Popular Transformers

Danna Gurari
University of Colorado Boulder
Fall 2022

https://home.cs.colorado.edu/~DrG/Courses/NeuralNetworksAndDeepLearning/AboutCourse.html

Review

- Last lecture:
  - Transformer overview
  - Self-attention
  - Multi-head attention
  - Common transformer ingredients
  - Pioneering transformer: machine translation
  - Programming tutorial
- Assignments (Canvas):
  - Problem set 3 due next Monday
  - Final project begins in a few weeks (quick discussion: finding partner and overview)
- Questions?

Today’s Topics

- Explosion of transformers
- GPT
- BERT
- ViT
- Limitations of transformer models

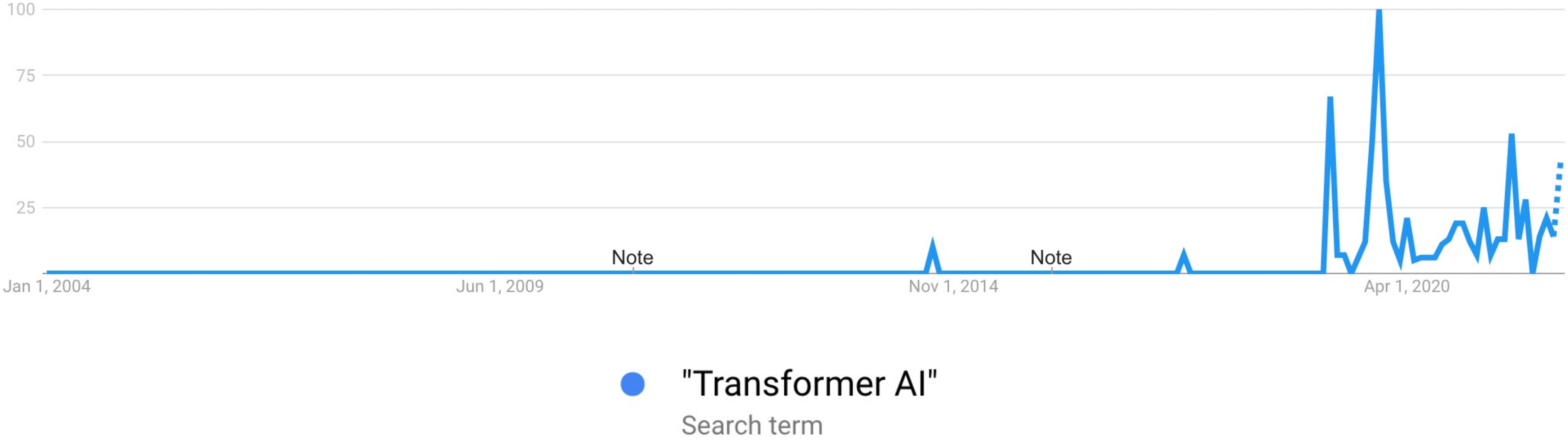
Today’s Topics

- Explosion of transformers
- GPT
- BERT
- ViT
- Limitations of transformer models

Explosion of Transformers in Society

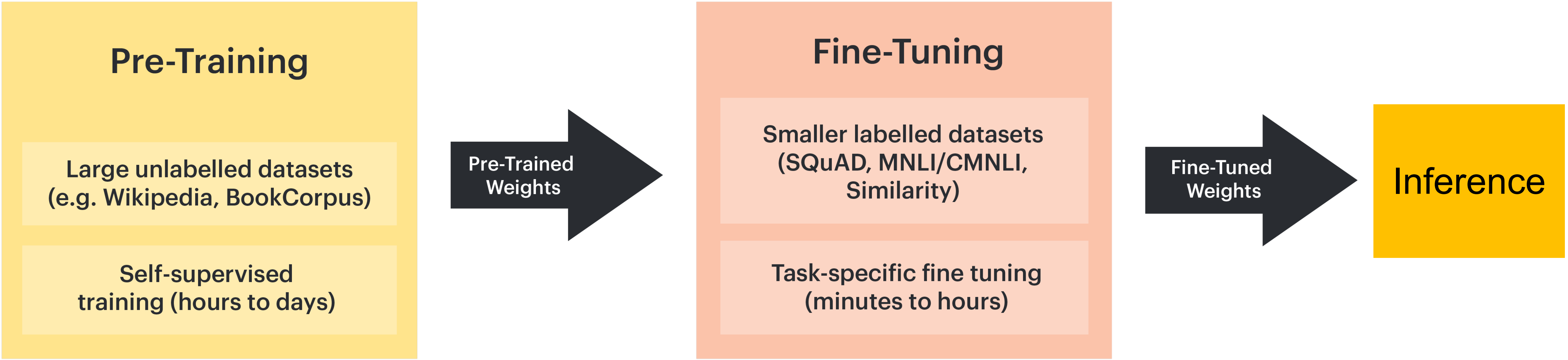
https://www.pinterest.com/pin/521784306804400819/

Popularity of Transformers in Society



https://trends.google.com/trends/explore

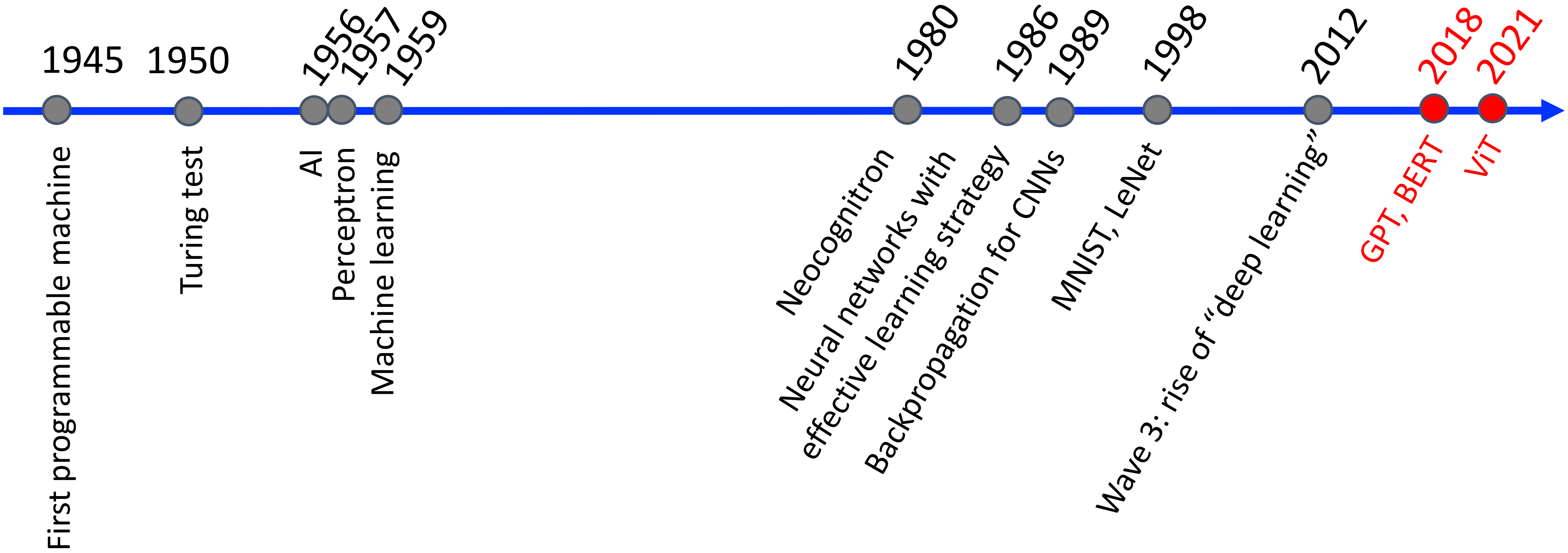
Today's Focus: Methods that Perform Pretraining and then Fine-tuning



Transformers can provide better embeddings for downstream tasks since they capture context (i.e., unlike context-free embeddings such as word2vec, the word embedding is different for a word used in different contexts… e.g., "I arrived at the bank after crossing the river/street")

https://docs.graphcore.ai/projects/bert-training/en/latest/bert.html

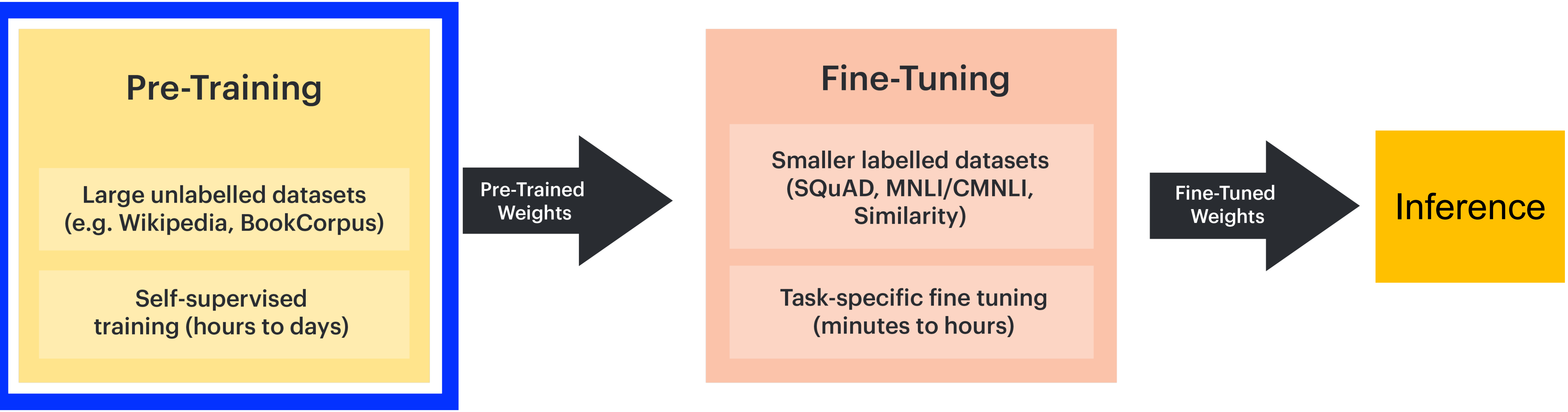
Today's Focus: Historical Context



Today’s Topics

- Explosion of transformers
- GPT
- BERT
- ViT
- Limitations of transformer models

GPT: Generative Pre-Training



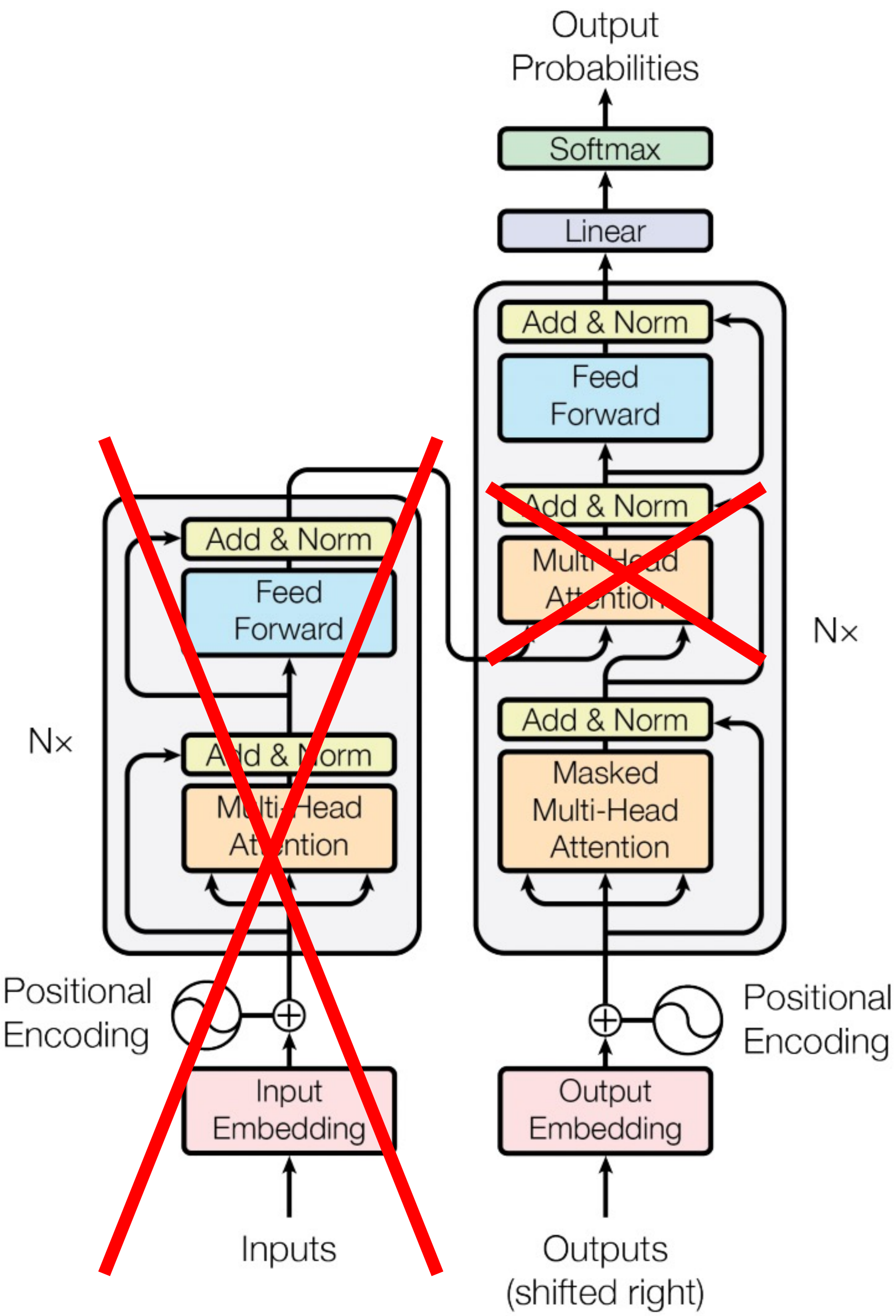
https://docs.graphcore.ai/projects/bert-training/en/latest/bert.html

Task: Predict Next Word Given Previous Ones

e.g.,

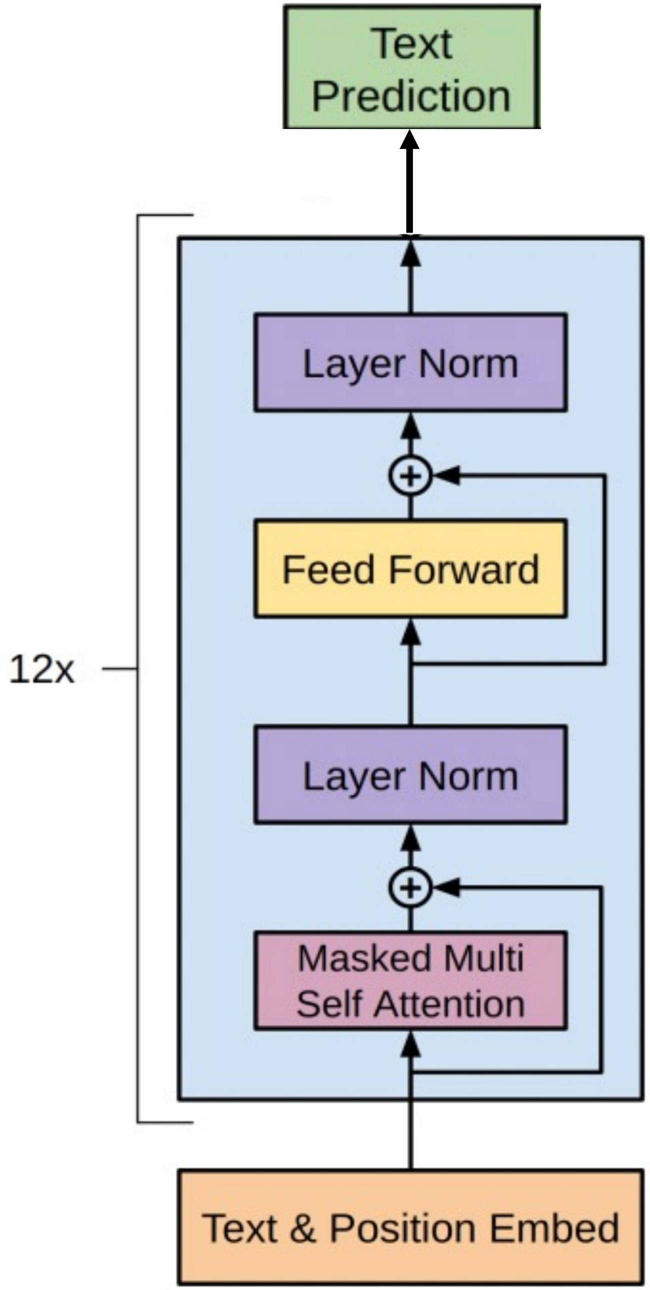
1. Background music from a ________
2. Many people danced around the ________
3. I practiced for many years to learn how to play the ________

Architecture: Decoder from Pioneering Transformer



Vaswani et al. Attention Is All You Need. Neurips 2017.

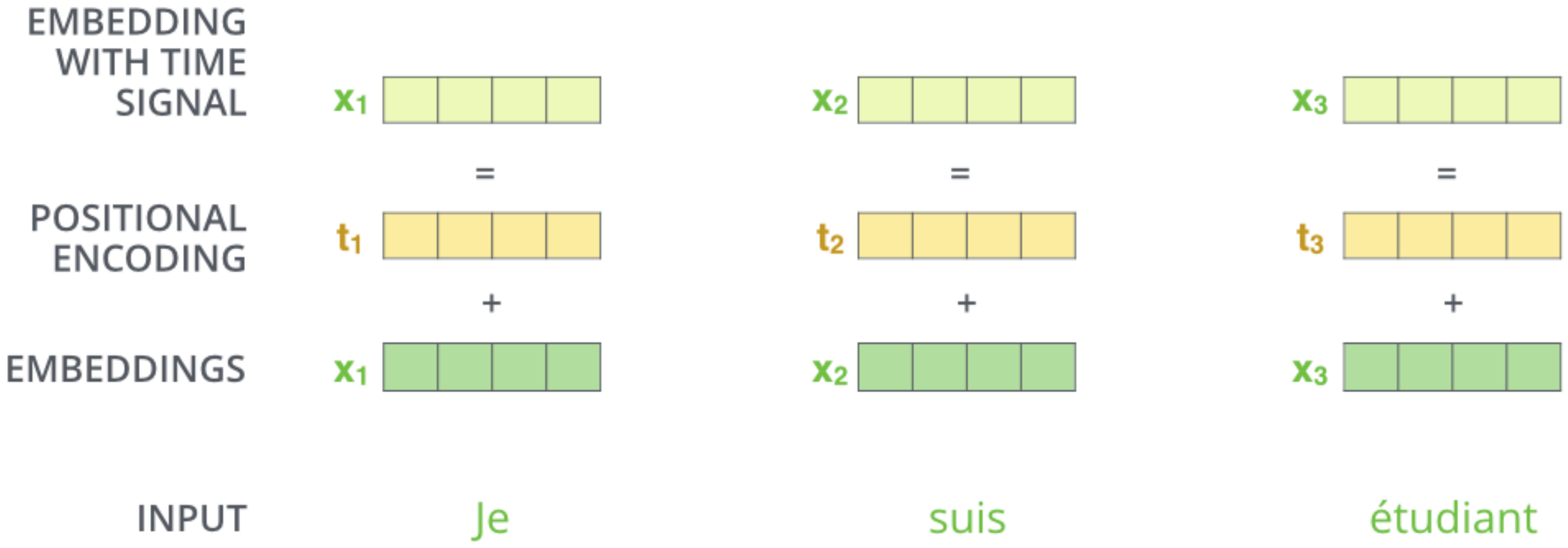
Architecture

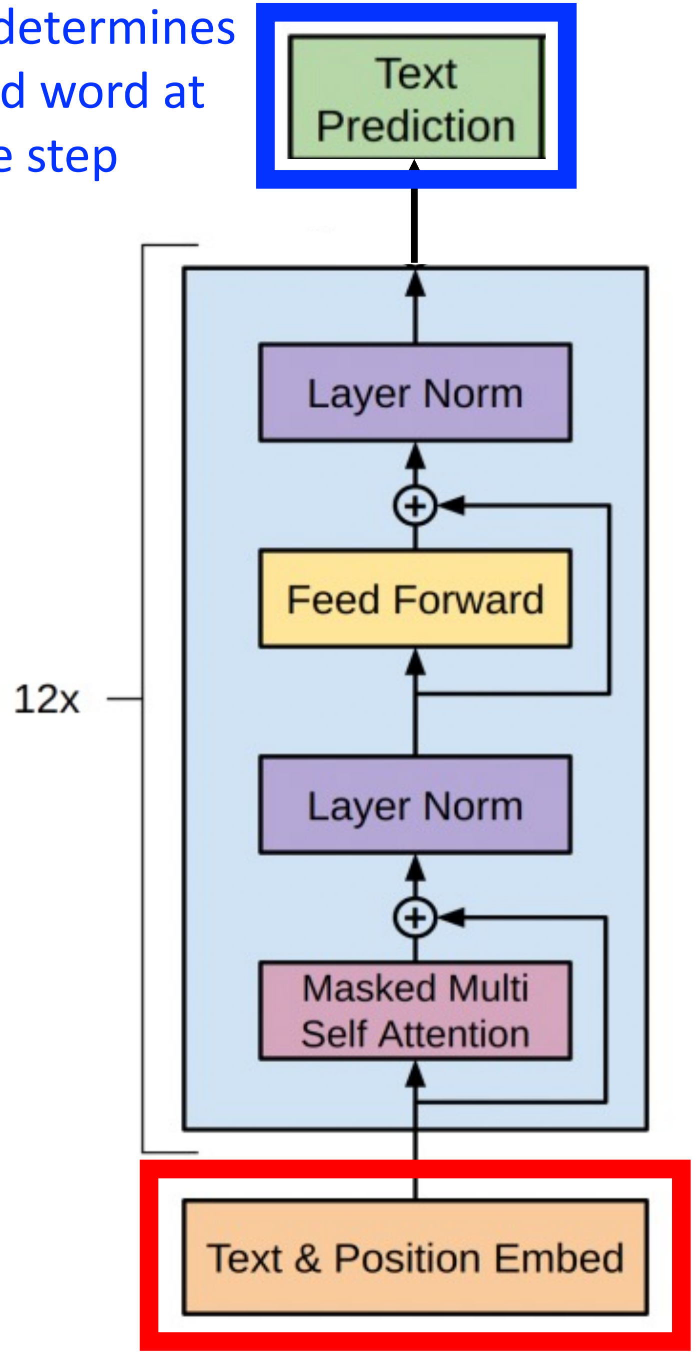


Radford et al. Improving Language Understanding by Generative Pre-Training. Technical Report 2018.

Architecture: Input & Output

Softmax layer determines next predicted word at each time step





http://jalammar.github.io/illustrated-transformer/

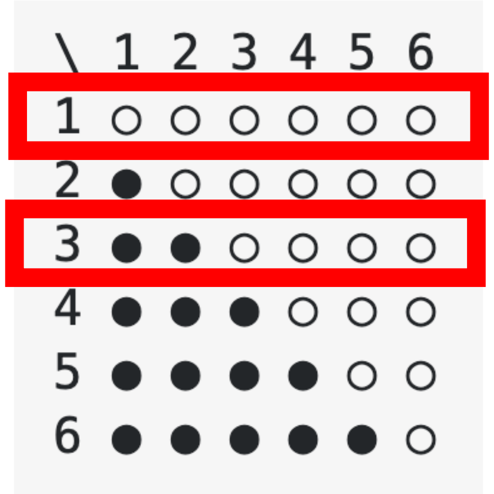
Radford et al. Improving Language Understanding by Generative Pre-Training. Technical Report 2018.

Architecture: Masked Attention

Limit each word's new representation to only reflect earlier words (mimics inference time when only previous tokens can be seen):

Query x Key

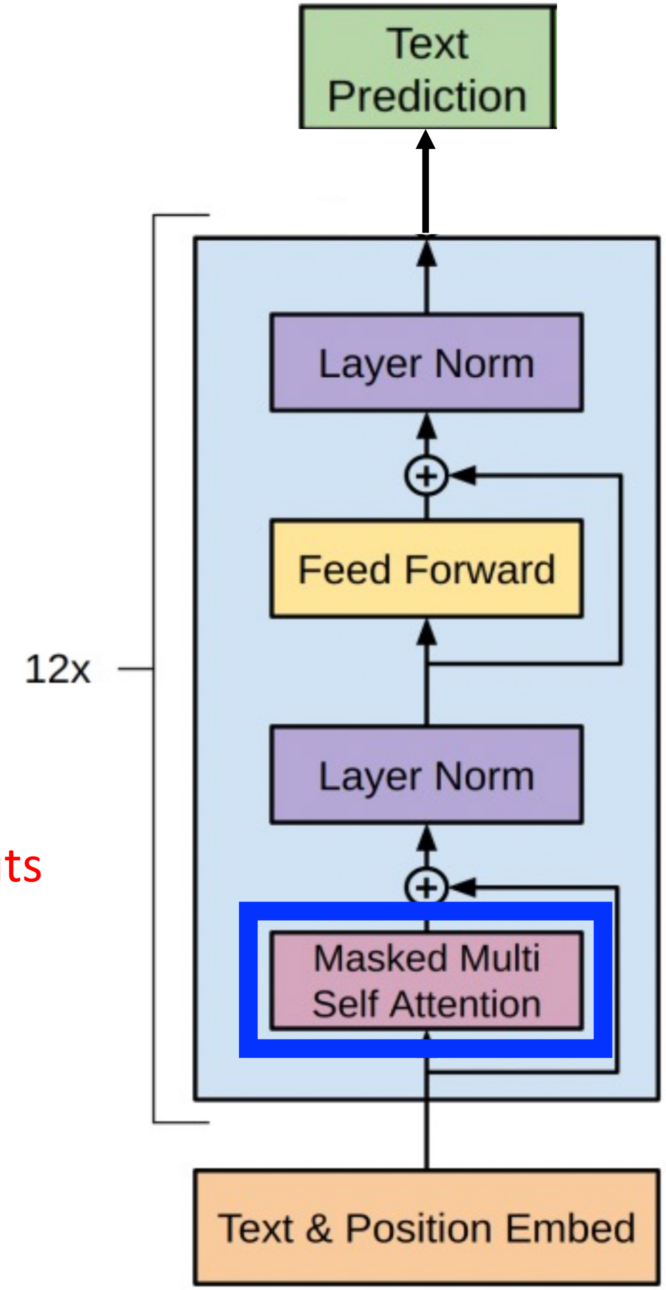
Query



e.g., at <sos> token, no previous inputs

e.g., at 3rd token, have two previous inputs

https://stackoverflow.com/questions/64799622/how-is-the-gpts-masked-self-attention-is-utilized-on-fine-tuning-inference



Radford et al. Improving Language Understanding by Generative Pre-Training. Technical Report 2018.

Architecture: Masked Attention

Limit each word's new representation to only reflect earlier words (mimics inference time when only previous tokens can be seen):

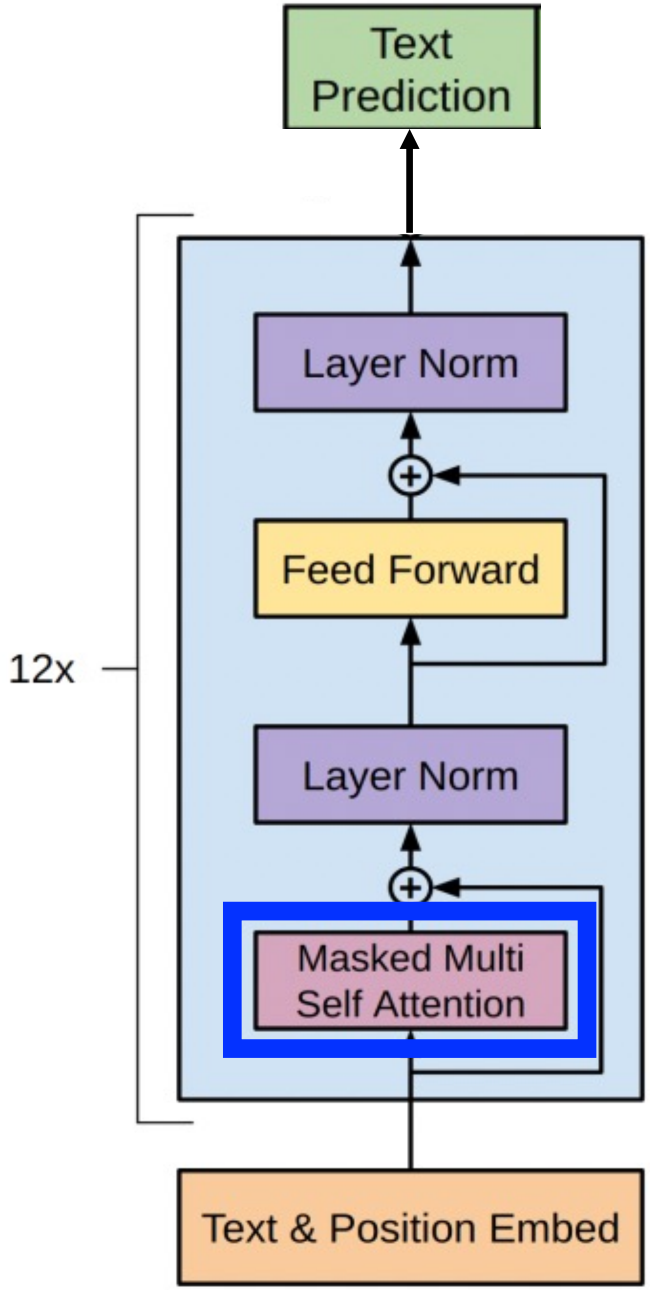
Query x Key

Query

\	1	2	3	4	5	6
1	○	○	○	○	○	○
2	●	○	○	○	○	○
3	●	●	○	○	○	○
4	●	●	●	○	○	○
5	●	●	●	●	○	○
6	●	●	●	●	●	○

Masked out scores are represented as negative infinity so the softmax result (i.e., attention weight) returns 0

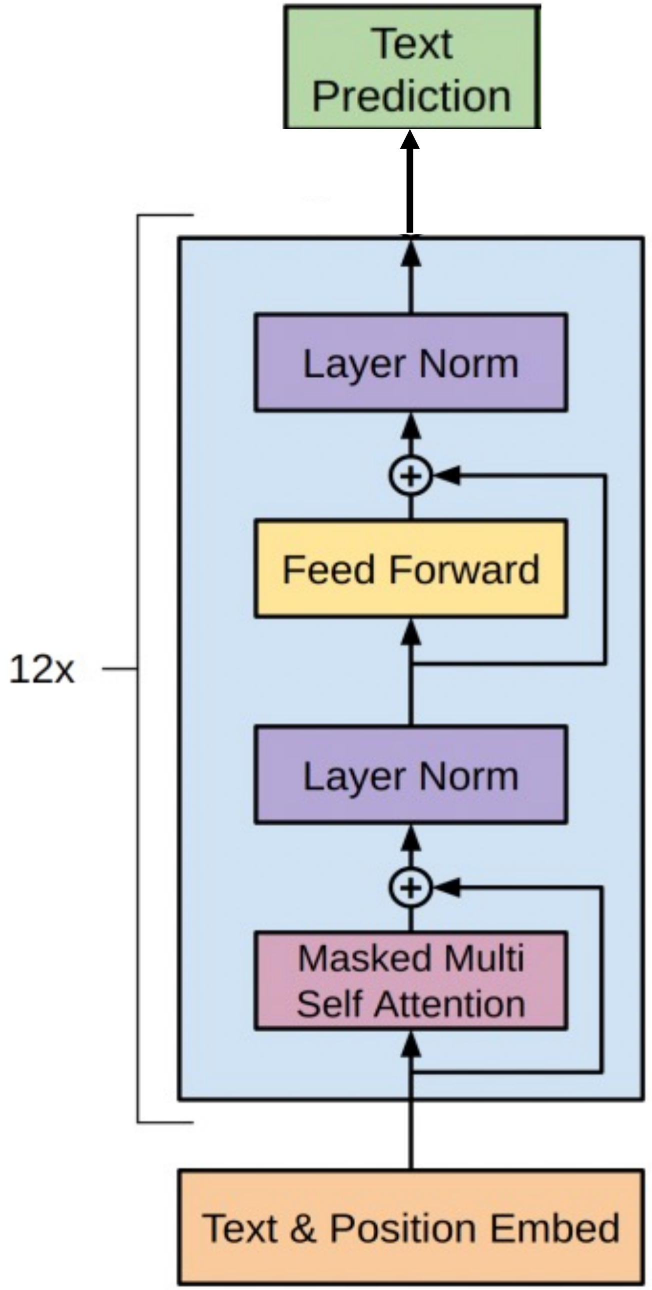
https://stackoverflow.com/questions/64799622/how-is-the-gpts-masked-self-attention-is-utilized-on-fine-tuning-inference



Radford et al. Improving Language Understanding by Generative Pre-Training. Technical Report 2018.

Training

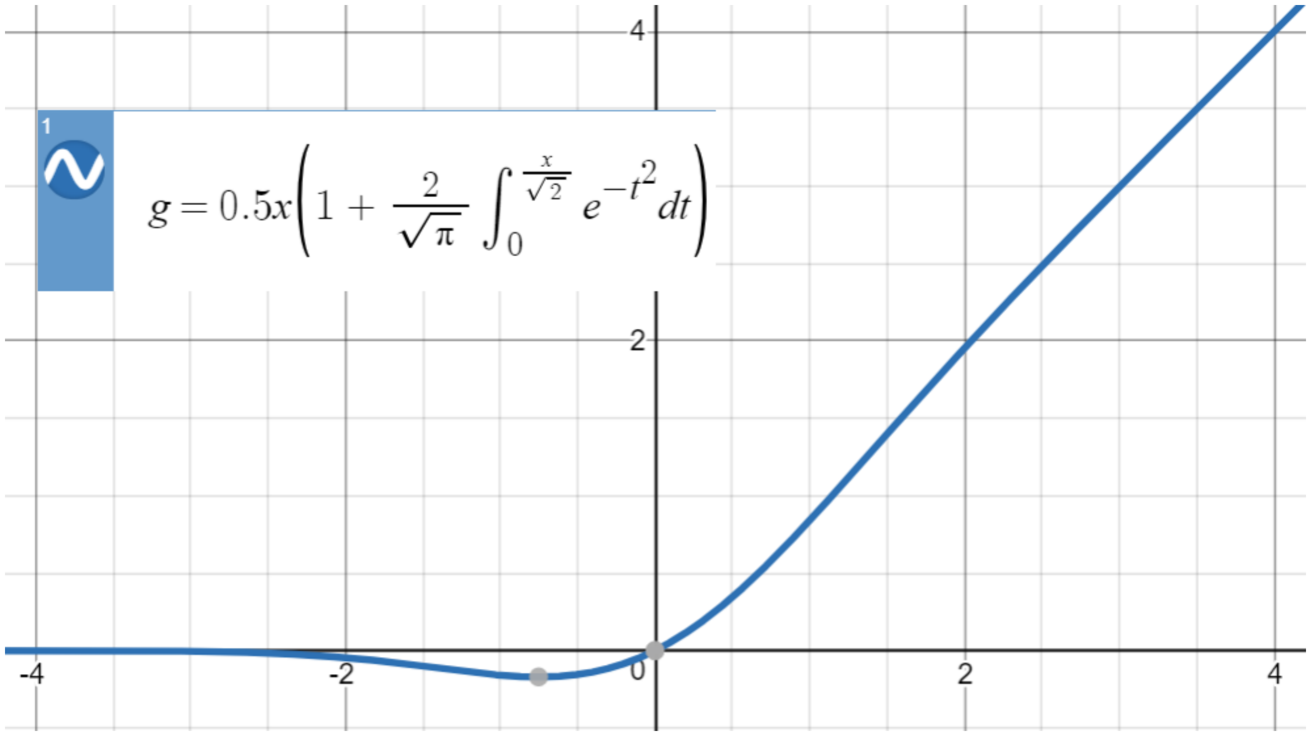
- Dataset: 800M words from BooksCorpus (>7,000 books)
- Optimizer: Adam
- Training duration: 100 epochs



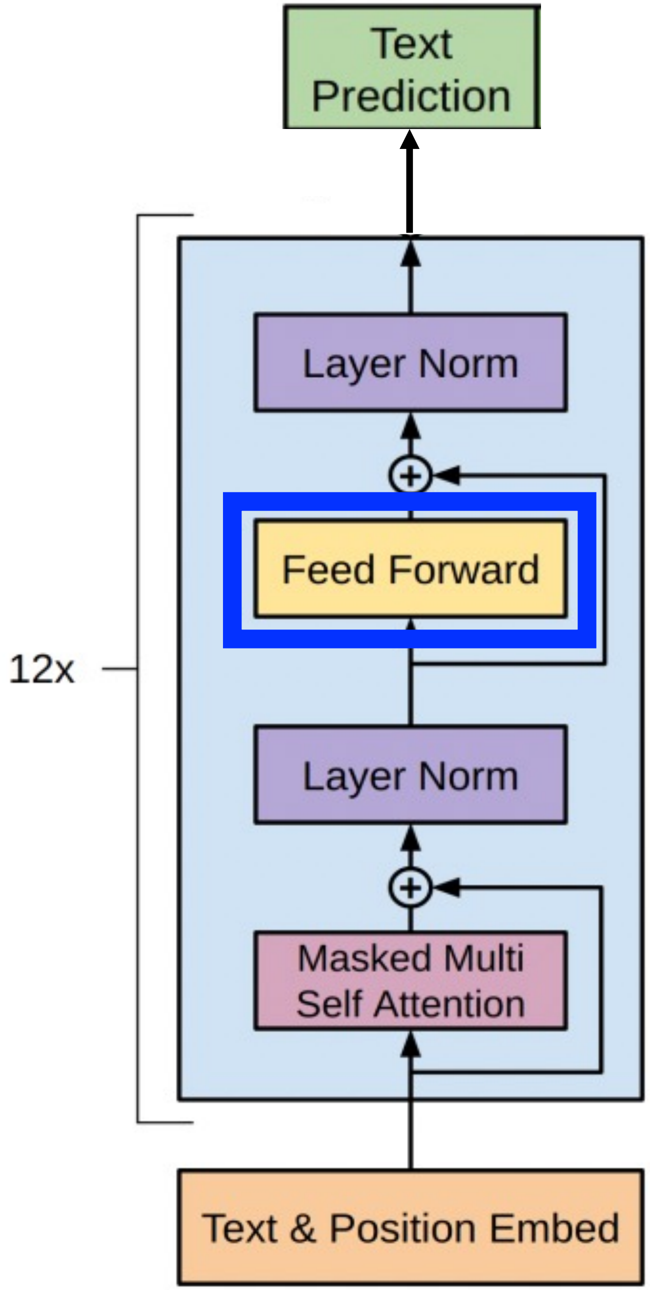
Radford et al. Improving Language Understanding by Generative Pre-Training. Technical Report 2018.

Implementation Details

Activation function: Gaussian error linear unit (GELU)



https://datascience.stackexchange.com/questions/49522/what-is-gelu-activation



Radford et al. Improving Language Understanding by Generative Pre-Training. Technical Report 2018.

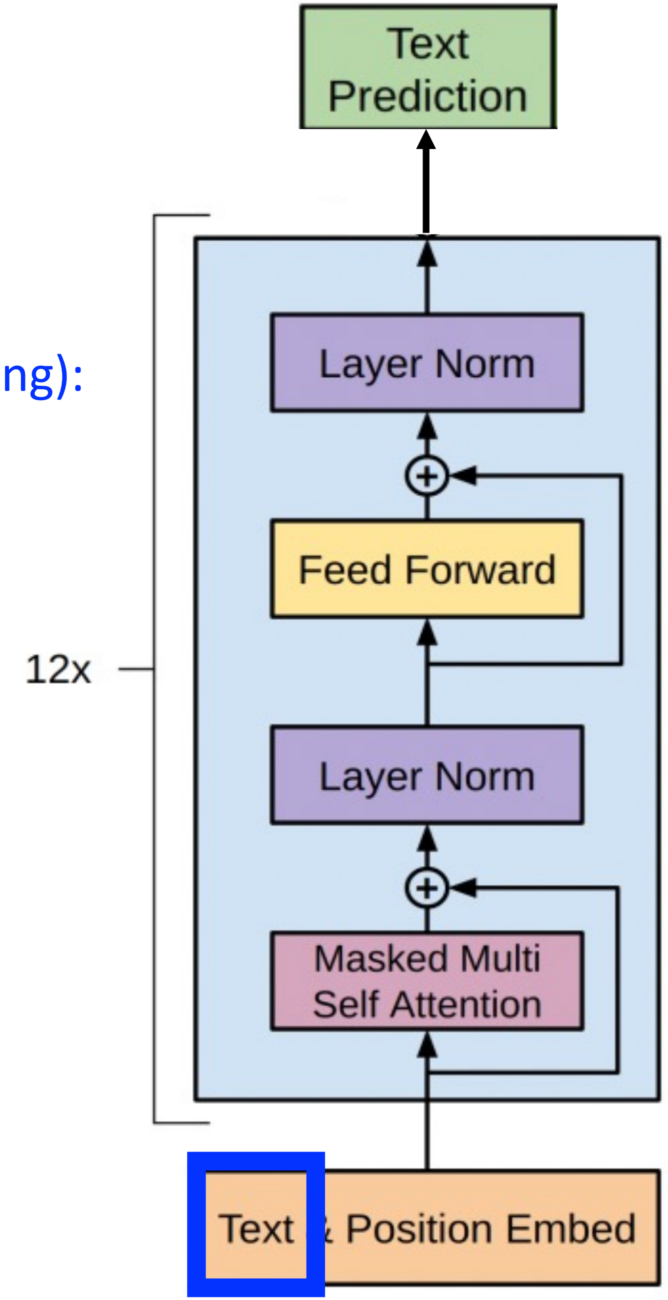
Implementation Details

Avoid out of vocabulary tokens with subword tokenization (byte pair encoding):

1. Identify all tokens in the training data with their frequency
2. Define vocabulary size; e.g., 14
3. Add all characters in the tokenized input to the vocabulary; e.g.,

Character sequence	Cost
C o s t	2
b e s t	2
m e n u	1
m e n	1
c a m e l	1

https://static.packt-cdn.com/downloads/9781838821593_ColorImages.pdf



Radford et al. Improving Language Understanding by Generative Pre-Training. Technical Report 2018.

Implementation Details

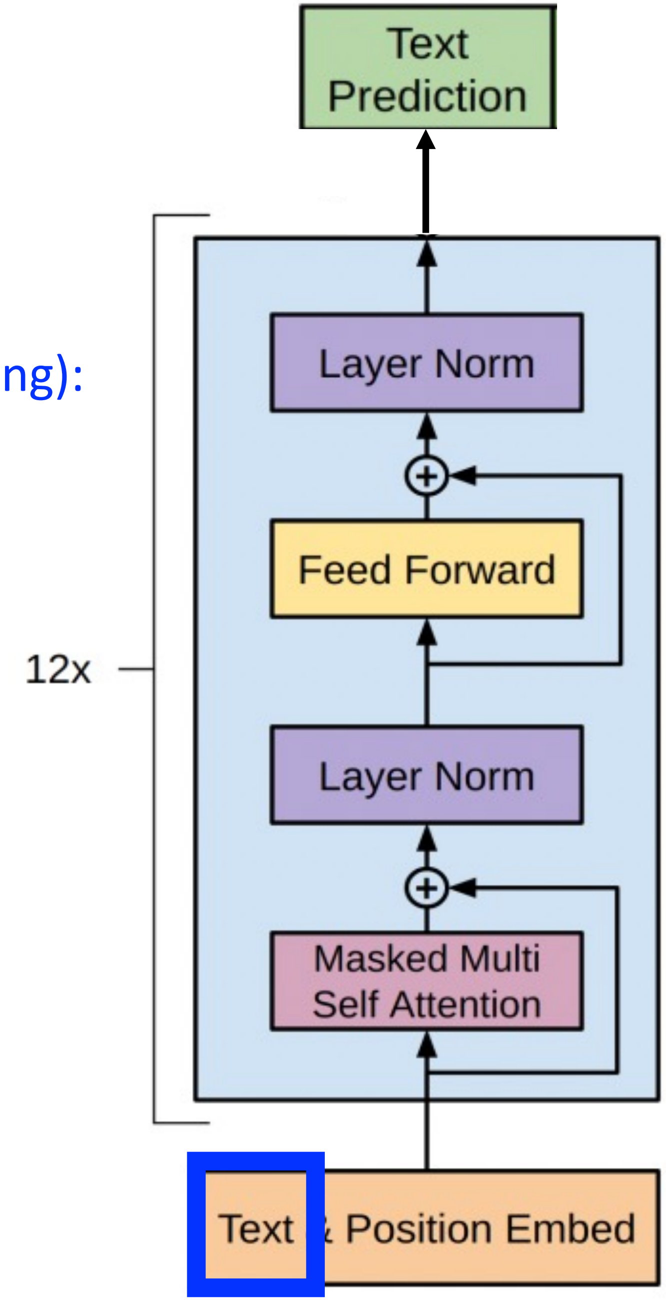
Avoid out of vocabulary tokens with subword tokenization (byte pair encoding):

1. Identify all tokens in the training data with their frequency
2. Define vocabulary size; e.g., 14
3. Add all characters in the tokenized input to the vocabulary; e.g.,
4. Until vocabulary is filled, add merged highest frequency symbol pairs

e.g., What are the highest frequency symbol pairs?

Character sequence	Cost	Vocabulary
c o s t	2	a, b, c, e, l, m, n, o, s, t, u
b e s t	2	
m e n u	1	
m e n	1	
c a m e l	1	

https://static.packt-cdn.com/downloads/9781838821593_ColorImages.pdf



Radford et al. Improving Language Understanding by Generative Pre-Training. Technical Report 2018.

Implementation Details

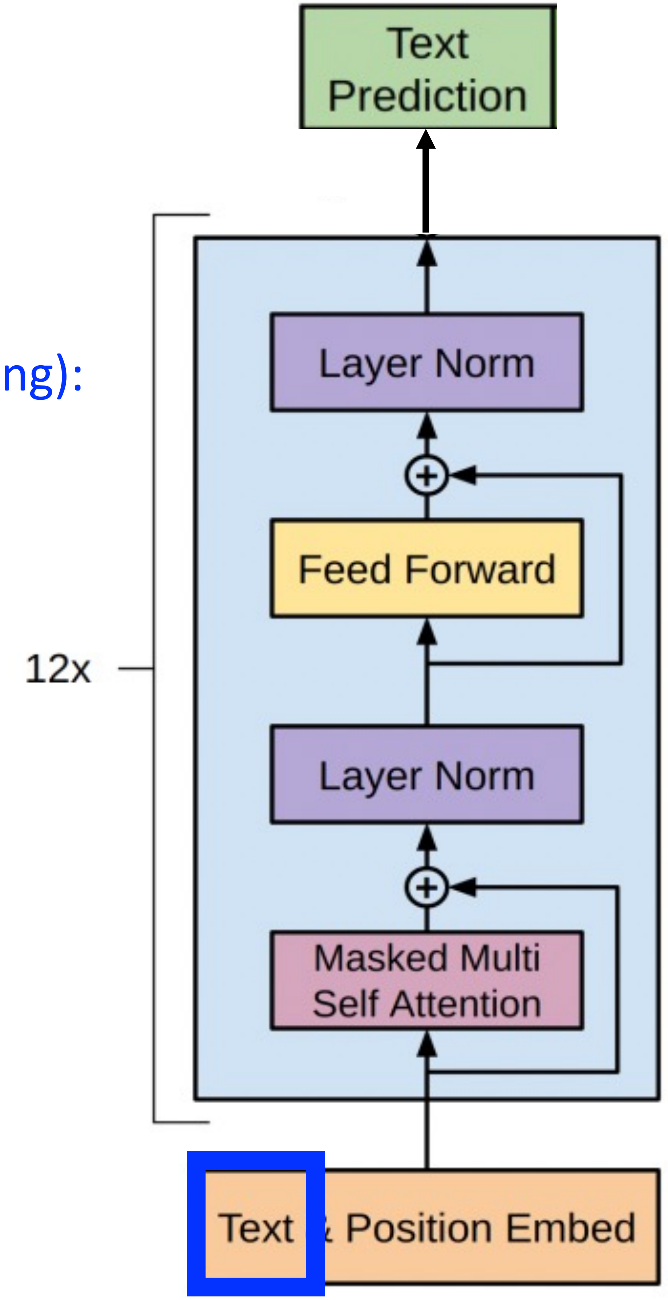
Avoid out of vocabulary tokens with subword tokenization (byte pair encoding):

1. Identify all tokens in the training data with their frequency
2. Define vocabulary size; e.g., 14
3. Add all characters in the tokenized input to the vocabulary; e.g.,
4. Until vocabulary is filled, add merged highest frequency symbol pairs

e.g., What are the highest frequency symbol pairs?

Character sequence	Cost	Vocabulary
C o s t	2	a, b, c, e, l, m, n, o, s, t, u, st
b e s t	2	
m e n u	1	
m e n	1	
c a m e l	1	

https://static.packt-cdn.com/downloads/9781838821593_ColorImages.pdf



Radford et al. Improving Language Understanding by Generative Pre-Training. Technical Report 2018.

Implementation Details

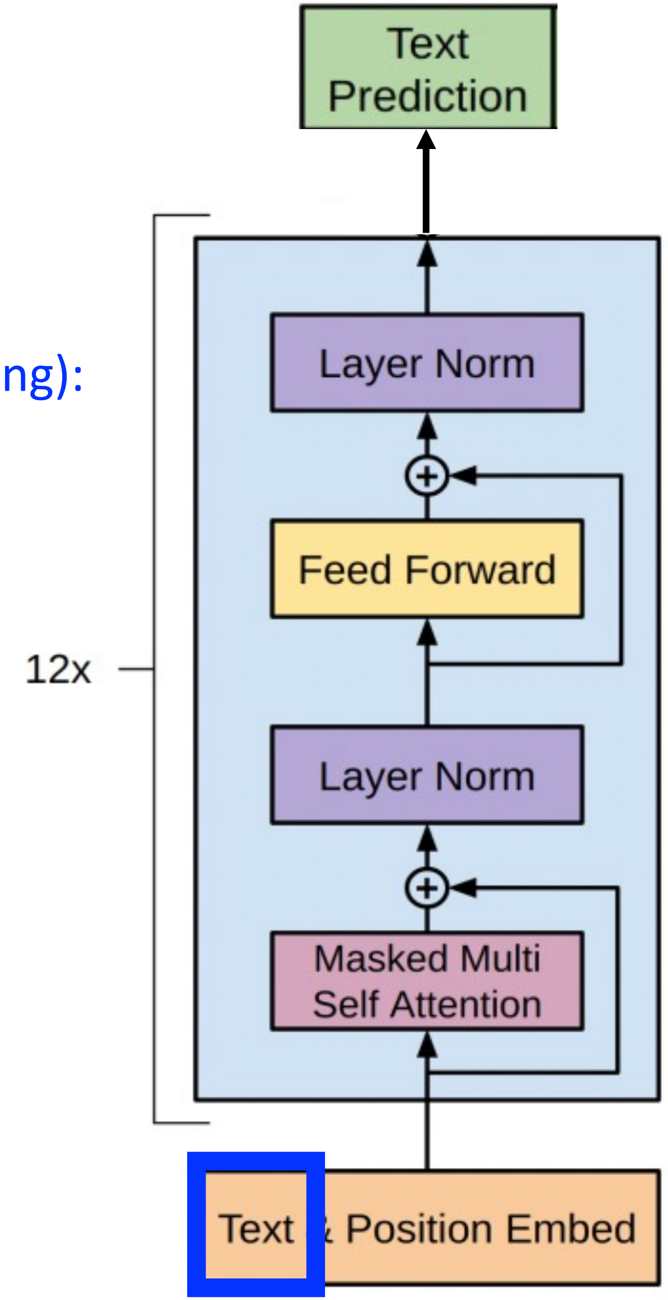
Avoid out of vocabulary tokens with subword tokenization (byte pair encoding):

1. Identify all tokens in the training data with their frequency
2. Define vocabulary size; e.g., 14
3. Add all characters in the tokenized input to the vocabulary; e.g.,
4. Until vocabulary is filled, add merged highest frequency symbol pairs

e.g., What are the highest frequency symbol pairs?

Character sequence	Cost	Vocabulary
C o s t	2	a, b, c, e, l, m, n, o, s, t, u, st, me, **men**
b e s t	2	
m e n u	1	
m e n	1	
c a m e l	1	

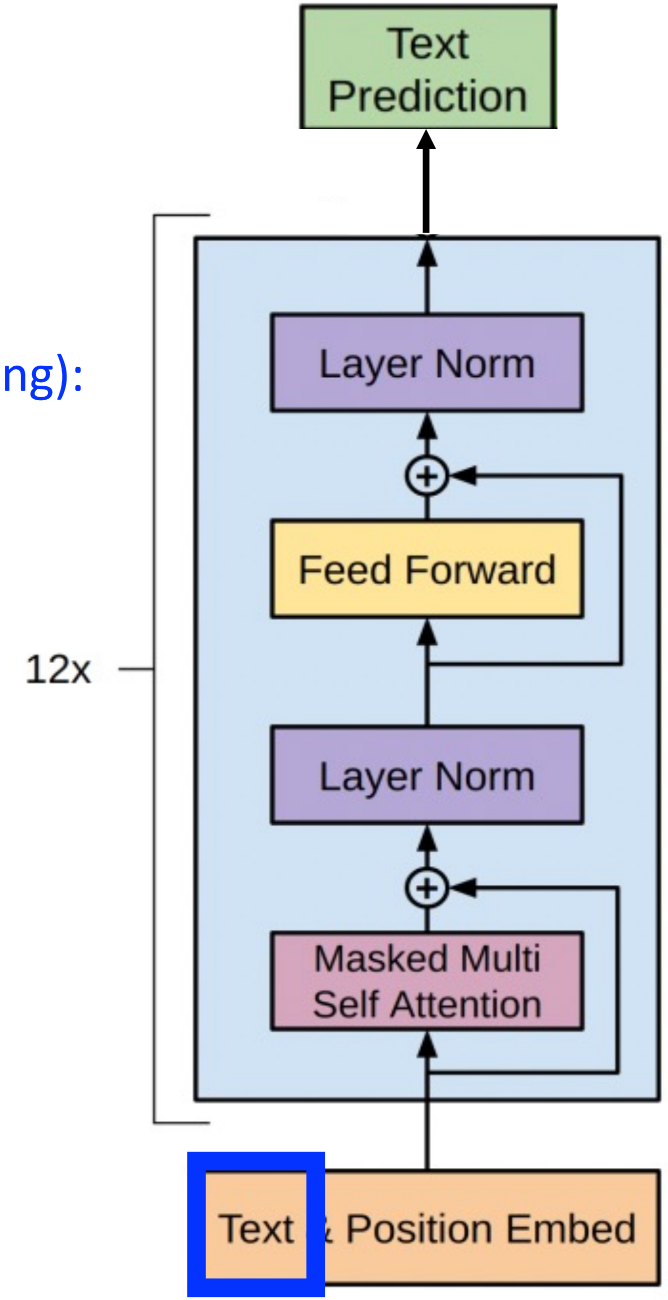
https://static.packt-cdn.com/downloads/9781838821593_ColorImages.pdf



Radford et al. Improving Language Understanding by Generative Pre-Training. Technical Report 2018.

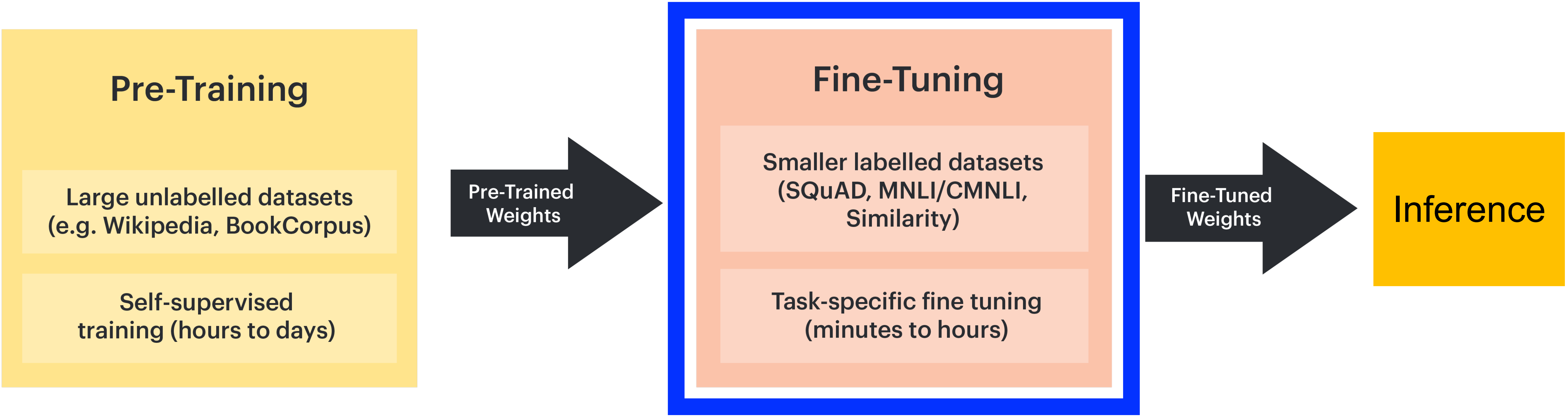
Implementation Details

Avoid out of vocabulary tokens with subword tokenization (byte pair encoding):
- 40,000 merges used



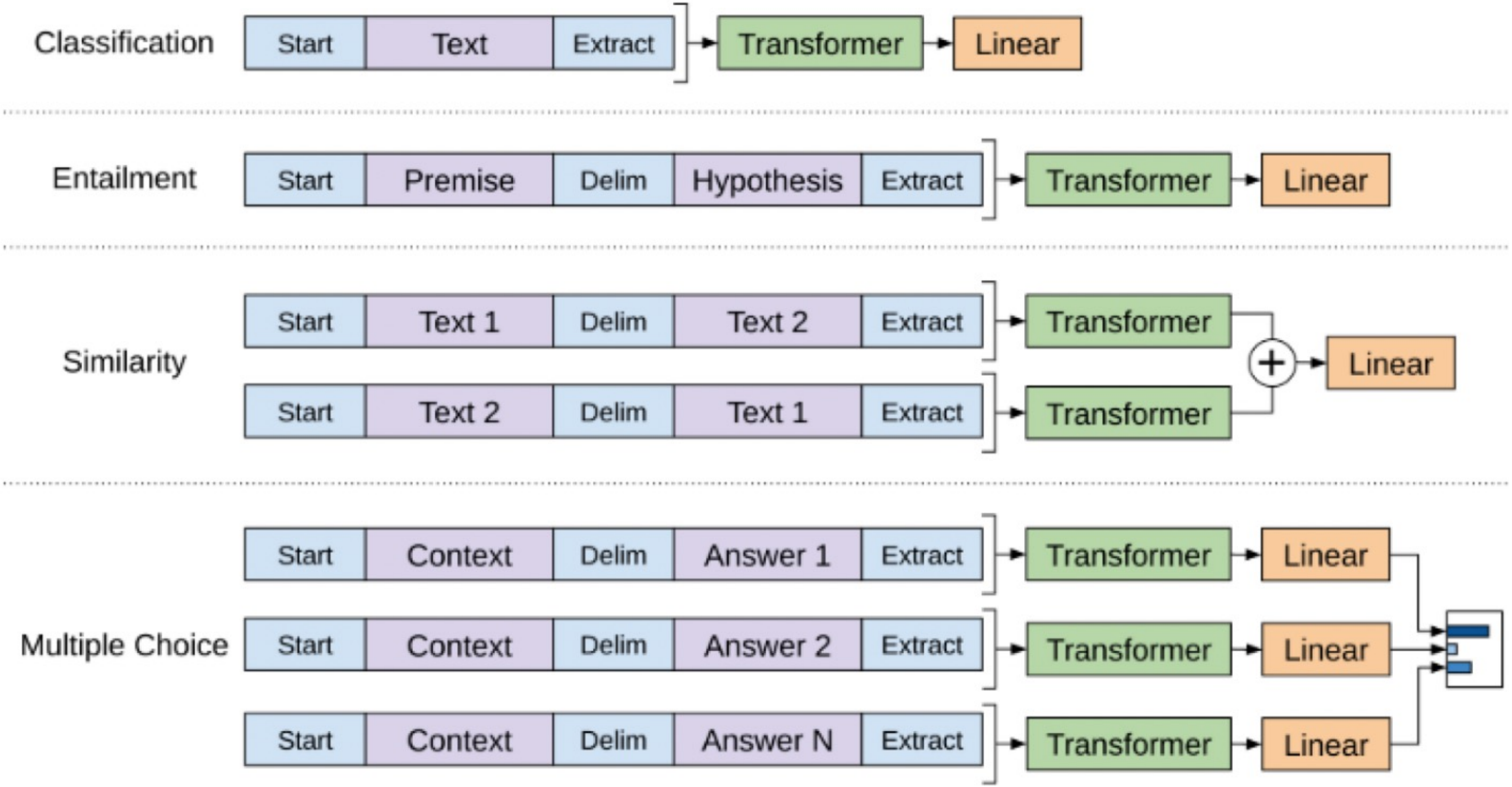
Radford et al. Improving Language Understanding by Generative Pre-Training. Technical Report 2018.

GPT: Generative Pre-Training



https://docs.graphcore.ai/projects/bert-training/en/latest/bert.html

Fine-Tuning to Target Task (Softmax Output); i.e.,



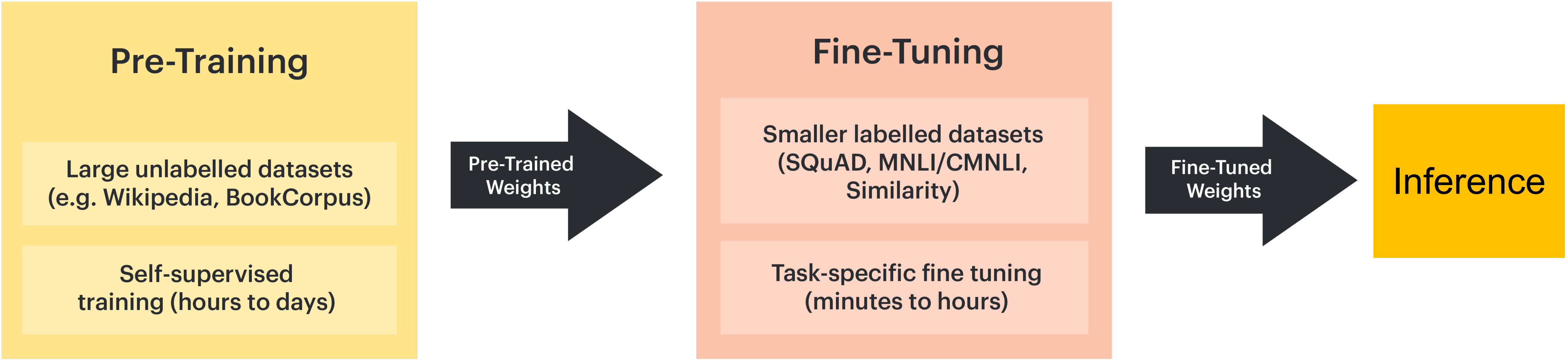
Radford et al. Improving Language Understanding by Generative Pre-Training. Technical Report 2018.

Experimental Findings

Achieved the best performance on 9 NLP dataset challenges

Radford et al. Improving Language Understanding by Generative Pre-Training. Technical Report 2018.

GPT: Generative Pre-Training



https://docs.graphcore.ai/projects/bert-training/en/latest/bert.html

Today's Topics

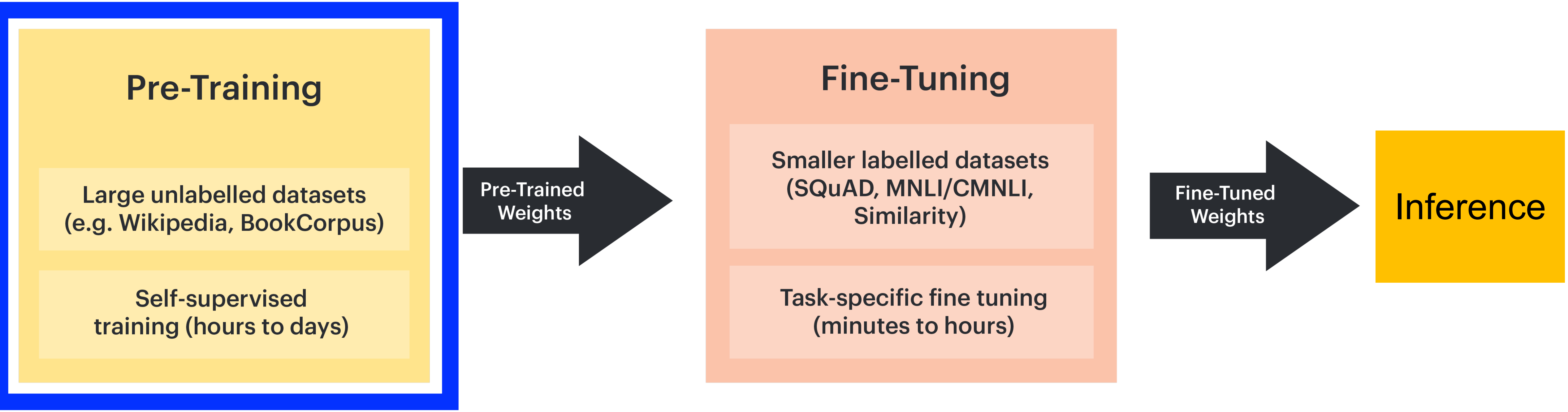
- Explosion of transformers
- GPT
- BERT
- ViT
- Limitations of transformer models

Motivation: Choose a Pretraining Task That Is Not Unidirectional

GPT's prediction of the next word given previous ones is unidirectional (left-to-right)

1. Background music from a ________
2. Many people danced around the ________
3. I practiced for many years to learn how to play the ________

BERT: Bidirectional Encoder Representation from Transformer



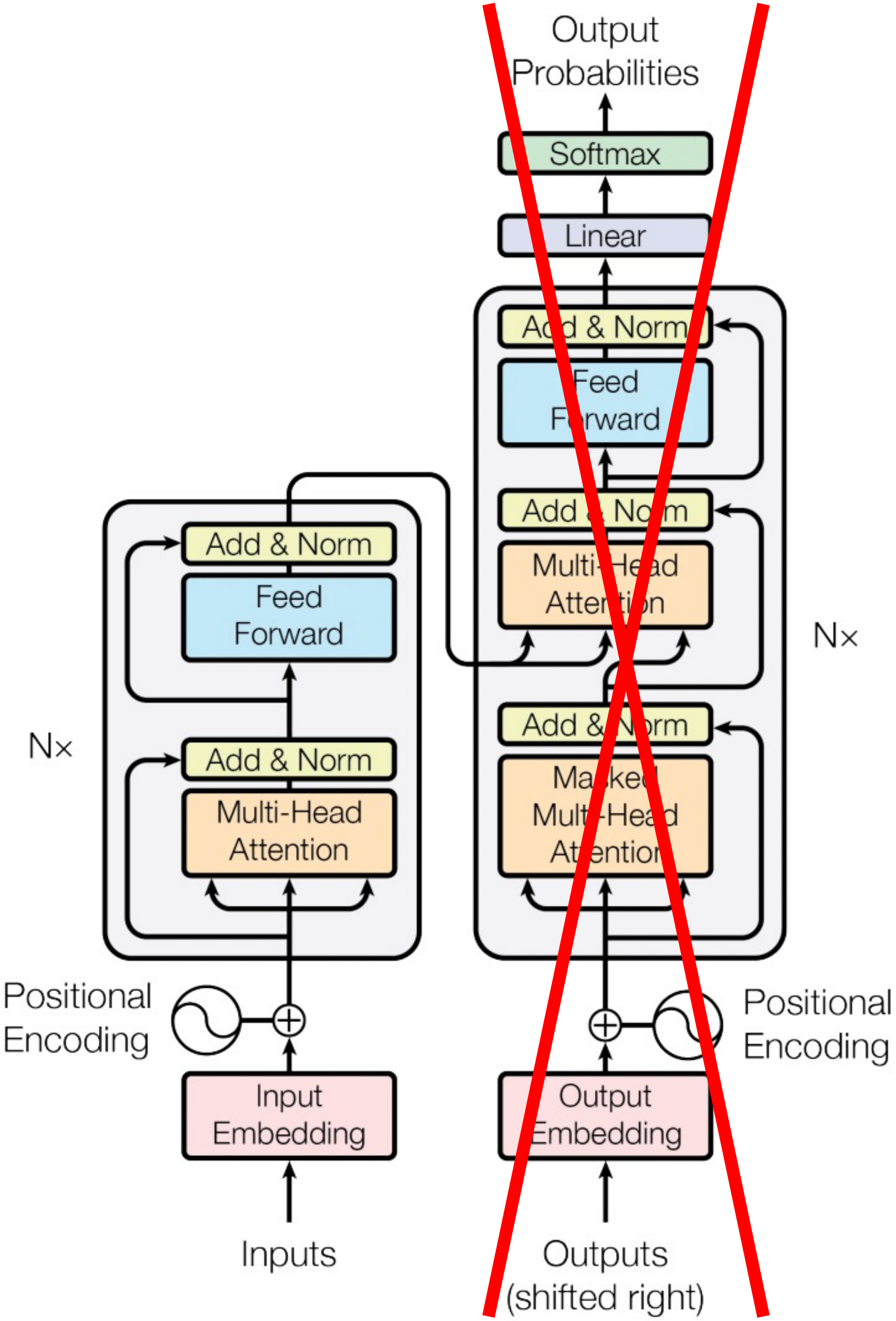
https://docs.graphcore.ai/projects/bert-training/en/latest/bert.html

Two Tasks

1. Predict masked token (key contribution)

2. Predict if one sentence follows a second sentence (augments understanding of how sentences relate)

Architecture: Encoder from Pioneering Transformer



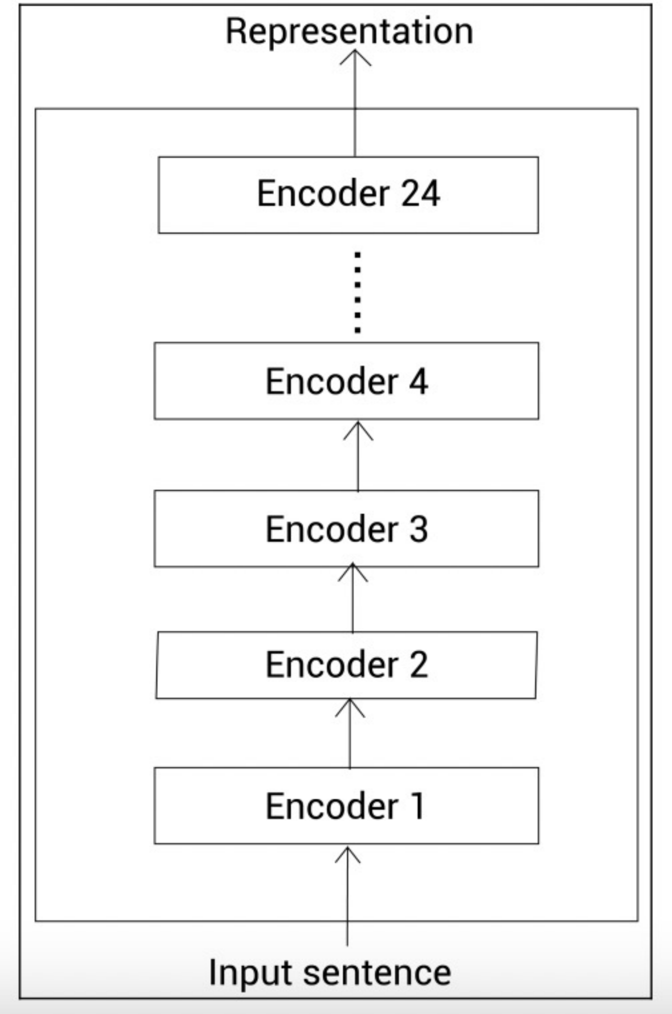
Vaswani et al. Attention Is All You Need. Neurips 2017.

Architecture: Variants

- L = number of stacked encoders
- H = number of hidden units in feedforward layer

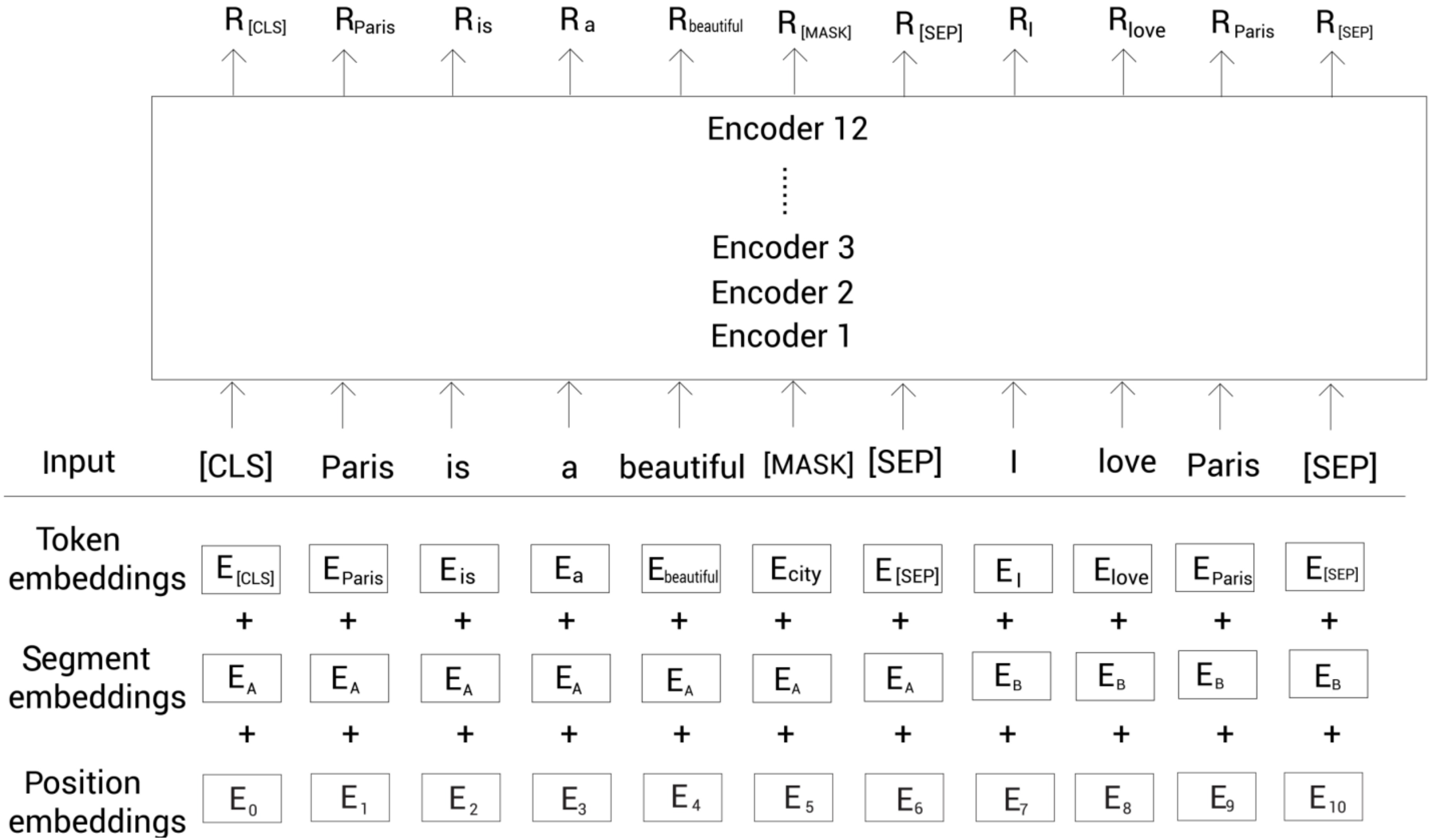
	H=128	H=256	H=512	H=768
L=2	2/128 (BERT-tiny)	2/256	2/512	2/768
L=4	4/128	4/256 (BERT-mini)	4/512 (BERT-small)	4/768
L=6	6/128	6/256	6/512	6/768
L=8	8/128	8/256	8/512 (BERT-medium)	8/768
L=10	10/128	10/256	10/512	10/768
L=12	12/128	12/256	12/512	12/768 (BERT-base)

BERT-large (H = 1024)



https://static.packt-cdn.com/downloads/9781838821593_ColorImages.pdf

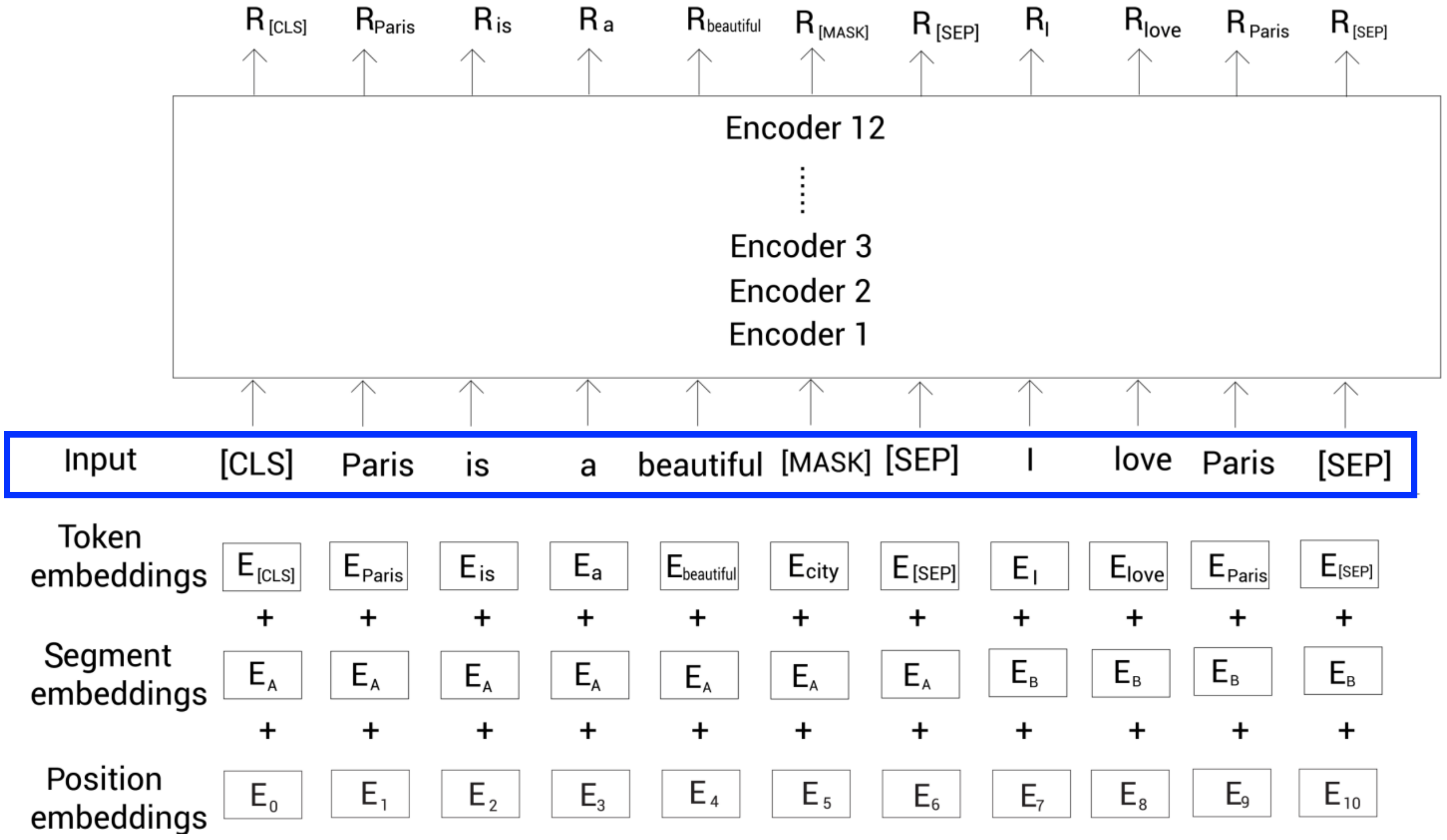
Architecture: BERT-Base (Matches Size of GPT)



https://static.packt-cdn.com/downloads/9781838821593_ColorImages.pdf

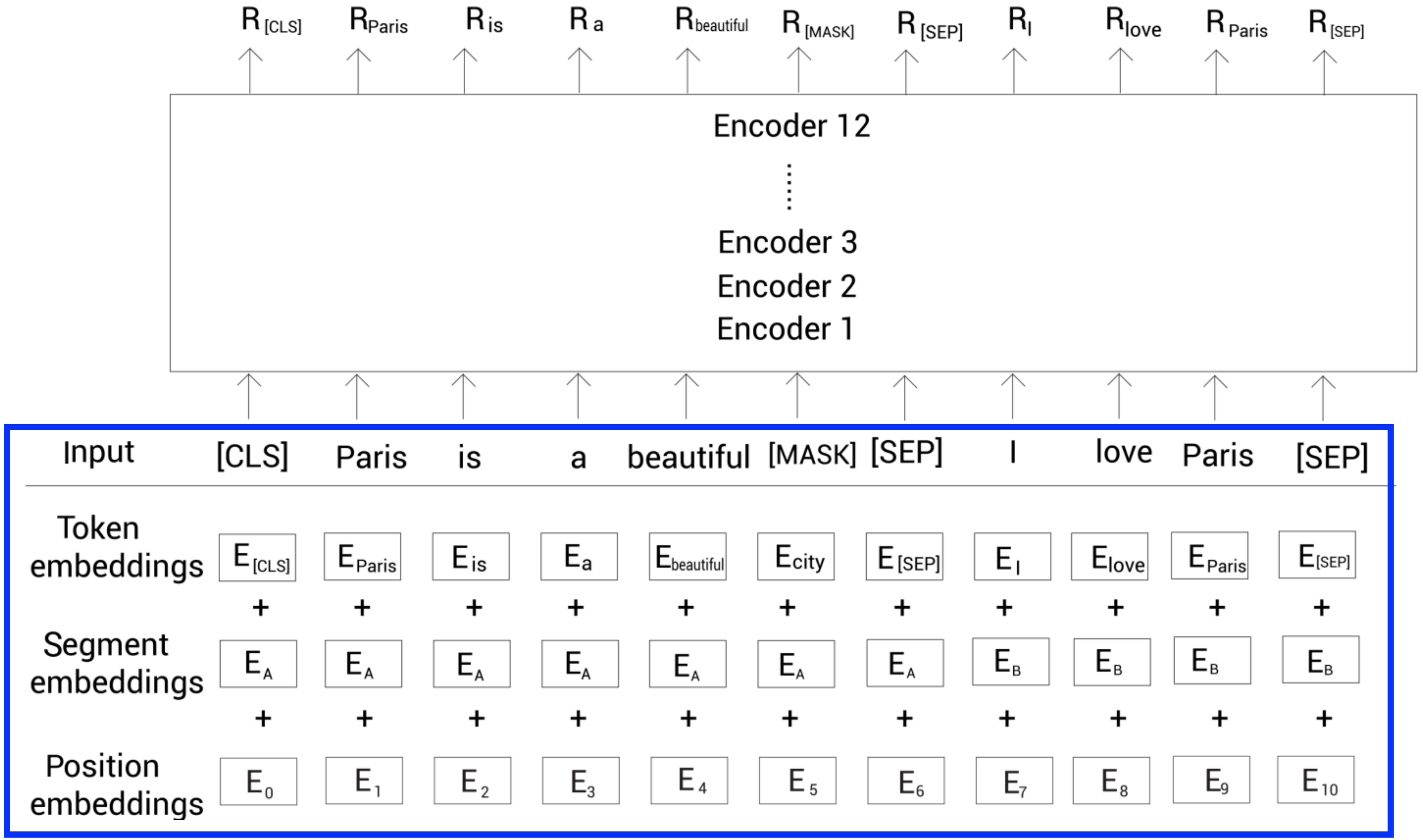
Architecture: Input

Two input sentences with [CLS] at the start and [SEP] between sentences



https://static.packt-cdn.com/downloads/9781838821593_ColorImages.pdf

Architecture: Input

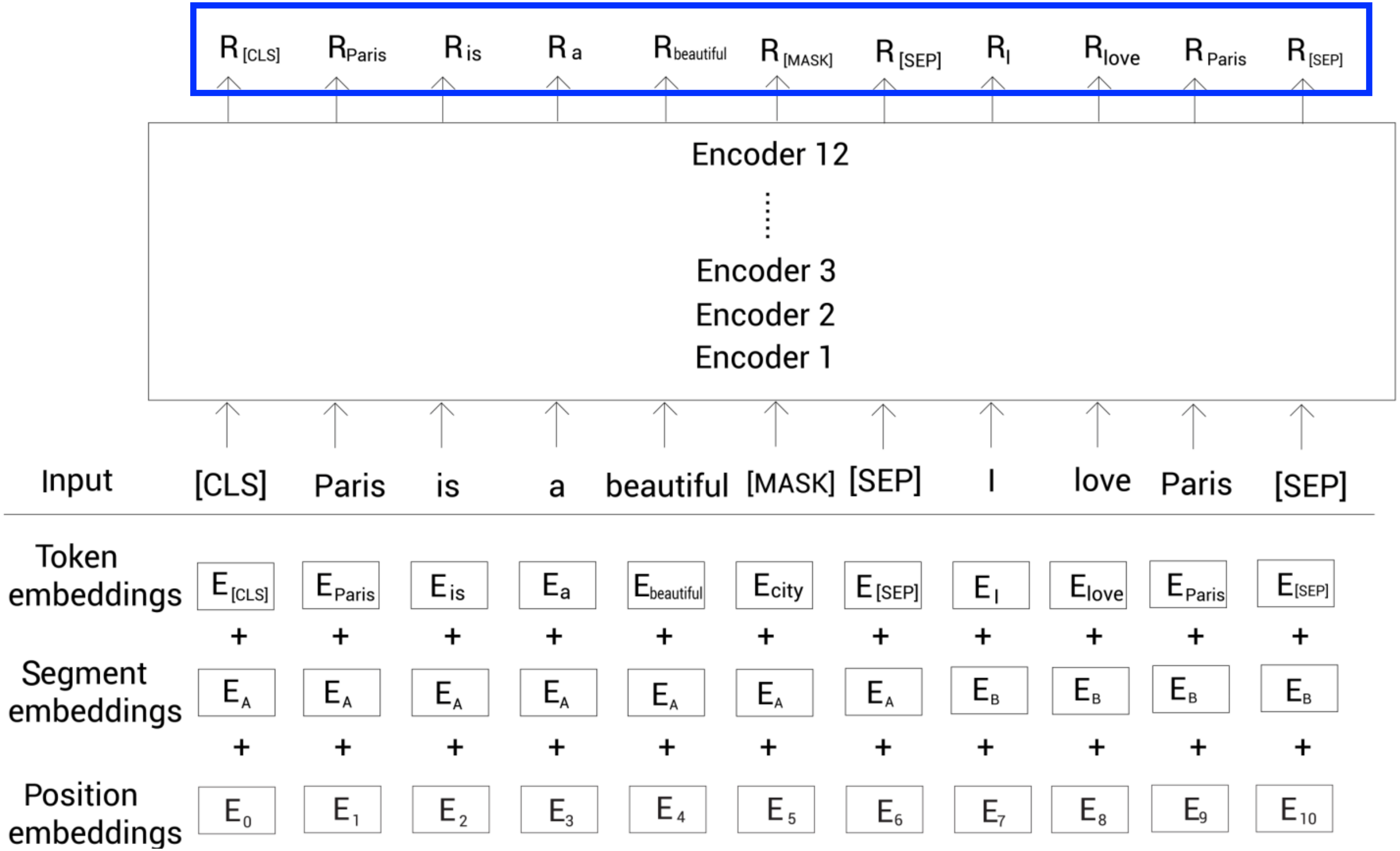


Input is addition of a segment embedding to the token and position embeddings (helps differentiate which tokens belong to which sentence)

https://static.packt-cdn.com/downloads/9781838821593_ColorImages.pdf

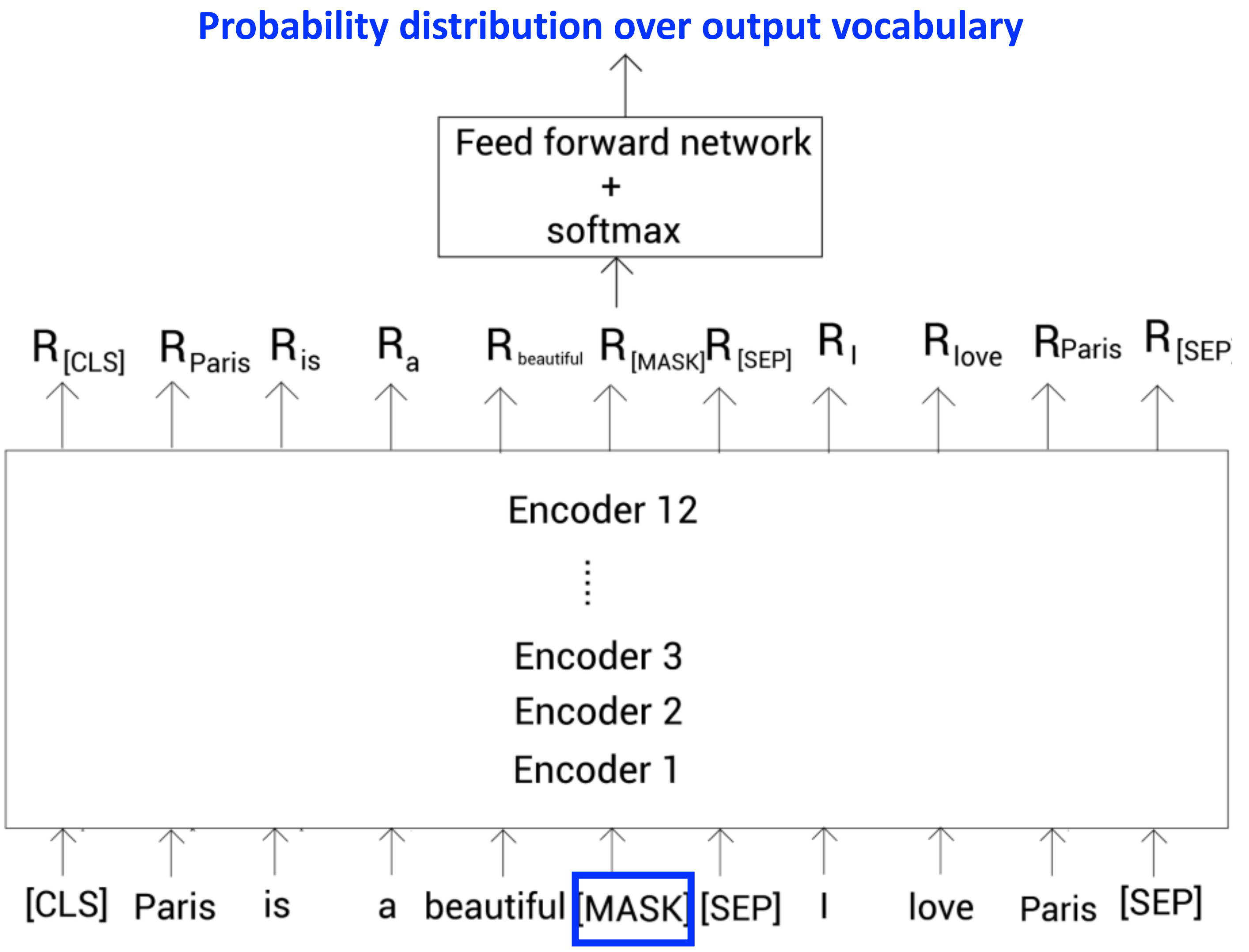
Architecture: Output

New representation of each input that accounts for context



https://static.packt-cdn.com/downloads/9781838821593_ColorImages.pdf

Architecture: Predicting Masked Token Task



https://static.packt-cdn.com/downloads/9781838821593_ColorImages.pdf

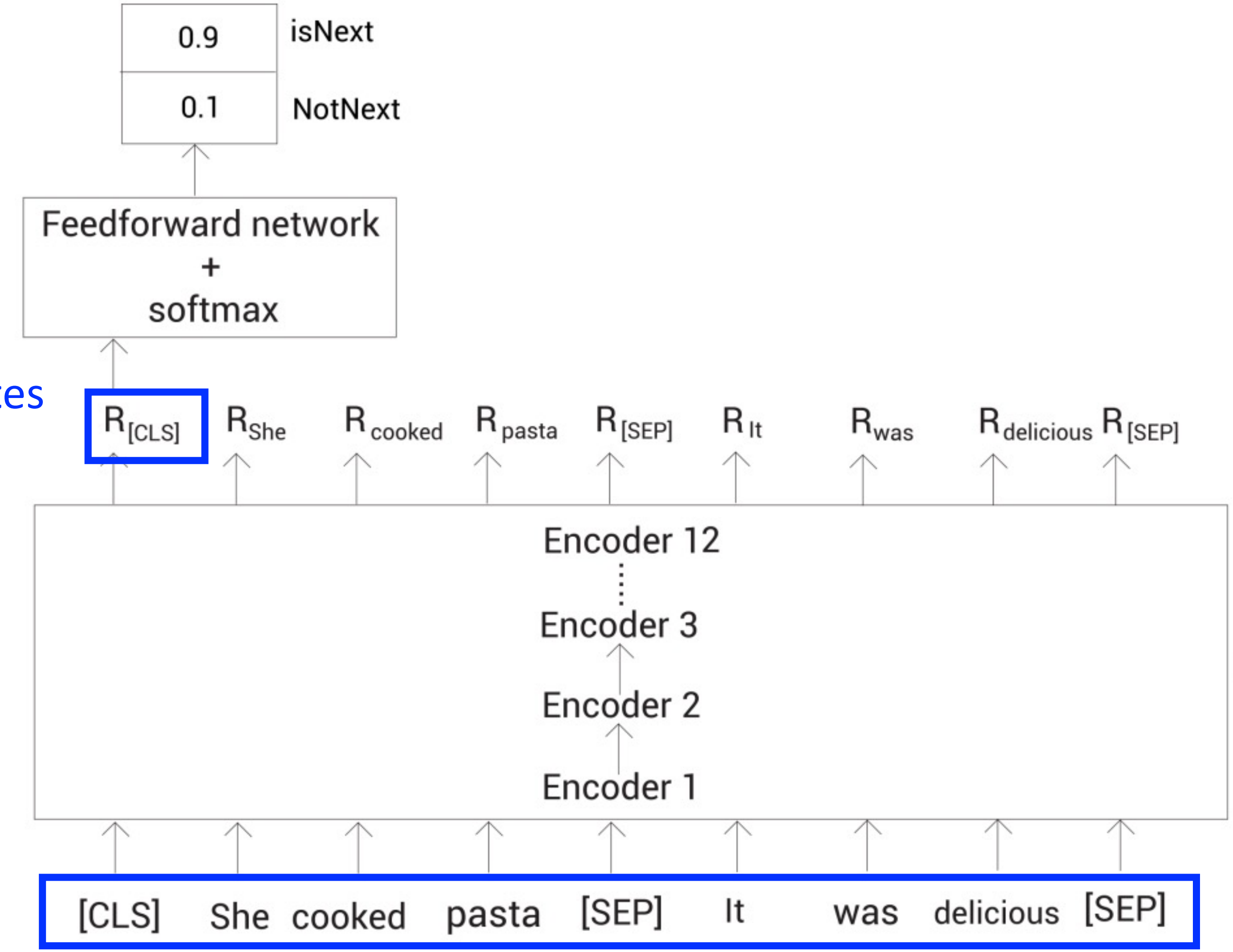
15% of random tokens from sequence masked

- 80% use [MASK]
- 10% use a random token
- 10% are unchanged

Multiple masking options encourage the model to pay attention to each token separately

Architecture: Predict if Next Sentence Task

Predict with token representation that aggregates representation of all tokens in both sentences



50% of 2nd sentences are the original next sentence and the rest are random

https://static.packt-cdn.com/downloads/9781838821593_ColorImages.pdf

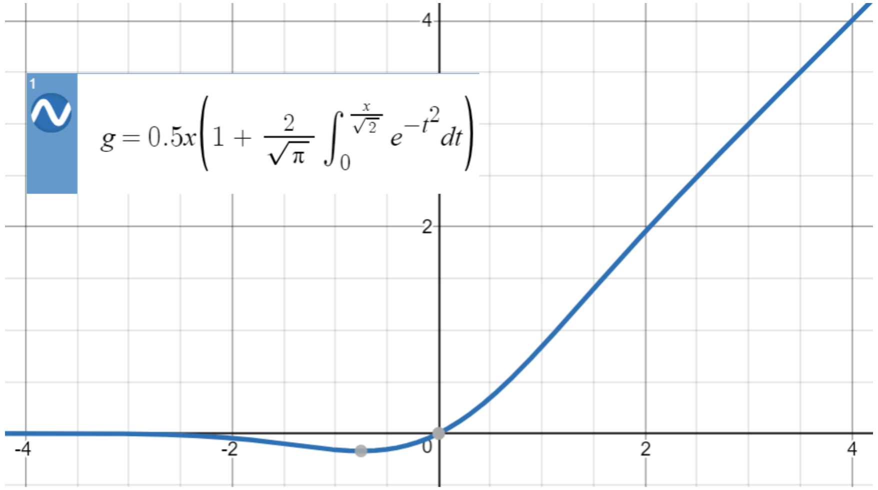
Training

- Dataset: 2,500M words in Wikipedia + 800M words in BooksCorpus used for GPT
- Optimizer: Adam
- Training loss: sums over losses from predicting masked words and if next sentence

Devlin et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv 2018.

Implementation Details: Mimics GPT

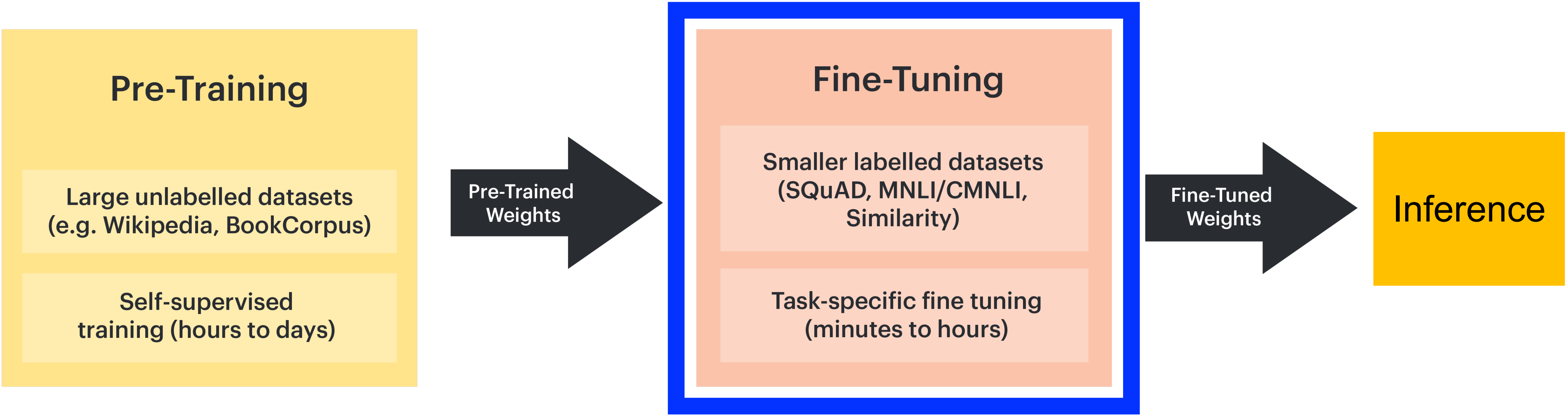
- Gaussian error linear unit (GELU) used as activation function in feedforward layers



https://datascience.stackexchange.com/questions/49522/what-is-gelu-activation

- Avoids out of vocabulary tokens by using subword tokenization, with a different variant called WordPiece Tokenization

BERT: Bidirectional Encoder Representation from Transformer



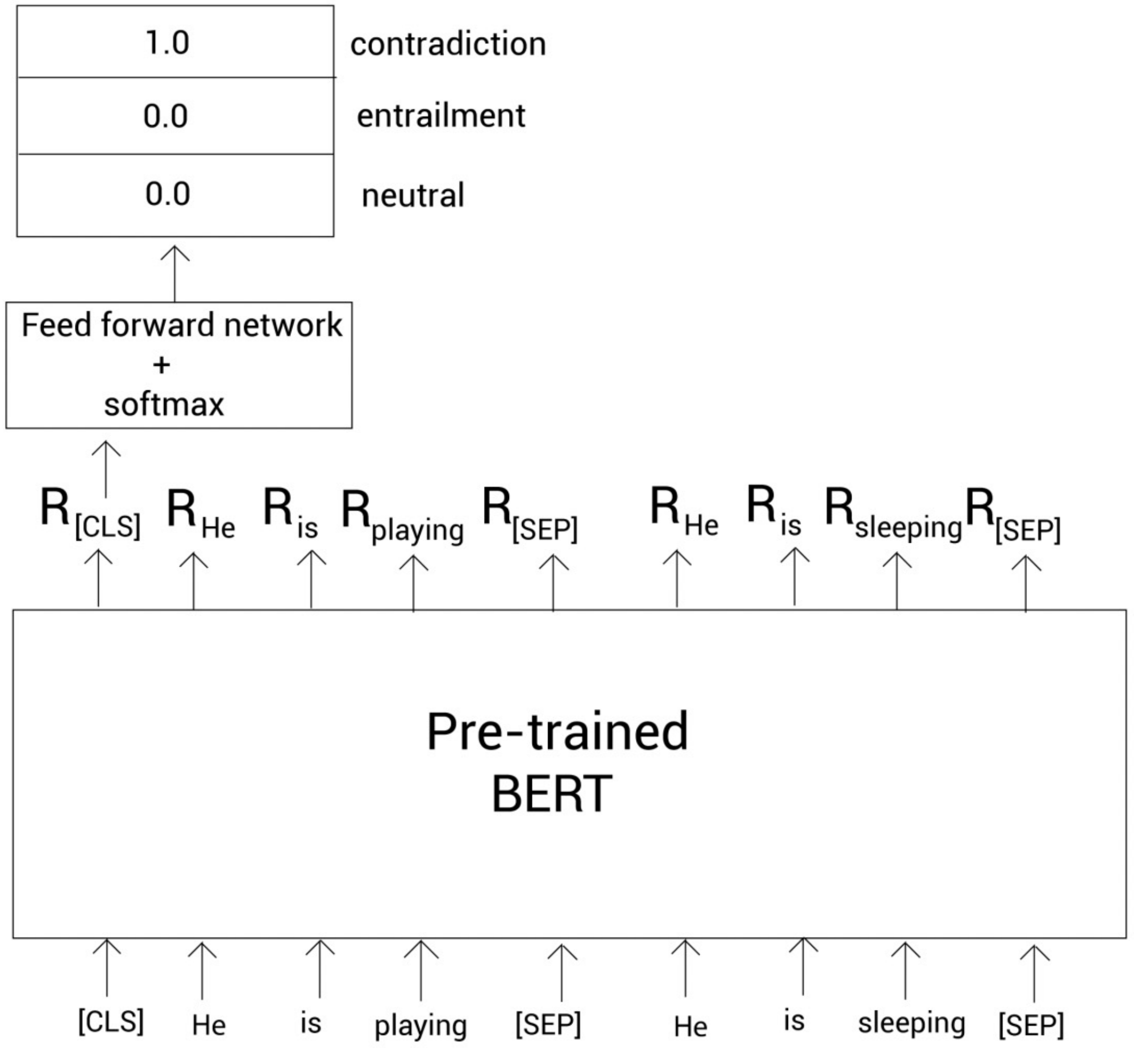
https://docs.graphcore.ai/projects/bert-training/en/latest/bert.html

Fine-Tuning for Natural Language Inference

Premise	Hypothesis	Label
He is playing	He is sleeping	Contradiction
A soccer game with multiple males playing	Some men are playing sport	Entailment
An older and a younger man smiling	Two men are smiling at the dogs playing on the floor	Neutral

https://static.packt-cdn.com/downloads/9781838821593_ColorImages.pdf

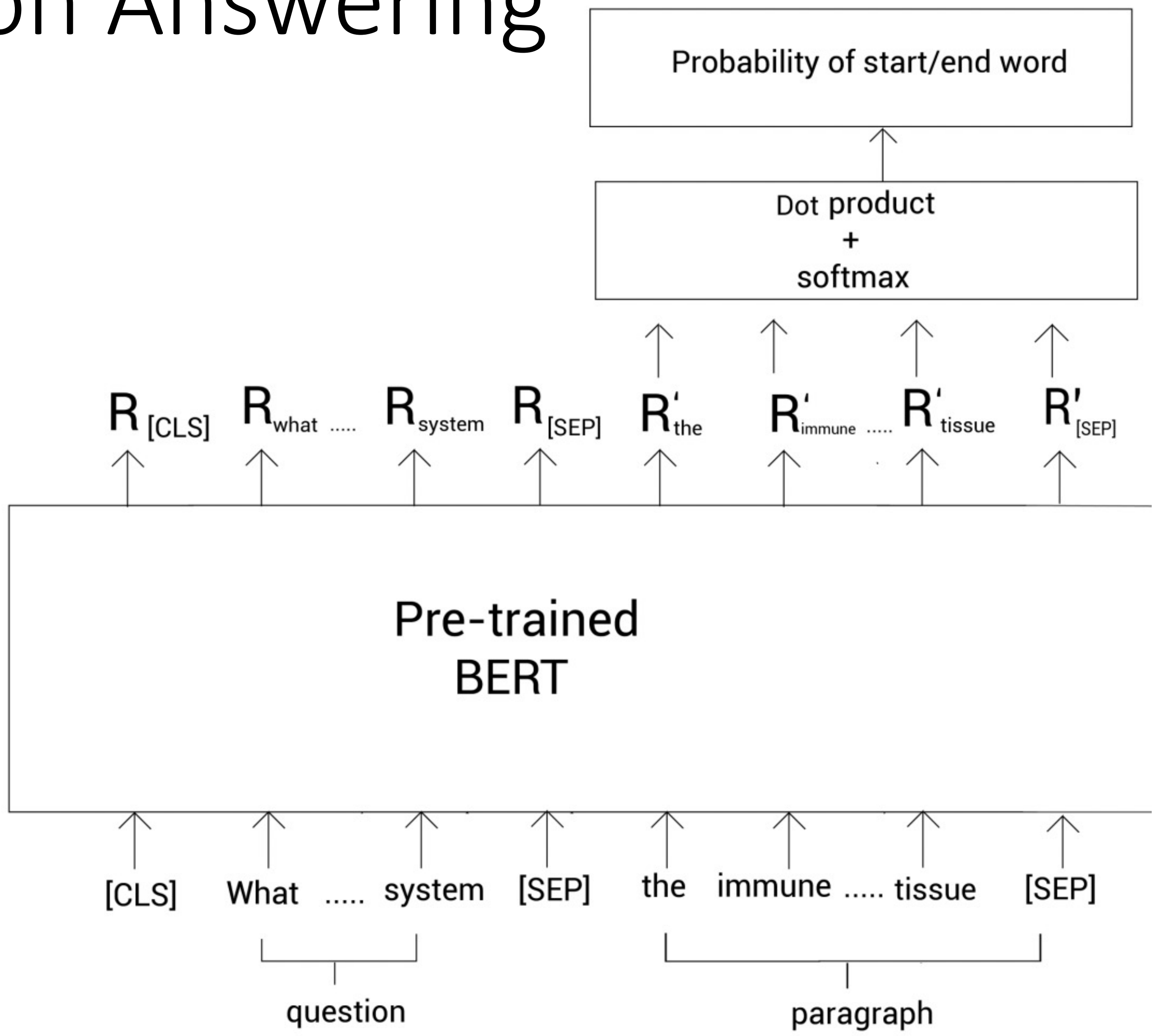
Fine-Tuning for Natural Language Inference



https://static.packt-cdn.com/downloads/9781838821593_ColorImages.pdf

Fine-Tuning for Question Answering

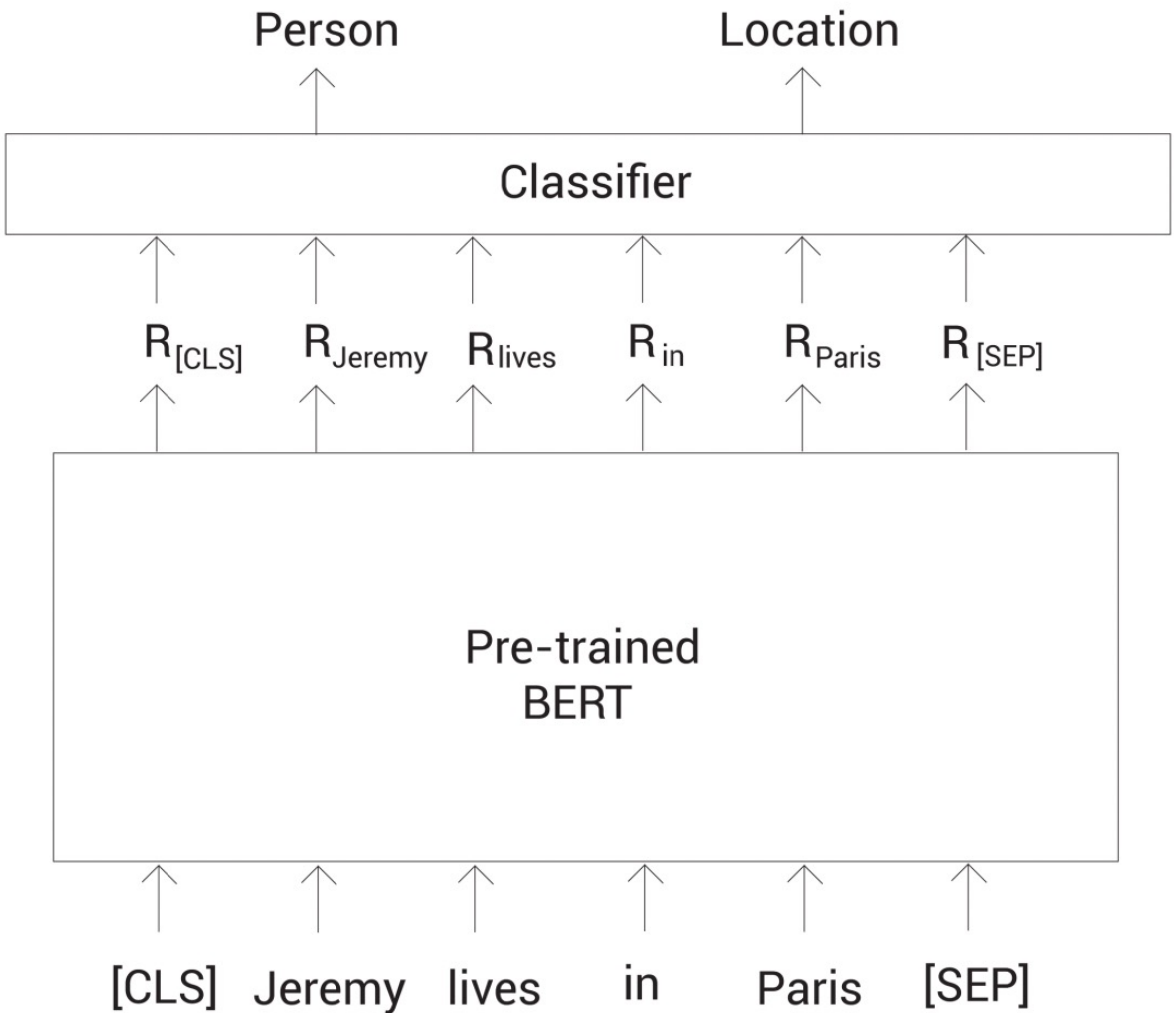
To find indices of the start and end words in the paragraph, two vector representations are learned that lead to the approximate softmax output when computing the dot product with each token



https://static.packt-cdn.com/downloads/9781838821593_ColorImages.pdf

Fine-Tuning for Named Entity Recognition

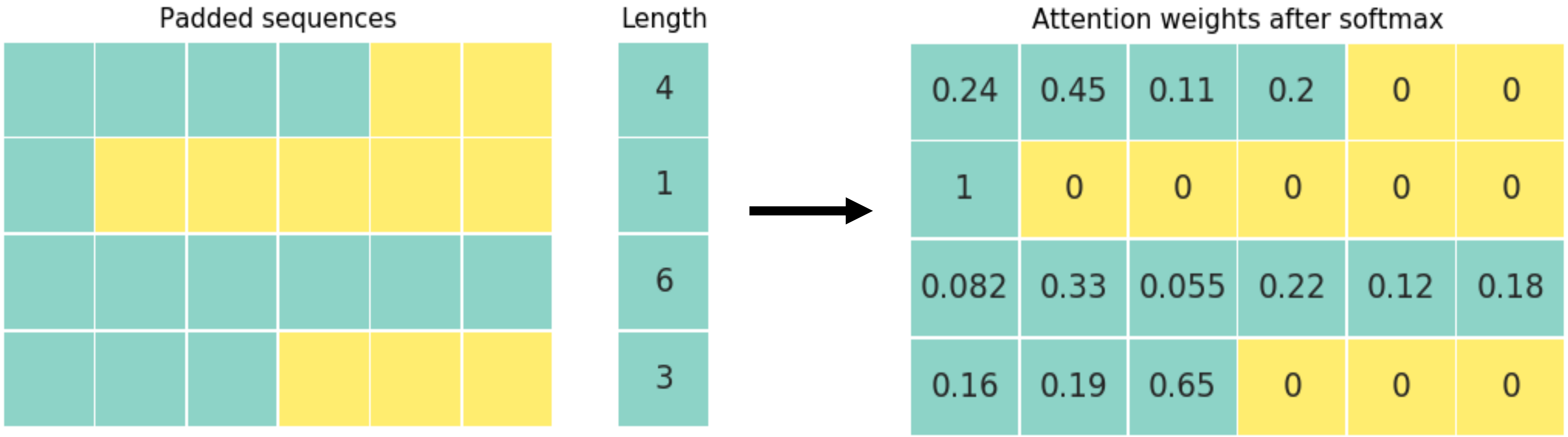
Each token's new representation is passed to a classifier



https://static.packt-cdn.com/downloads/9781838821593_ColorImages.pdf

Implementation Detail

- Padding supports the use of variable input length
  - Uses attention vector of 1s and 0s, with the latter at indices of [PAD] tokens



http://juditacs.github.io/2018/12/27/masked-attention.html

Experimental Findings

Achieved the best performance on 11 NLP dataset challenges

Devlin et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv 2018.

Experimental Findings

Tasks	MNLI-m (Acc)	QNLI (Acc)	MRPC (Acc)	SST-2 (Acc)	SQuAD (F1)
	Dev Set				
$\text{BERT}_{\text{BASE}}$	84.4	88.4	86.7	92.7	88.5
No NSP	83.9	84.9	86.5	92.6	87.9
LTR & No NSP	82.1	84.3	77.5	92.1	77.8
+ BiLSTM	82.1	84.1	75.7	91.6	84.9

Next sentence prediction (NSP) supports slight improvements

Devlin et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv 2018.

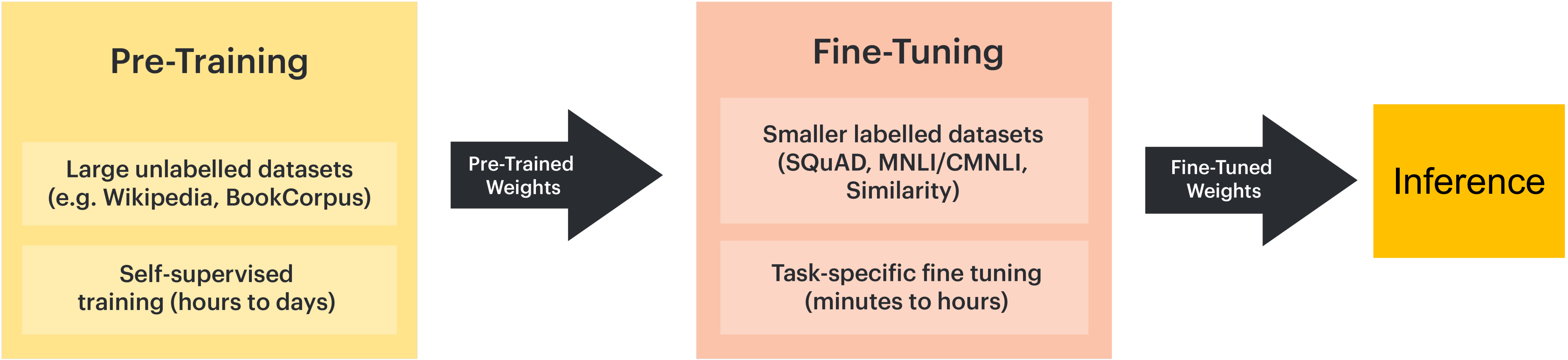
Experimental Findings

Tasks	MNLI-m (Acc)	QNLI (Acc)	MRPC (Acc)	SST-2 (Acc)	SQuAD (F1)
	Dev Set				
BERT$_{BASE}$	84.4	88.4	86.7	92.7	88.5
No NSP	83.9	84.9	86.5	92.6	87.9
LTR & No NSP	82.1	84.3	77.5	92.1	77.8
+ BiLSTM	82.1	84.1	75.7	91.6	84.9

We observe a performance boost when using bidirectional pretraining instead of unidirectional pretraining (LTR)

Devlin et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv 2018.

BERT: Bidirectional Encoder Representation from Transformer

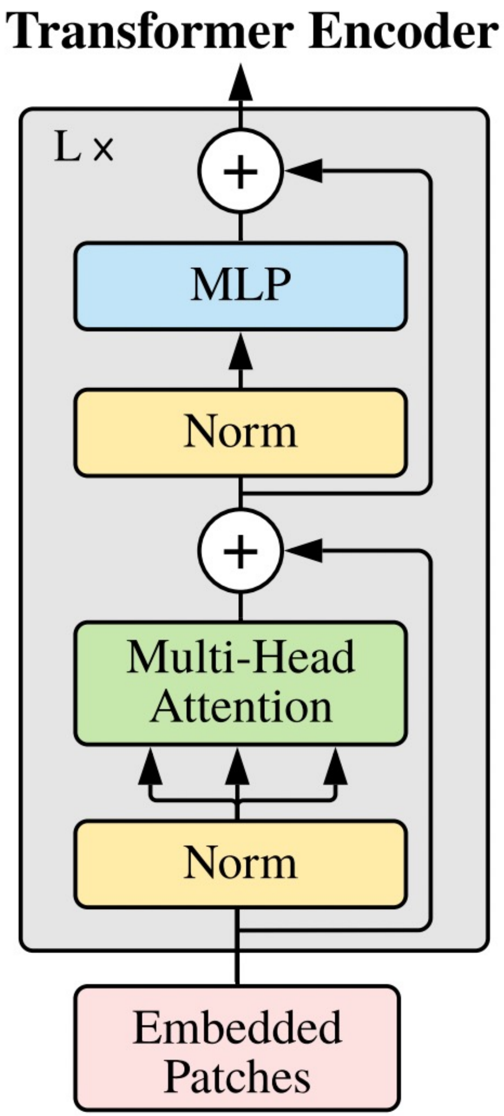


https://docs.graphcore.ai/projects/bert-training/en/latest/bert.html

Today’s Topics

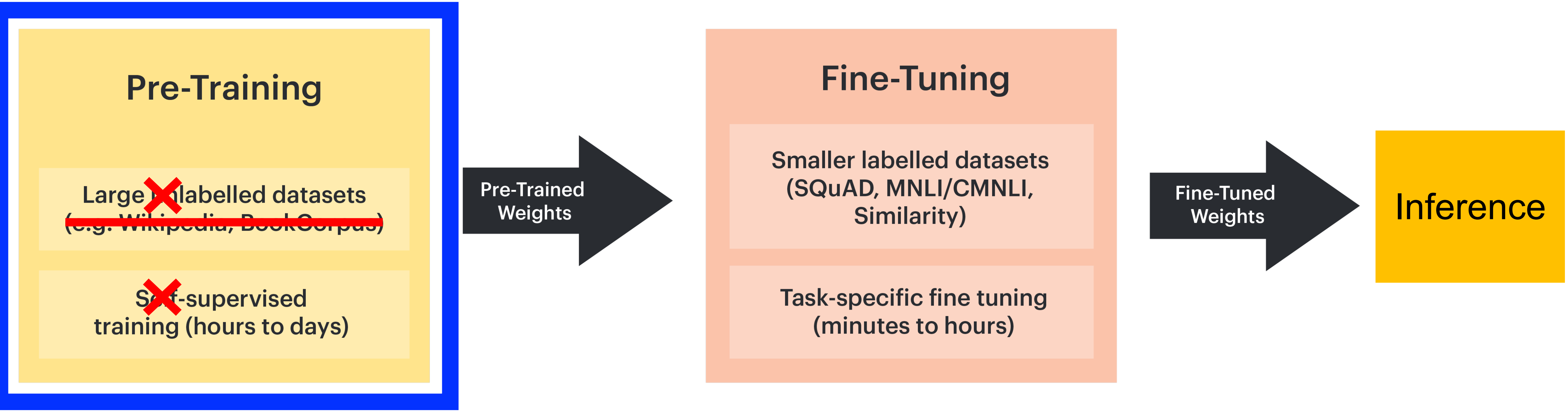
- Explosion of transformers
- GPT
- BERT
- ViT
- Limitations of transformer models

Motivation: Transformers for Image Classification (Repurpose BERT)



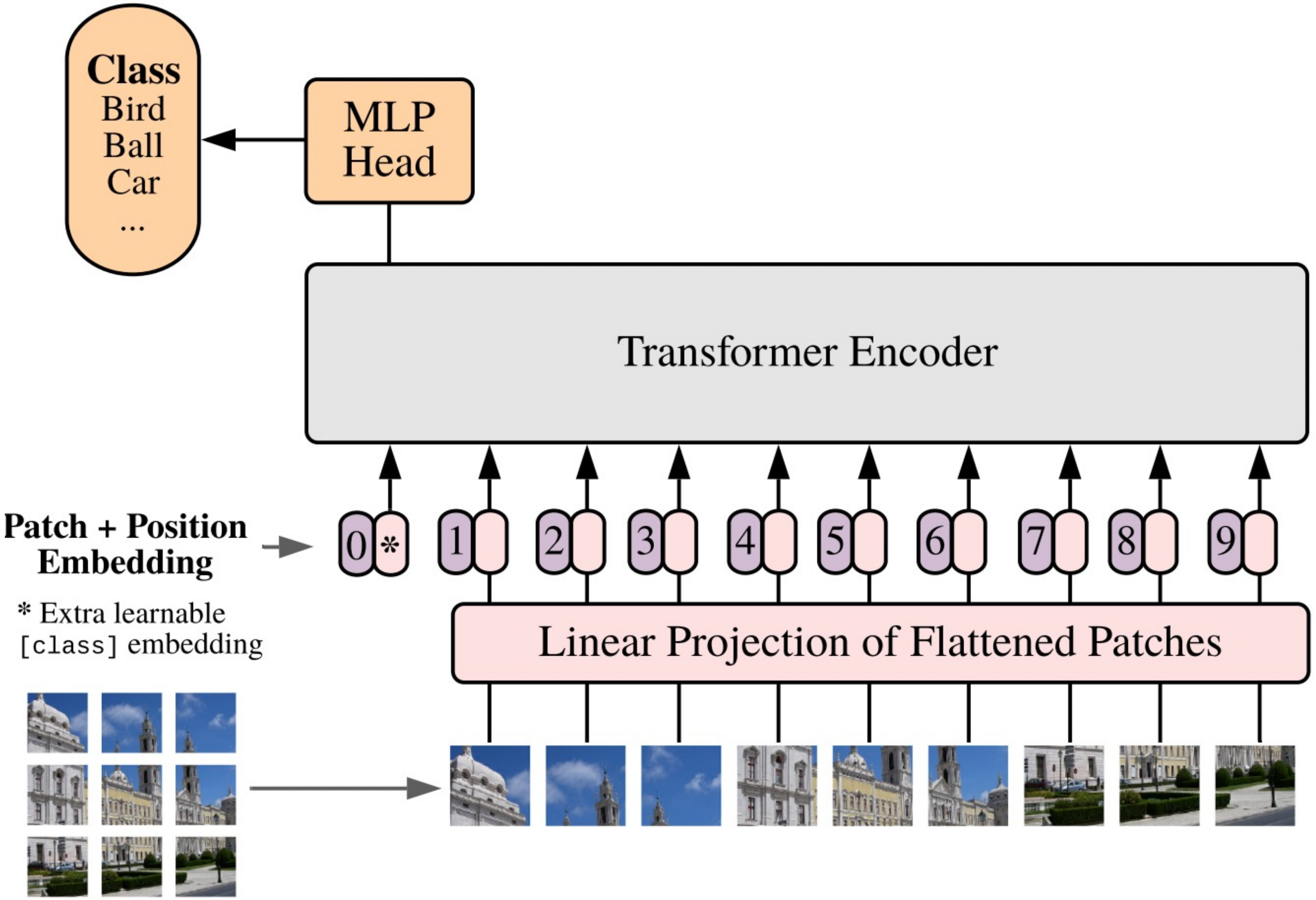
Dosovitskiy et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR 2021.

ViT: Vision Transformer



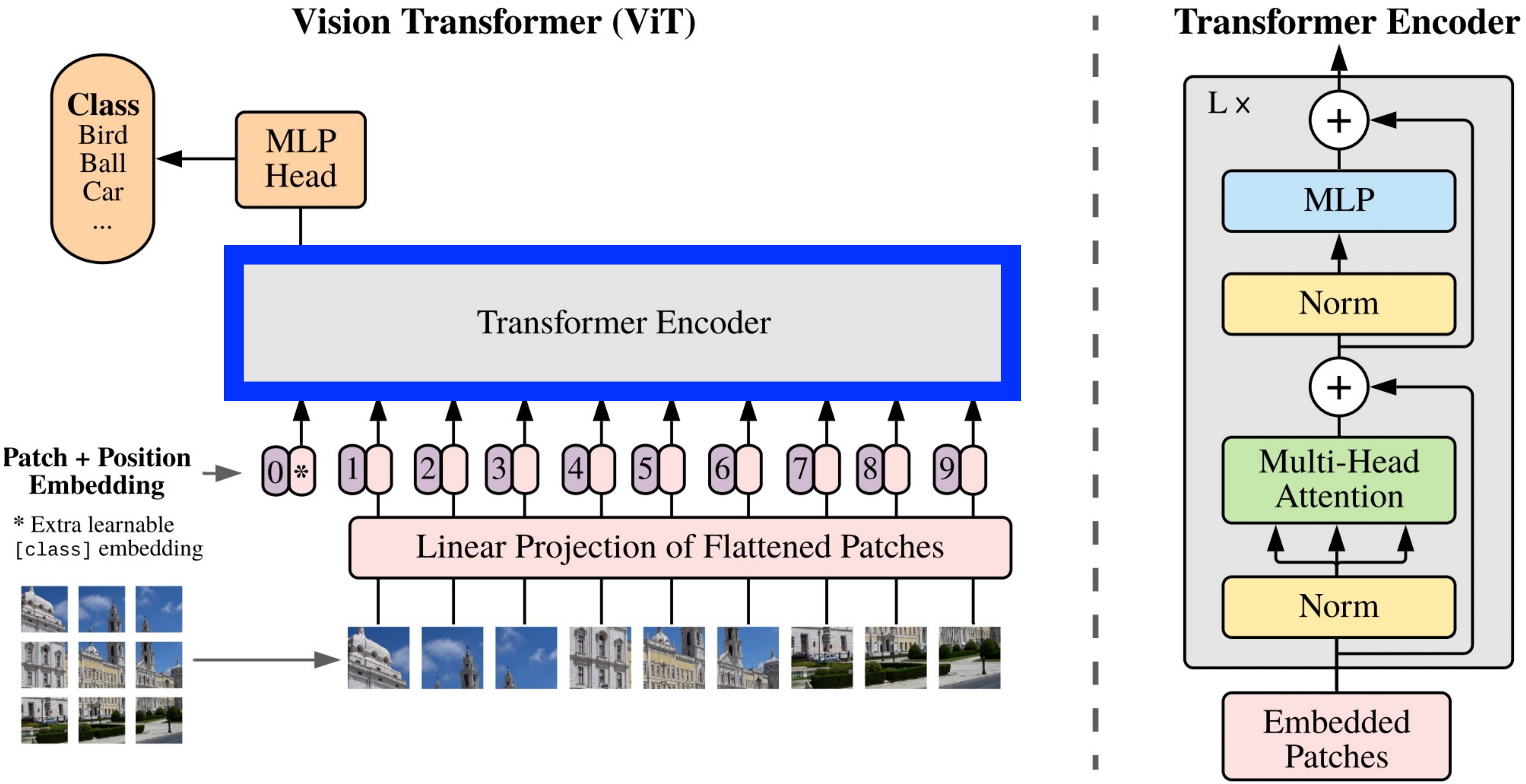
https://docs.graphcore.ai/projects/bert-training/en/latest/bert.html

Architecture



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

Architecture: BERT



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

Architecture: Input

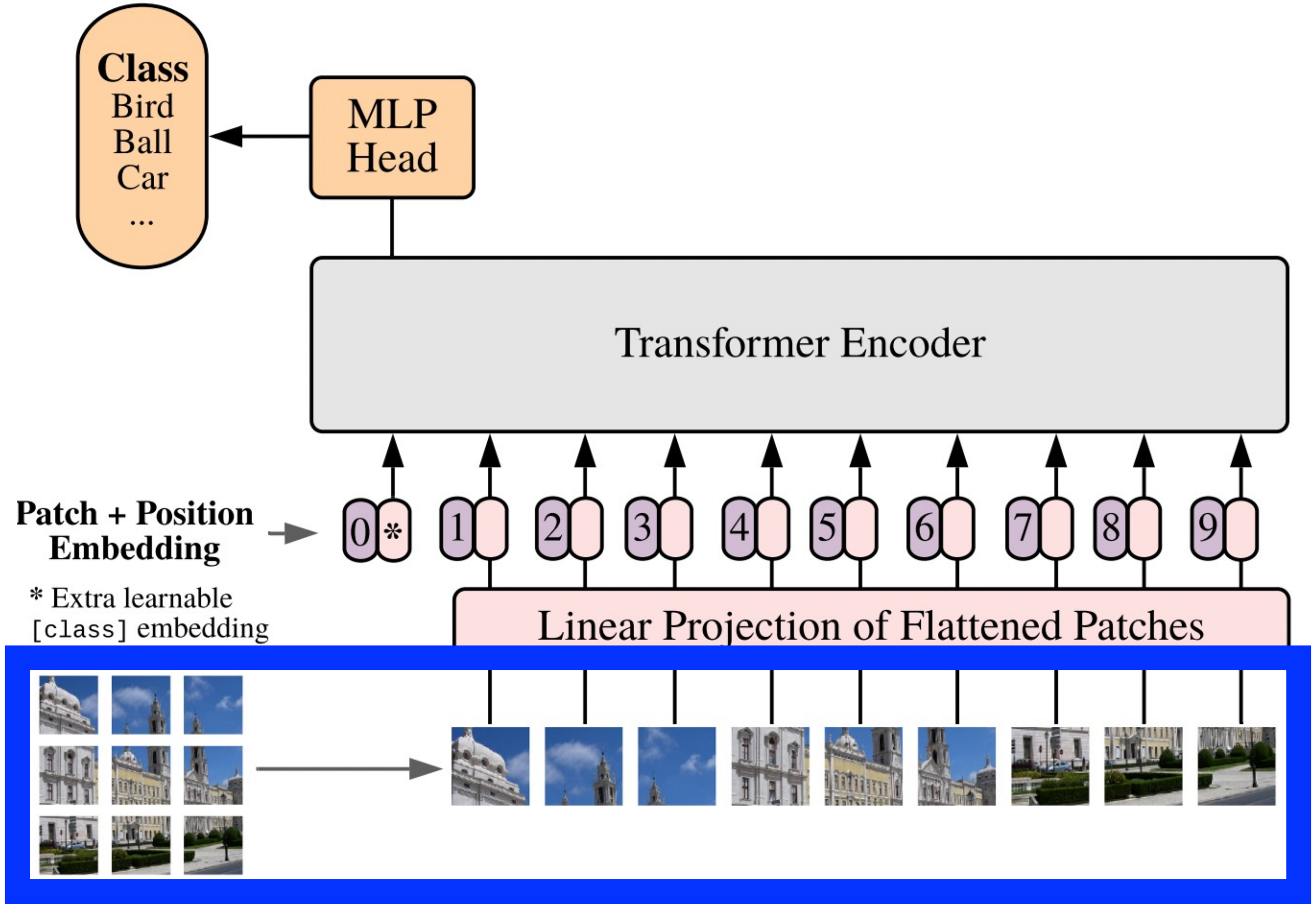
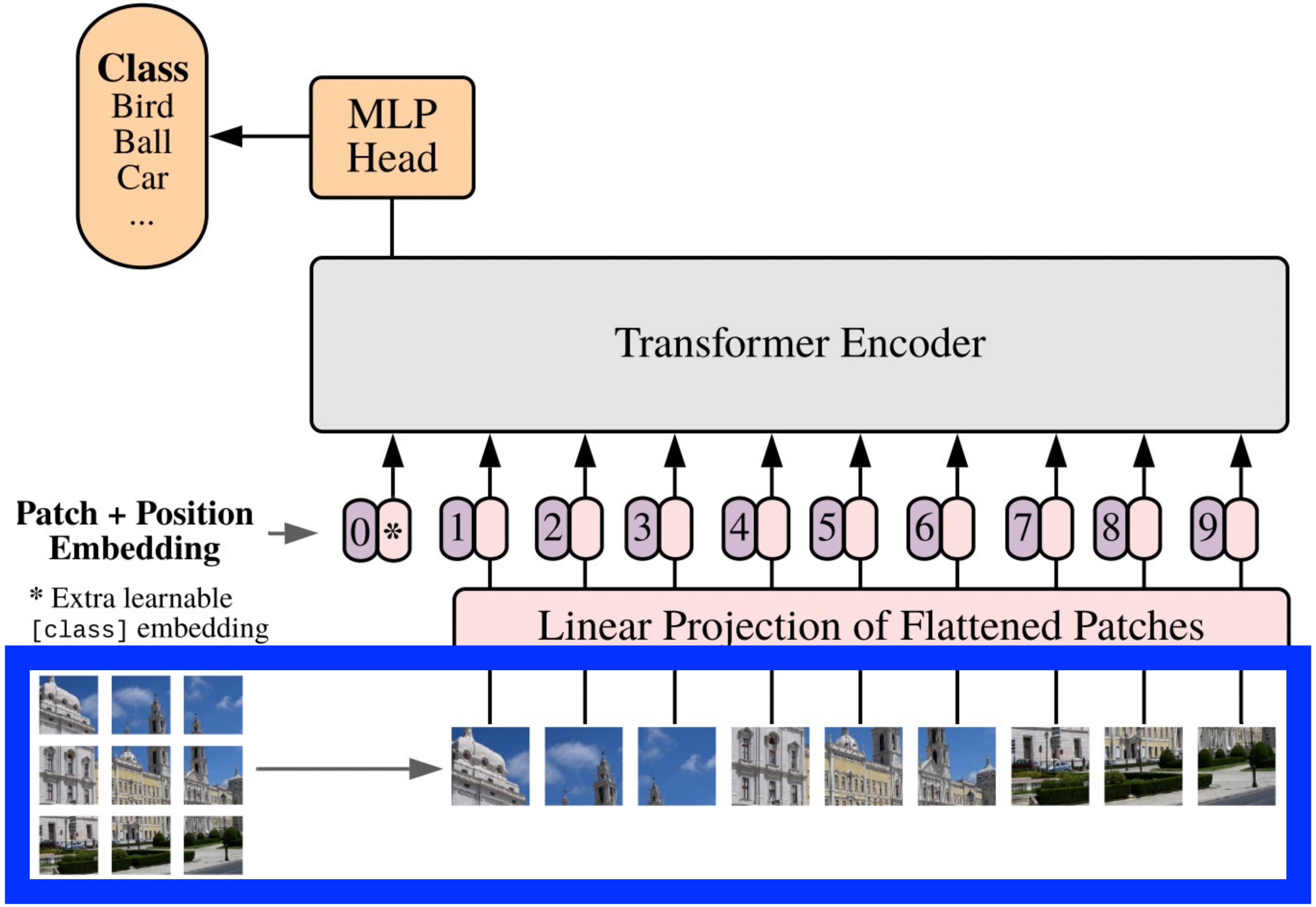


image decomposed into 16x16 patches (example simplified); representations include "flattened" and ResNet features

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

Architecture: Input

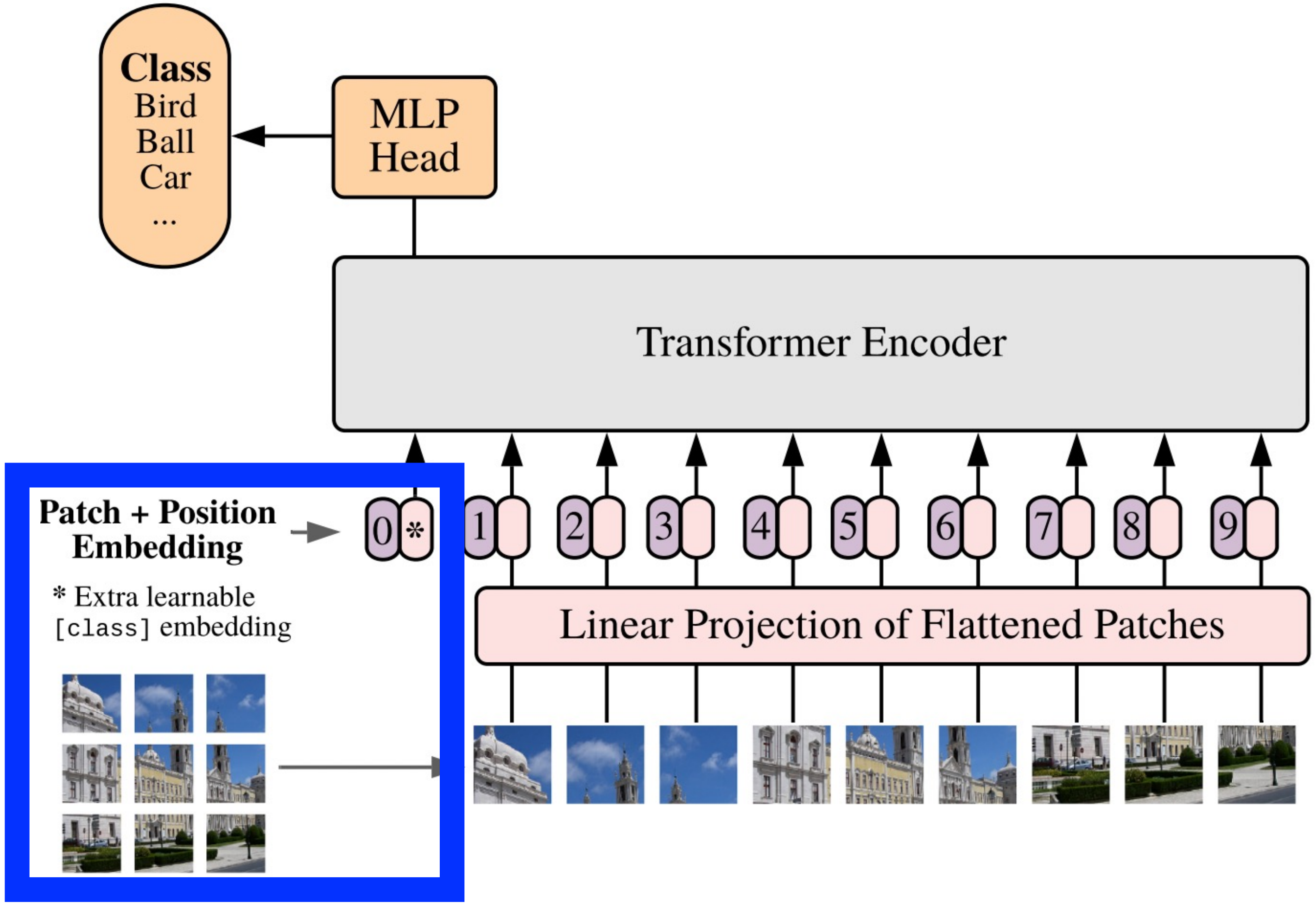


Why not use the raw pixels as input?

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

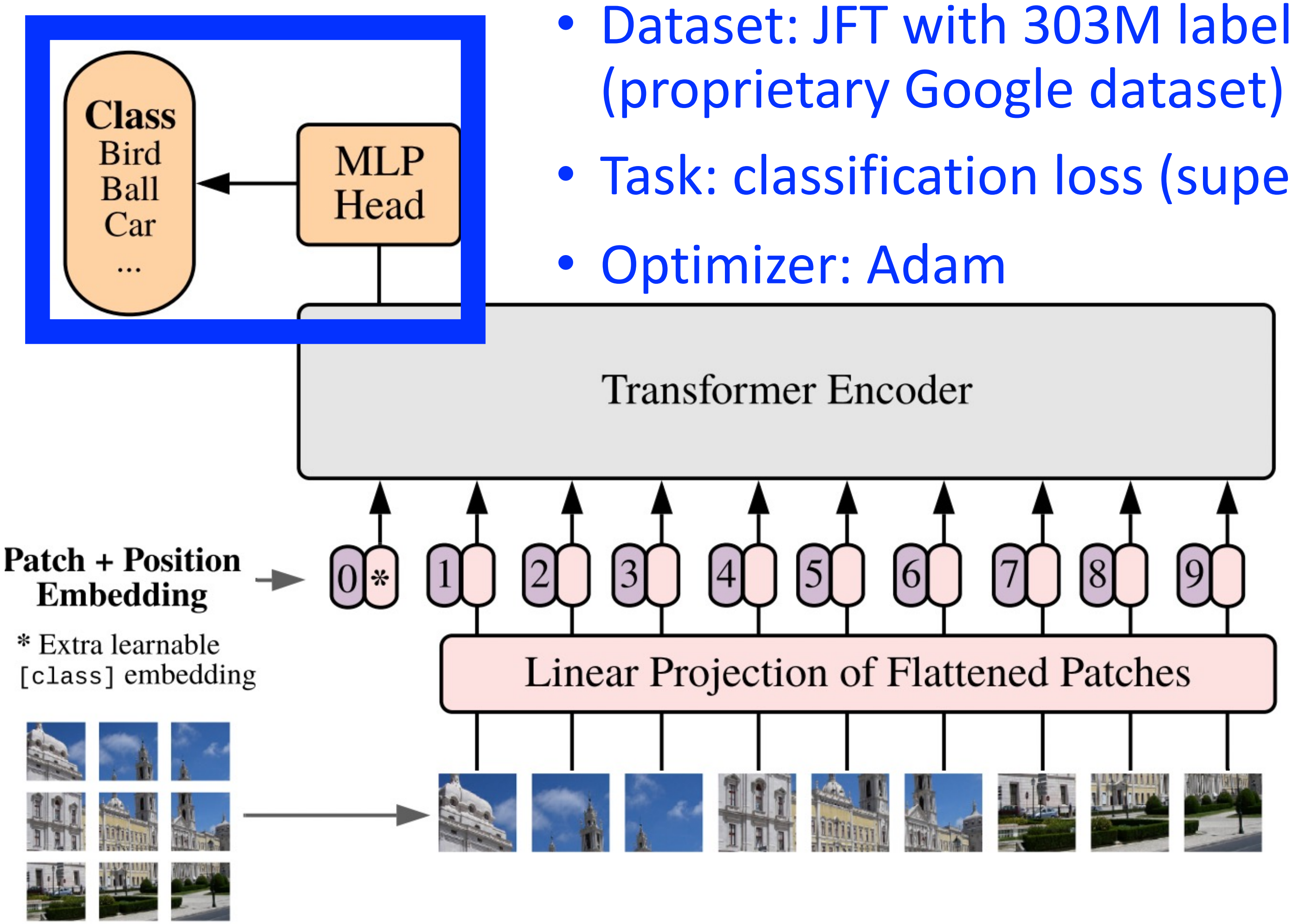
Architecture: Input

[CLS] token prepended to input and positional embeddings added to patch embeddings



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

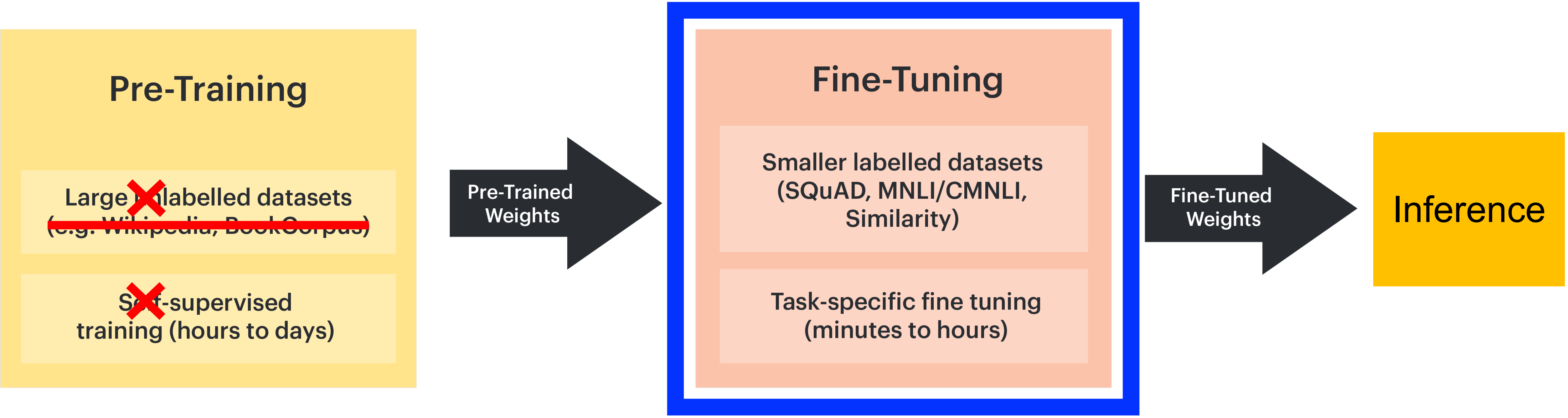
Architecture: Training



- Dataset: JFT with 303M labeled images (proprietary Google dataset)
- Task: classification loss (supervised)
- Optimizer: Adam

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

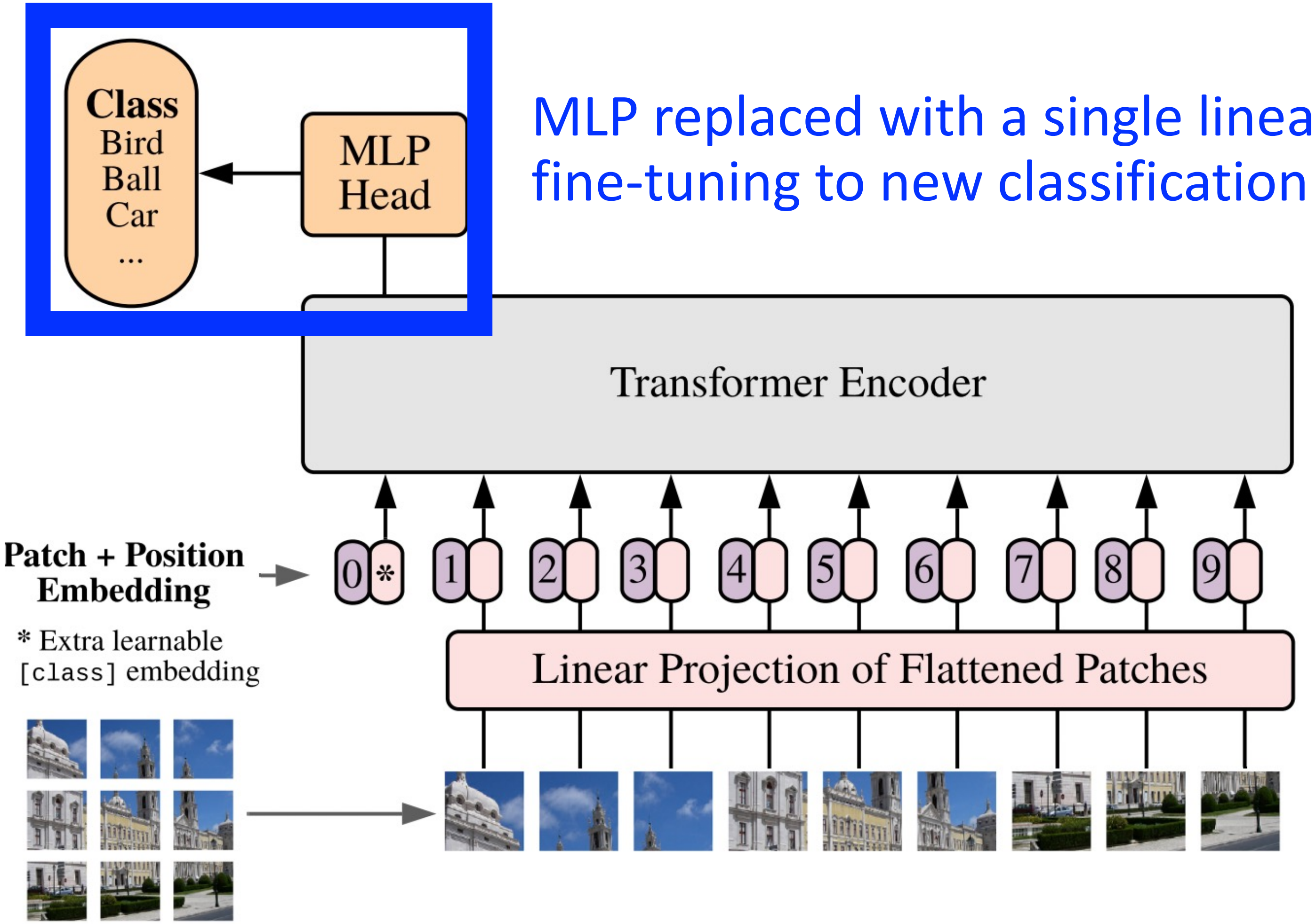
ViT: Vision Transformer



https://docs.graphcore.ai/projects/bert-training/en/latest/bert.html

Fine-Tuning for Other Image Classification Tasks

MLP replaced with a single linear layer when fine-tuning to new classification categories

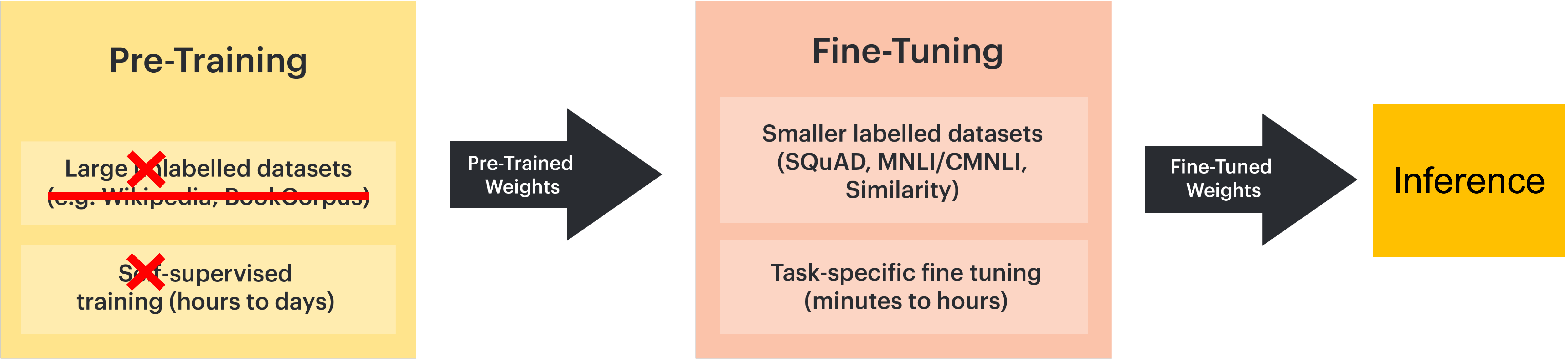


Experimental Findings

Achieved strong results on five image classification datasets

Devlin et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv 2018.

ViT: Vision Transformer



https://docs.graphcore.ai/projects/bert-training/en/latest/bert.html

Today’s Topics

- Explosion of transformers
- GPT
- BERT
- ViT
- Limitations of transformer models

On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜

Emily M. Bender*
ebender@uw.edu
University of Washington
Seattle, WA, USA

Timnit Gebru*
timnit@blackinai.org
Black in AI
Palo Alto, CA, USA

Angelina McMillan-Major
aymm@uw.edu
University of Washington
Seattle, WA, USA

Shmargaret Shmitchell
shmargaret.shmitchell@gmail.com
The Aether

Context: original Transformer paper and BERT published by Google

https://www.wired.com/story/google-timnit-gebru-ai-what-really-happened/

Transformers' Financial Cost; e.g., To Train BERT, How Much Do You Think it Costed in US Dollars?

THE COST OF TRAINING NLP MODELS
A CONCISE OVERVIEW

Or Sharir
AI21 Labs
ors@ai21.com

Barak Peleg
AI21 Labs
barakp@ai21.com

Yoav Shoham
AI21 Labs
yoavs@ai21.com

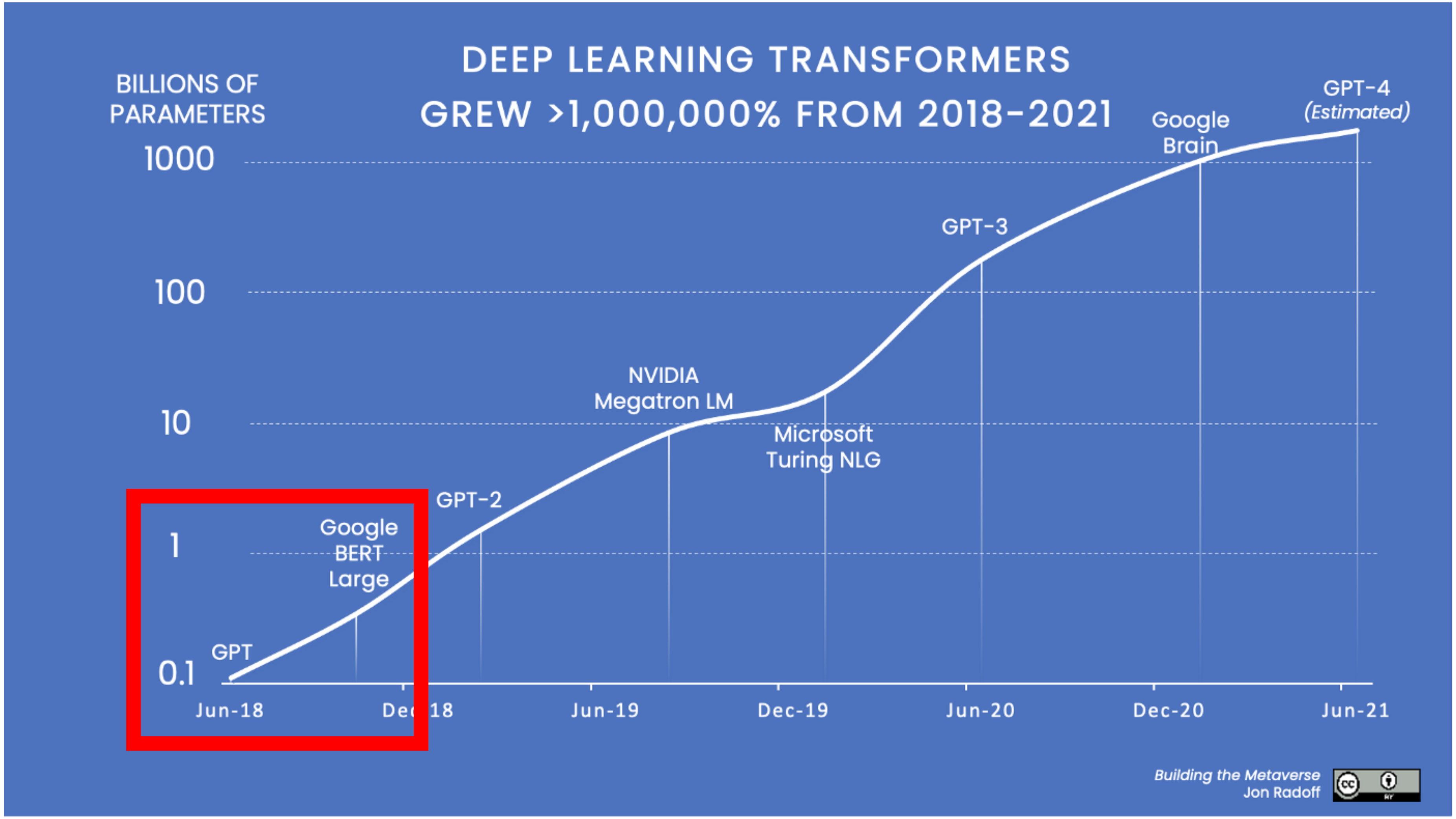
- $2.5k - $50k (110 million parameter model)
- $10k - $200k (340 million parameter model)
- $80k - $1.6m (1.5 billion parameter model)

Transformers' Environmental Cost

- Does training a BERT base model require as much energy as:
  a) Microwaving food for 7 minutes
  b) Heating a home for a day
  c) Driving 100 miles
  d) A trans-American flight

Bender et al. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big. FAT 2021.

Transformers: Huge and Growing in Size



https://medium.com/building-the-metaverse/the-metaverse-and-artificial-intelligence-ai-577343895411

Transformers' Societal Cost; e.g., BERT

- Influence of training data: 2,500M words in Wikipedia + 800M words in BooksCorpus
  - Who does and who does not contribute to such data repositories?
    - e.g., "recent surveys of Wikipedians find that only 8.8–15% are women or girls"
    - e.g., "Internet access itself is not evenly distributed, resulting in Internet data overrepresenting younger users and those from developed countries"
  - What kind of biases might be found in such data repositories?
    - e.g., "BERT associates phrases referencing persons with disabilities with more negative sentiment words, and that gun violence, homelessness, and drug addiction are overrepresented in texts discussing mental illness"
  - Given that "unsupervised pre-training is an integral part of many language understanding systems" (BERT paper: Devlin et al. arXiv 2018), how do we do this responsibly?

Bender et al. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big. FAT 2021.

Today's Topics

- Explosion of transformers
- GPT
- BERT
- ViT
- Limitations of transformer models

The End