miro.com/app/board/uXjVKENK1Qc=/

miro | Convolutional Neural Networks | Upgrade | Present | Share

Lecture4: Basic Intro to Transformers

(1) Most modern LLMs rely on the transformer architecture → deep neural network architecture introduced in 2017 paper

Attention is All you Need →

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

24%

miro | Convolutional Neural Networks | Upgrade | Present | Share

introduced in 2017 paper

Attention is All you Need

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com
Noam Shazeer* Google Brain noam@google.com
Niki Parmar* Google Research nikip@google.com
Jakob Uszkoreit* Google Research usz@google.com
Llion Jones* Google Research llion@google.com
Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu
Łukasz Kaiser* Google Brain lukaszkaiser@google.com
Illia Polosukhin* † illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

Original Transformer: Developed for machine translation

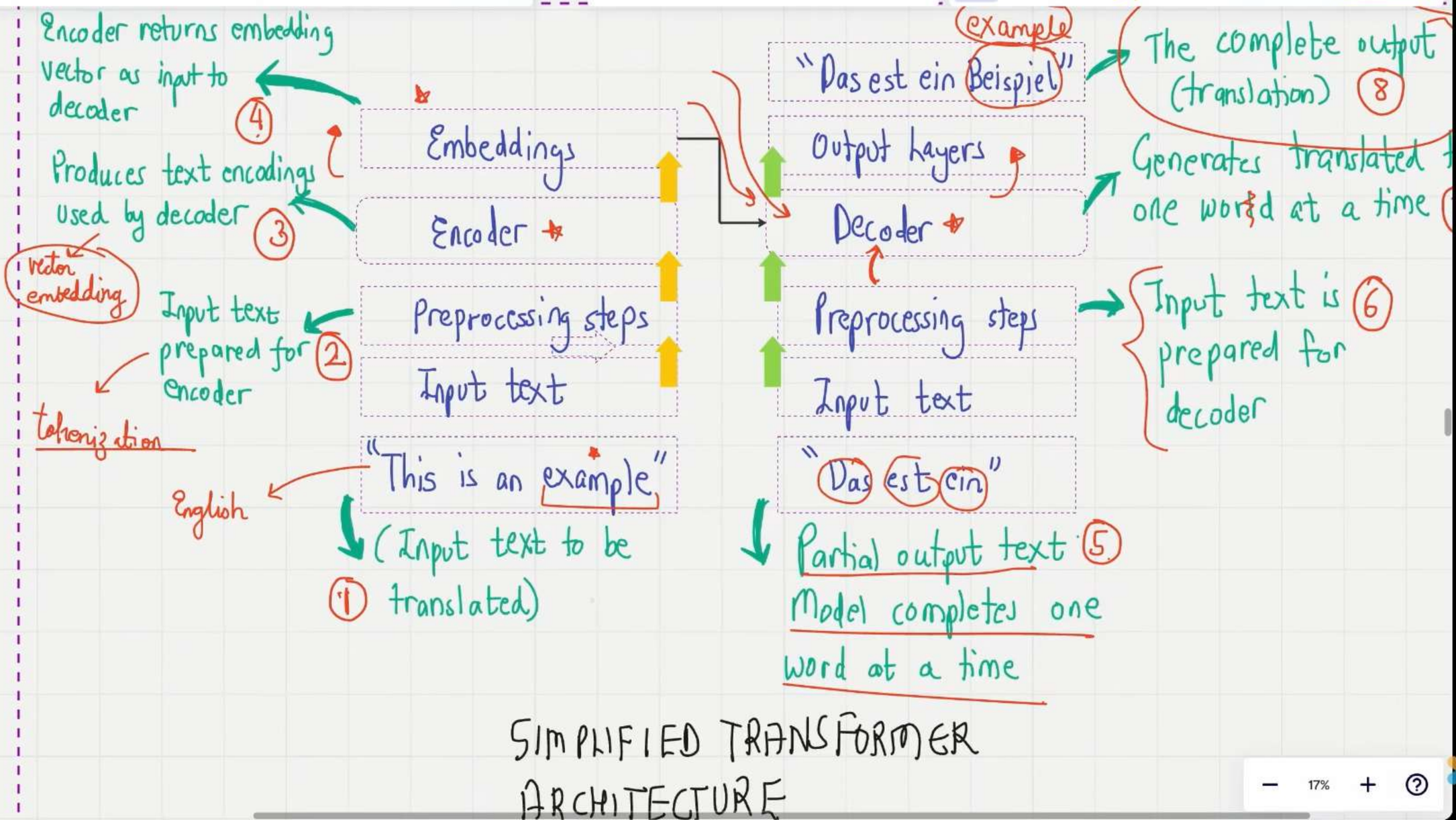
→ translating English texts to German and French.

(2)

Encoder returns embedding

example

21%

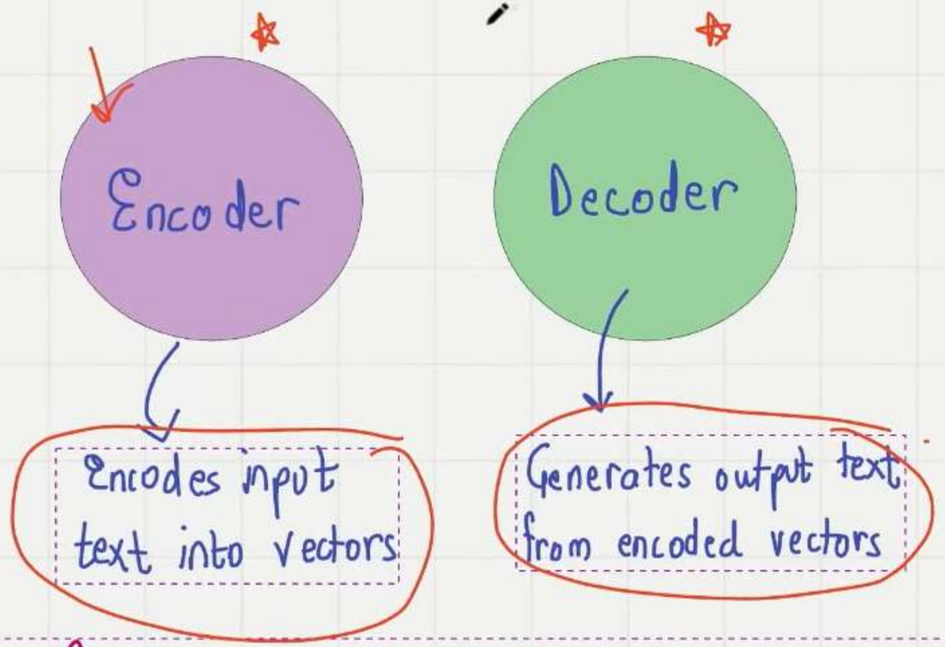


SIMPLIFIED TRANSFORMER ARCHITECTURE

word at a time

SIMPLIFIED TRANSFORMER ARCHITECTURE

(3) Transformer architecture consists of:



A note on self-attention mechanism:

Encodes input text into vectors

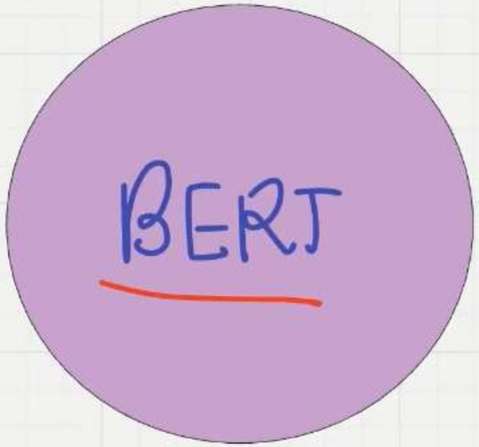
Generates output text from encoded vectors

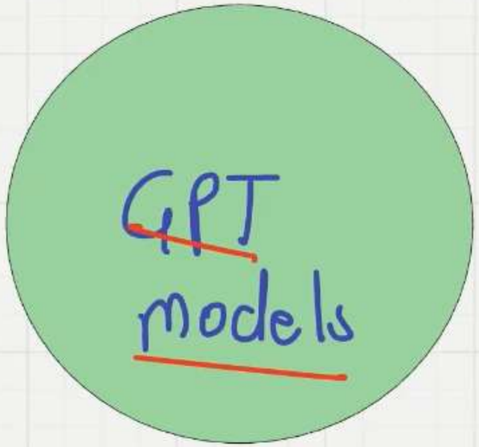
A note on self-attention mechanism:

- Key part of transformers. Allows model to weigh importance of different words / tokens relative to each other.
- Enables model to capture long range dependencies.
- We will look at this in detail later.

- We will look at this in detail later.

(4) Later variations of transformer architecture. → 2017





* "Bidirectional encoder representations

"Generative pretrained

BERT

GPT models

"Bidirectional encoder representations from transformers"

"Generative pretrained Transformers"

Predicts hidden words

Generates new

"Bidirectional encoder representations from transformers"

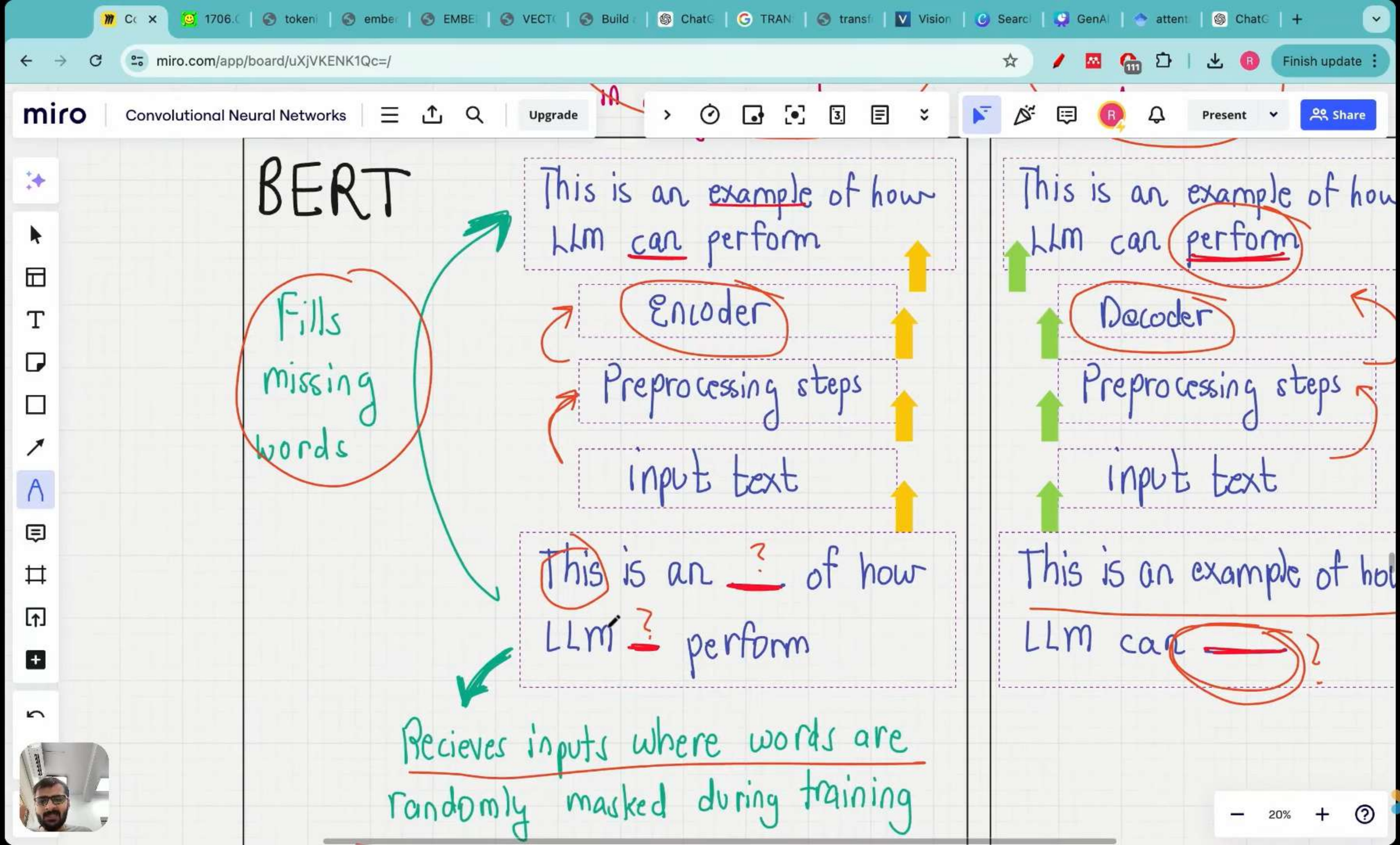
Predicts hidden words in a given sentence

"Generative pretrained Transformers"

Generates new word

This is an example of how LLM can perform

This is an example o
LLM can perform


miro.com/app/board/uXjVKENK1Qc=/
Finish update
miro
Convolutional Neural Networks
Upgrade
Present
Share
BERT
Fills missing words
This is an example of how LLM can perform
Encoder
Preprocessing steps
input text
This is an ? of how LLM ? perform
Recieves inputs where words are randomly masked during training
This is an example of how LLM can perform
Decoder
Preprocessing steps
input text
This is an example of hou
LLM can ?
20%

a given sentence

word

This is an example of how LLM can perform

Enloder

Preprocessing steps

input text

This is an ? of how LLM ? perform

ecieves inputs where words are

This is an example of how LLM can perform

Decoder

Preprocessing steps

input text

This is an example of how LLM can ?

learns to one word a time.

Recieves

This is an example of how LLM can perform

Encoder

Preprocessing steps

input text

This is an ? of how LLM ? perform

puts where words are masked during training

This is an example of how LLM can perform

Decoder

Preprocessing steps

input text

This is an example of how LLM can —?

learns to generate one word at a time.

Recieves incomplete text

5 Transformers vs LLMs

→ Not all transformers are LLMs

→ Transformers can also be used for Computer Vision

→ Not all LLMs are transformers

→ LMs can be based on recurrent or convolutional architectures as well.