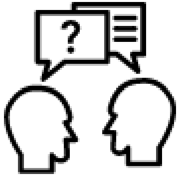
Copyright Notice

deeplearning.ai

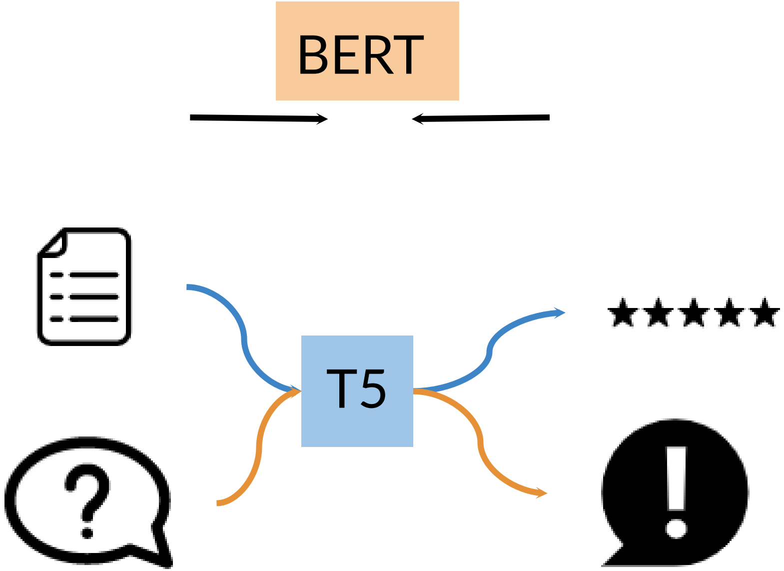
Week 3 Overview

Week 3

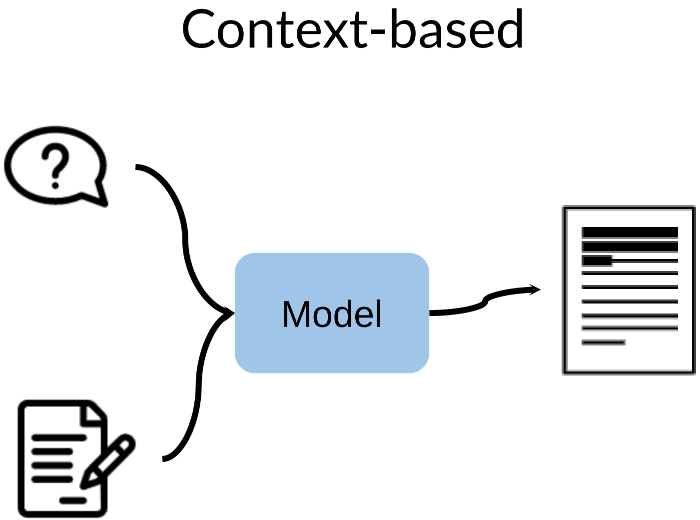
Question Answering

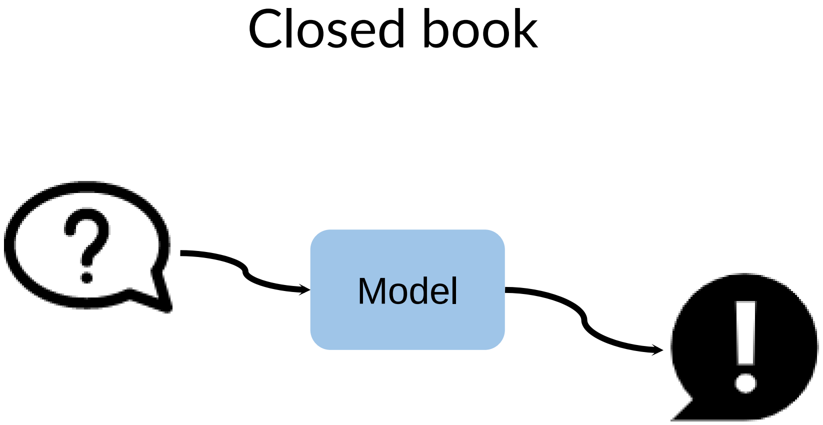


Transfer learning

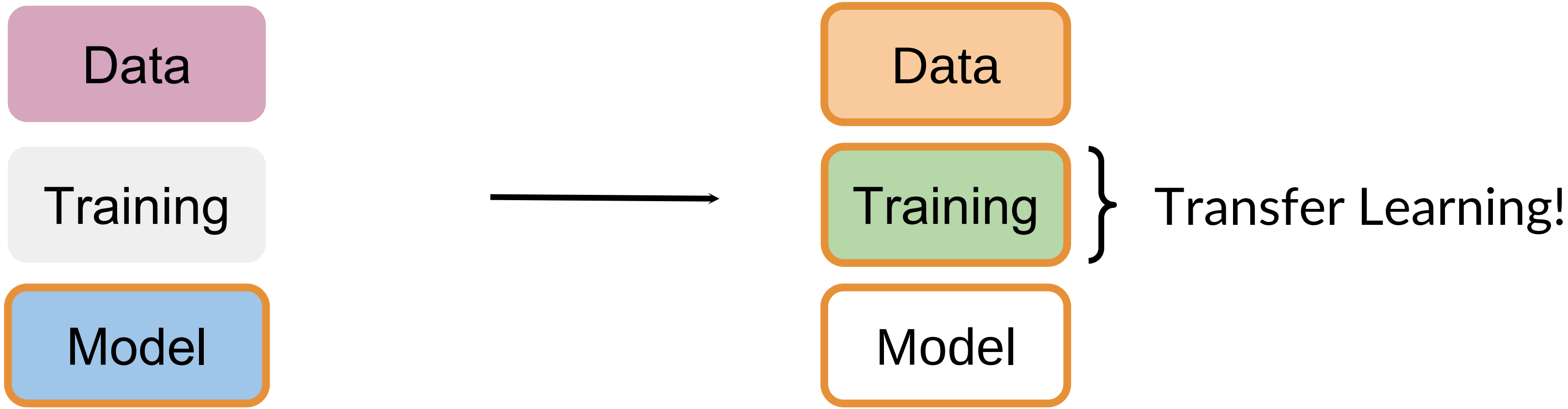


Question Answering

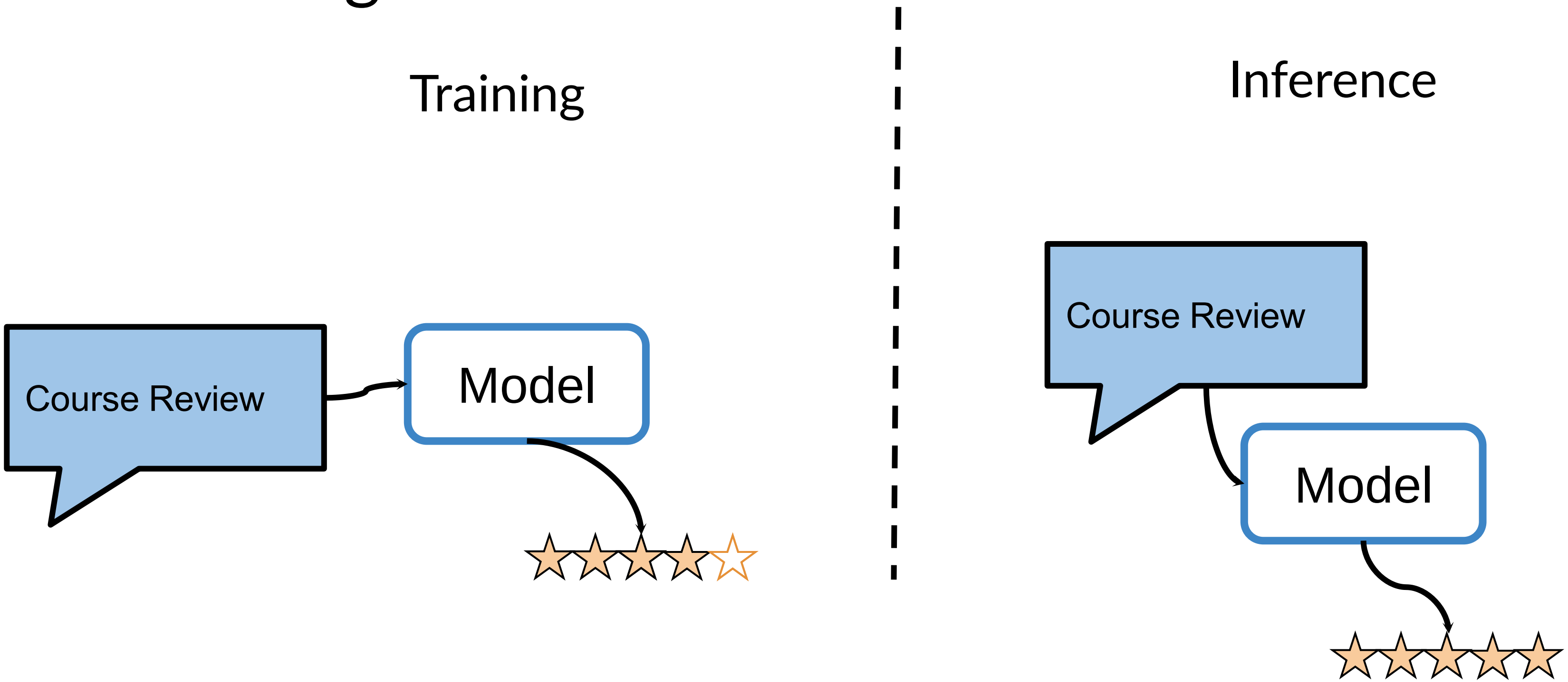




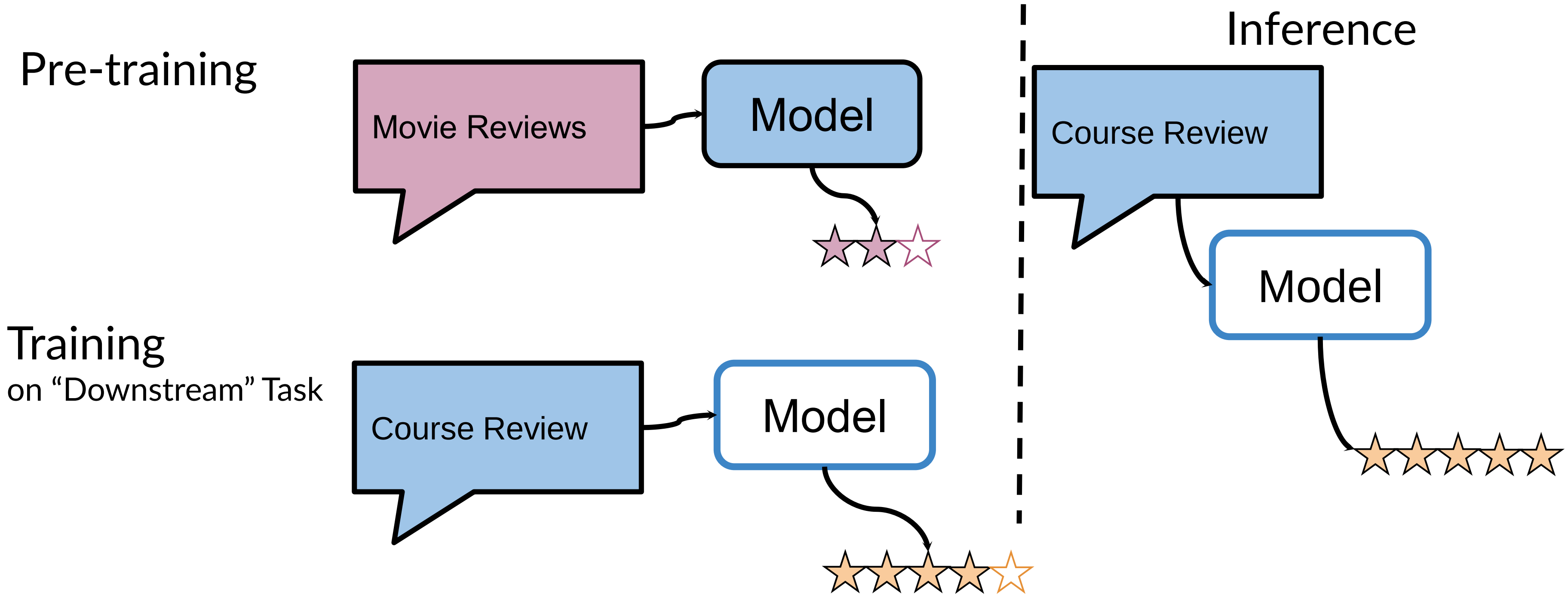
Not just the model



Classical training



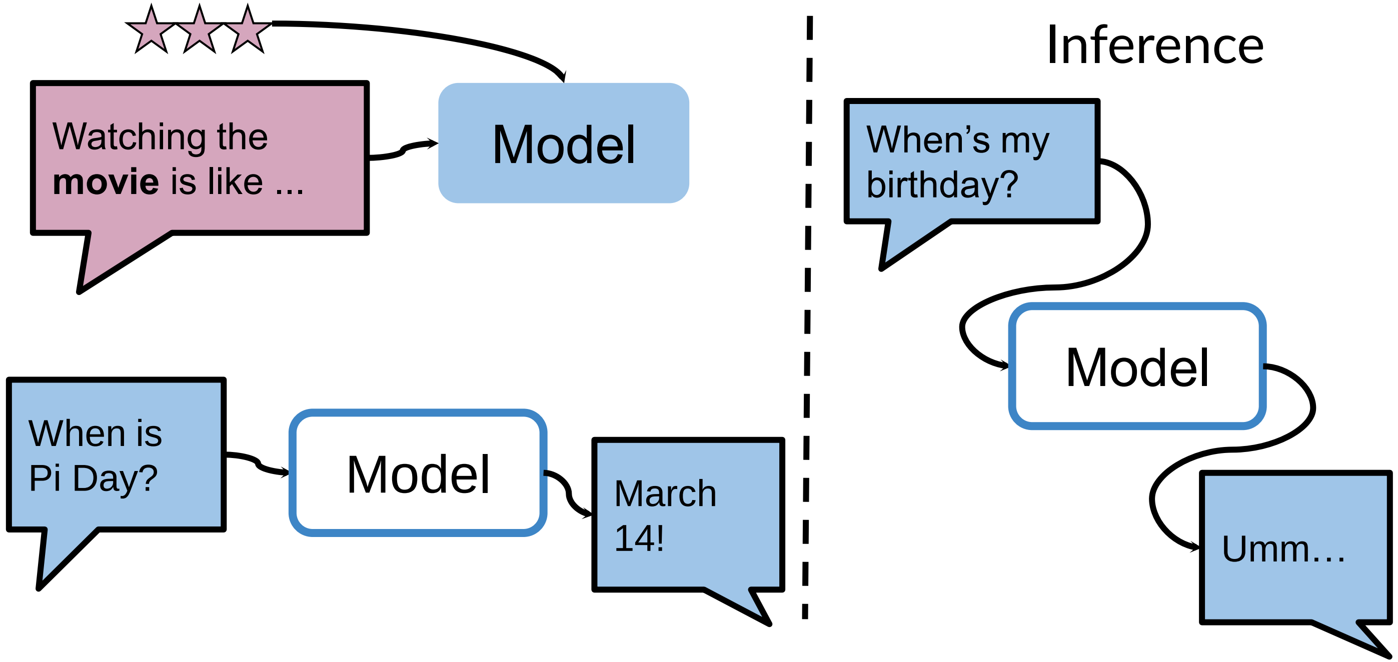
Transfer learning



Transfer Learning: Different Tasks

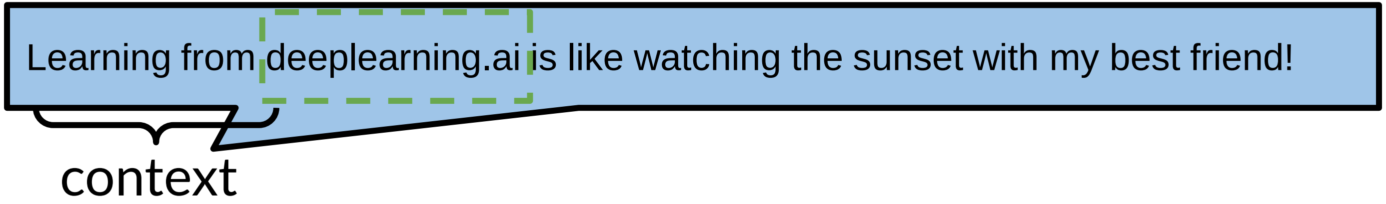
Pre-
Training
Sentiment
Classification

Training
Downstream task:
Question Answering

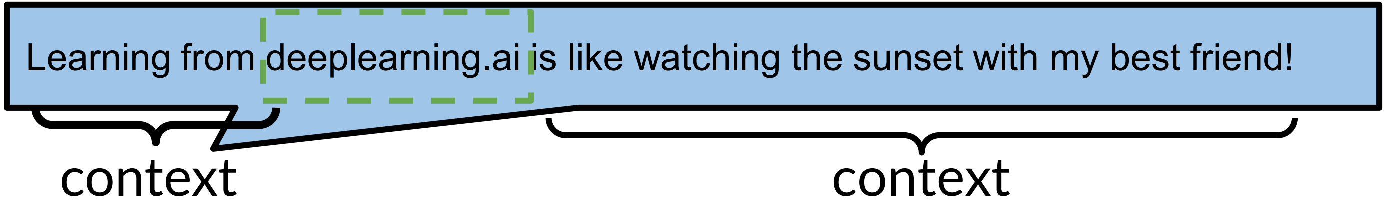


BERT: Bi-directional Context

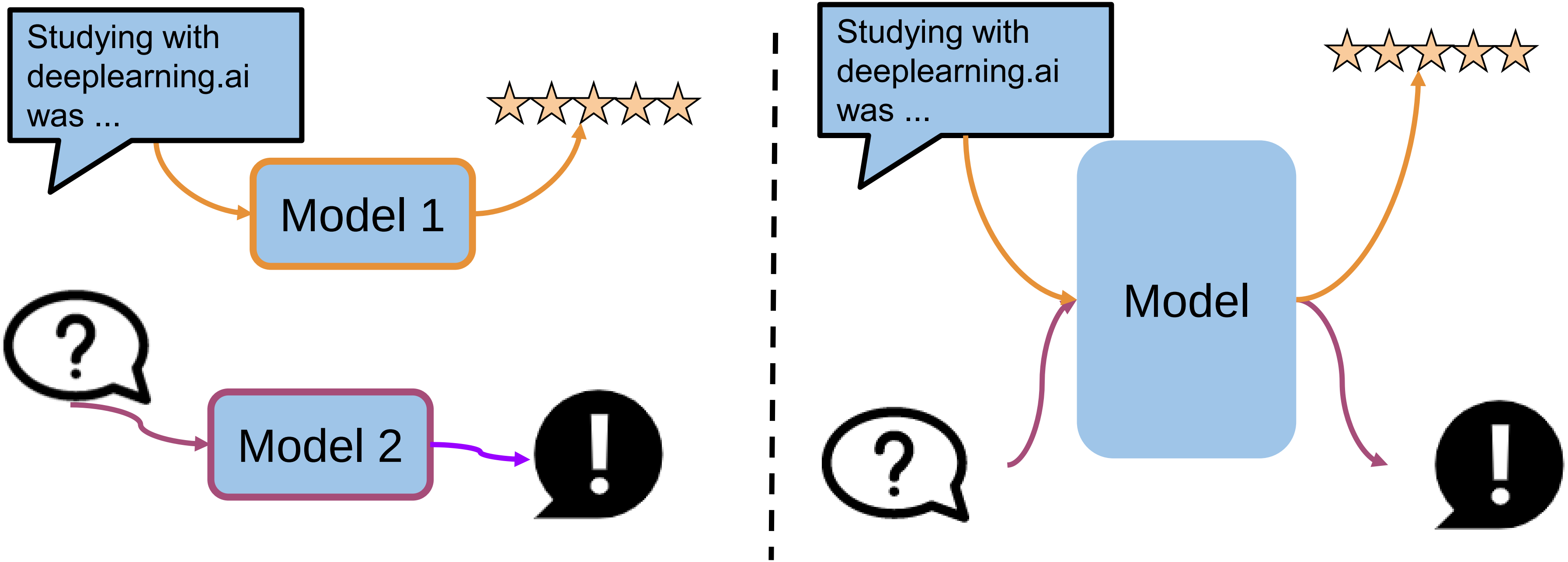
Uni-directional



Bi-directional



T5: Single task vs. Multi task



T5: more data, better performance

English wikipedia
~13 GB

C4
Colossal **C**lean **C**rawled
Corpus
~800 GB

deeplearning.ai

Transfer Learning in NLP

Desirable Goals

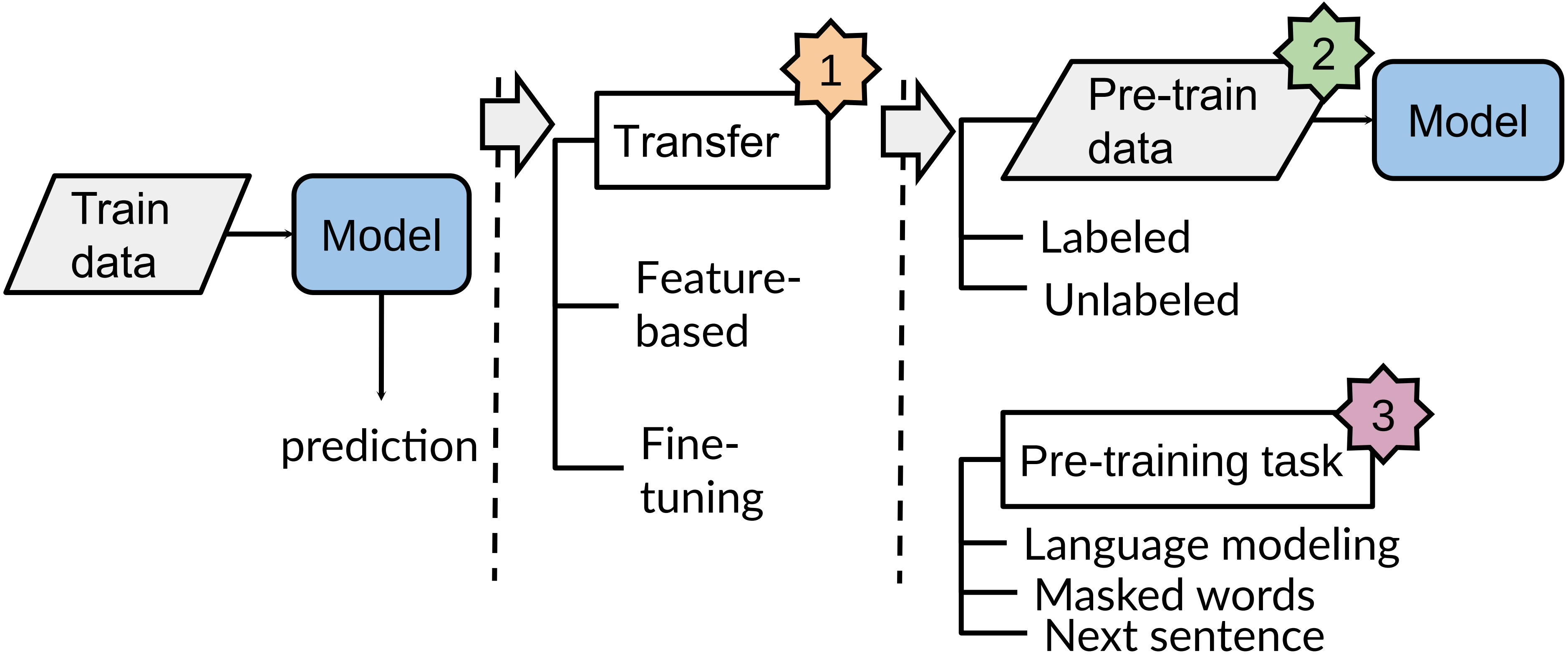
- Reduce training time

- Improve predictions

- Small datasets

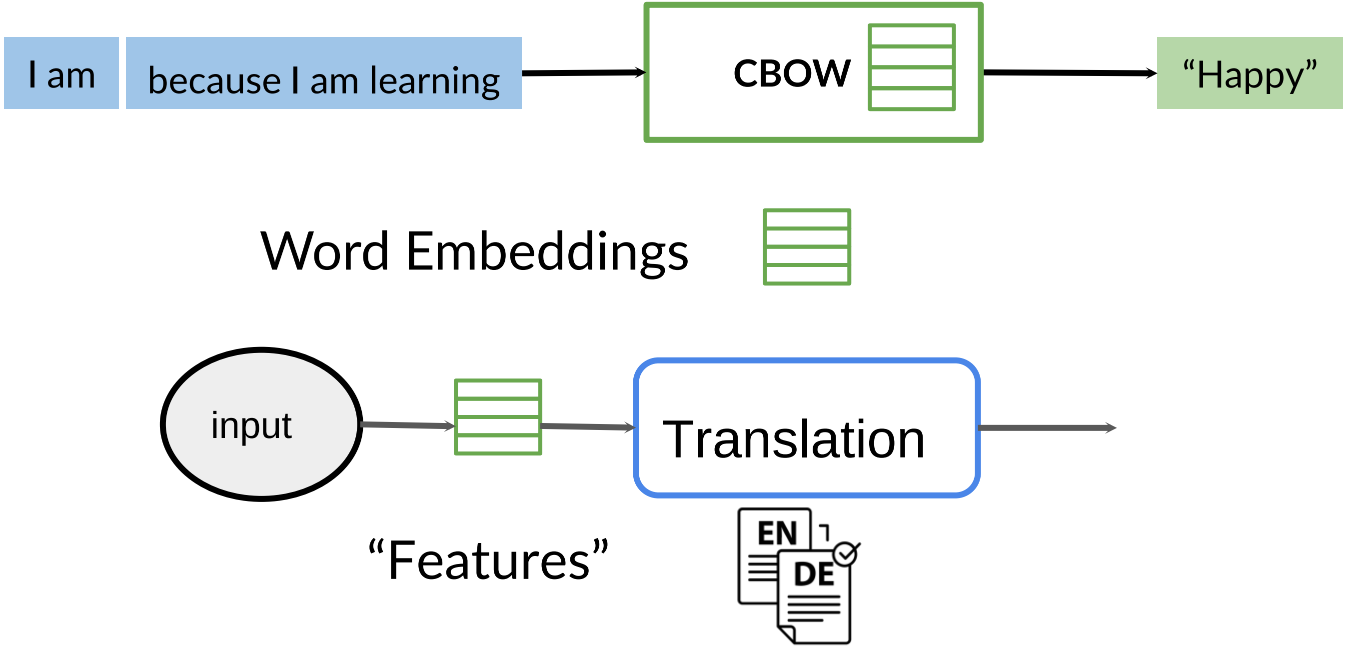
Transfer Learning!

Transfer learning options

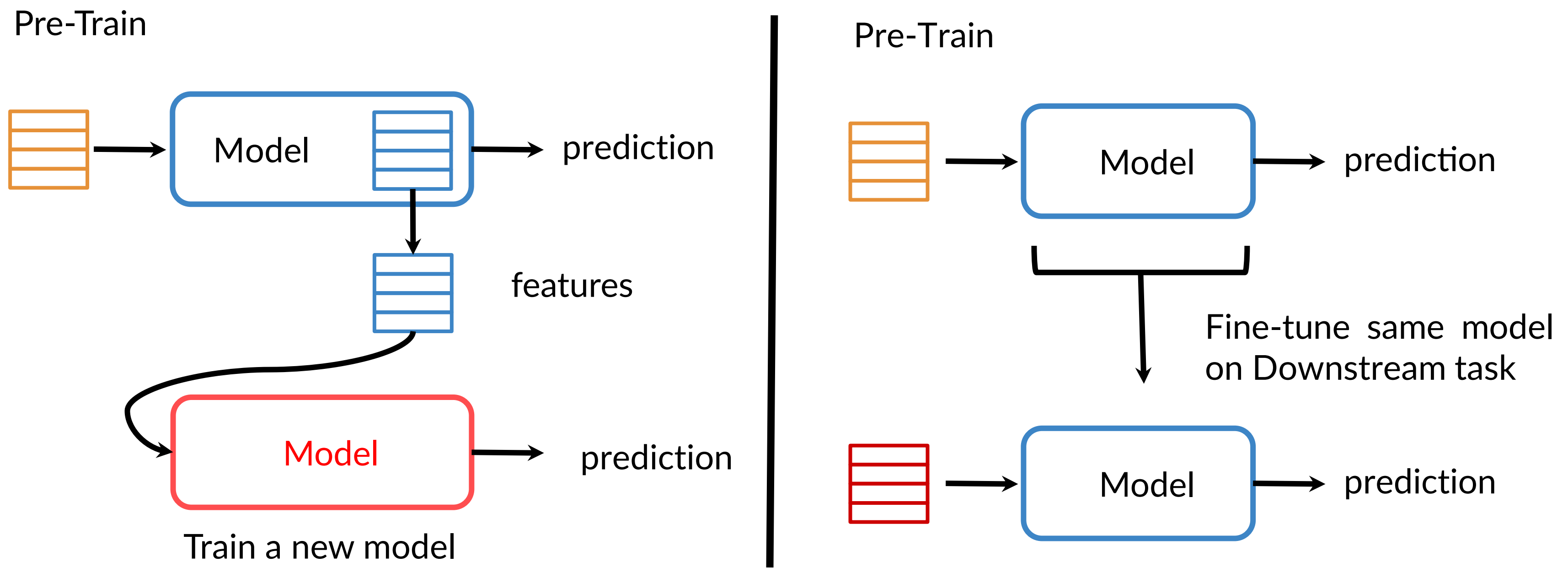


General purpose learning

Transfer 1

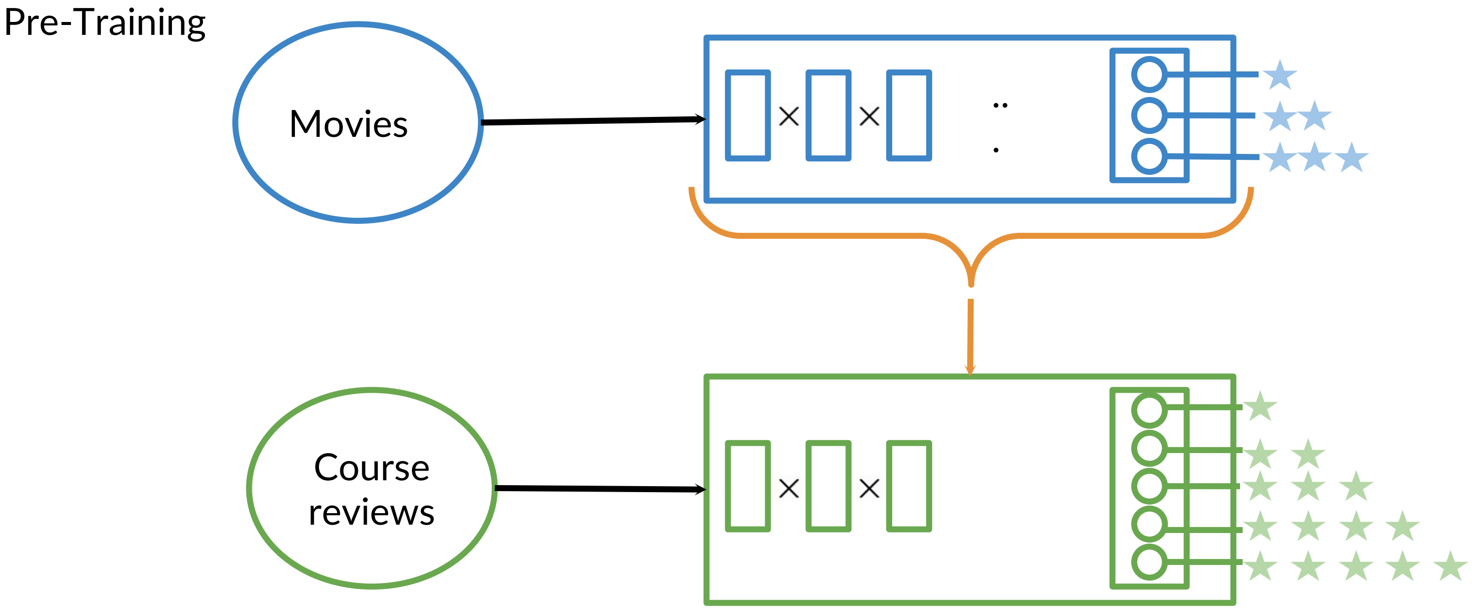


Feature-based vs. Fine-Tuning

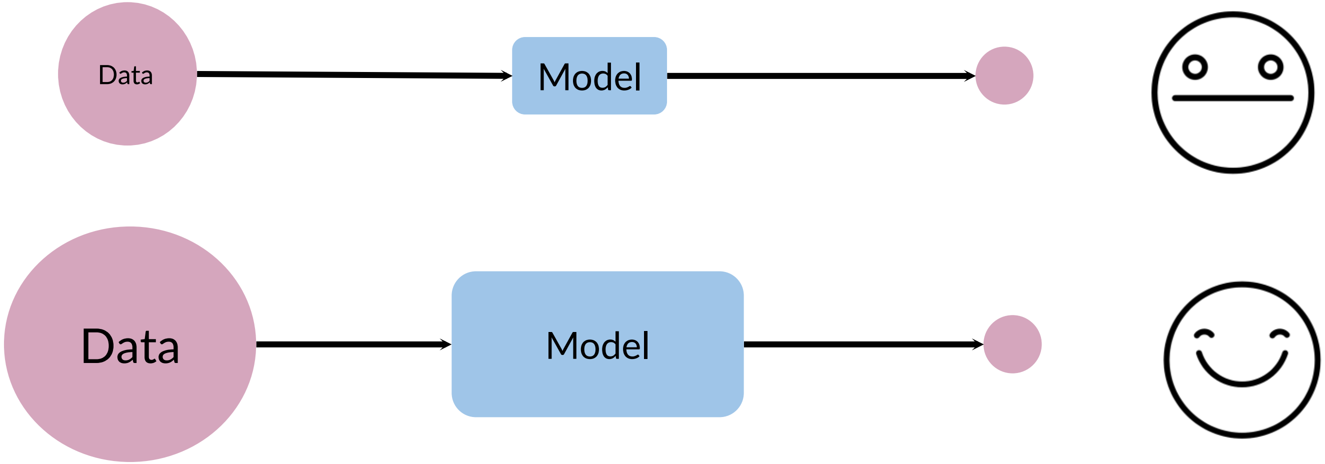


Fine-tune: adding a layer

Transfer 1



Data and performance



Labeled vs Unlabeled Data

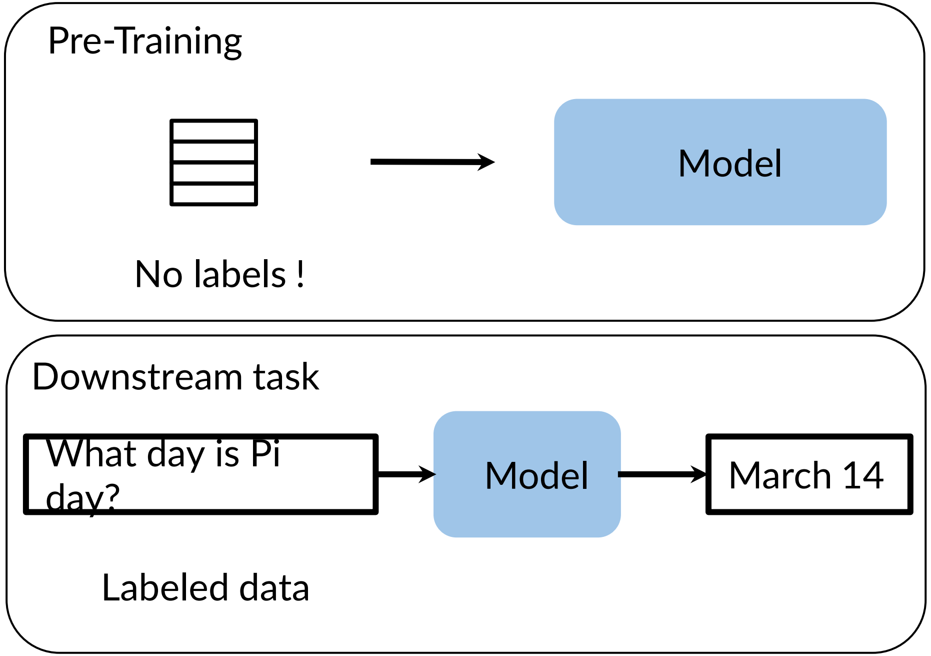
Pre-train data

2

Labeled text data

Unlabeled text data

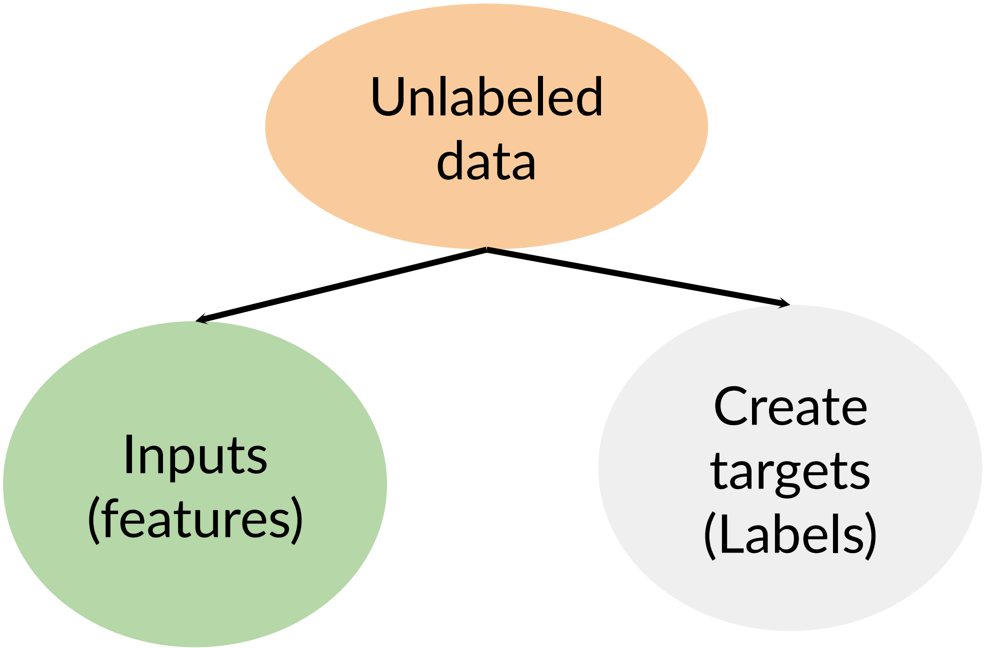
Transfer learning with unlabeled data



Which tasks work with **unlabeled** data?

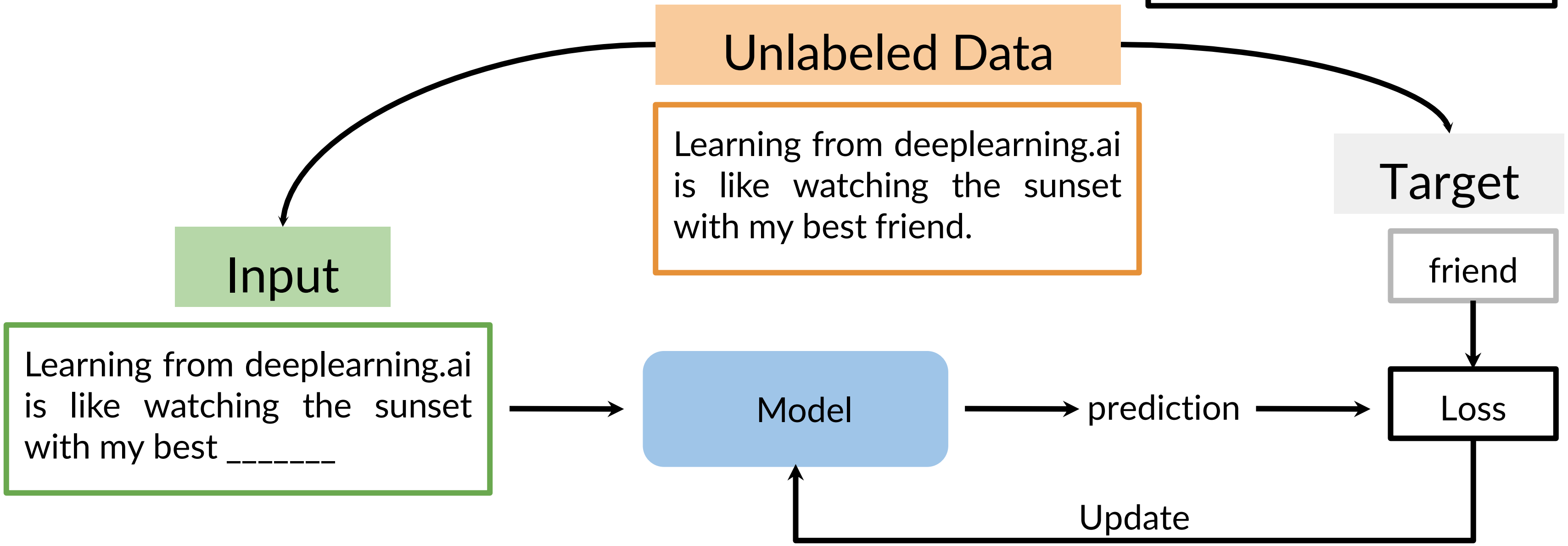
Self-supervised task

Pre-training task 3



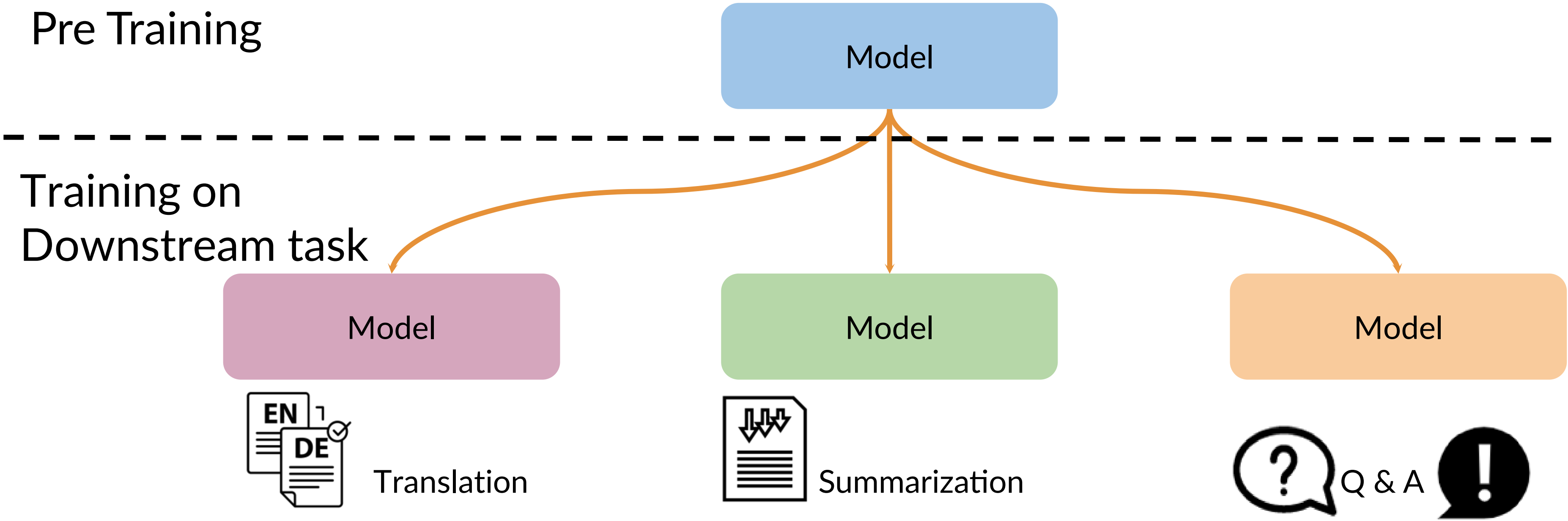
Self-supervised tasks

Pre-training task 3

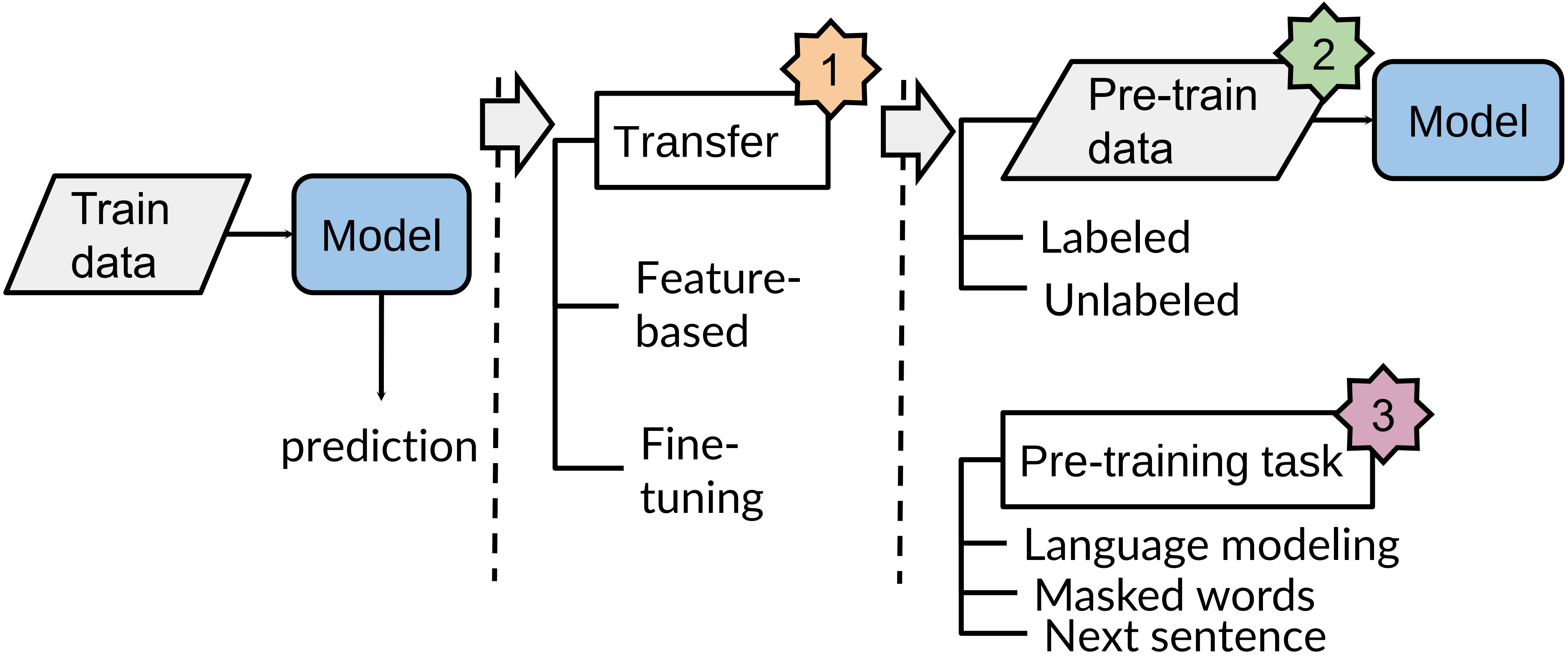


Language modeling

Fine-tune a model for each downstream task



Summary



Train data
Model
prediction
Transfer
1
Feature-based
Fine-tuning
Pre-train data
2
Model
Labeled
Unlabeled
Pre-training task
3
Language modeling
Masked words
Next sentence

deeplearning.ai

ELMo, GPT, BERT, T5

Outline

CBOW ELMo GPT BERT T5

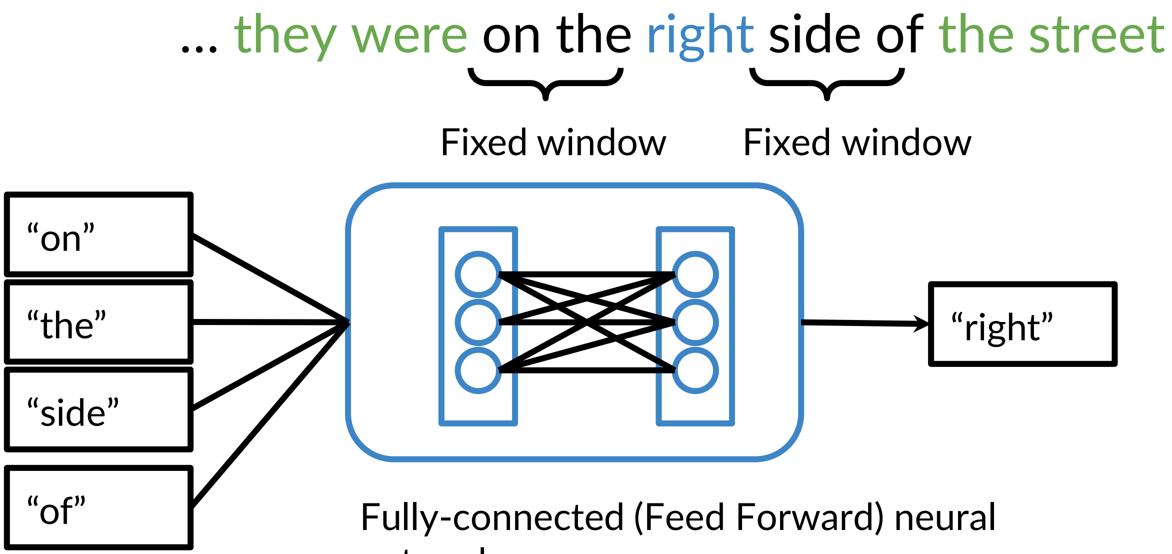
Context

… right …

… they were on the right …

… they were on the right side of the street

Continuous Bag of Words



Fully-connected (Feed Forward) neural network

Need more context?

... they were $\underbrace{\text{on the}}_{\text{Fixed window}}$ right $\underbrace{\text{side of}}_{\text{Fixed window}}$ the street.

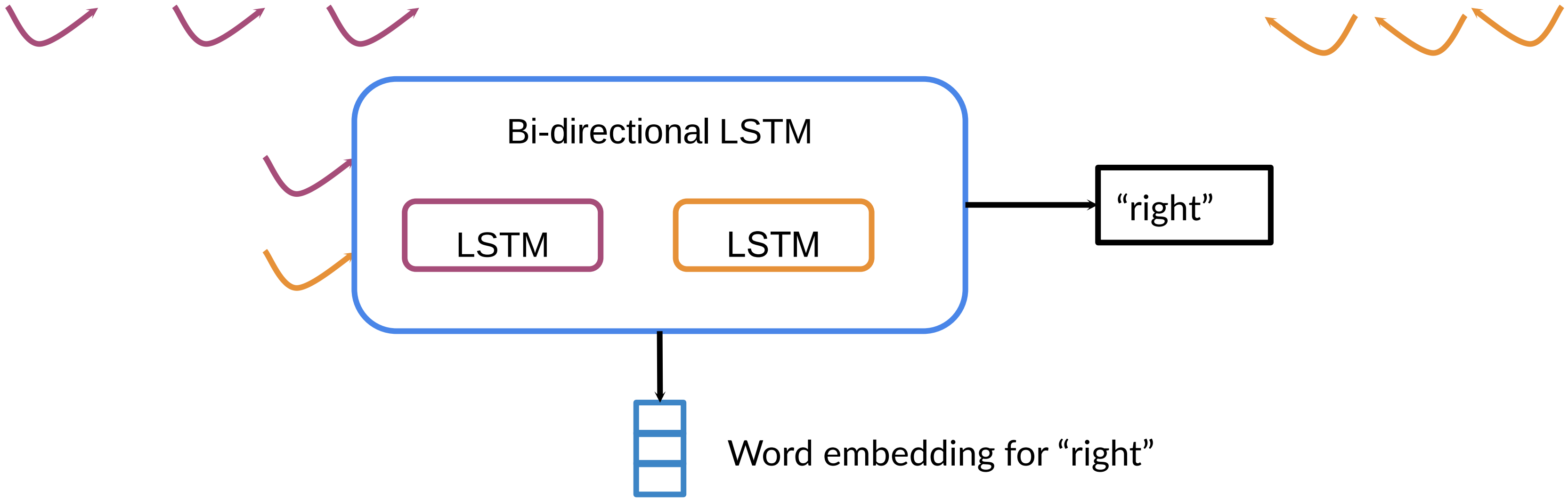
... they were on the right side of history.

Use all context words

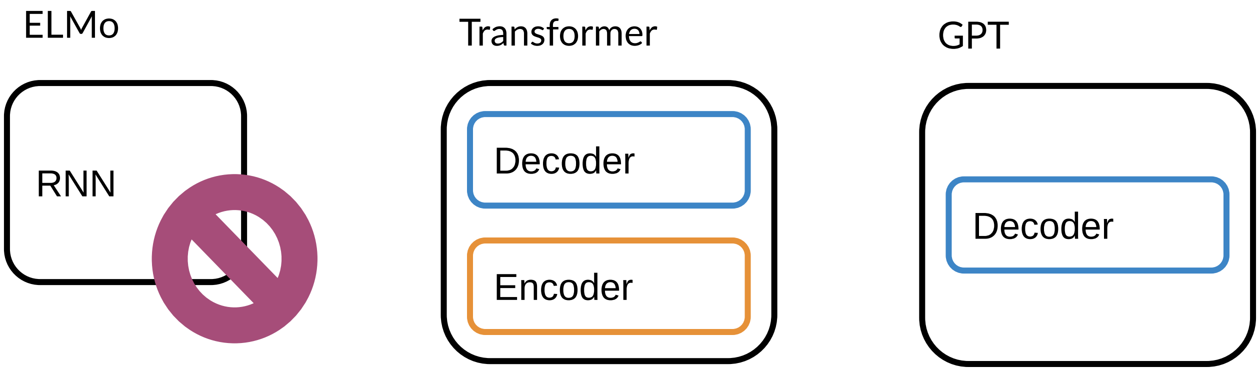
The legislators believed that they were on the right side of history, so they changed the law.

ELMo: Full context using RNN

The legislators believed that they were on the _____ side of history so they changed the law.



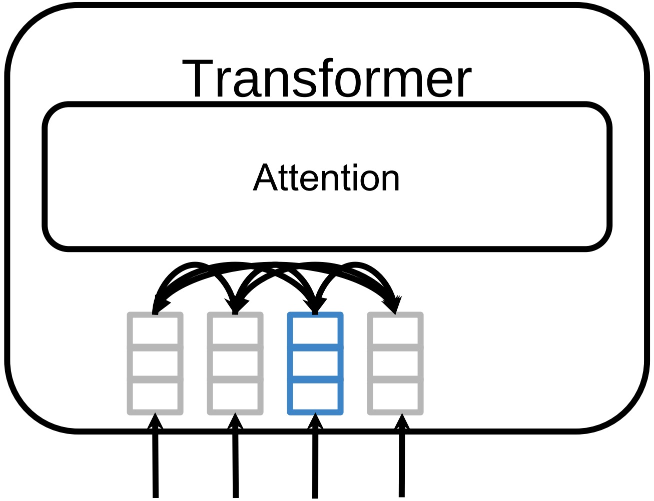
Open AI GPT



The legislators believed that they were on the _____

Uni-directional

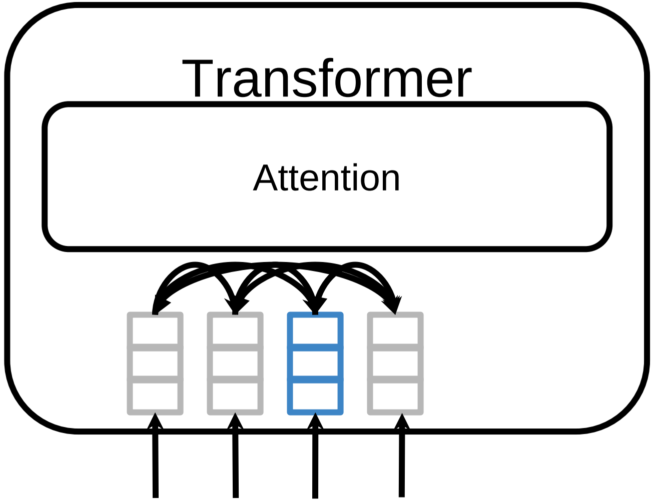
Why not bi-directional?



... on the right side...

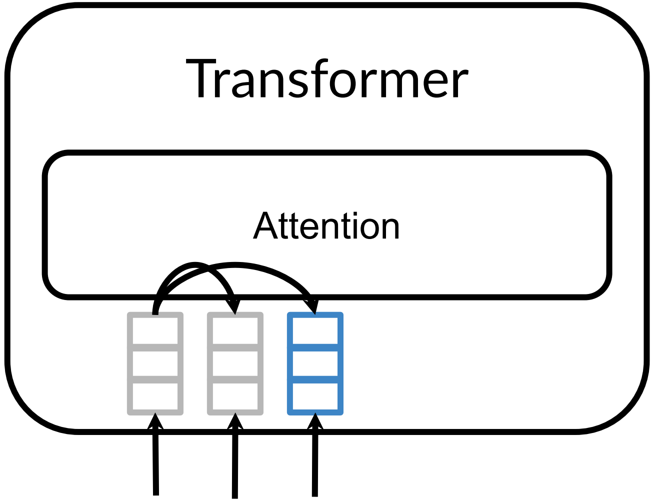
Each word can peek at itself!

GPT: Uni-directional



... on the right side...

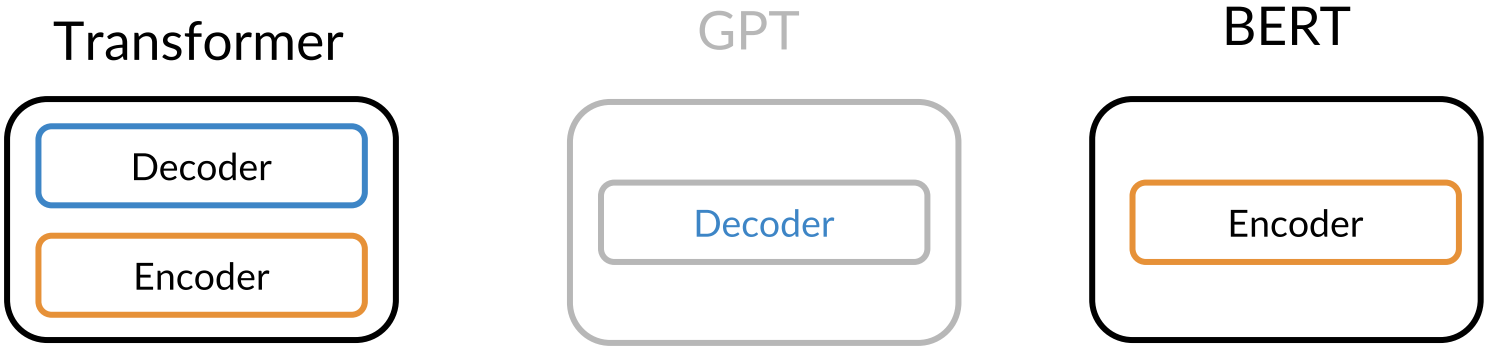
Each word can peek at itself!



... on the right

No peeking!

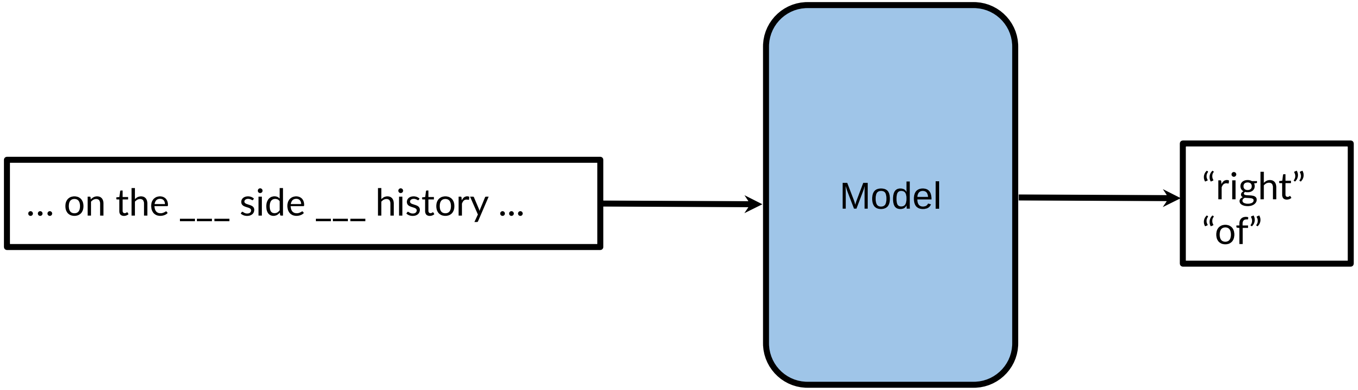
BERT



The legislators believed that they were on the _____ side of history, so they changed the law.

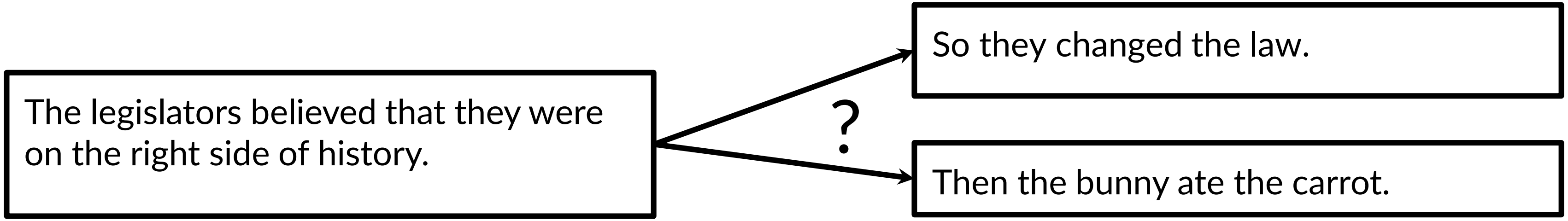
Bi-directional

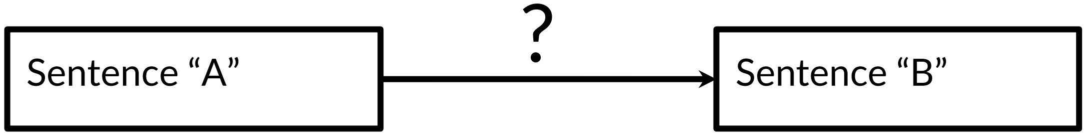
Transformer + Bi-directional Context



Multi-Mask Language Modeling

BERT: Words to Sentences





Next Sentence Prediction

BERT Pre-training Tasks

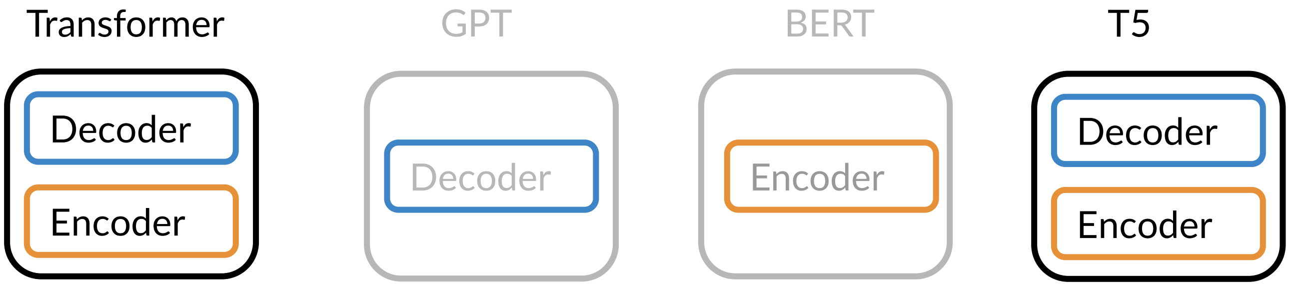
Multi-Mask Language Modeling

... on the ___ side ___ history ... → Model → "right" "of"

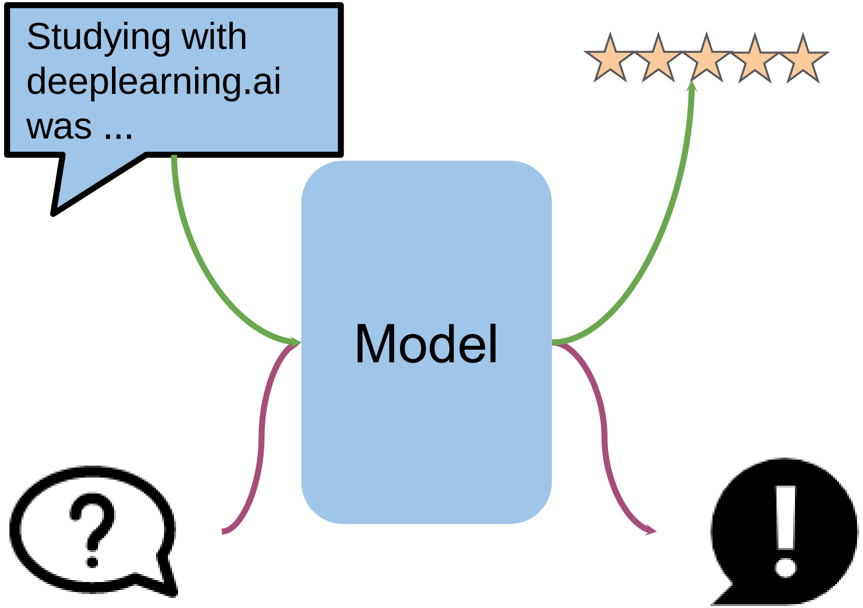
Next Sentence Prediction

Sentence "A" $\xrightarrow{?}$ Sentence "B"

T5: Encoder vs. Encoder-Decoder

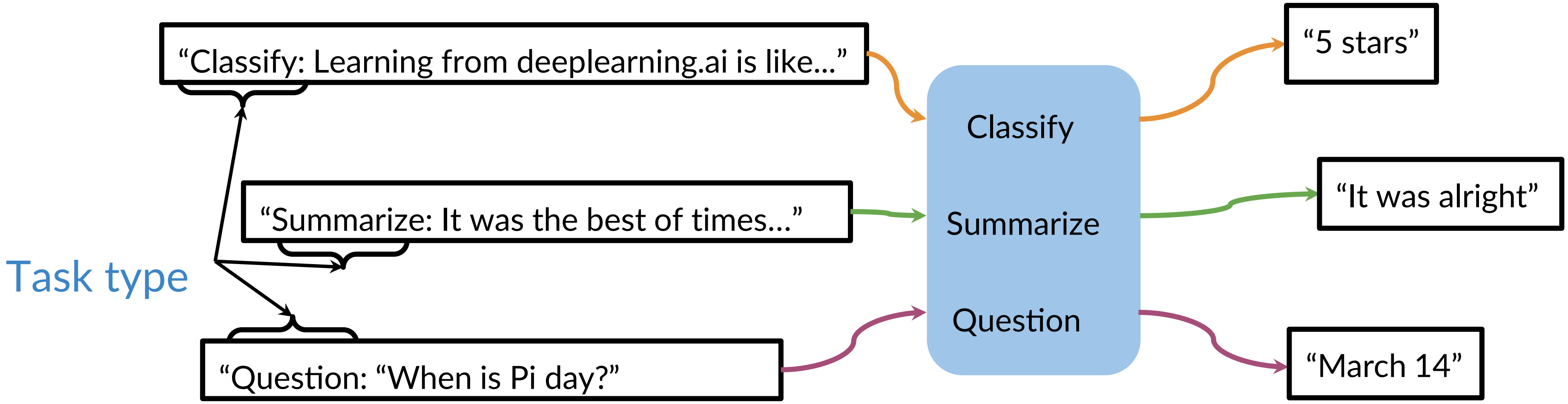


T5: Multi-task



How?

T5: Text-to-Text



Summary

More details next!

CBOW	ELMo	GPT	BERT	T5
Context window	Full sentence	Transformer: Decoder	Transformer: Encoder	Transformer: Encoder - Decoder
FFNN	Bi-directional Context	Uni-directional Context	Bi-directional Context	Bi-directional Context
	RNN		Multi-Mask	Multi-Task
			Next Sentence Prediction	

deeplearning.ai

Bidirectional Encoder Representations from Transformers (BERT)

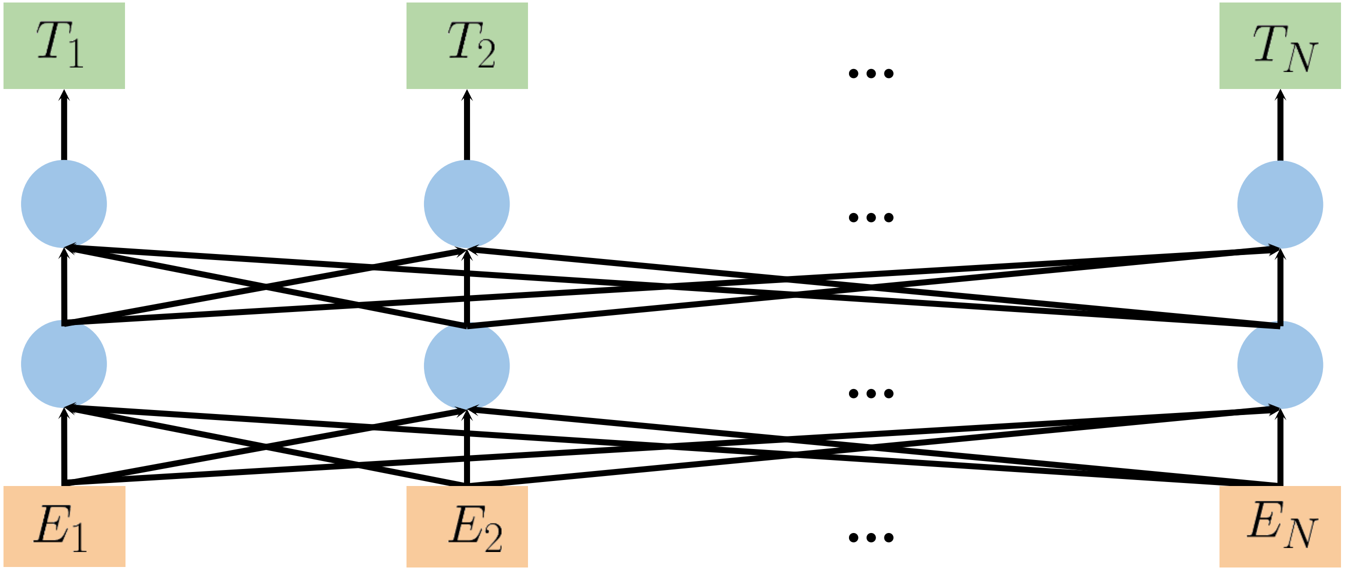
Outline

- Learn about the BERT architecture

- Understand how BERT pre-training works

BERT

- Makes use of transfer learning/pre-training:



BERT

- A multi layer bidirectional transformer
- Positional embeddings
- BERT_base:

  12 layers (12 transformer blocks)

  12 attentions heads

  110 million parameters

BERT pre-training

After school Lukasz does his ______________ in the library.

- Masked language modeling (MLM)

BERT pre-training

After school Lukasz does his homework in the library.

After school ______ his homework in the ______ .

Summary

- Choose 15% of the tokens at random: mask them 80% of the time, replace them with a random token 10% of the time, or keep as is 10% of the time.
- There could be multiple masked spans in a sentence
- Next sentence prediction is also used when pre-training.

deeplearning.ai

BERT Objective

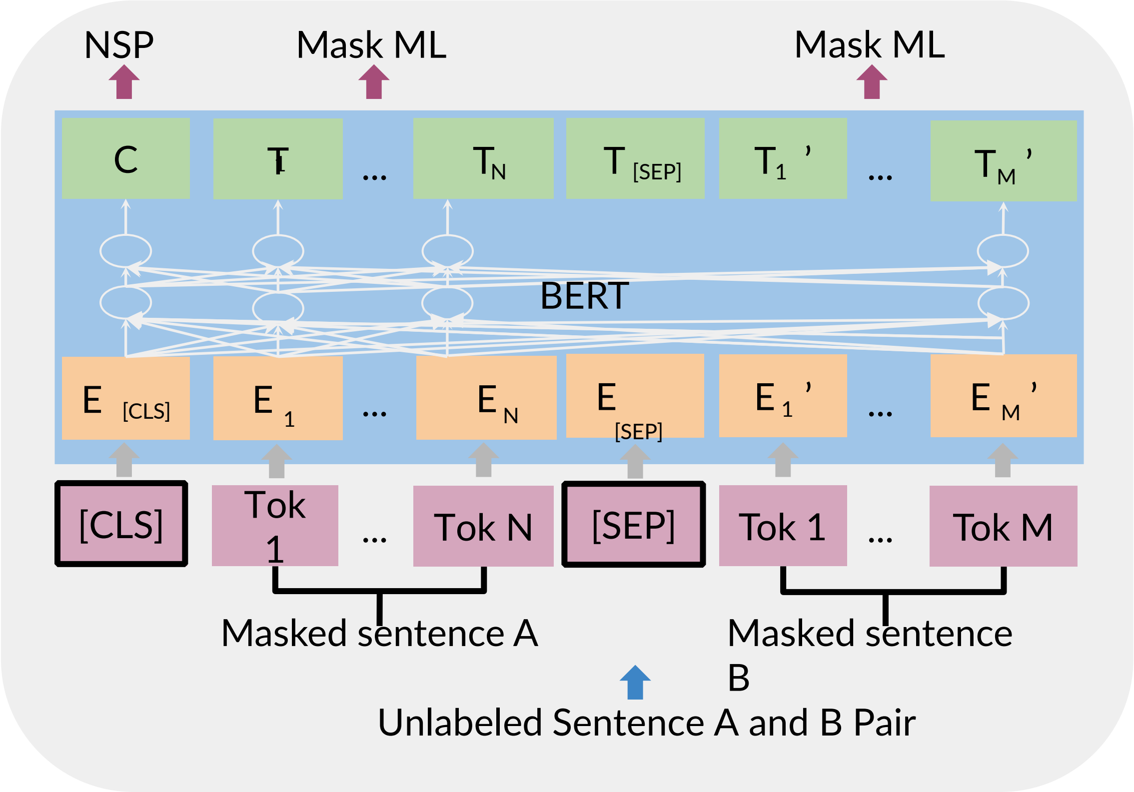
Outline

- Understand how BERT inputs are fed into the model
- Visualize the output
- Learn about the BERT objective

Formalizing the input

Input	[CLS]	my	dog	is	cute	[SEP]	he	likes	play	##ing	[SEP]
Token Embeddings	$E_{[CLS]}$	E_{my}	E_{dog}	E_{is}	E_{cute}	$E_{[SEP]}$	E_{he}	E_{likes}	E_{play}	$E_{\#\#ing}$	$E_{[SEP]}$
	+	+	+	+	+	+	+	+	+	+	+
Segment Embeddings	E_A	E_A	E_A	E_A	E_A	E_A	E_B	E_B	E_B	E_B	E_B
	+	+	+	+	+	+	+	+	+	+	+
Position Embeddings	E_0	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8	E_9	E_{10}

Visualizing the output

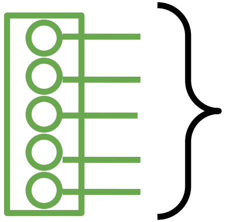


- [CLS]: a special classification symbol added in front of every input
- [SEP]: a special separator token

BERT Objective

Objective 1:
Multi-Mask LM

Loss: Cross Entropy Loss

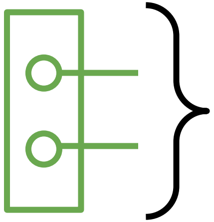


V

+

Objective 2:
Next Sentence Prediction

Loss: Binary Loss



2

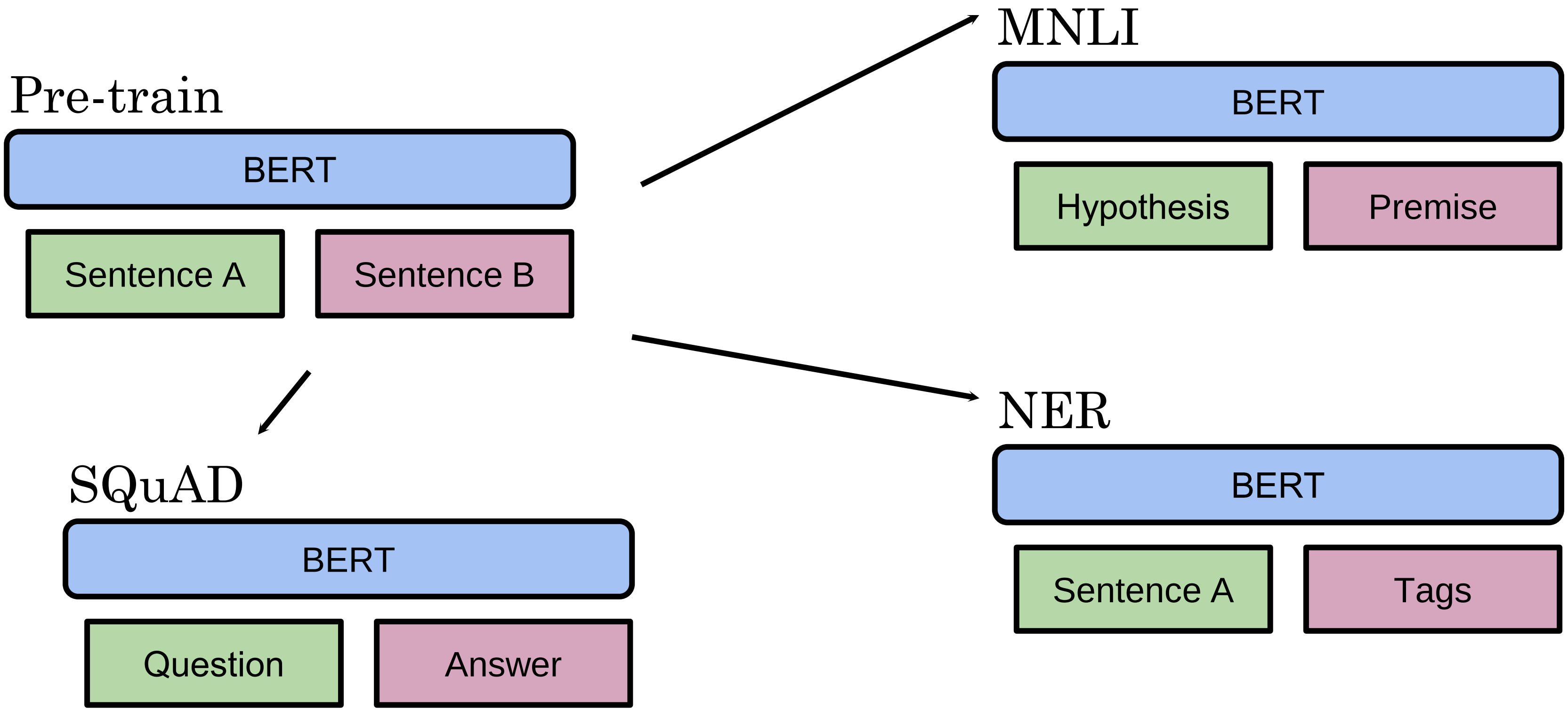
Summary

- BERT objective
- Model inputs/outputs

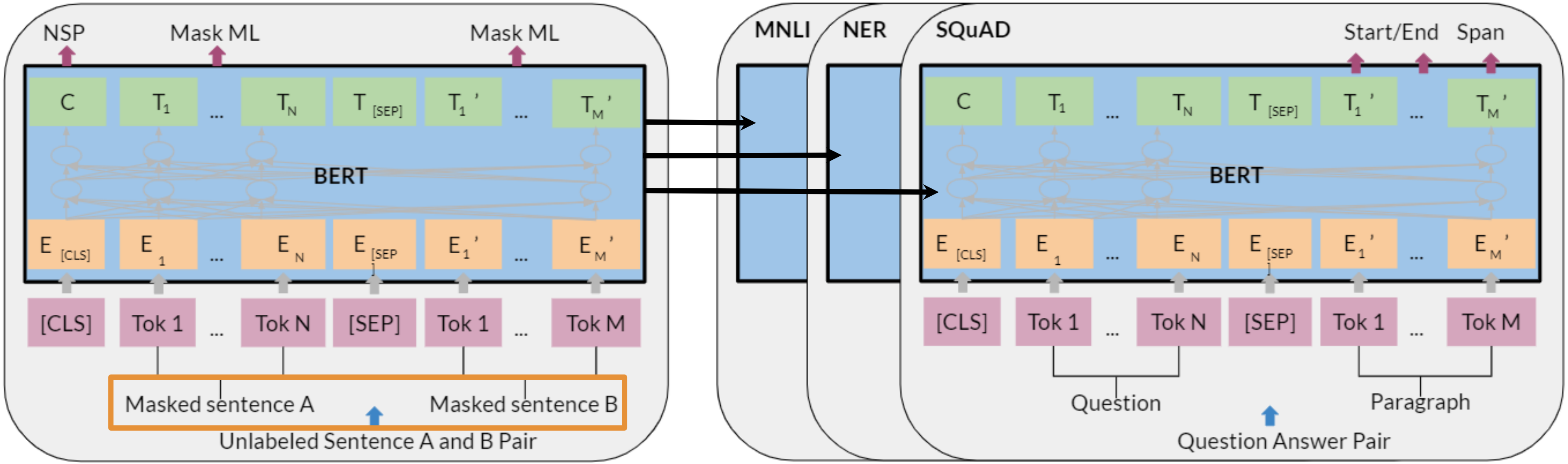
Fine-tuning BERT

deeplearning.ai

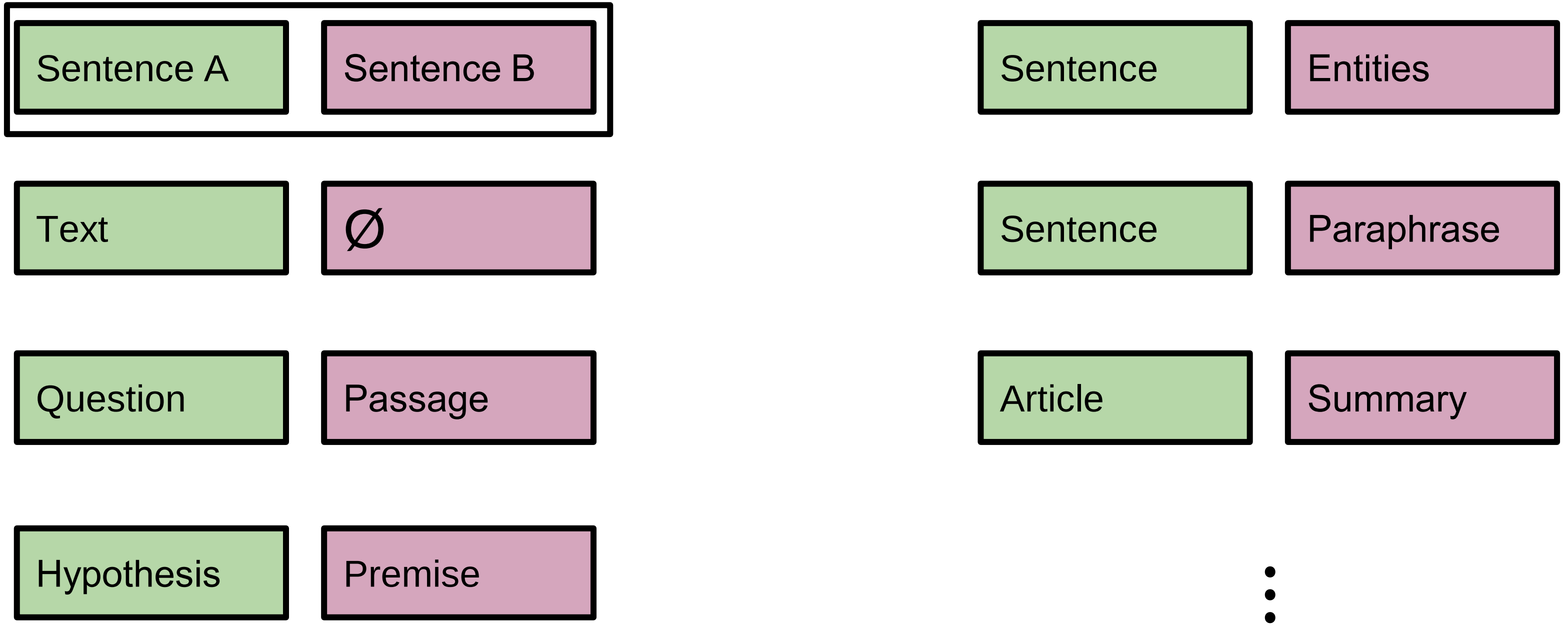
Fine-tuning BERT: Outline



Inputs


NSP
Mask ML
Mask ML
C
T_1
T_N
$T_{[SEP]}$
T_1'
T_M'
BERT
$E_{[CLS]}$
E_1
E_N
$E_{[SEP]}$
E_1'
E_M'
[CLS]
Tok 1
Tok N
[SEP]
Tok 1
Tok M
Masked sentence A
Masked sentence B
Unlabeled Sentence A and B Pair
MNLI
NER
SQuAD
Start/End
Span
C
T_1
T_N
$T_{[SEP]}$
T_1'
T_M'
BERT
$E_{[CLS]}$
E_1
E_N
$E_{[SEP]}$
E_1'
E_M'
[CLS]
Tok 1
Tok N
[SEP]
Tok 1
Tok M
Question
Paragraph
Question Answer Pair

Summary


Sentence A
Sentence B
Sentence
Entities
Text
Ø
Sentence
Paraphrase
Question
Passage
Article
Summary
Hypothesis
Premise

deeplearning.ai

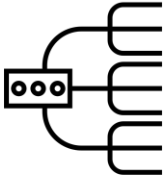
Transformer T5

Outline

- Understand how T5 works
- Recognize the different types of attention used
- Overview of model architecture

Transformer - T5 Model

Text to Text

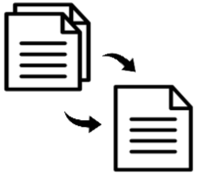


Classification

Question Answering (Q&A)

Machine Translation

Summarization



Sentiment

Transformer - T5 Model

Original text

Thank you ~~for inviting~~ me to your party ~~last~~ week.

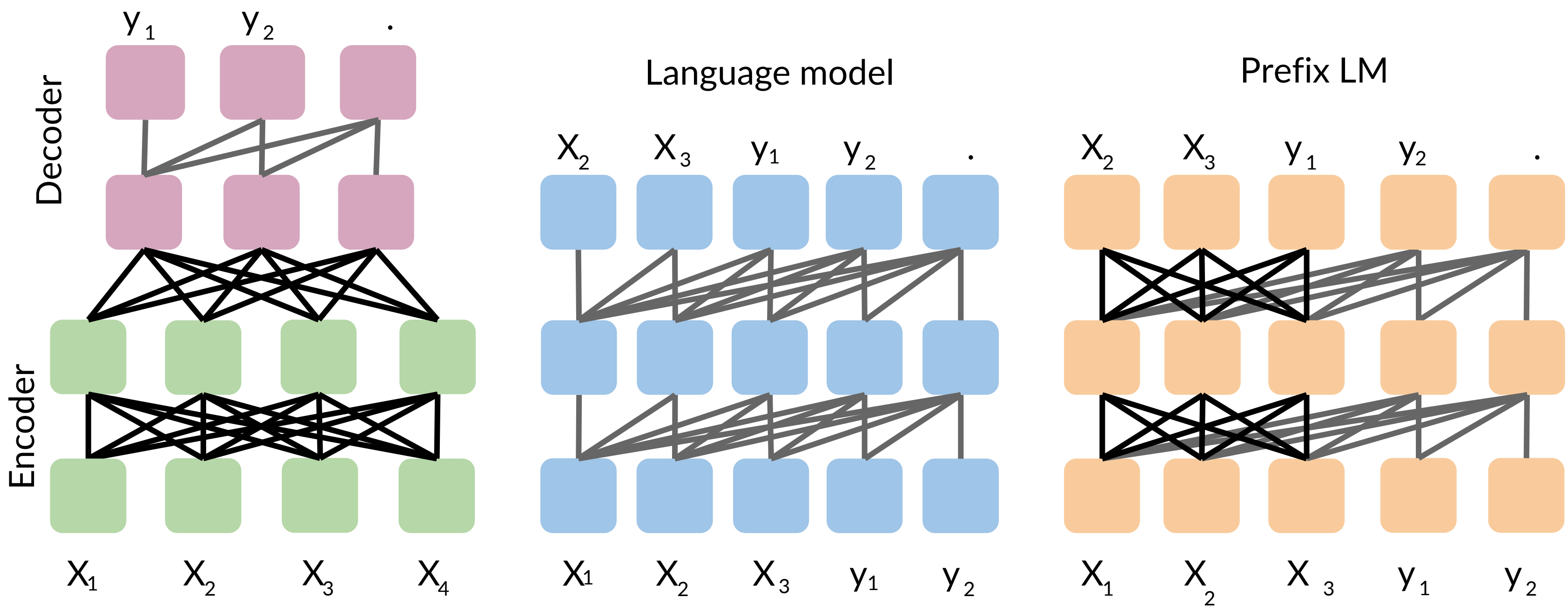
Inputs

Thank you <X> me to your party <Y> week

Targets

<X> for inviting <Y> last <Z>

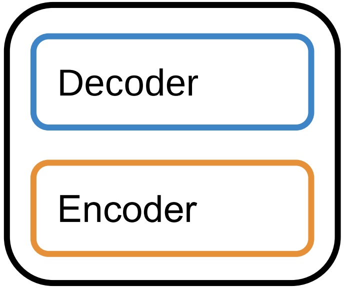
Model Architecture



©Exploring the Limits of Transfer learning with a unified text to Text Transformer. Raffel et. al. 2020

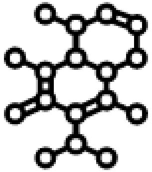
Model Architecture

- Encoder/decoder



- 12 transformer blocks each

- 220 million parameters



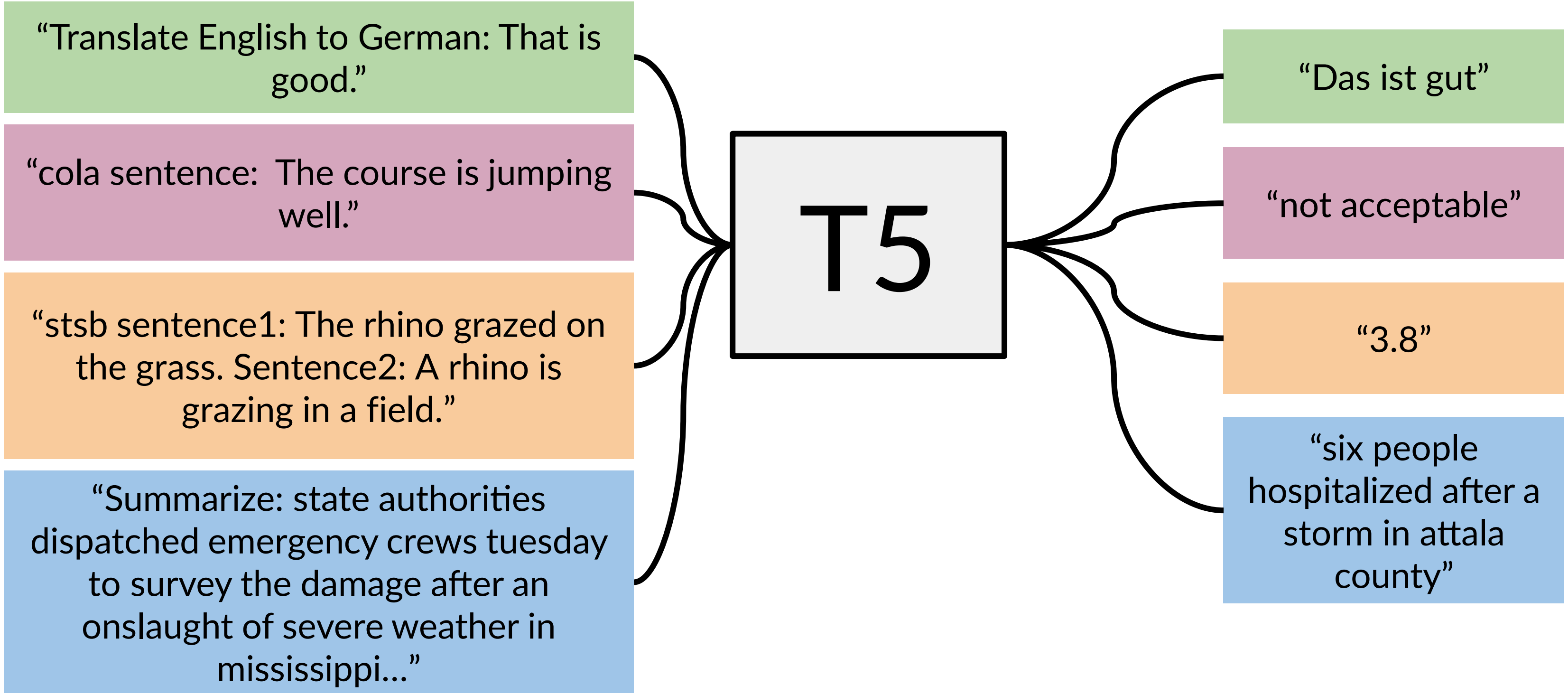
Summary

- Prefix LM attention
- Model architecture
- Pre-training T5 (MLM)

deeplearning.ai

Multi-task Training Strategy

Multi-task training strategy



Input and Output Format

Machine translation:

- translate English to German: That is good.

- Predict entailment, contradiction , or neutral
  - mnli premise: I hate pigeons hypothesis: My feelings towards pigeons are filled with animosity. target: entailment
- Winograd schema
  - The city councilmen refused the demonstrators a permit because *they* feared violence

Multi-task Training Strategy

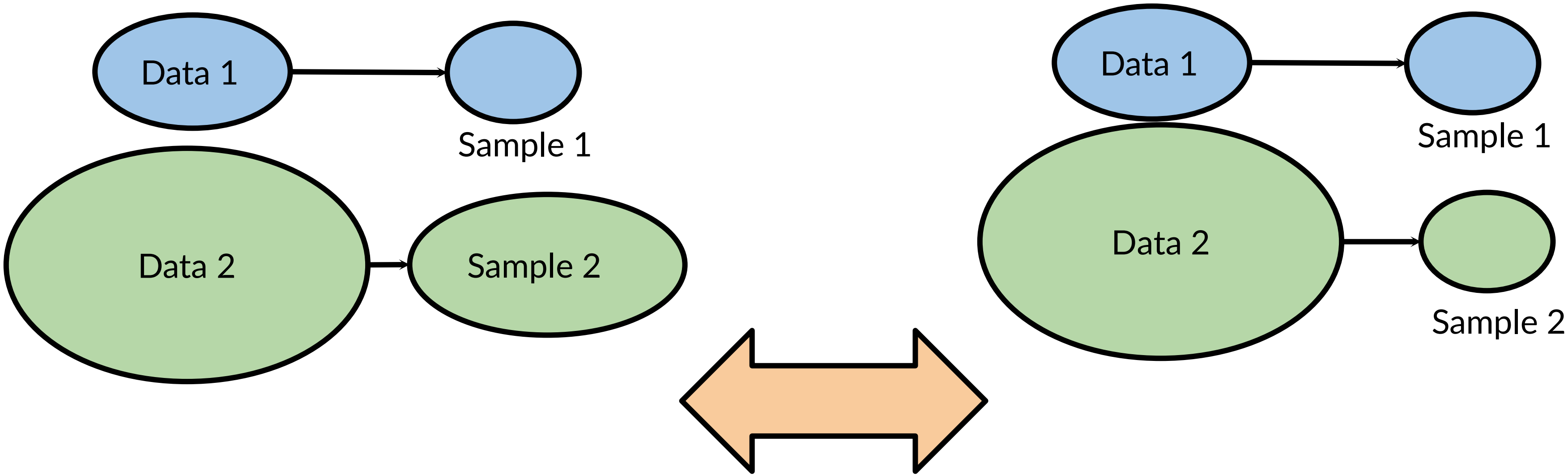
Fine-tuning method	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
* All parameters	**83.28**	**19.24**	**80.88**	**71.36**	**26.98**	**39.82**	**27.65**
Adapter layers, $d = 32$	80.52	15.08	79.32	60.40	13.84	17.88	15.54
Adapter layers, $d = 128$	81.51	16.62	79.47	63.03	19.83	27.50	22.63
Adapter layers, $d = 512$	81.54	17.78	79.18	64.30	23.45	33.98	25.81
Adapter layers, $d = 2048$	81.51	16.62	79.47	63.03	19.83	27.50	22.63
Gradual unfreezing	82.50	18.95	79.17	**70.79**	26.71	39.02	26.93

How much data from each task to train on?

Data Training Strategies

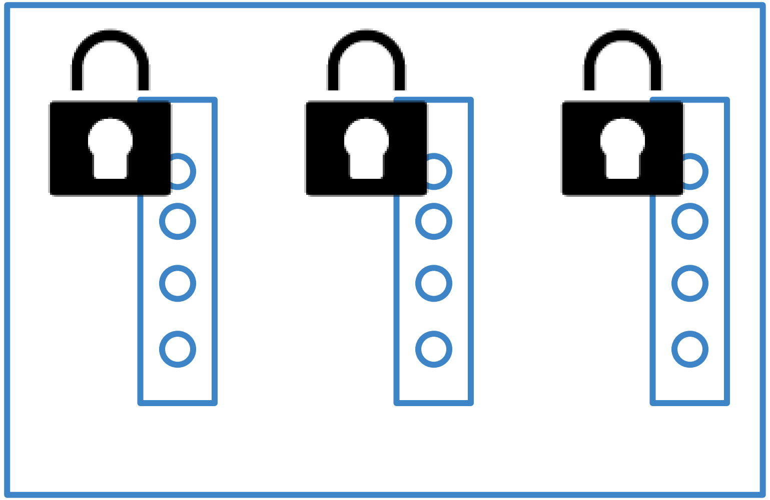
Examples-proportional mixing

Equal mixing

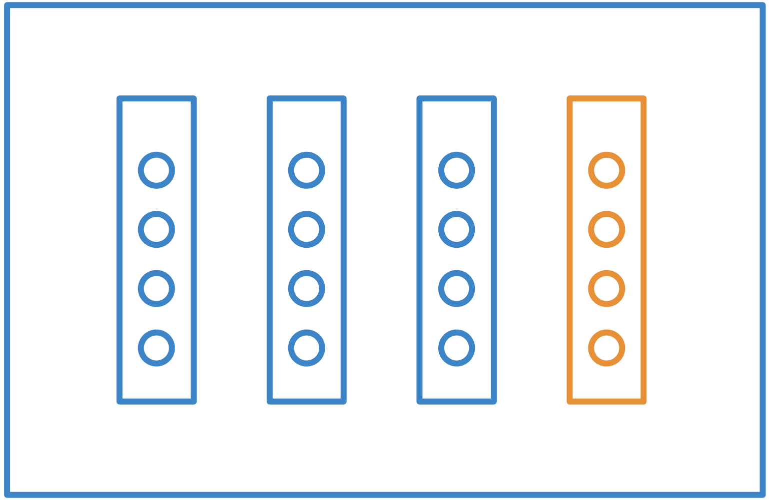


Temperature-scaled mixing

Gradual unfreezing vs. Adapter layers



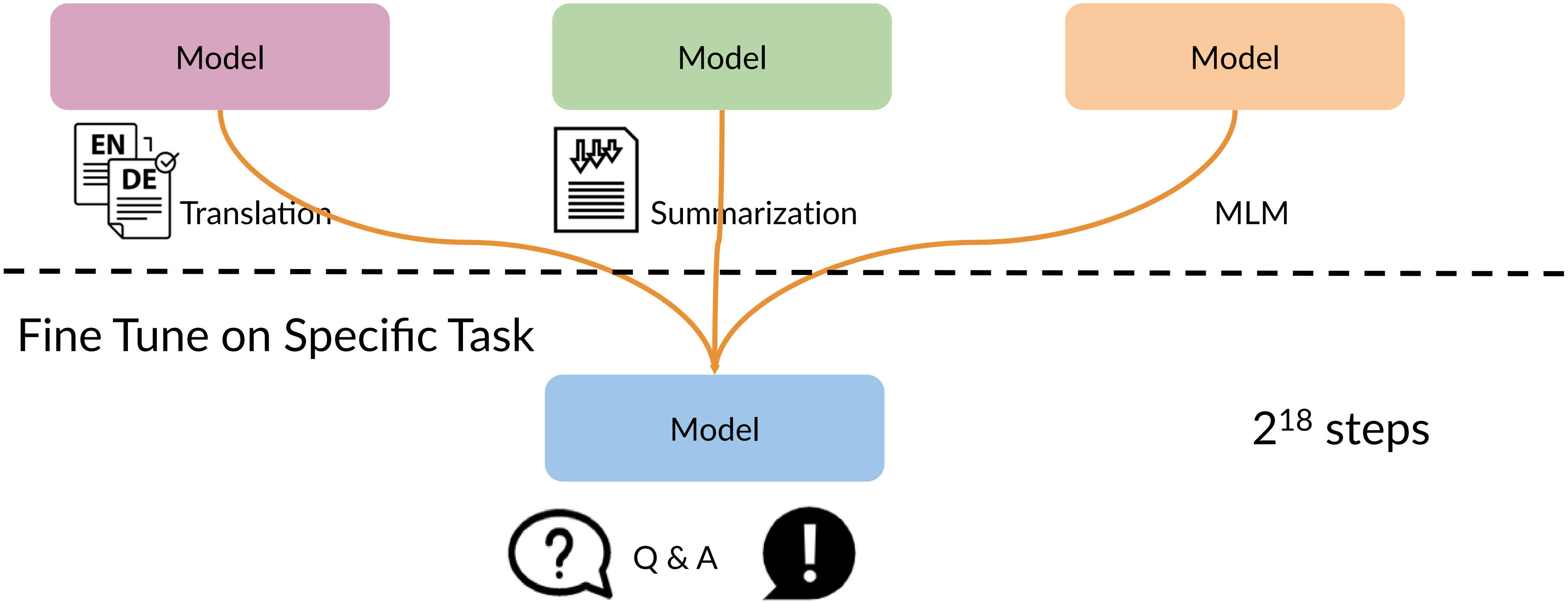
Gradual unfreezing



Adapter layers

Fine-tuning

Pre Training



2^{18} steps

deeplearning.ai

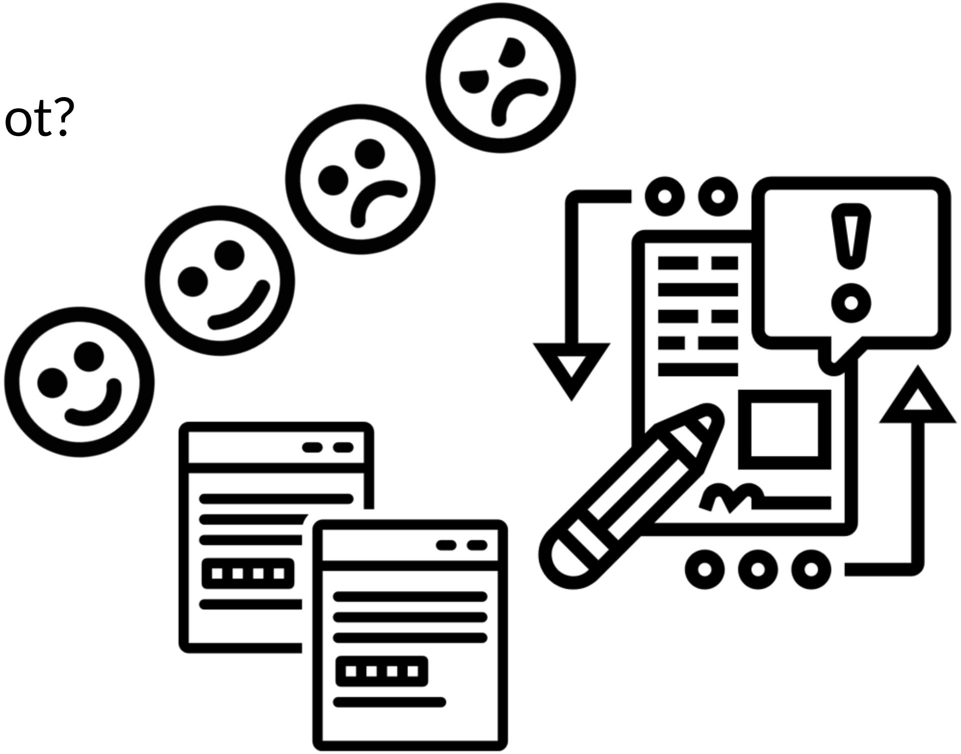
GLUE Benchmark

General Language Understanding Evaluation

- A collection used to train, evaluate, analyze natural language understanding systems
- Datasets with different genres, and of different sizes and difficulties
- Leaderboard

Tasks Evaluated on

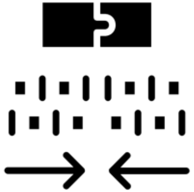
- Sentence grammatical or not?
- Sentiment
- Paraphrase
- Similarity
- Questions duplicates
- Answerable
- Contradiction
- Entailment
- Winograd (co-ref)



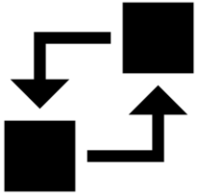
General Language Understanding Evaluation

- Drive research

- Model agnostic



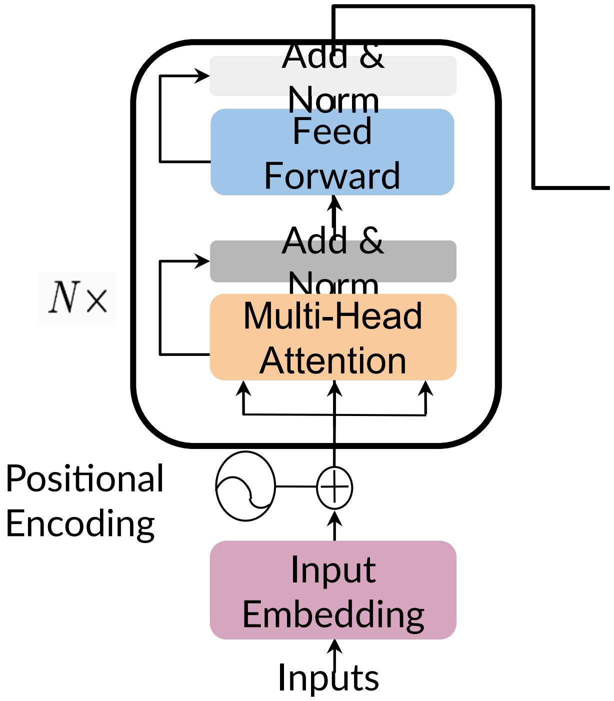
- Makes use of transfer learning



deeplearning.ai

Question Answering

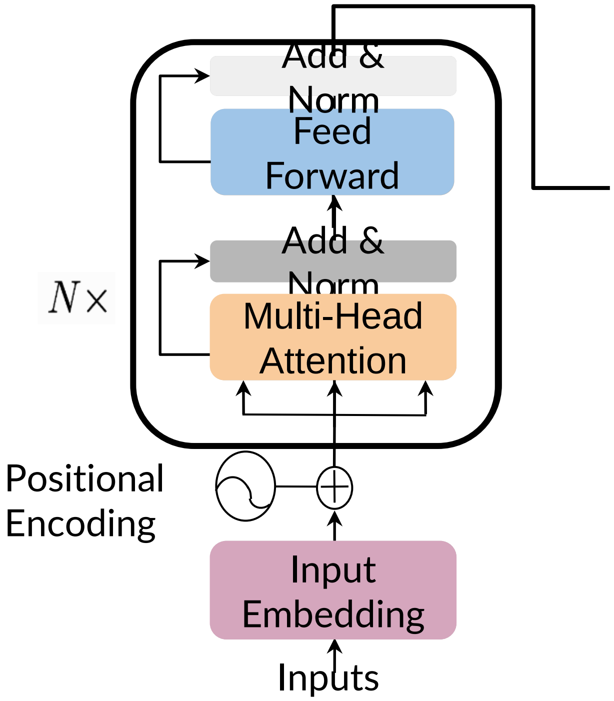
Transformer encoder



Feedforward:

```
[
    LayerNorm,
    dense,
    activation,
    dropout_middle,
    dense,
    dropout_final
]
```

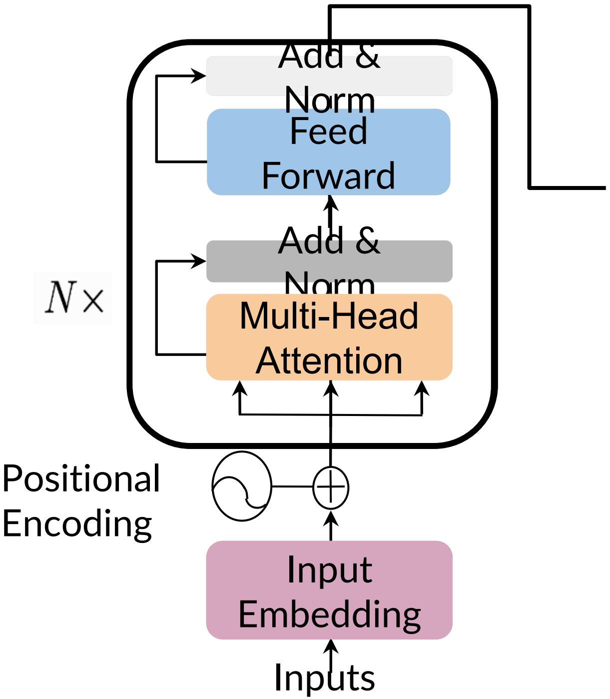
Transformer encoder



Encoder block:

```
[
      Residual(
            LayerNorm,
            attention,
            dropout_,
      ),
      Residual(
            feed_forward,
      ),
 ]
```

Transformer encoder



Feedforward:

```
[
      LayerNorm,
      dense,
      activation,
      dropout_middle,
      dense,
      dropout_final
 ]
```

Encoder block:

```
[
    Residual(
          LayerNorm,
          attention,
          dropout_,
    ),
    Residual(
          feed_forward,
    )
 ]
```

Data examples

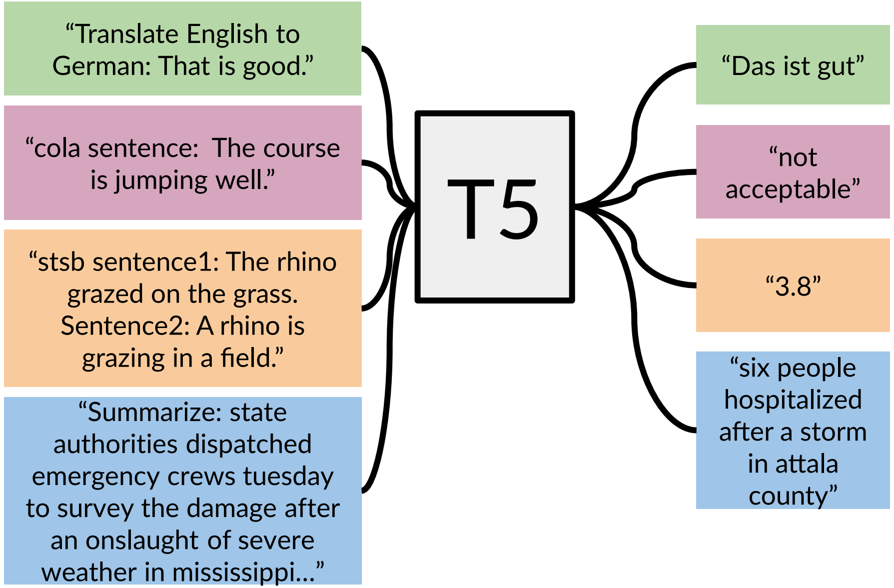
Question: What percentage of the French population today is non - European ?

Context: Since the end of the Second World War , France has become an ethnically diverse country . Today , **approximately five percent** of the French population is non - European and non - white . This does not approach the number of non - white citizens in the United States (roughly 28 – 37 % , depending on how Latinos are classified ; see Demographics of the United States) . Nevertheless , it amounts to at least three million people , and has forced the issues of ethnic diversity onto the French policy agenda . France has developed an approach to dealing with ethnic problems that stands in contrast to that of many advanced , industrialized countries . Unlike the United States , Britain , or even the Netherlands , France maintains a " color - blind " model of public policy . This means that it targets virtually no policies directly at racial or ethnic groups . Instead , it uses geographic or class criteria to address issues of social inequalities . It has , however , developed an extensive anti - racist policy repertoire since the early 1970s . Until recently , French policies focused primarily on issues of hate speech — going much further than their American counterparts — and relatively less on issues of discrimination in jobs , housing , and in provision of goods and services .

Target: Approximately five percent

Implementing Q&A with T5

- Load a pre-trained model
- Process data to get the required inputs and outputs: "question: Q context: C" as input and "A" as target
- Fine tune your model on the new task and input
- Predict using your own model



deeplearning.ai

Hugging Face: Introduction

Outline

- What is Hugging Face?
- How you can use the Hugging Face ecosystem

Hugging Face

Transformers library

Use it with

TensorFlow

PyTorch

Use it for

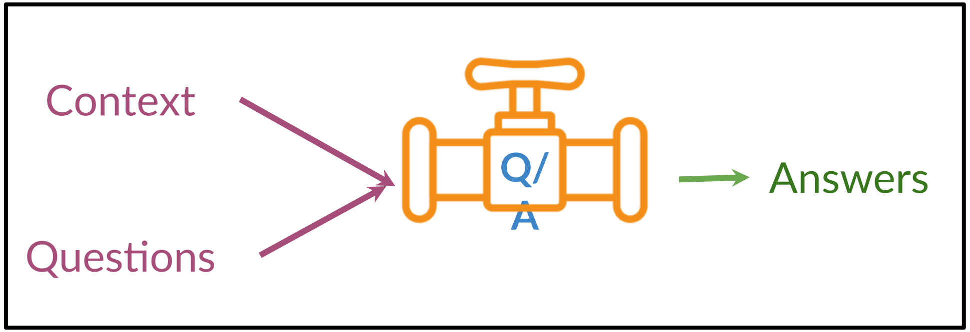
Applying state of the art transformer models

Fine-tuning pretrained transformer models

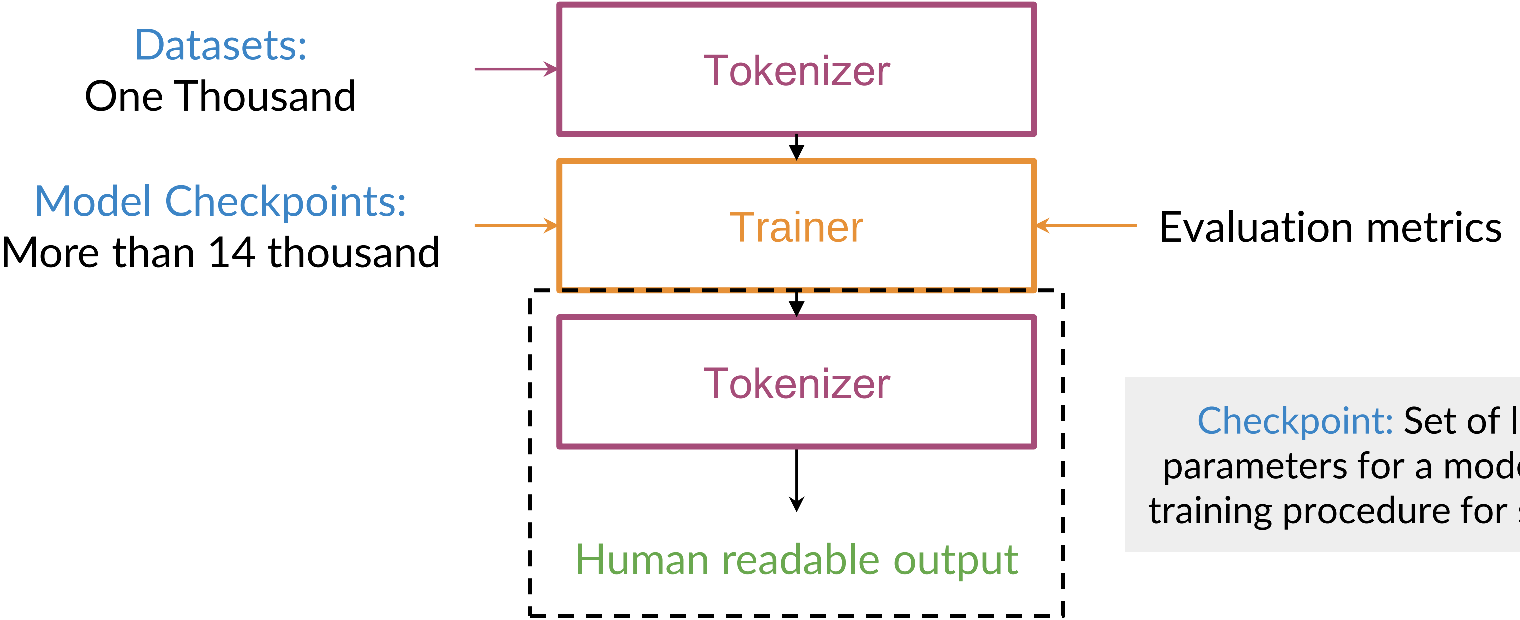
Hugging Face: Using Transformers

Pipelines

1. **Pre-processing** your inputs
2. **Running** the model
3. **Post-processing** the outputs



Hugging Face: Fine-Tuning Transformers



Checkpoint: Set of learned parameters for a model using a training procedure for some task

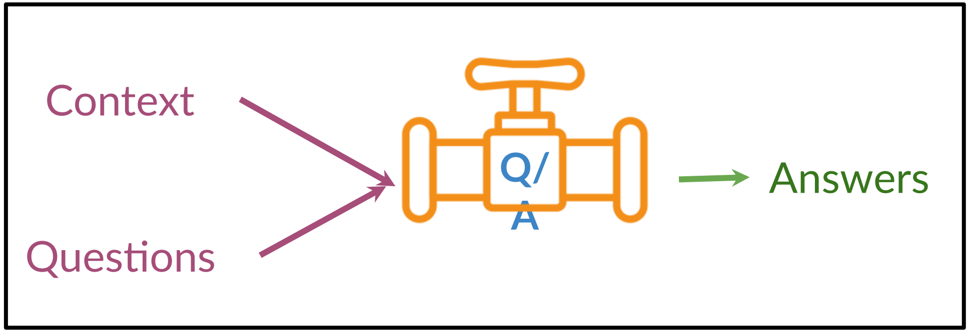
deeplearning.ai

Hugging Face: Using Transformers

Using Transformers

Pipelines

1. **Pre-processing** your inputs
2. **Running** the model
3. **Post-processing** the outputs



Tasks

Pipelines

Initialization → Task

Initialization → Model Checkpoint

Use → Inputs for the task

Sentiment Analysis	Question Answering	Fill-Mask
Sequence	Context and questions	Sentence and position

Checkpoints

Huge number of model checkpoints that you can use in your pipelines.

But **beware**, not every checkpoint would be suitable for your task.

Model Hub

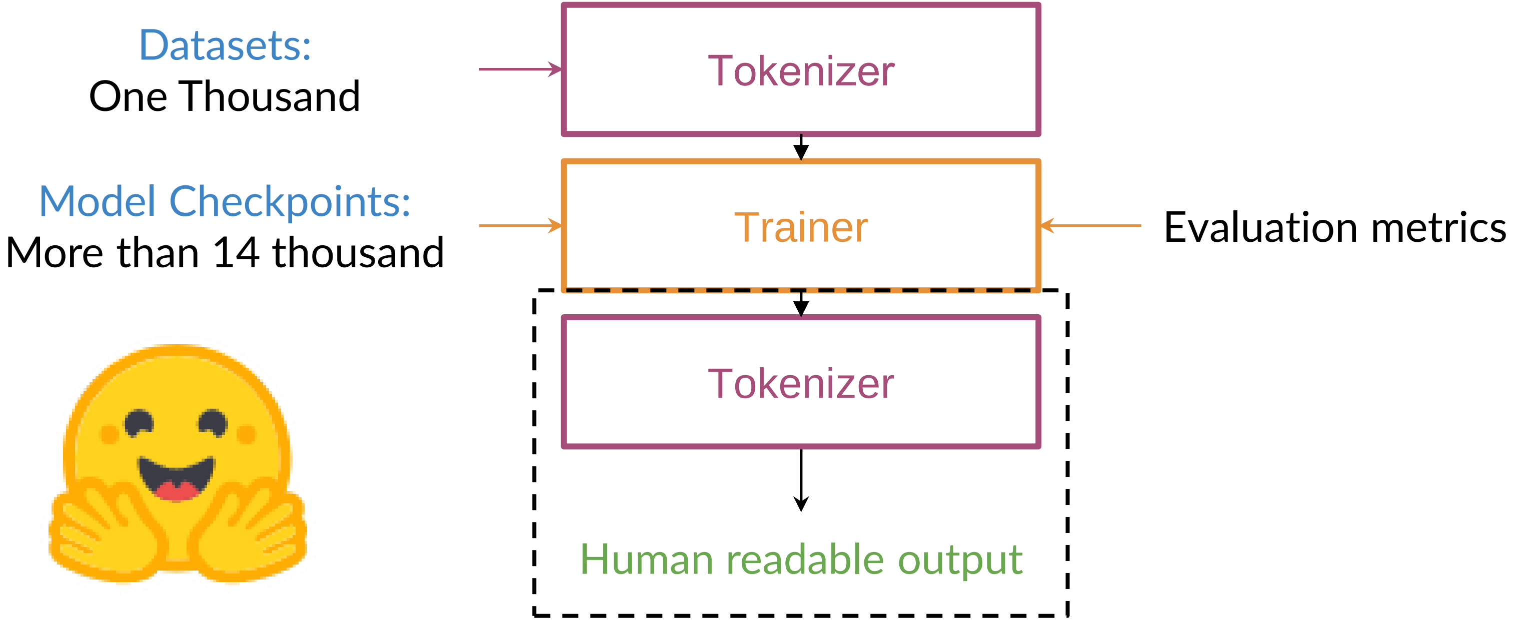
Hub containing models that you can use in your pipelines according to the task you need: https://huggingface.co/models

Model Card shows a description of your selected model and useful information such as code snippet examples.

deeplearning.ai

Hugging Face: Fine-Tuning Transformers

Fine-Tuning Tools



Model Checkpoints

Model Checkpoints:
More than 15 thousand
(and increasing)

Upload the architecture and weights with 1 line of code!

Model	Dataset	Name in 🤗
DistilBERT	Stanford Question Answering Dataset (SQuAD)	distilbert-base-cased-distilled-squad
BERT	Wikipedia and Book Corpus	bert-base-cased
...	...	...

Datasets

Datasets:
One Thousand

Load them using just one function

Optimized to work with massive amounts of data!

Tokenizers

"What well-known superheroes were introduced between 1939 and 1941 by Detective Comics?"

Tokenizer

[101, 1327, 1218, 118,
1227, 18365, 1279, 1127,
2234, 1206, 3061, 1105,
3018, 1118, 9187, 7452,
136, 102]

Depending on the use case, you might need to run additional steps.

Trainer and Evaluation Metrics

Trainer object let's you define the training procedure

Number of epochs

Warm-up steps

Weight decay

...

Train using one line of code!

Pre-defined evaluation metrics, like BLEU and ROUGE