Transformers (2)

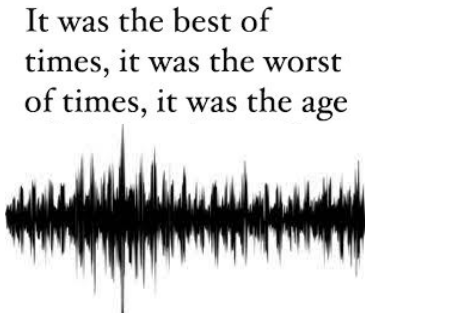
2022/03/09

CS 182/282A Lecture 13

Transformers review

Setup

features $\mathbf{x}$



- sequential data
- may be variable length

label $\mathbf{y}$

could correspond to…

- sentiment analysis,
  translation to another language,
  …
- audio transcription,
  speaker identification,
  …
- activity identification,
  video captioning,
  …

or there could be no label!

- unsupervised learning /
  generative modeling

model

~~???~~

transformers

Self-attention: the building block of transformers

The goal of self-attention is to handle sequential features as the input

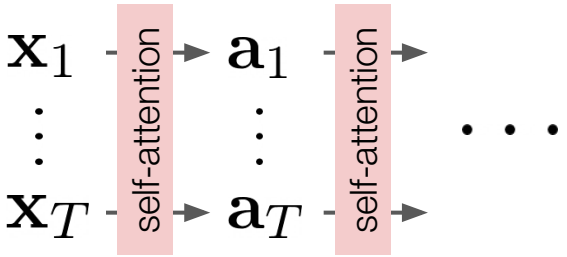
Think of it as a neural network layer that allows for processing the whole sequence

"key-value-query" system:

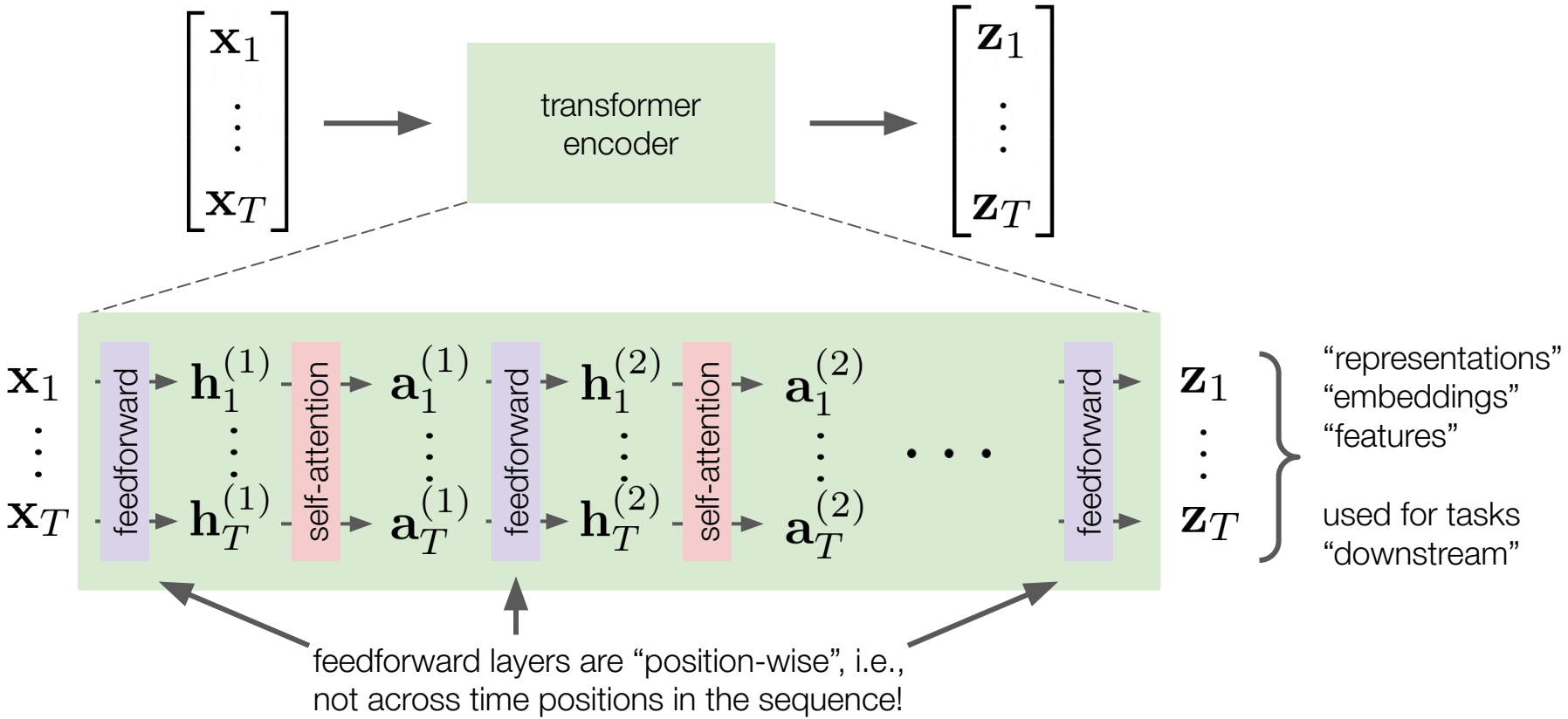
$$\mathbf{q}_t = q(\mathbf{x}_t)$$
$$\mathbf{k}_t = k(\mathbf{x}_t) \text{ or } \mathbf{h}_t$$
$$\mathbf{v}_t = v(\mathbf{x}_t)$$

These functions are learned and can be, e.g., simple linear layers

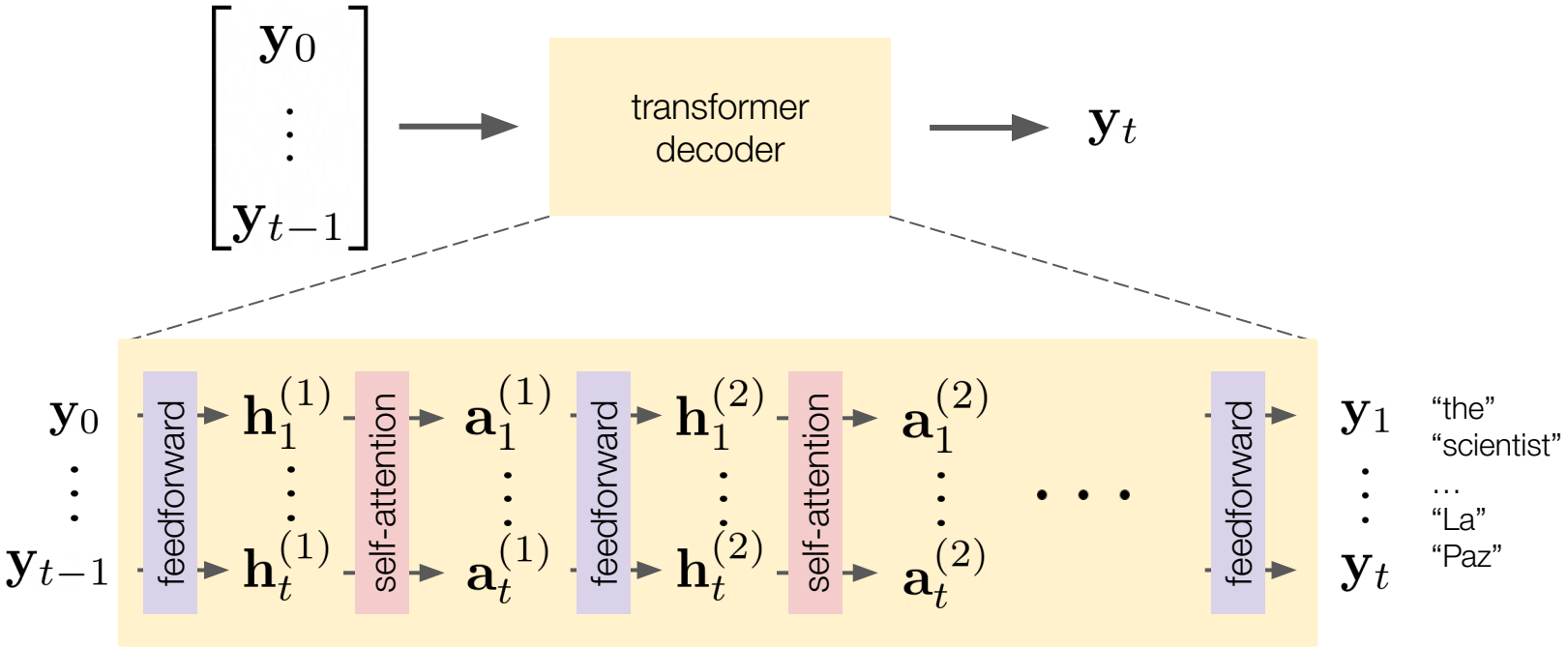
$$\mathbf{k}_1 \cdot \mathbf{q}_2 = e_{1,2} \xrightarrow{\text{softmax}} \alpha_{1,2}$$
$$\mathbf{k}_T \cdot \mathbf{q}_2 = e_{T,2} \xrightarrow{\text{softmax}} \alpha_{T,2}$$
$$\mathbf{a}_2 = \sum_t \alpha_{t,2} \mathbf{v}_t$$



Transformers for “encoding”

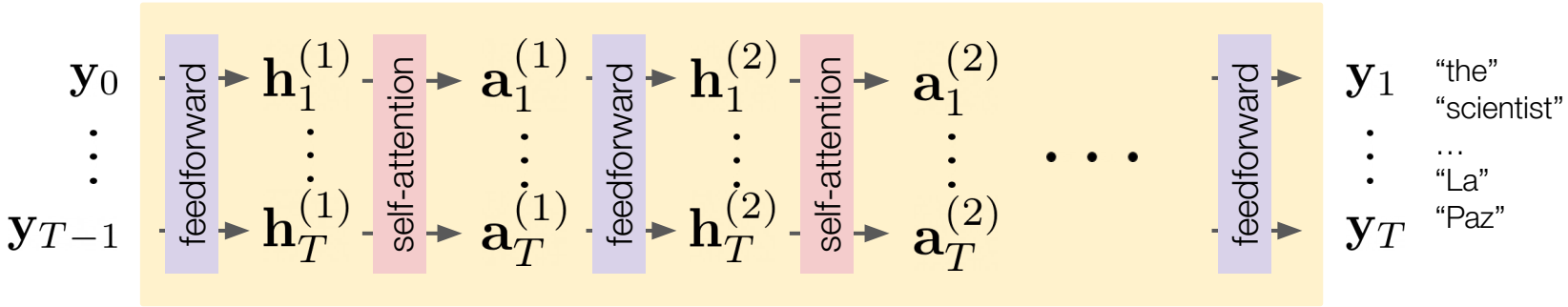


Transformers for “decoding” (generation)



A note about training transformer decoders

At training time, we pass in entire sequences to the decoder



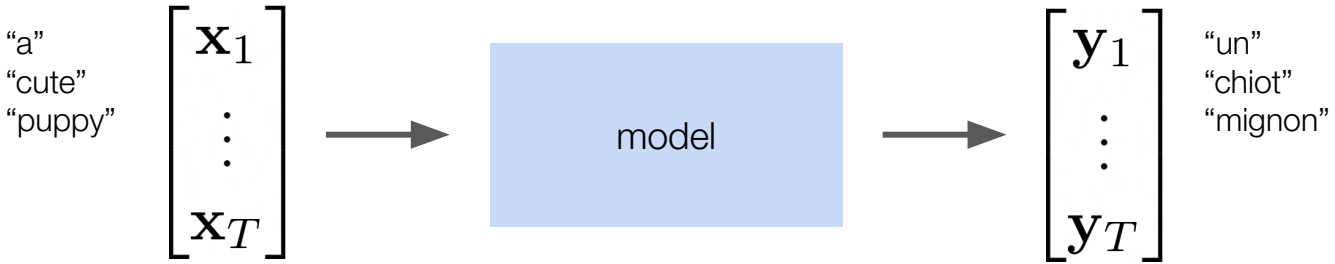
This is why the self-attention in the decoder must be *masked*

In this way, the decoder is simultaneously trained on T next token (word) predictions

Transformers in action

The original transformer

The original transformer is a *sequence-to-sequence* model for translation



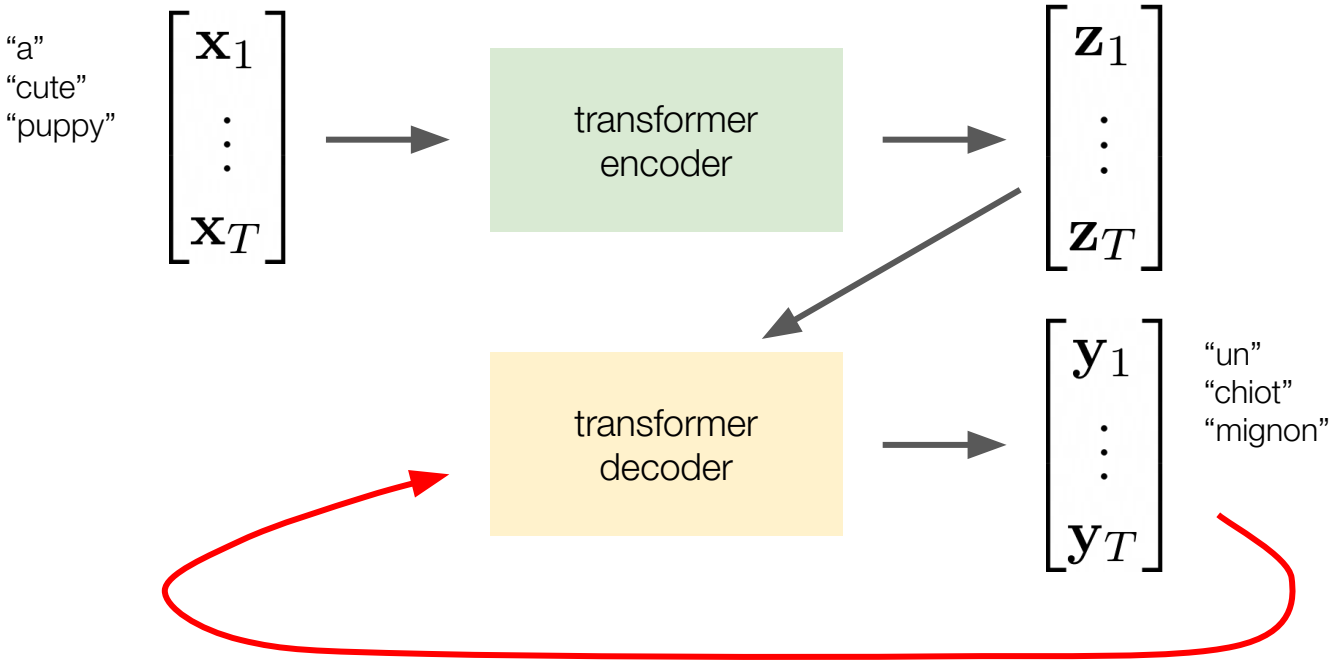
Sequence-to-sequence models typically follow an *encoder-decoder* architecture, and the transformer does as well

Next week, we will talk in greater detail about sequence-to-sequence models

Vaswani et al, “Attention is All You Need”. NIPS 2017.

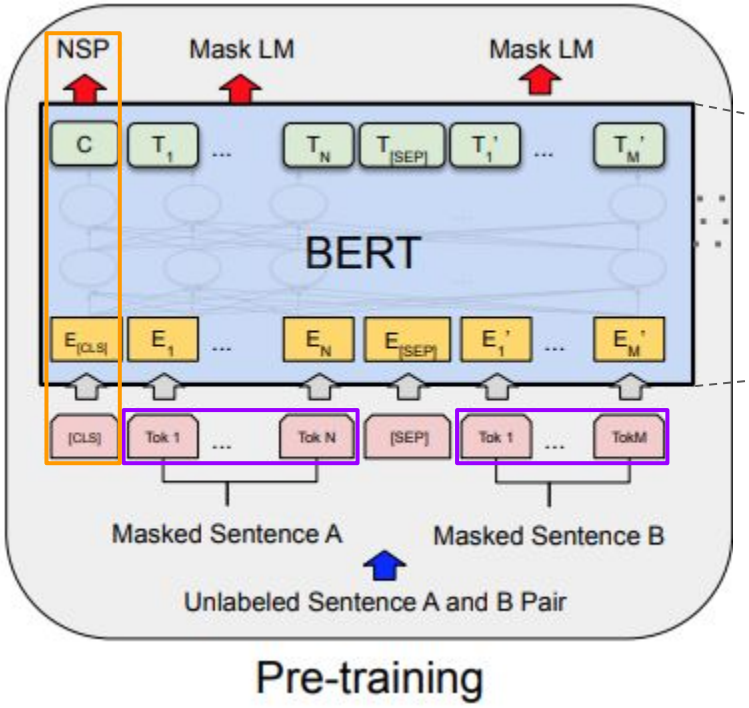
The original transformer

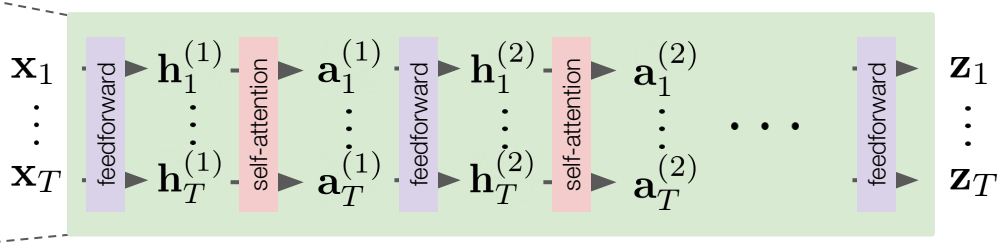
The original transformer follows an *encoder-decoder* architecture



Vaswani et al, “Attention is All You Need”. NIPS 2017.

Just the encoder: BERT





During training:

- ➔ Input tokens (words) are randomly masked with 15% probability, the model must predict the tokens
  - word level representations
- ➔ Pairs of sentences are passed in, the model must predict which sentence follows the other
  - sentence level representations

Devlin et al, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. NAACL 2019.

https://gluebenchmark.com/

The GLUE benchmark

GLUE Tasks

Name	Download	More Info	Metric
The Corpus of Linguistic Acceptability			Matthew's C
The Stanford Sentiment Treebank			Accuracy
Microsoft Research Paraphrase Corpus			F1 / Accuracy
Semantic Textual Similarity Benchmark			P
Quora Question Pairs			F
MultiNLI Matched			A
MultiNLI Mismatched			A
Question NLI			A
Recognizing Textual Entailment			A
Winograd NLI			A
Diagnostics Main			M

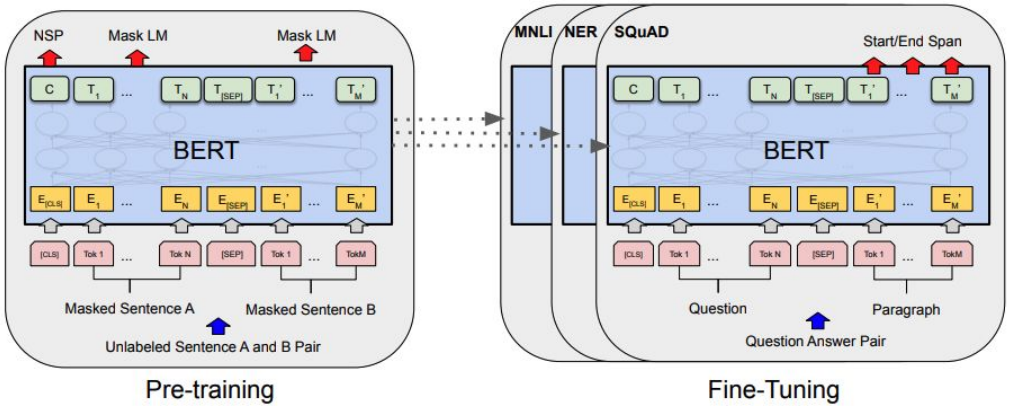
0 * Books were sent to each other by the students.

1 She voted for herself.

1 I saw that gas can explode.

Premise	Label	Hypothesis
Fiction		
The Old One always comforted Ca'daan, except today.	*neutral*	Ca'daan knew the Old One very well.
Letters		
Your gift is appreciated by each and every student who will benefit from your generosity.	*neutral*	Hundreds of students will benefit from your generosity.
Telephone Speech		
yes now you know if if everybody like in August when everybody's on vacation or something we can dress a little more casual or	*contradiction*	August is a black out month for vacations in the company.
9/11 Report		
At the other end of Pennsylvania Avenue, people began to line up for a White House tour.	*entailment*	People formed a line at the end of Pennsylvania Avenue.

“Fine tuning” BERT

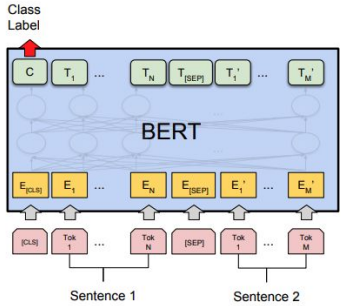


Up to billions of sentences

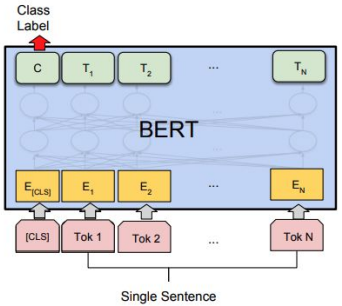
Much less data

Mask / next sentence prediction

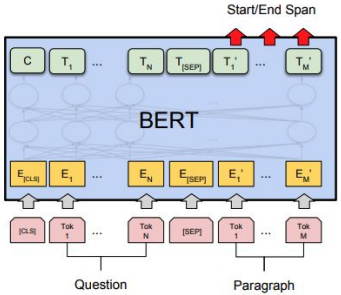
Specific to the task at hand



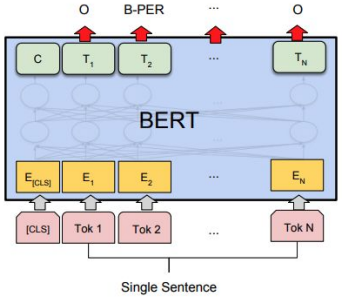
(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG



(b) Single Sentence Classification Tasks: SST-2, CoLA



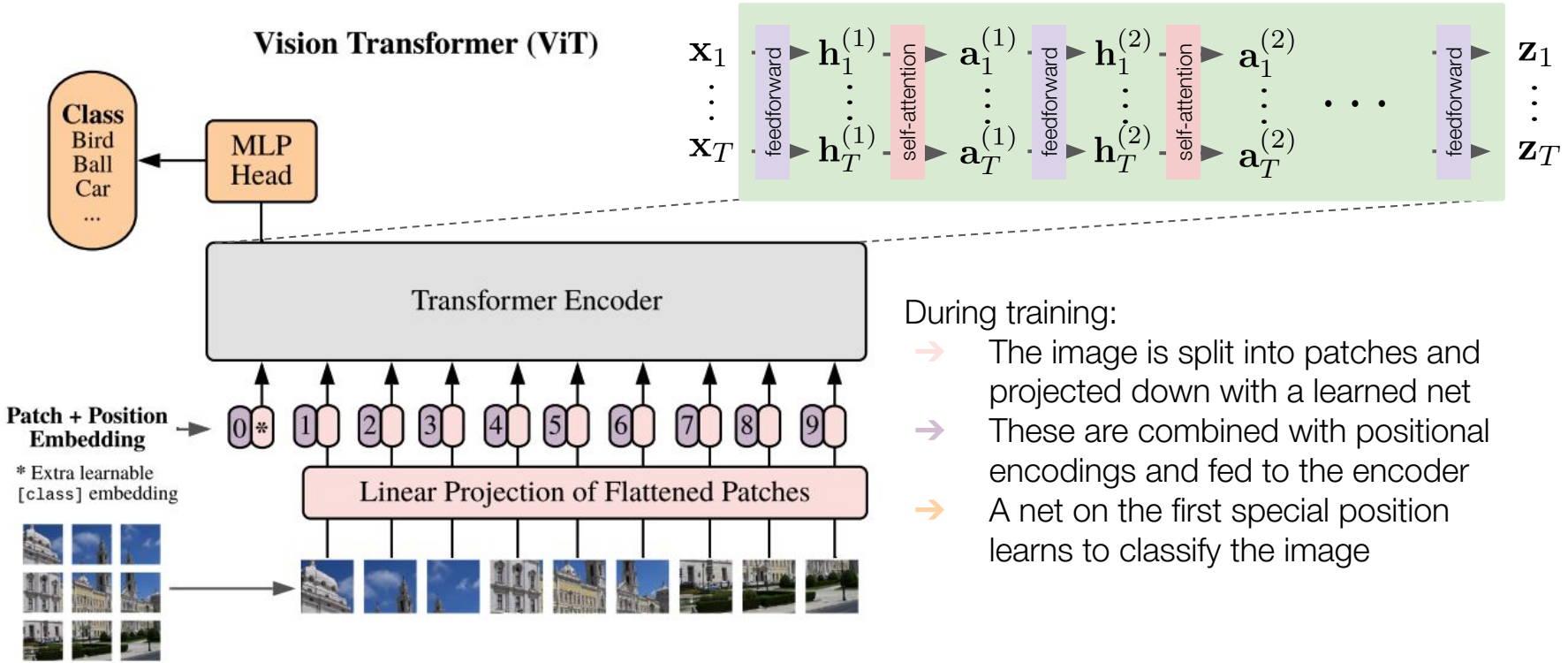
(c) Question Answering Tasks: SQuAD v1.1



(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

Devlin et al, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. NAACL 2019.

Just the encoder: vision transformers (ViTs)



During training:

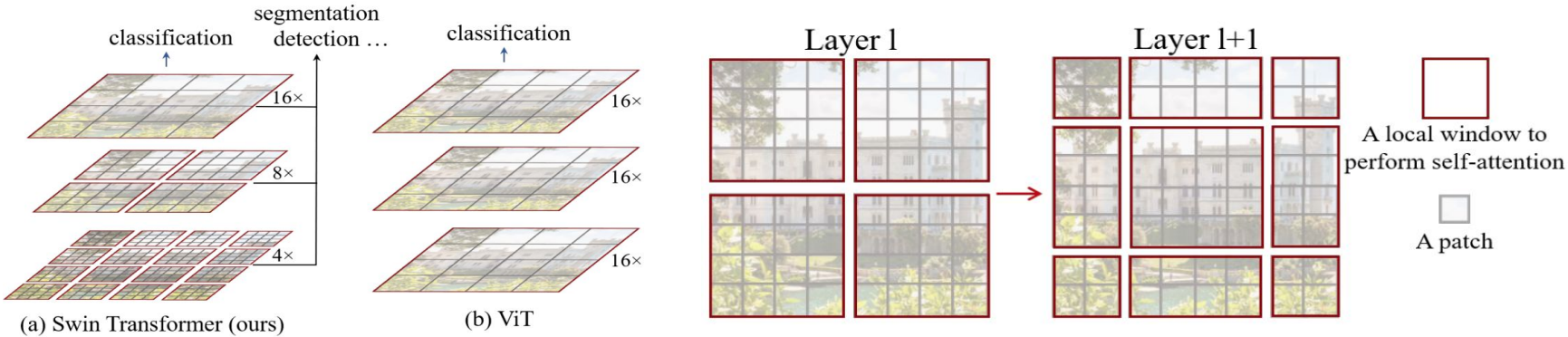
- The image is split into patches and projected down with a learned net
- These are combined with positional encodings and fed to the encoder
- A net on the first special position learns to classify the image

Dosovitskiy et al, “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale”. ICLR 2021.

An even more recent ViT

Swin Transformer: Hierarchical Vision Transformer using Shifted Windows

Ze Liu[†*] Yutong Lin[†*] Yue Cao[*] Han Hu[*‡] Yixuan Wei[†]
Zheng Zhang Stephen Lin Baining Guo
Microsoft Research Asia



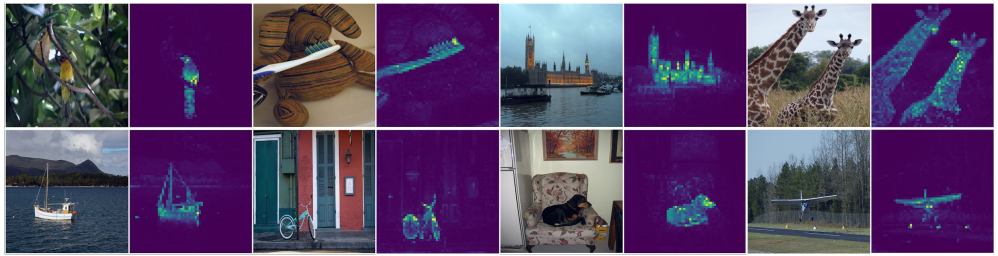
Liu et al, “Swin Transformer: Hierarchical Vision Transformer using Shifted Windows”. ICCV 2021.

Unsupervised training of ViTs

Emerging Properties in Self-Supervised Vision Transformers

Mathilde Caron[1,2] Hugo Touvron[1,3] Ishan Misra[1] Hervé Jegou[1]
Julien Mairal[2] Piotr Bojanowski[1] Armand Joulin[1]

[1] Facebook AI Research [2] Inria* [3] Sorbonne University



Caron et al, “Emerging Properties in Self-Supervised Vision Transformers”. ICCV 2021.

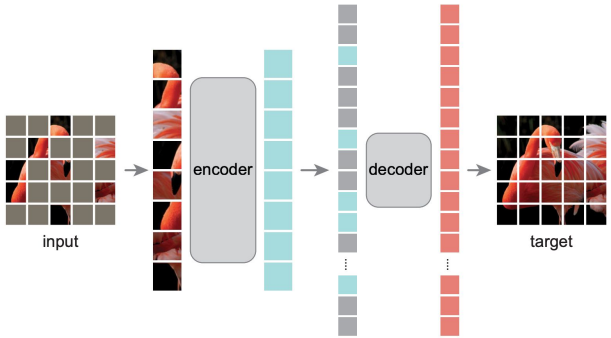
Unsupervised training of ViTs

Masked Autoencoders Are Scalable Vision Learners

Kaiming He*,† Xinlei Chen* Saining Xie Yanghao Li Piotr Dollár Ross Girshick

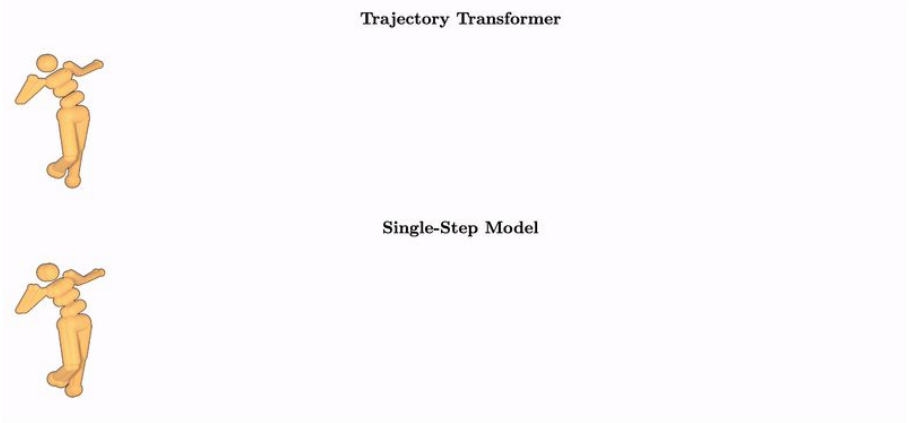
*equal technical contribution †project lead

Facebook AI Research (FAIR)



He et al, “Masked Autoencoders are Scalable Vision Learners”. arXiv 2111.05377.

Just the decoder: reinforcement learning



Chen et al, “Decision Transformer: Reinforcement Learning via Sequence Modeling”. NeurIPS 2021.
Janner et al, “Reinforcement Learning as One Big Sequence Modeling Problem”. NeurIPS 2021.

https://lacker.io/ai/2020/07/06/giving-gpt-3-a-turing-test.html

Just the decoder: GPT

Now we're getting into surreal territory. GPT-3 knows how to have a
know how to say "Wait a moment… your question is nonsense." It a
know."

```
Q: How do you sporgle a morgle?
A: You sporgle a morgle by using a sporgle.

Q: How many bonks are in a quoit?
A: There are three bonks in a quoit.

Q: How many rainbows does it take to jump from Hawaii t
A: It takes two rainbows to jump from Hawaii to seventee

Q: Which colorless green ideas sleep furiously?
A: Ideas that are colorless, green, and sleep furiously
furiously.

Q: Do you understand these questions?
A: I understand these questions.
```

This is a conversation between a human and a brilliant AI. If a question is "normal" the AI answers it. If the question is "nonsense" the AI says "yo be real"

Q: What is human life expectancy in the United States?
A: Human life expectancy in the United States is 78 years.

Q: How do you sporkle a morgle?
A: yo be real

Q: Who was president of the United States before George W. Bush?
A: Bill Clinton was president of the United States before George W. Bush.

Q: How many rainbows does it take to jump from Hawaii to seventeen?
A: yo be real

Q: How does an umbrella work
A: An umbrella works by using a series of spokes to keep the rain from falling on you.

Q: How many bonks are in a quoit?
A: yo be real

Q: Which colorless green ideas speak furiously
A: yo be real

OpenAI, "Language Models are Few-Shot Learners". arXiv 2005.14165.

Summary

- Transformers are the latest and greatest model for *sequential data*

- The fundamental building block is *self-attention*, which allows for processing the whole sequence all at once
    - Compared to previous sequence models, long range dependencies are captured much better with self-attention

- Even data that is not obviously sequential has benefited from transformers

- What other great architecture ideas are still out there?