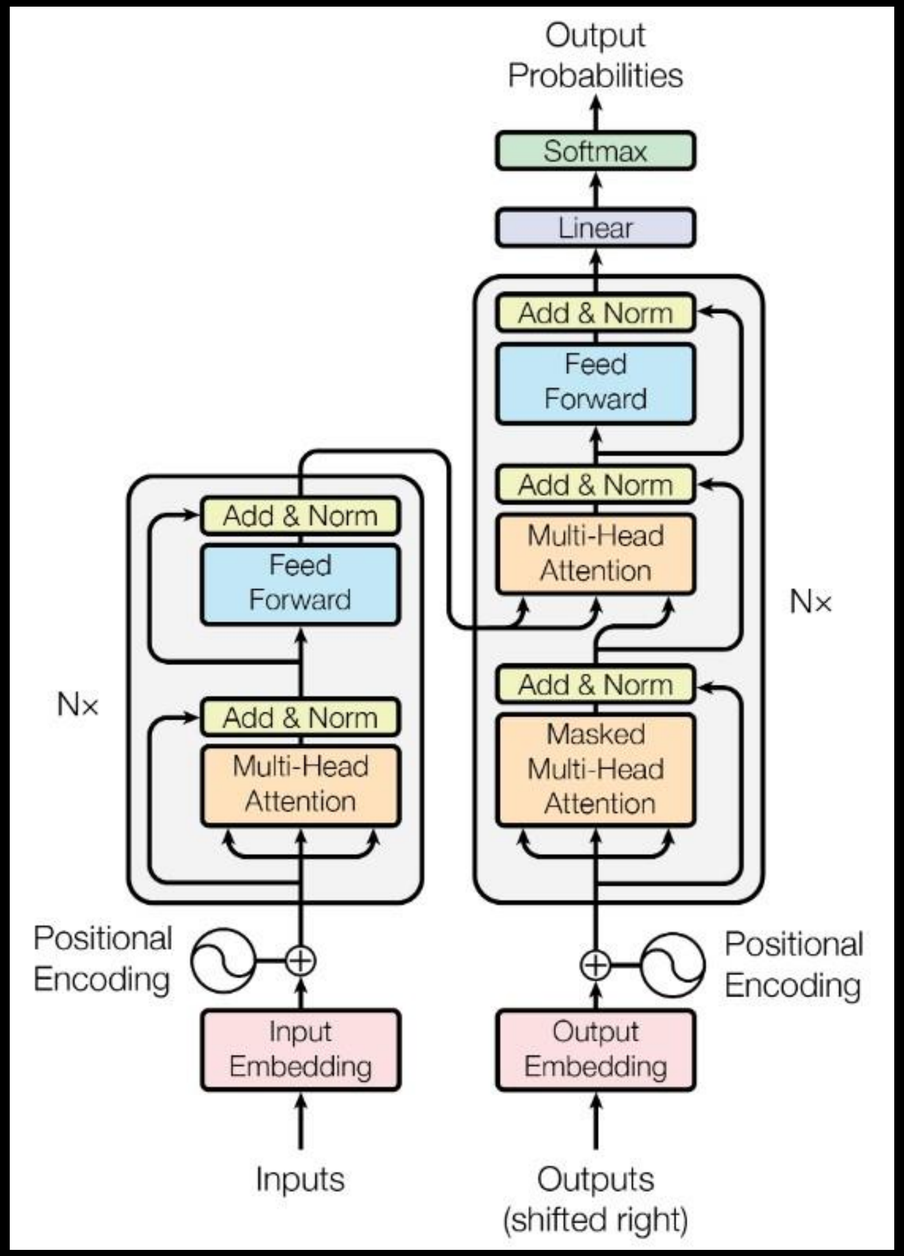
- Logits to probs
- Probs to classes
- (Classes to tokens/text)



Tokenized subwords (integers)

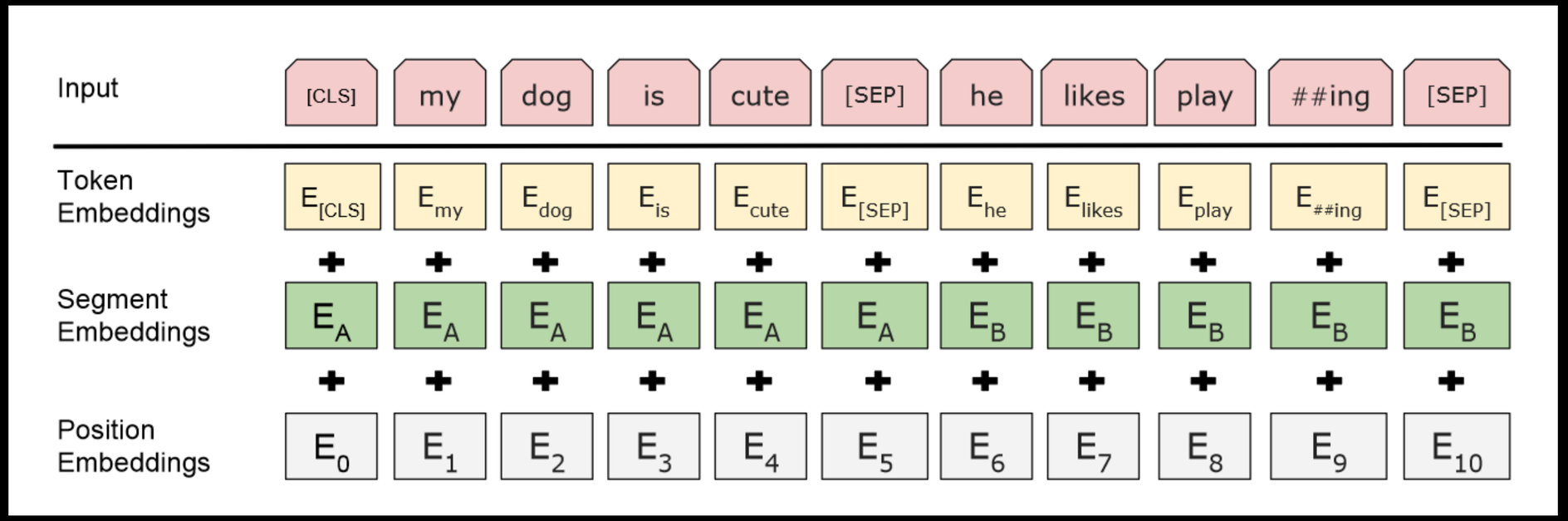
Contextualized embeddings for each token (high dimensional vectors of floats)

Logits (one float for each trained class)



Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). Attention Is All You Need. *ArXiv:1706.03762 [Cs]*. Retrieved from http://arxiv.org/abs/1706.03762

BERT EMBEDDINGS



Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv:1810.04805 [Cs]*. Retrieved from http://arxiv.org/abs/1810.04805

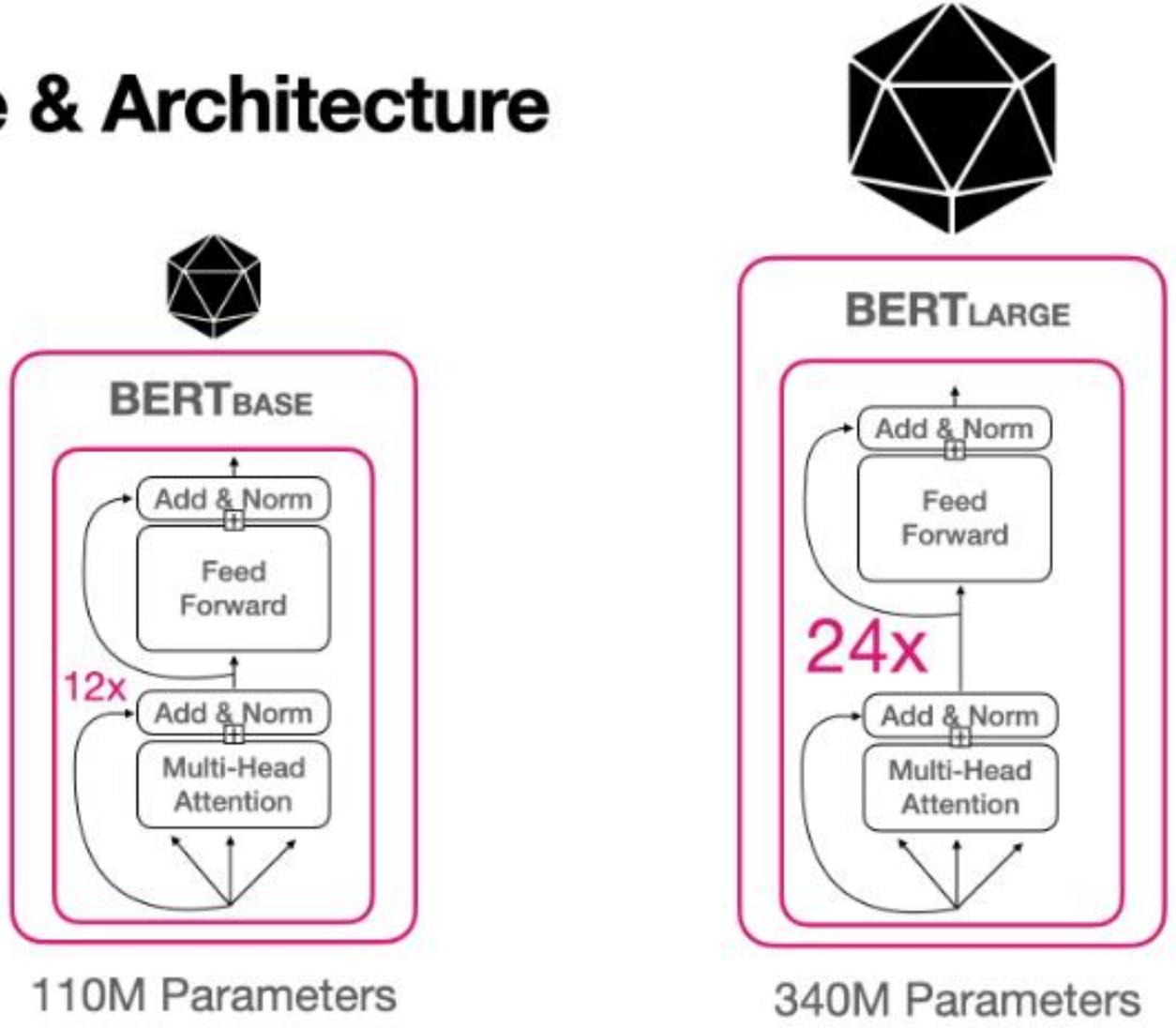
ENCODER-ONLY MODELS

- Auto-Encoding Models
- Training:
  - Masked Language Modeling
  - Next sentence prediction

- Typical Models:
  - ALBERT
  - BERT
  - DistillBERT
  - ELECTRA
  - RoBERTa

- Typical Tasks:
  - Understanding sentences
  - Text classification
  - Extractive QA
  - NER

BERT Size & Architecture



https://huggingface.co/blog/bert-101

DECODER-ONLY MODELS

- Auto-Regressive Models
- Training:
  - Only access to the past
  - Next word prediction

- Typical Models:
  - CTRL
  - GPT
  - GPT-2
  - Transformer XL

- Typical Tasks:
  - Text generation

ENCODER-DECODER MODELS

- Seq2Seq Models
- Training:
  - Like encoder only
  - Like decoder only
  - More complex tasks
- Typical Models:
  - BART
  - mBART
  - MARIAN
  - T5
- Typical Tasks:
  - Summarization
  - Translation
  - Generative QA

HOMEWORK NOTEBOOK ON EMBEDDINGS

- What did you learn when experimenting with the notebook?
- How did the results change with the embeddings?

Archive

EVENTS

Coding.Waterkant 2023

Prototyping Week

PROJECTS

How to Start, Complete, and Submit Your Project

Possible Projects

Past Projects

ADDITIONAL RESOURSES

Glossary

Coursera

Selecting the Optimizer

Choosing the Learning Rate

Learning Linear Algebra

Learning Python

Support Vector Machines

ML Statistics

TOOLS

Git

RStudio

Google Colab

How to Start, Complete, and Submit Your Project

ON THIS PAGE

Starting Your Project

Working on Your Project

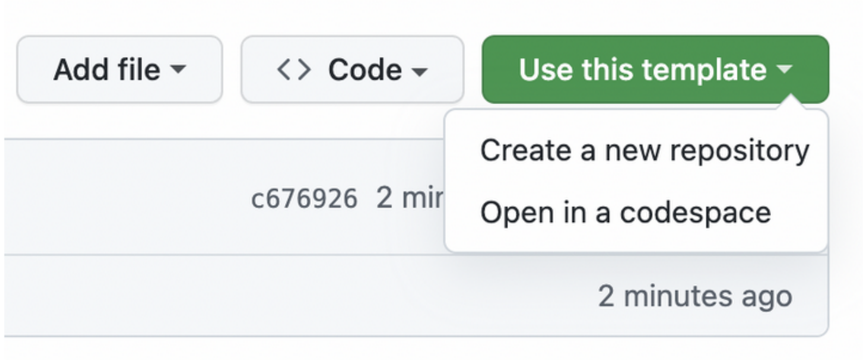
Submitting Your Project

In all Machine Learning courses you have:

- to complete a machine learning project in a team of up to 4 participants,
- attend at least all but 2 sessions of the course, and
- use the provided project template repository for documentation (unless otherwise instructed).

Starting Your Project

1. **Navigate to the** Template Repository
2. **Use this Template**: Above the file list, click the "Use this template" button.

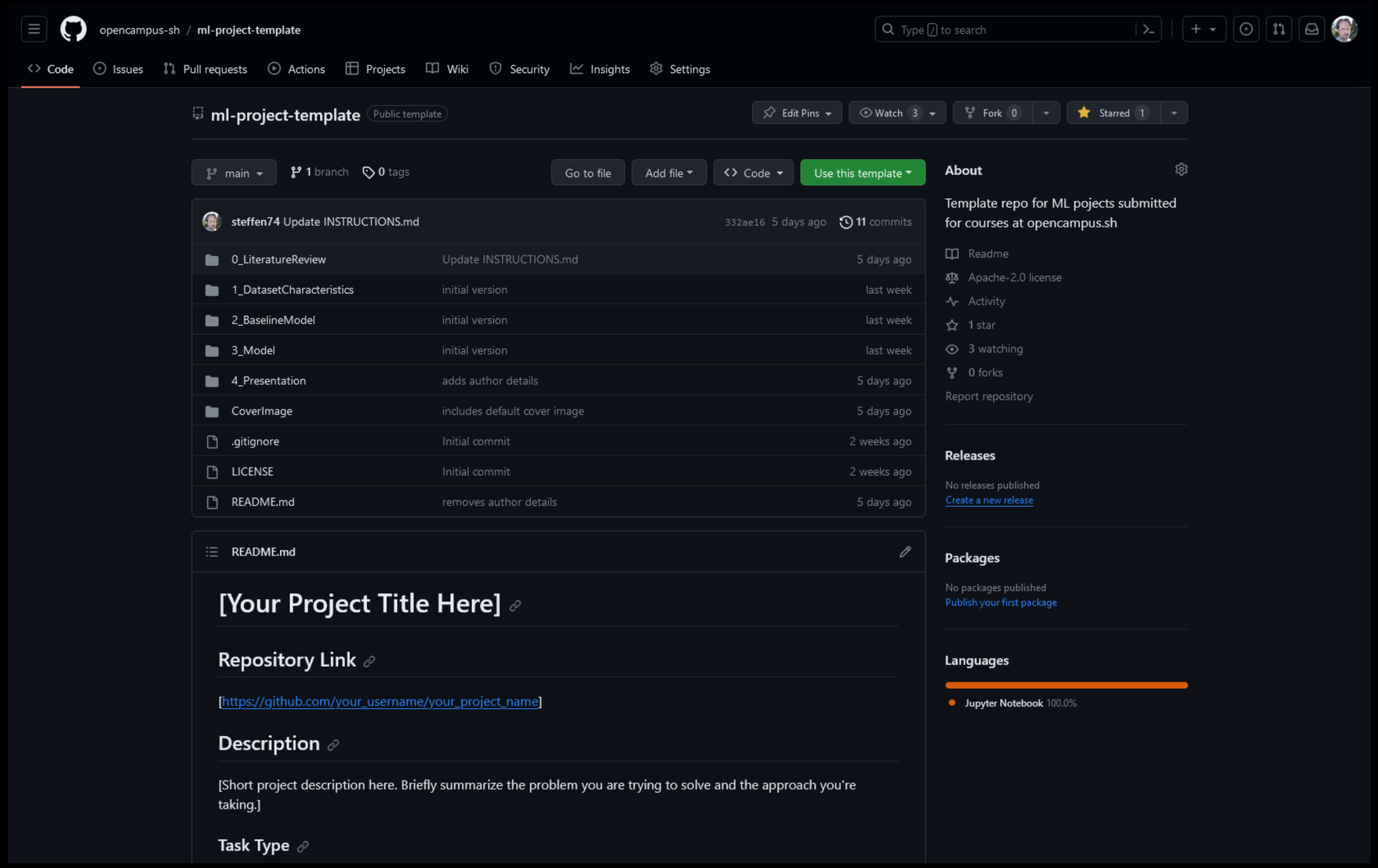


Use this template button

3. **Create Repository from Template**: You'll be prompted to name your new repository and you can choose whether it should be public or private. You'll also have the option to include all branches in the template repository, if there are more than one.
4. **Create Repository**: Click "Create repository from template" to create the new repository.
5. **Clone the New Repository**: You can now clone the new repository to your local machine using `git clone` and start working on your project.

https://opencampus.gitbook.io/opencampus-machine-learning-program/projects/requirements

PROJECT TEMPLATE REPO



Literature Review

Overview

This project milestone requires you to review the literature related to your project. The objective is to gain a deeper understanding of the problem domain, as well as to identify similar approaches or solutions that have been tried before. You might want to answer the following questions:

- Which are the models commonly used for my problem?
- Which format must the training data have?
- How much training data is typically used in similar problems?
- Are there pretrained models I can use for my problem?

Guidelines

1. **Minimum Number of Sources**: You are required to review at least two or three papers, blogs, or authoritative sources related to your project topic.
2. **Summary**: For each work, provide a brief summary that includes the objective of the work, methods used, and the outcomes. One sentence on each point is sufficient.

Submission

Complete the template provided in the README of this folder.

TASKS UNTIL NEXT WEEK

- Create your project repo
- Complete the literature review for your project
- Create a ChatGPT account
- Completion of week 1 of the LLM course on Coursera