deeplearning.ai

Week 3 Overview

Week 3

Week 3

Question
Answering

Week 3

Question Answering

Transfer learning

Week 3

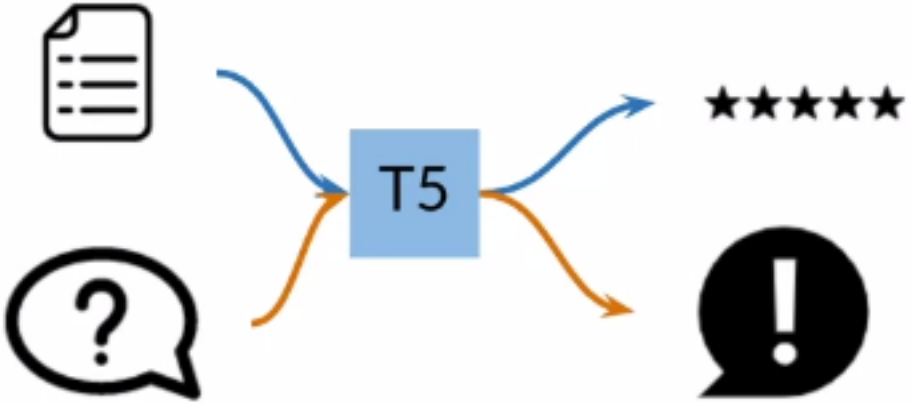
Question Answering

Transfer learning

Week 3

Question Answering

Transfer learning



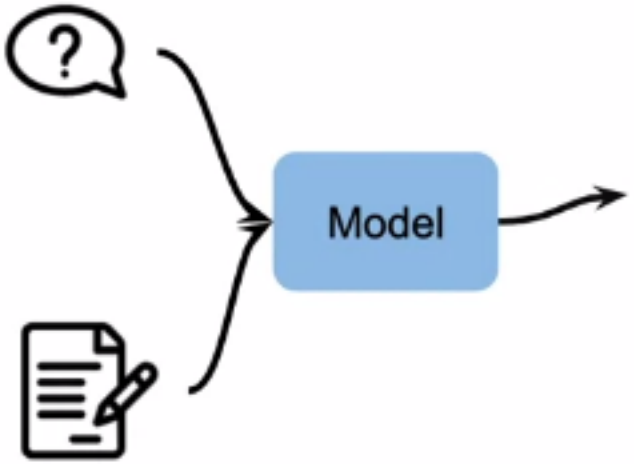
Question Answering

Context-based

Closed book

Question Answering

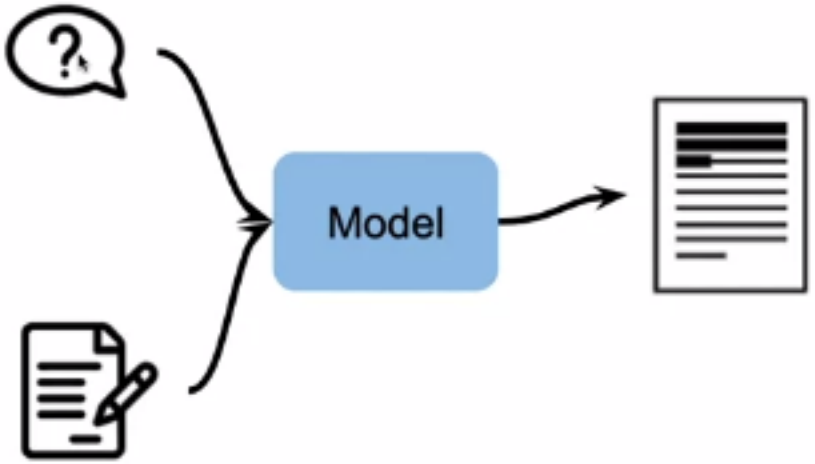
Context-based



Closed book

Question Answering

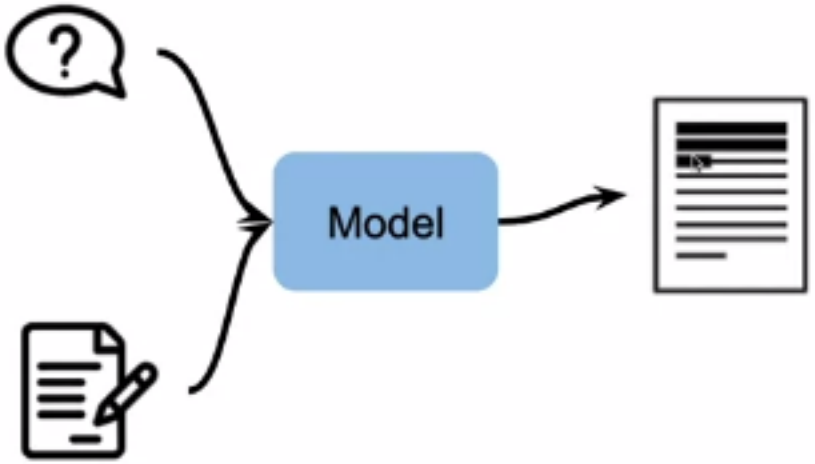
Context-based



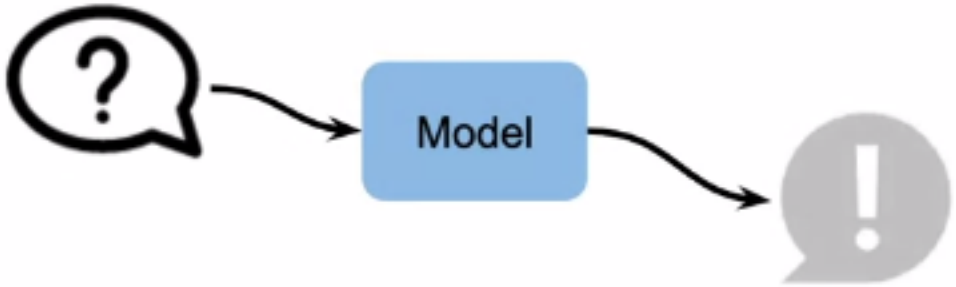
Closed book

Question Answering

Context-based



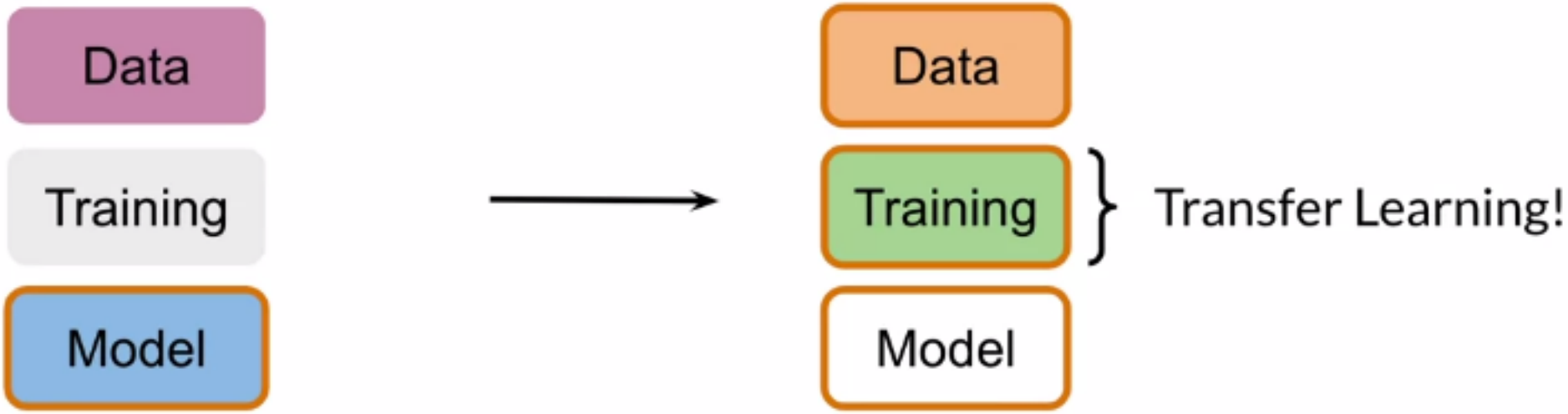
Closed book



Not just the model

Not just the model

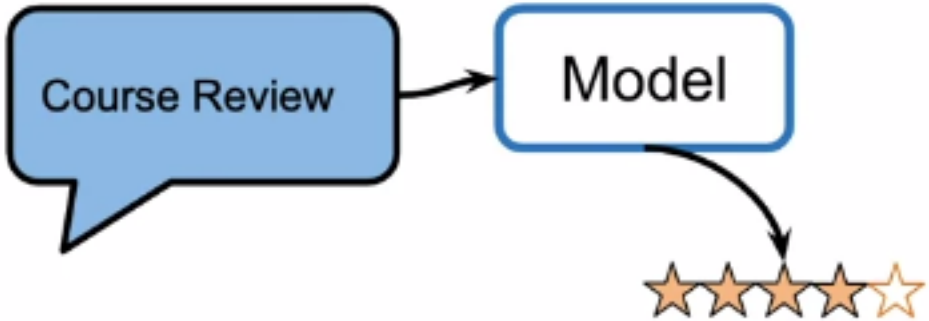
Not just the model



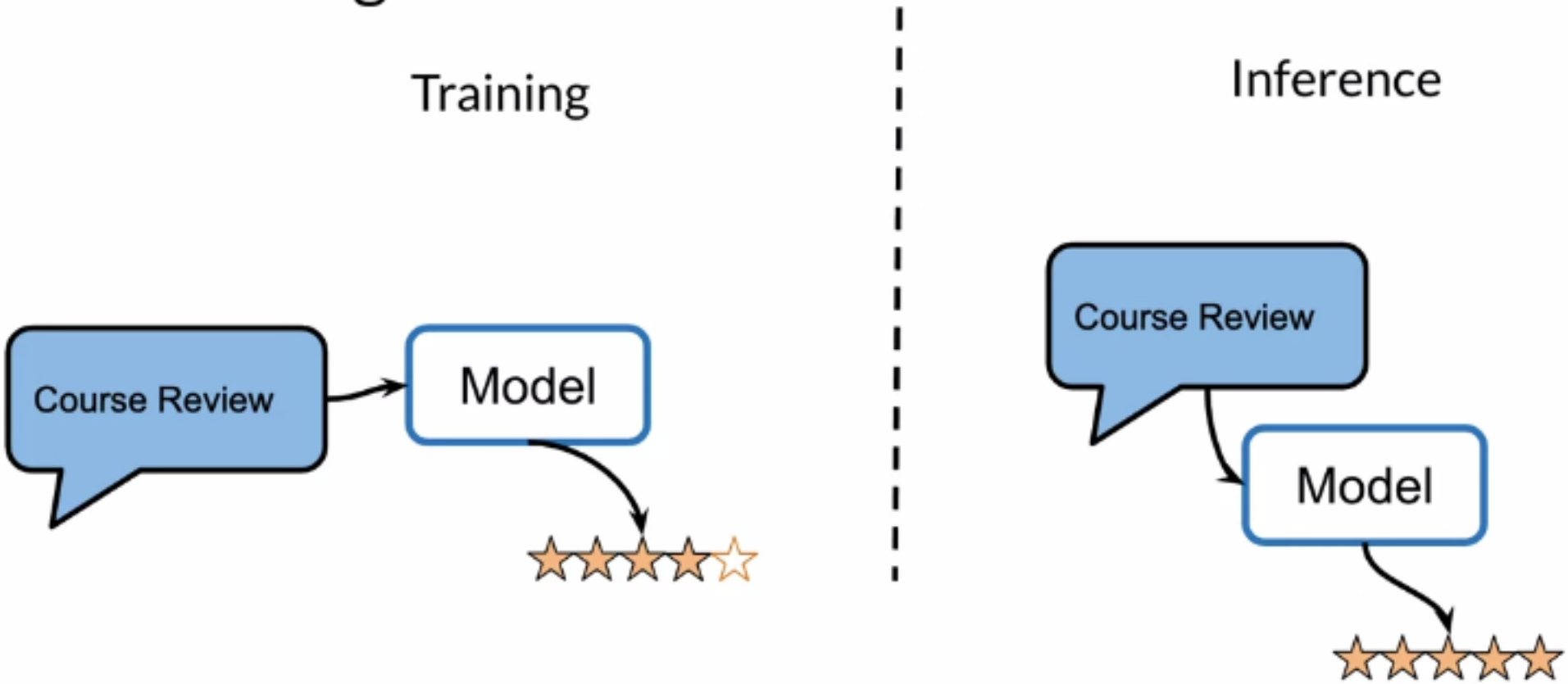
Classical training

Training

Inference



Classical training



Transfer learning

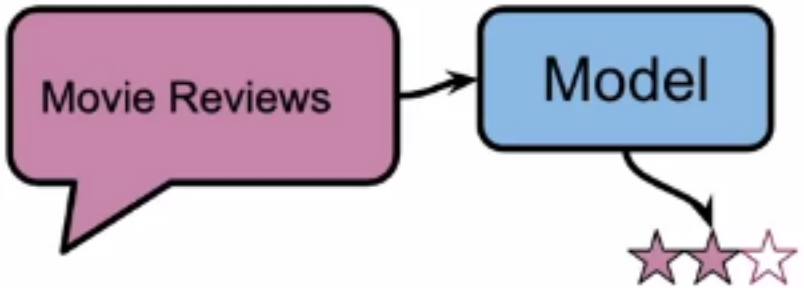
Pre-training

Inference

Training
on “Downstream” Task

Transfer learning

Pre-training



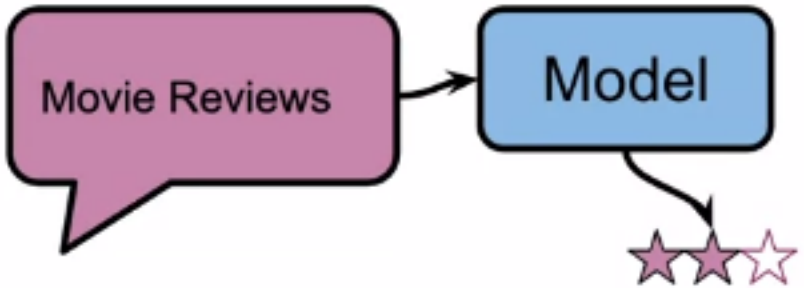
Inference

Training
on “Downstream” Task

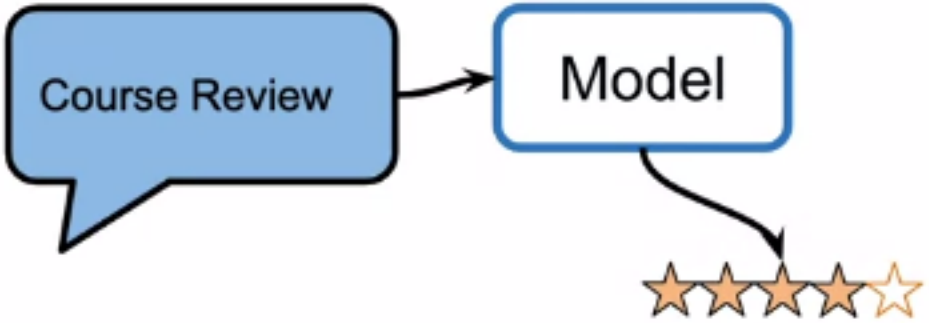
Transfer learning

Inference

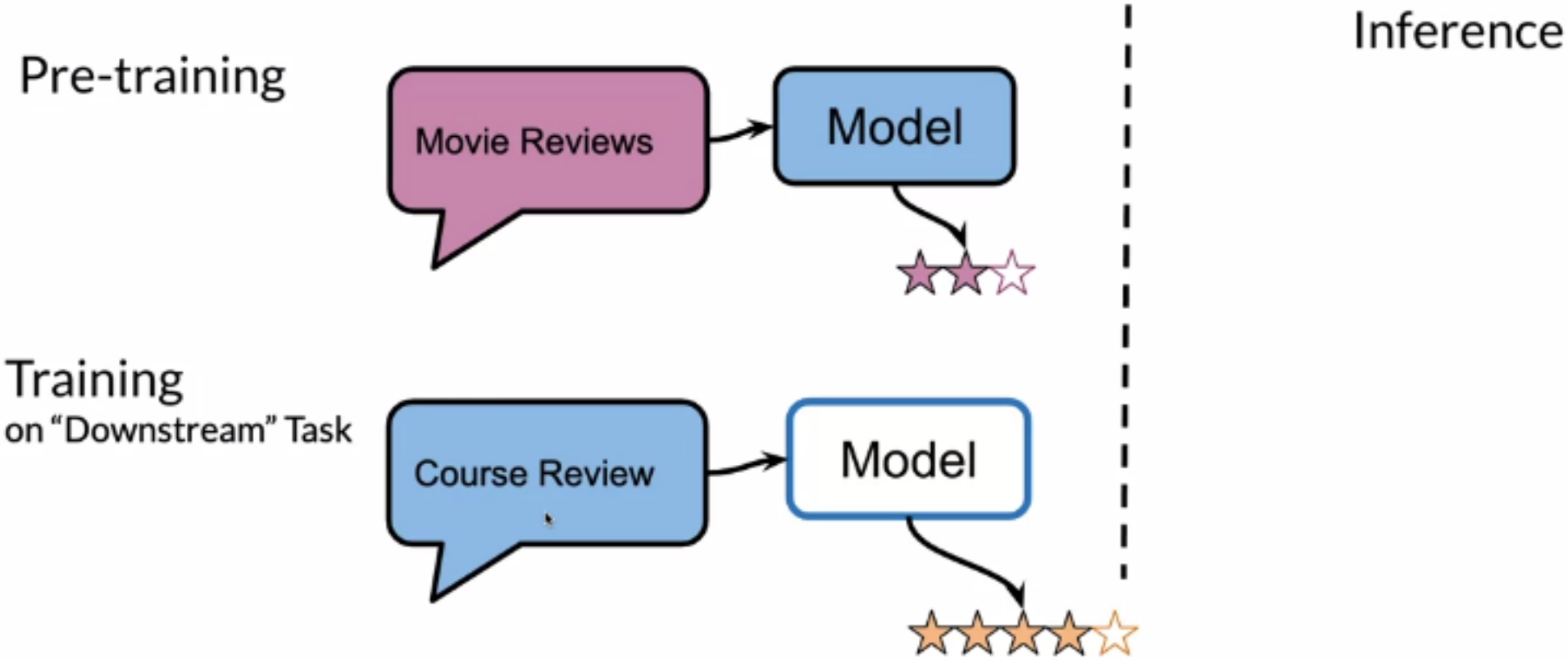
Pre-training



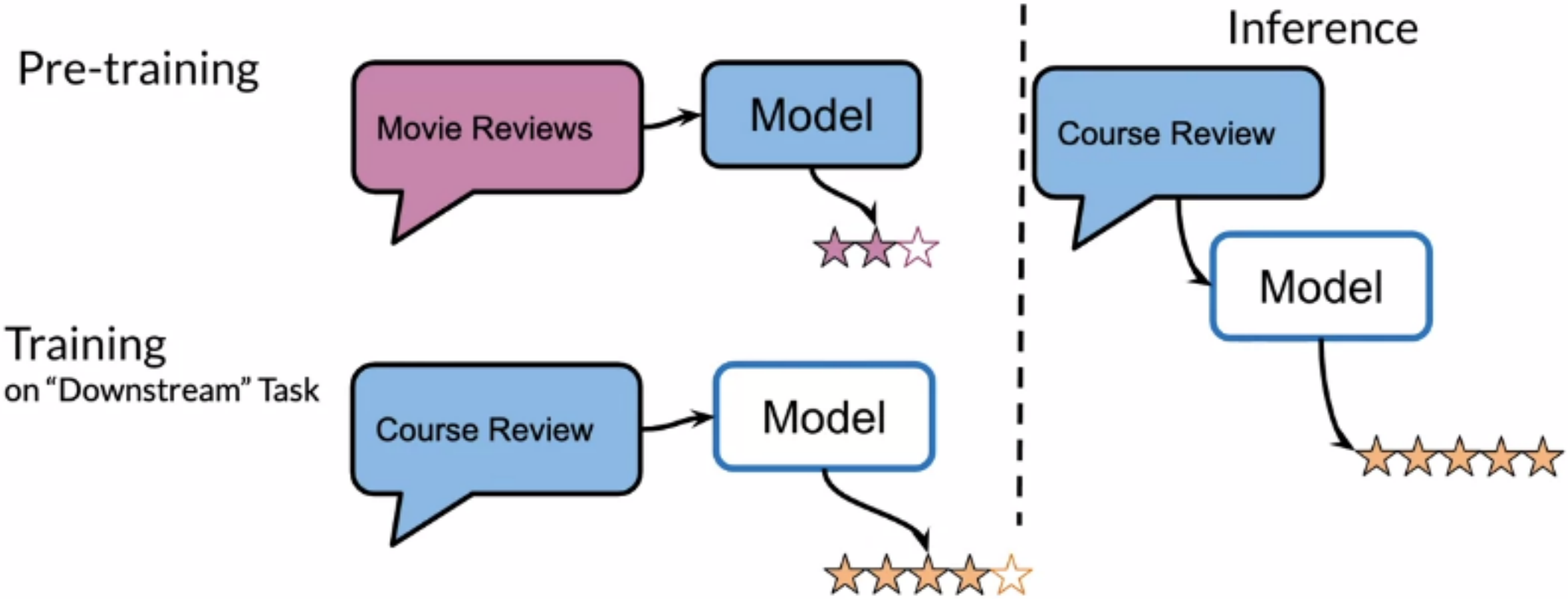
Training
on "Downstream" Task



Transfer learning



Transfer learning



Transfer Learning: Different Tasks

Inference

Pre-Training

Sentiment
Classification

Training

Downstream task:
Question Answering

Transfer Learning: Different Tasks

Inference

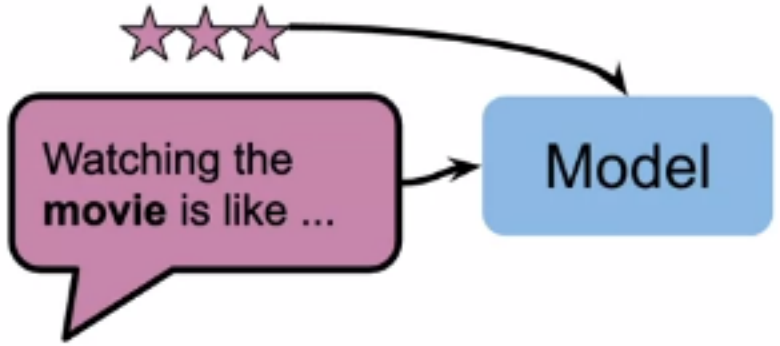
Pre-Training

Sentiment
Classification

Training

Downstream task:
Question Answering

Transfer Learning: Different Tasks

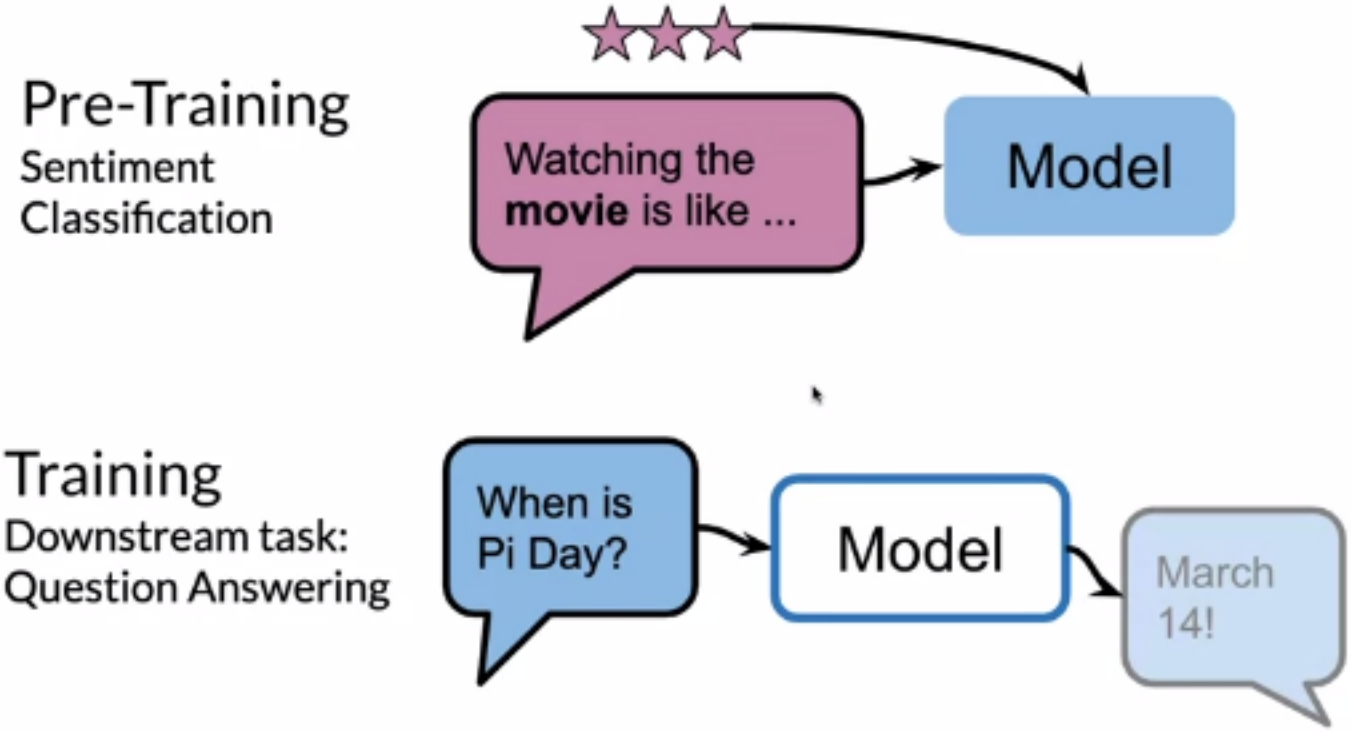


Pre-Training
Sentiment
Classification

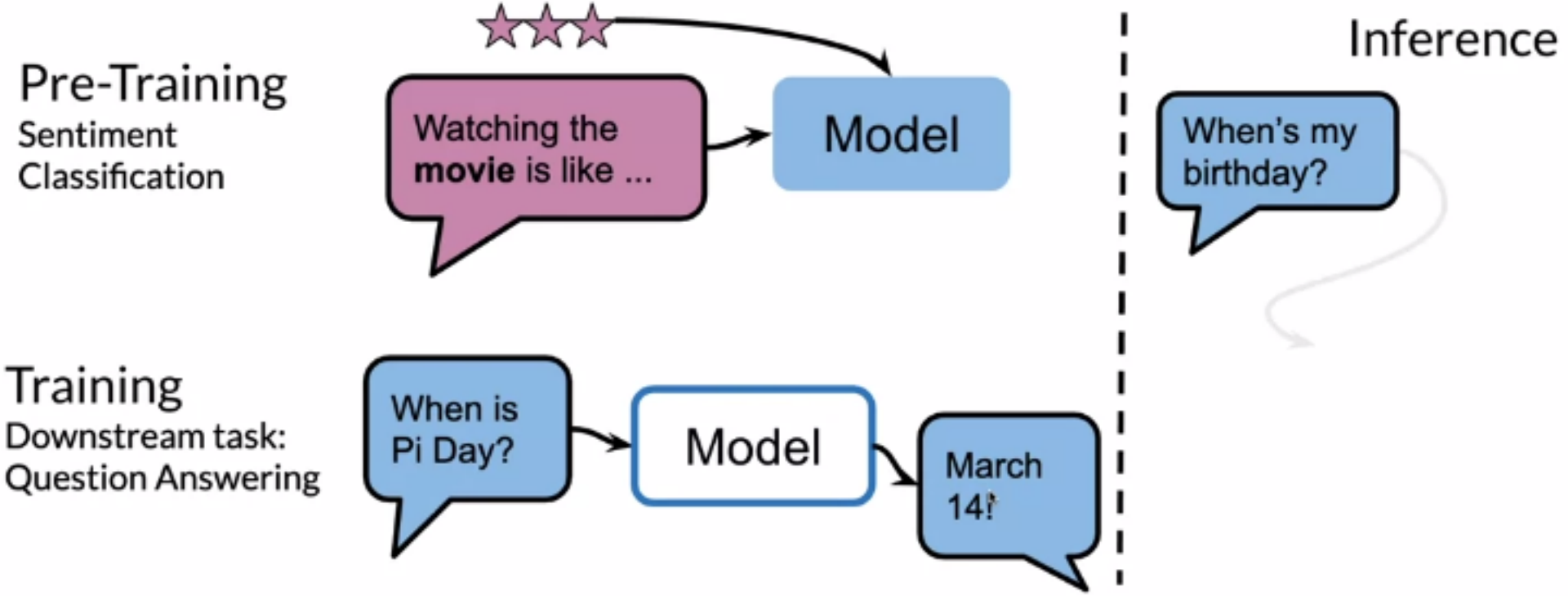
Inference

Training
Downstream task:
Question Answering

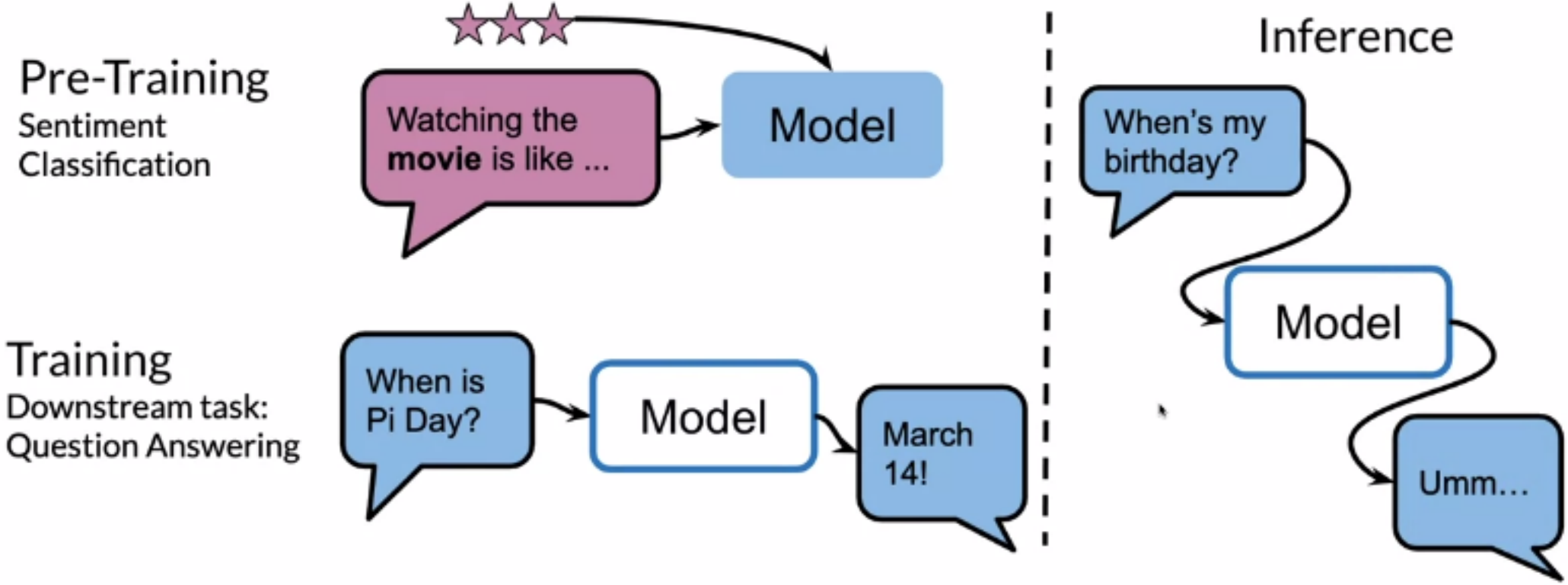
Transfer Learning: Different Tasks



Transfer Learning: Different Tasks



Transfer Learning: Different Tasks



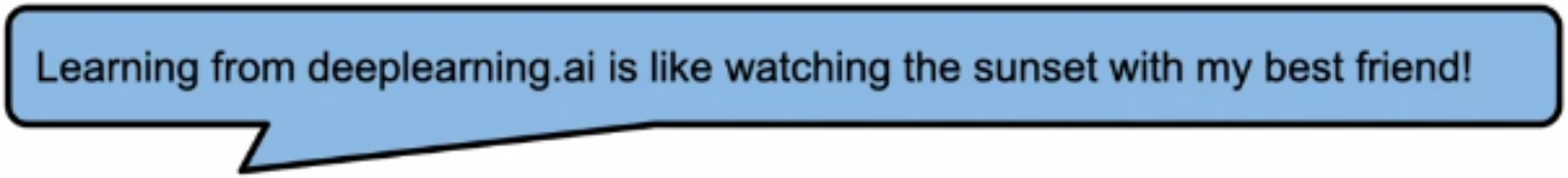
BERT: Bi-directional Context

Uni-directional

Bi-directional

BERT: Bi-directional Context

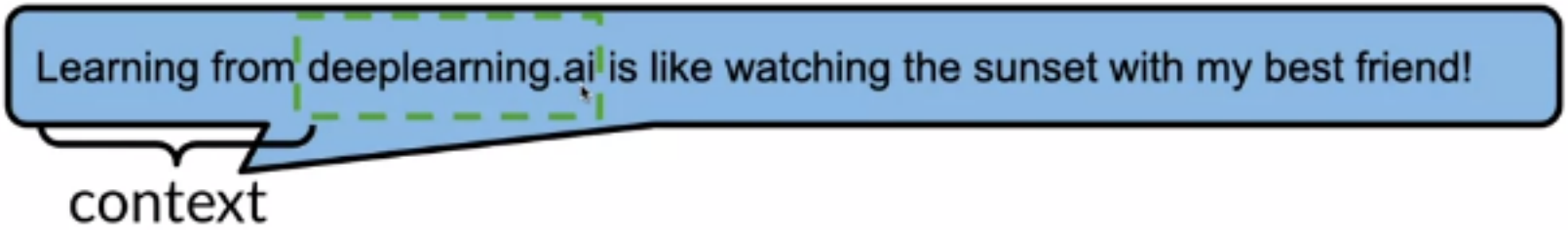
Uni-directional



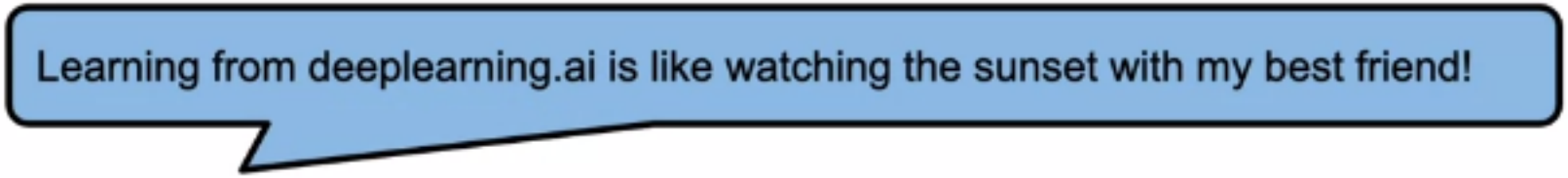
Bi-directional

BERT: Bi-directional Context

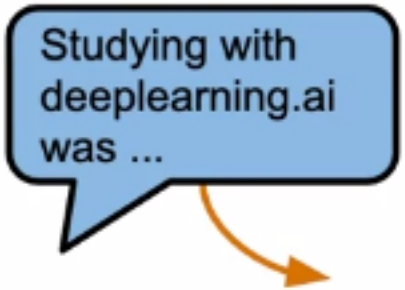
Uni-directional



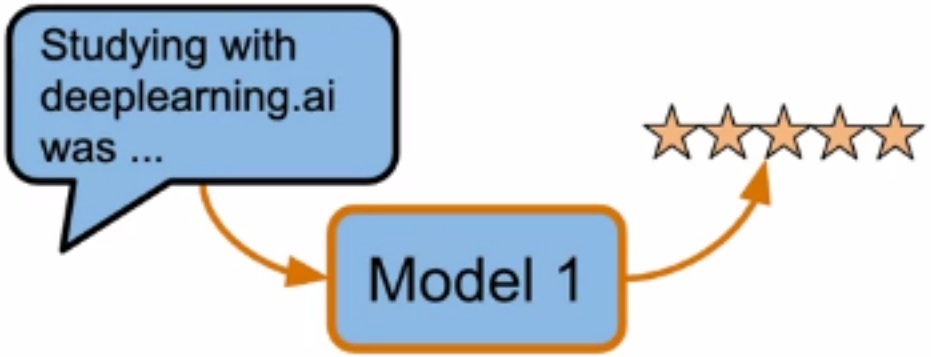
Bi-directional



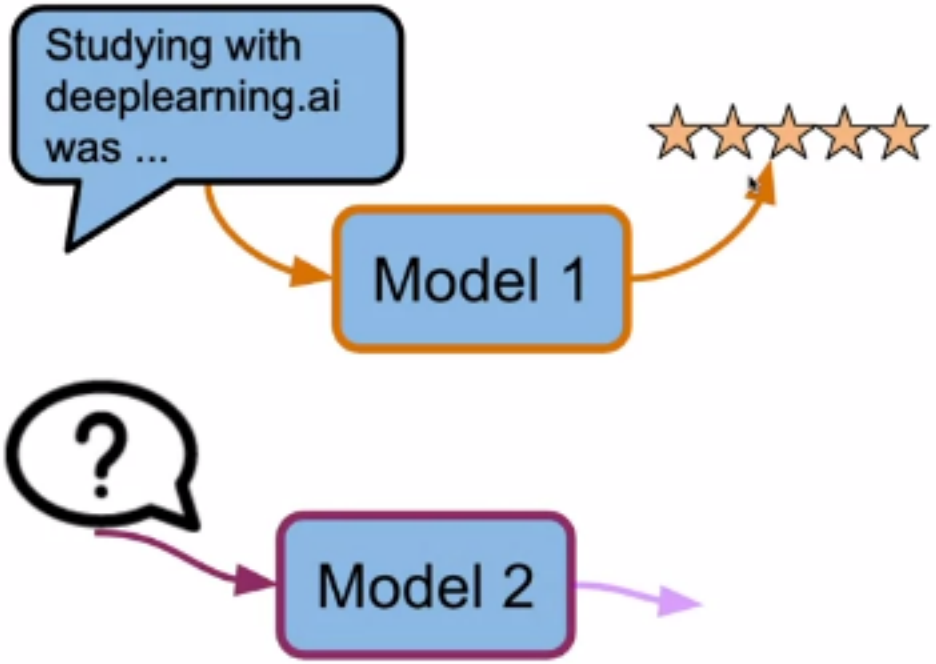
T5: Single task vs. Multi task



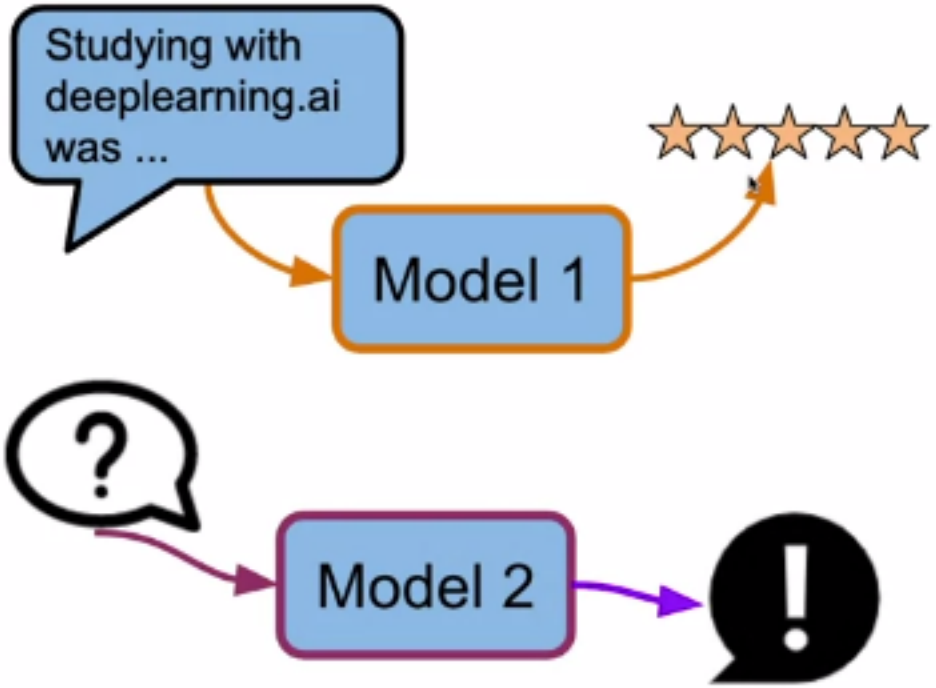
T5: Single task vs. Multi task



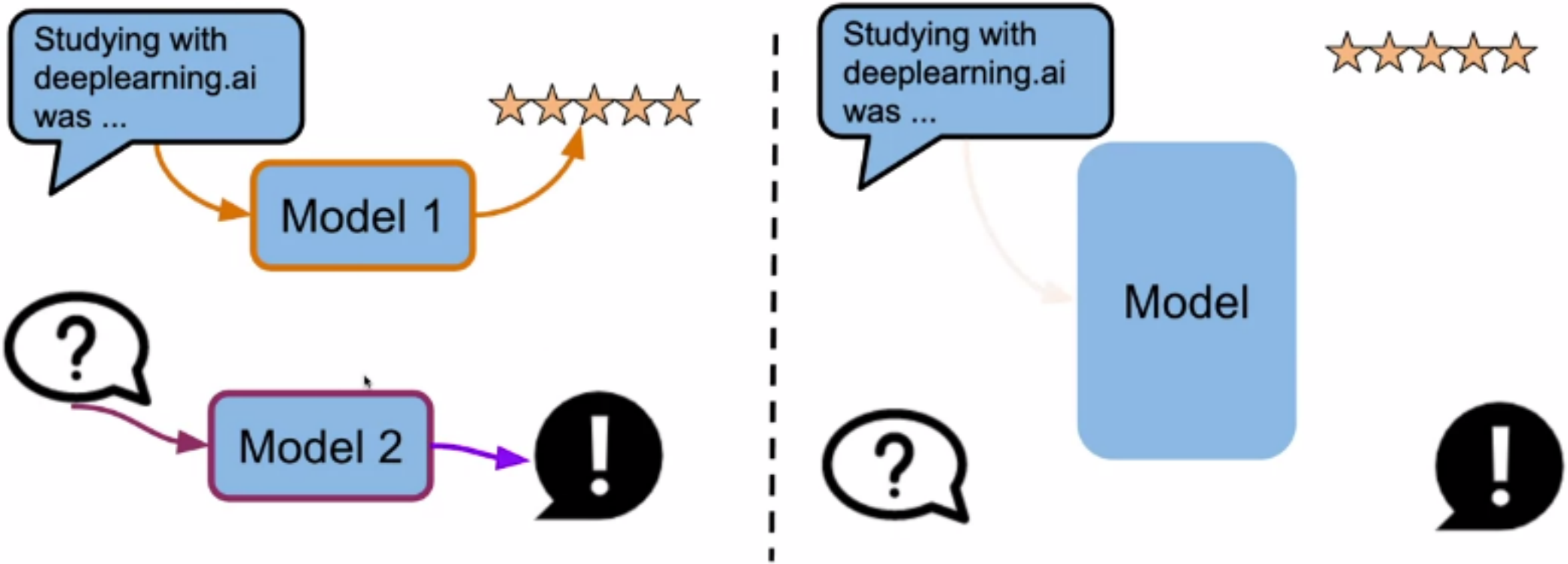
T5: Single task vs. Multi task



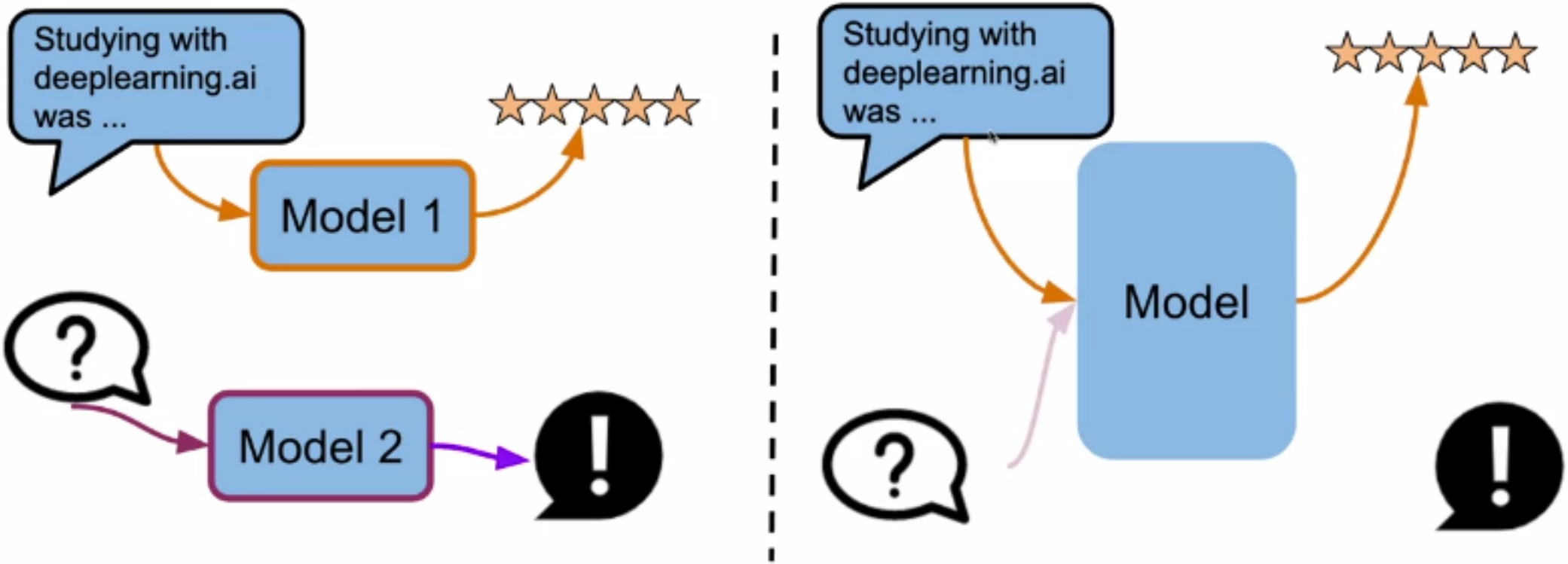
T5: Single task vs. Multi task



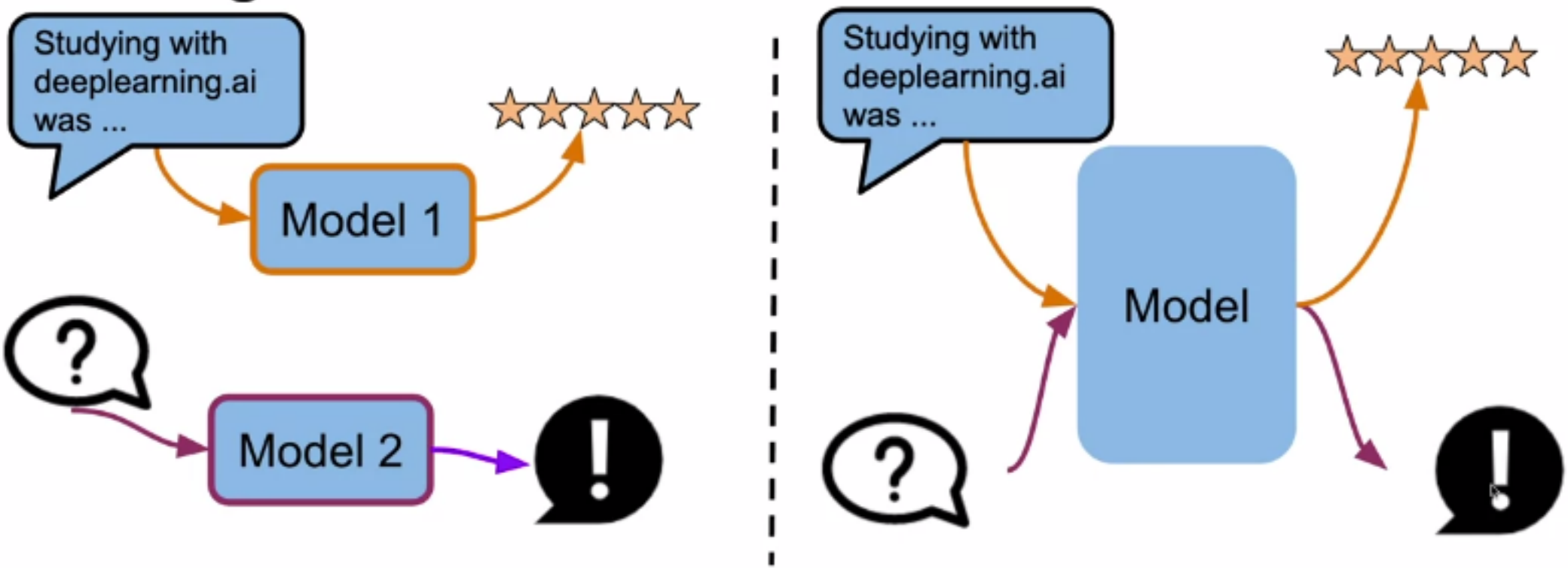
T5: Single task vs. Multi task



T5: Single task vs. Multi task



T5: Single task vs. Multi task



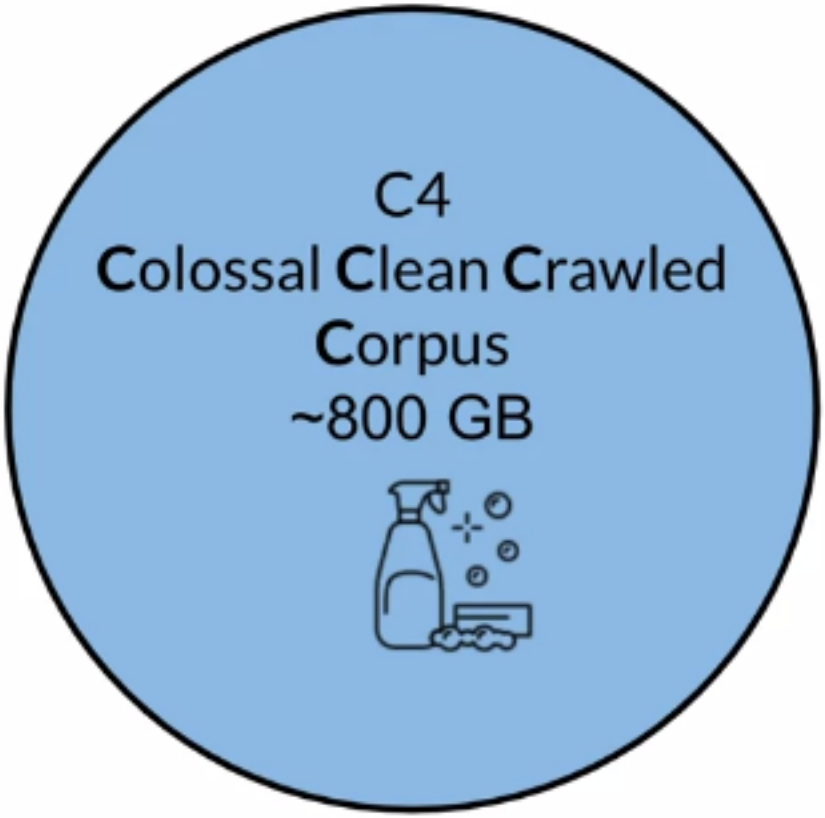
T5: more data, better performance

English wikipedia
~13 GB

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

T5: more data, better performance

English wikipedia
~13 GB



Desirable Goals

Transfer Learning!

Desirable Goals

- Reduce training time

Transfer Learning!

Desirable Goals

- Reduce training time

- Improve predictions

Transfer Learning!

Desirable Goals

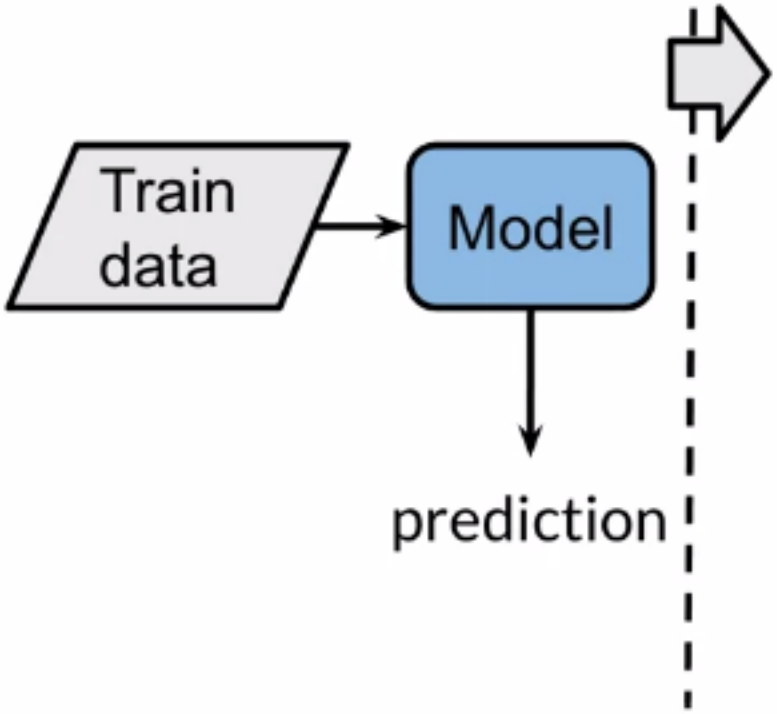
- Reduce training time

- Improve predictions

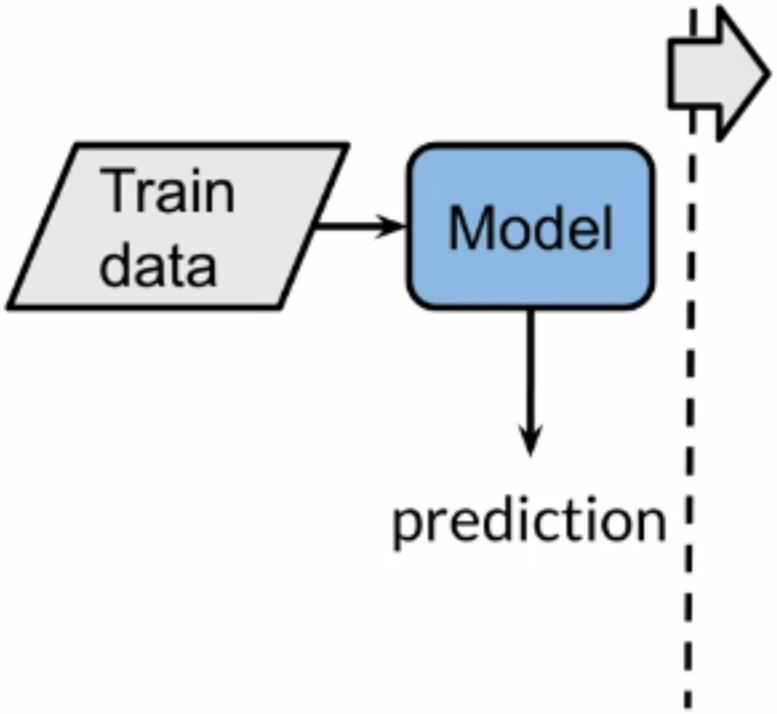
- Small datasets

Transfer Learning!

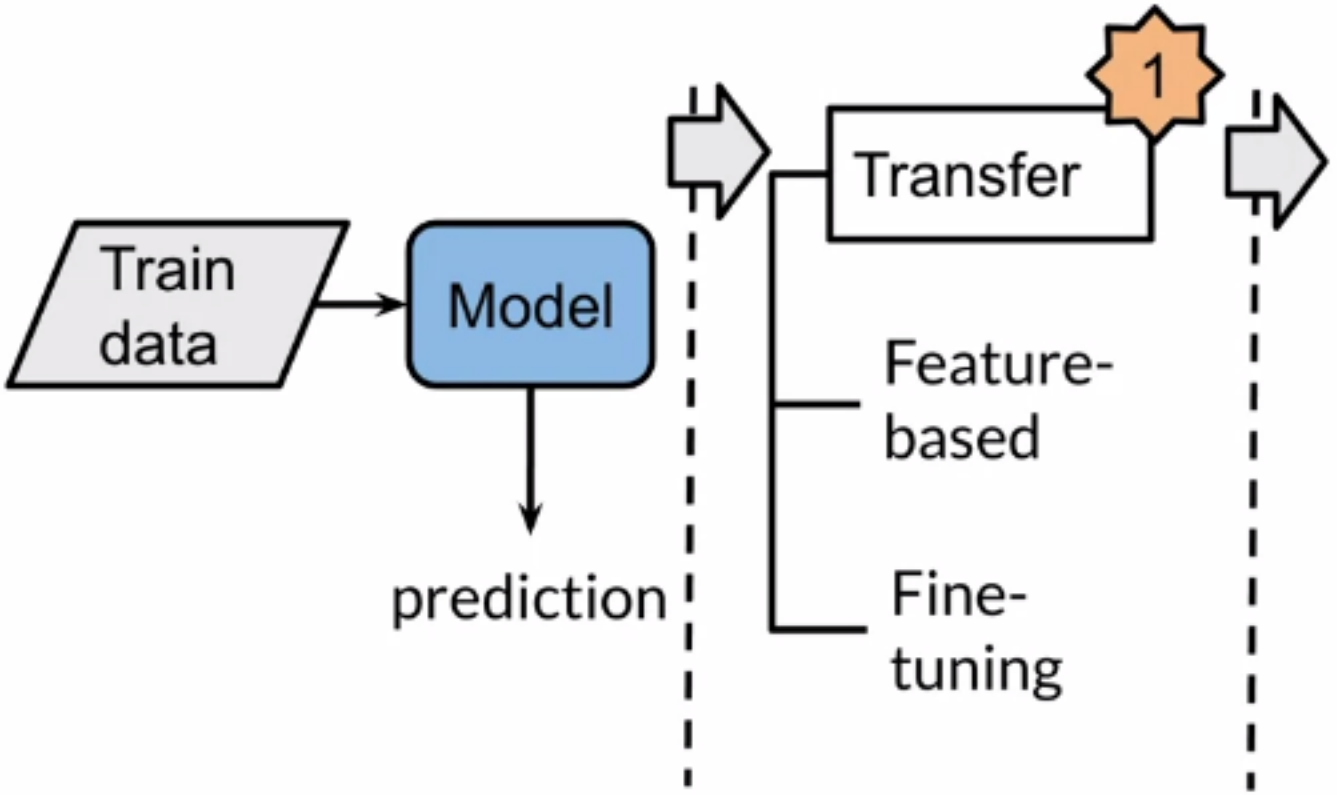
Transfer learning options



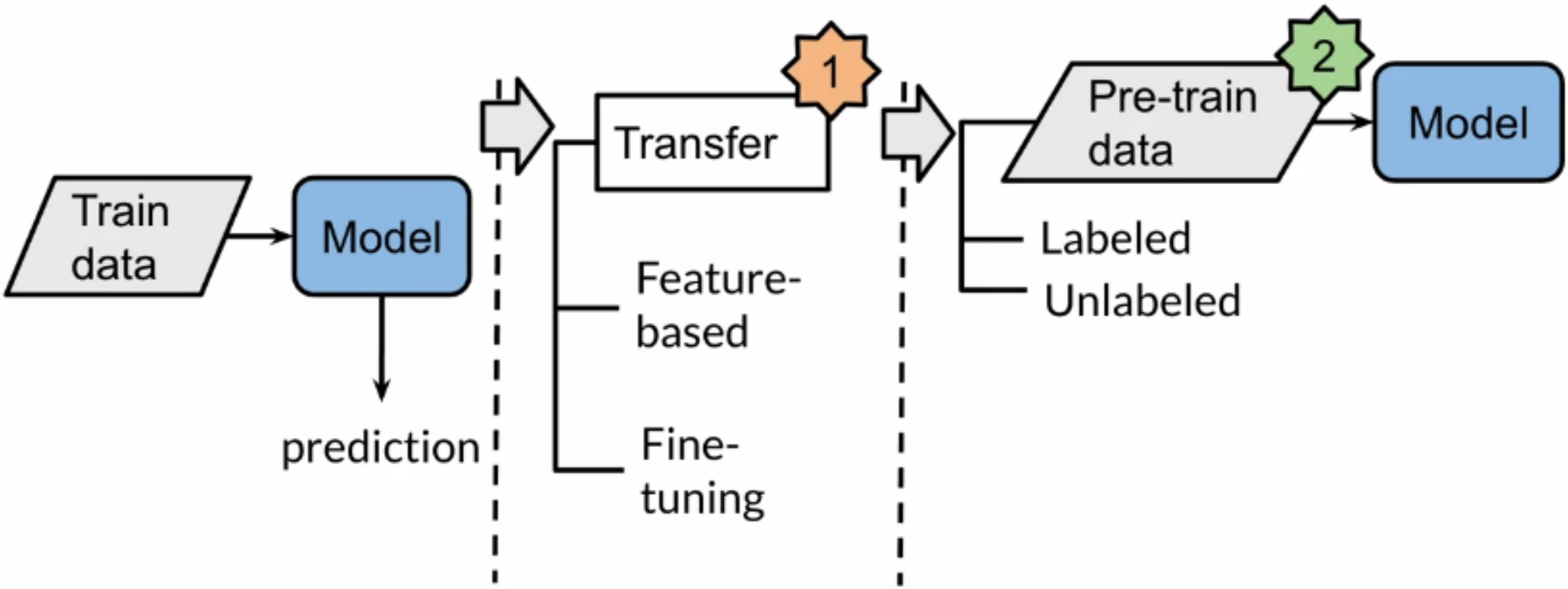
Transfer learning options



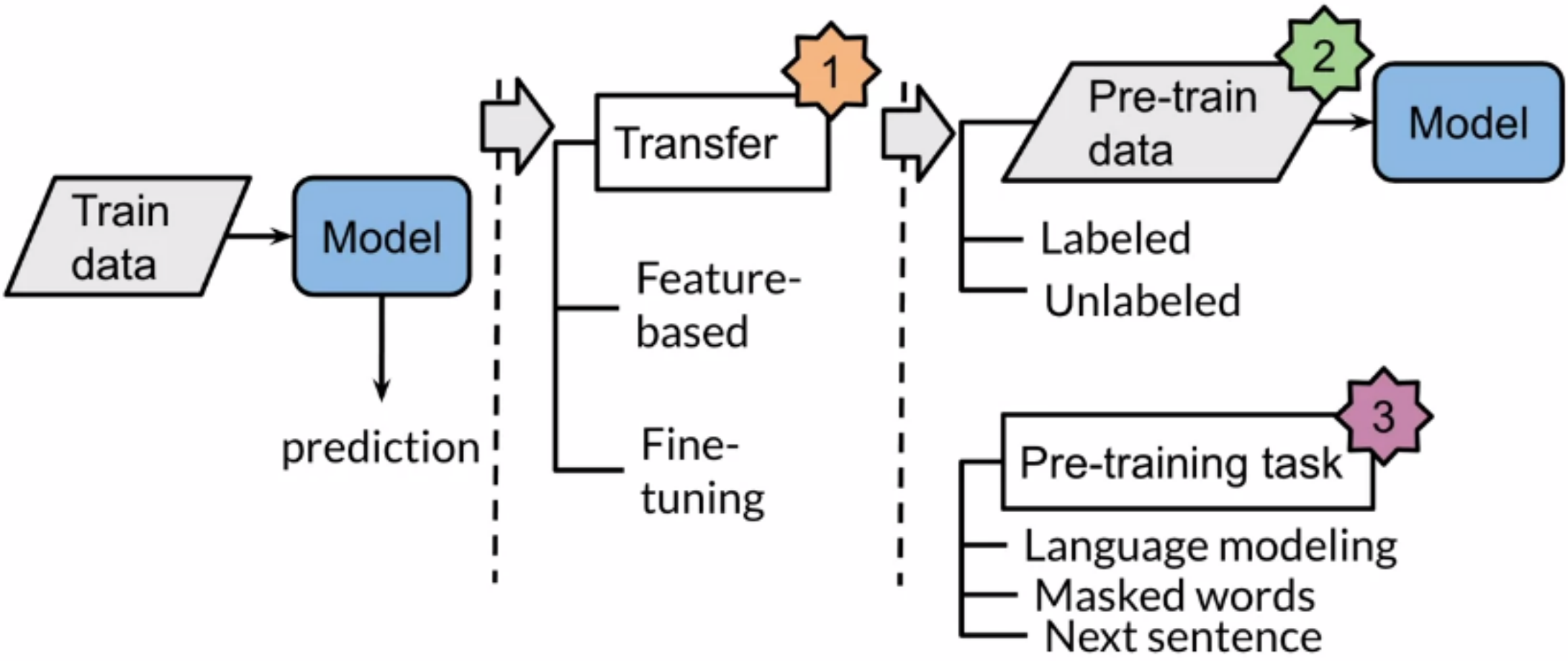
Transfer learning options



Transfer learning options



Transfer learning options



General purpose learning

Transfer 1

I am | because I am learning

"Happy"

Transfer 1

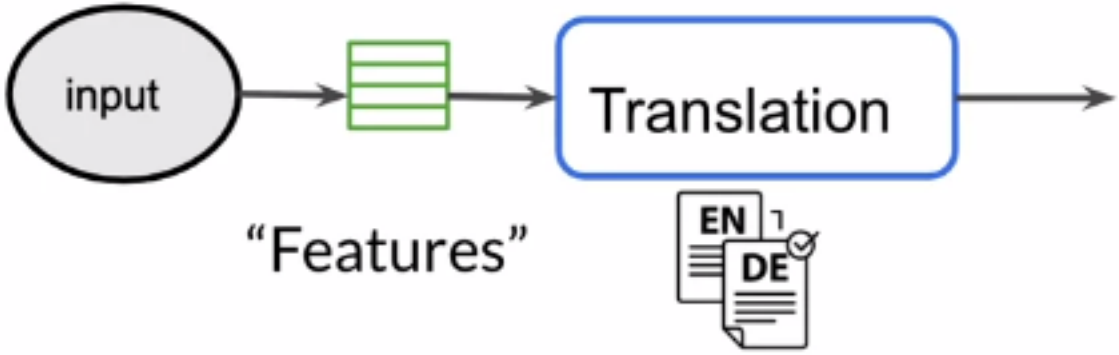
General purpose learning

Word Embeddings 

Transfer 1

General purpose learning

Word Embeddings



Feature-based vs. Fine-Tuning

Pre-Train

Pre-Train

Feature-based vs. Fine-Tuning

Pre-Train

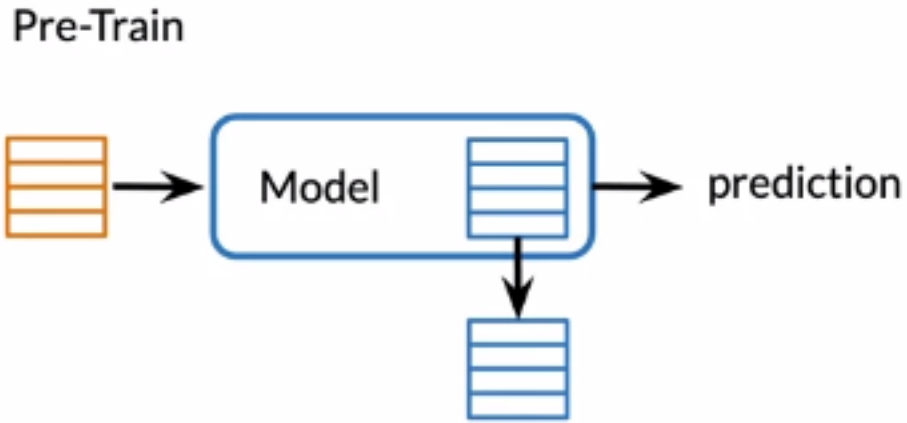
Pre-Train

Feature-based vs. Fine-Tuning

Pre-Train

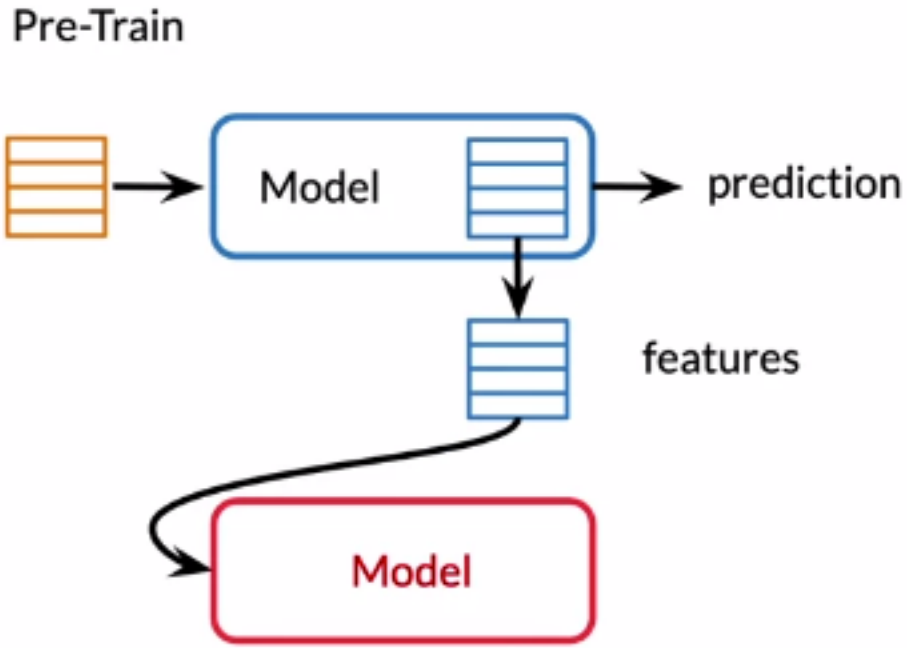
Pre-Train

Feature-based vs. Fine-Tuning



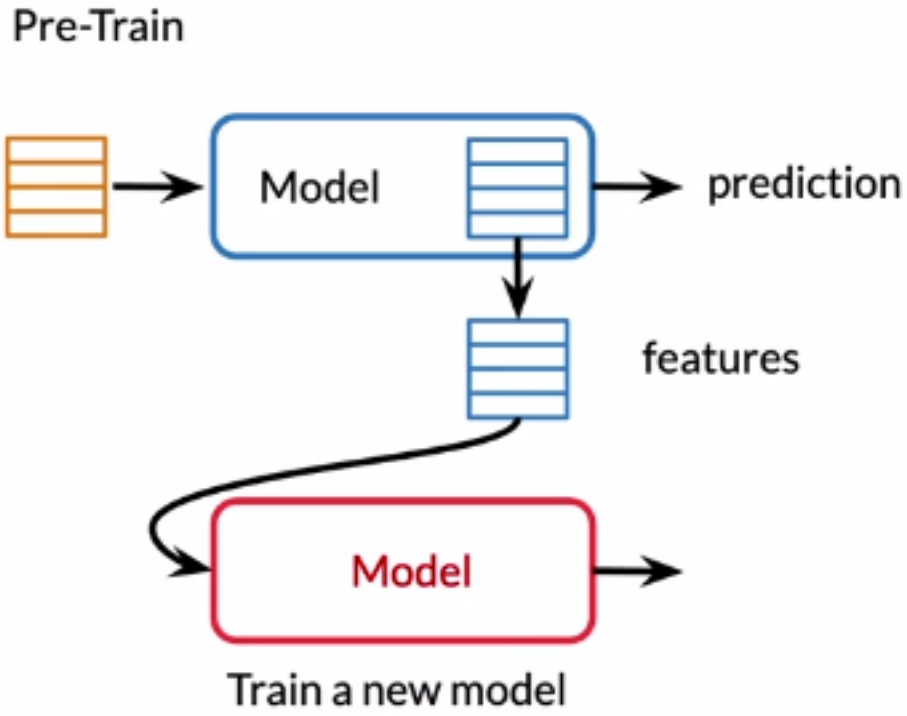
Pre-Train

Feature-based vs. Fine-Tuning



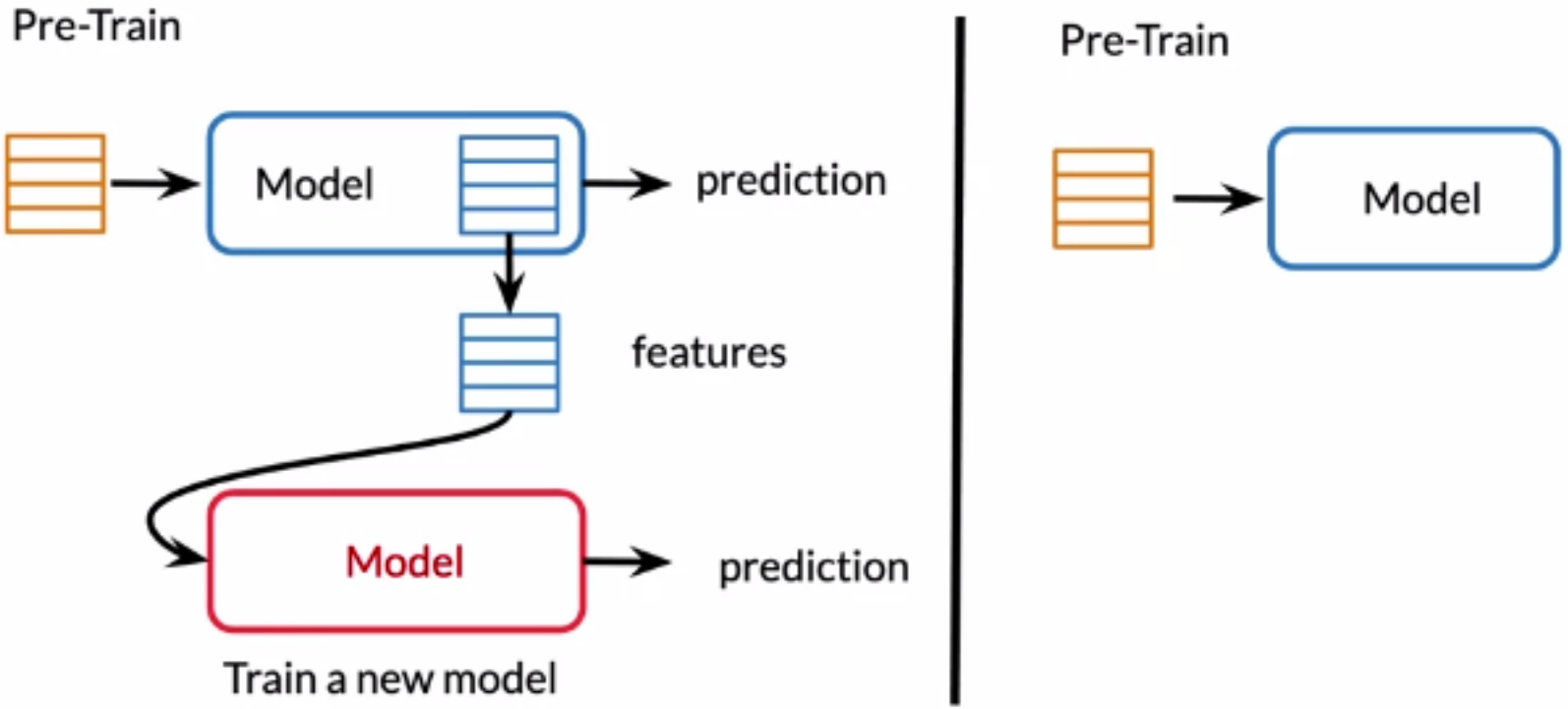
Pre-Train

Feature-based vs. Fine-Tuning

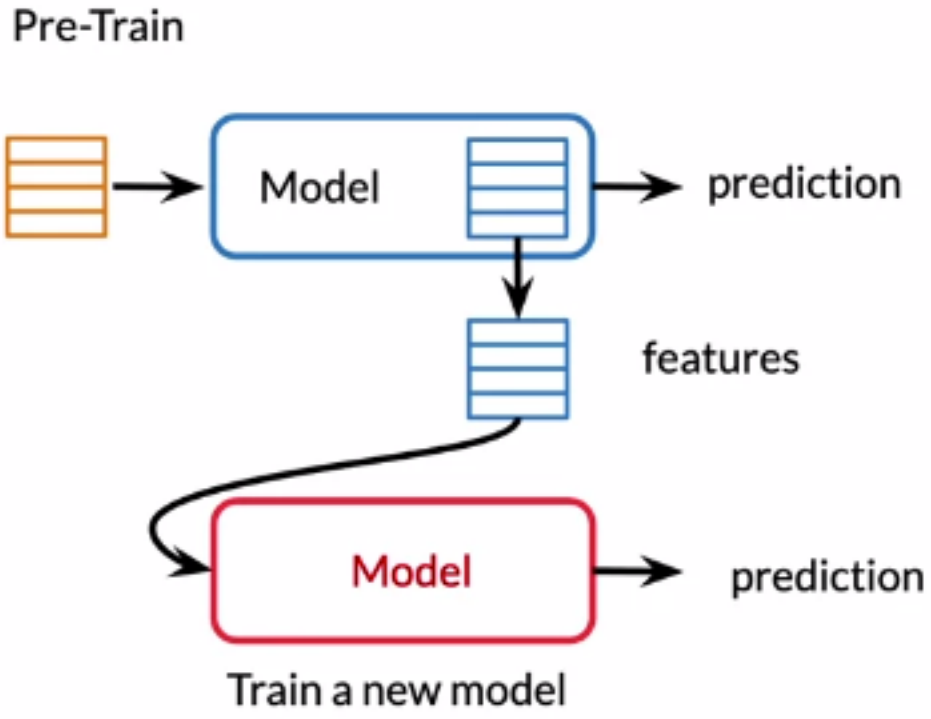


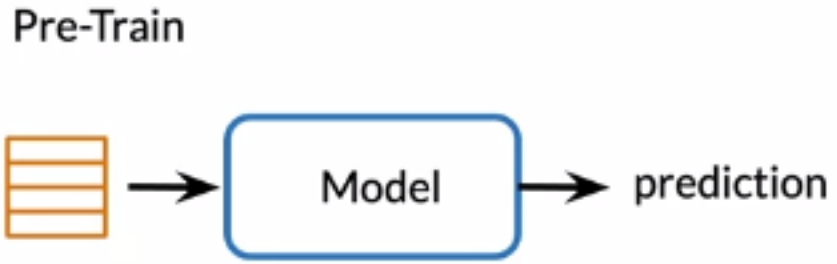
Pre-Train

Feature-based vs. Fine-Tuning

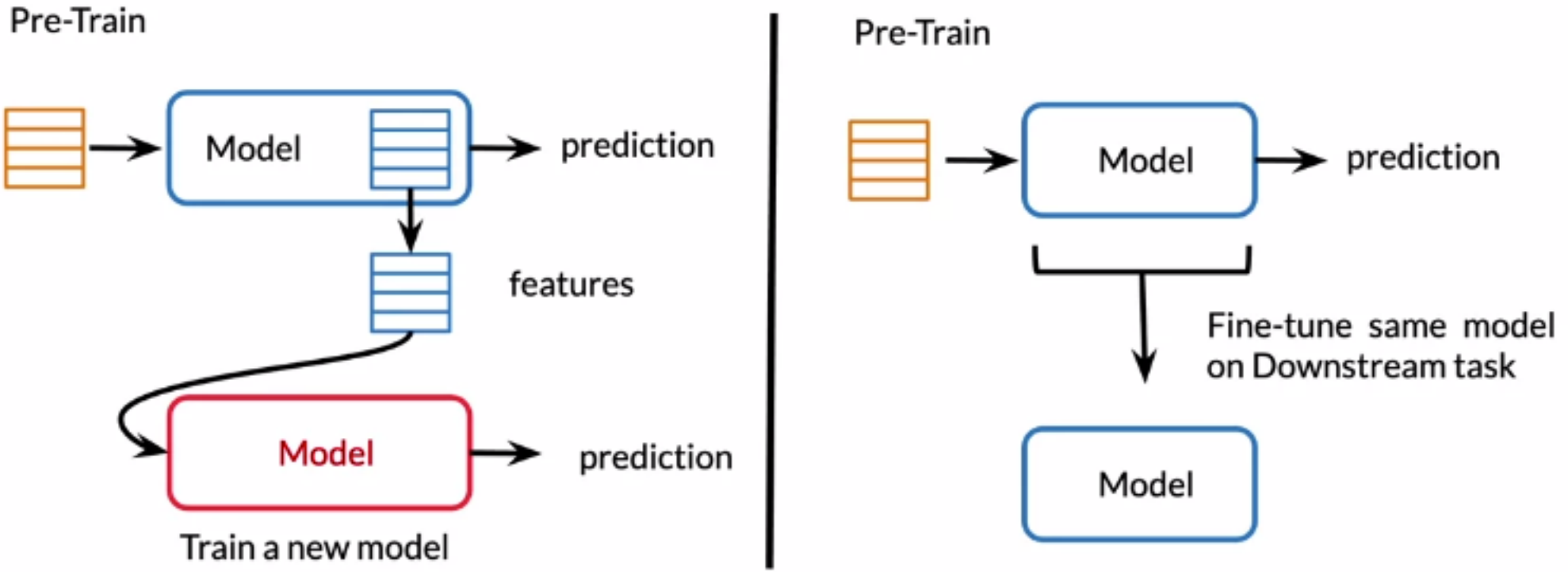


Feature-based vs. Fine-Tuning

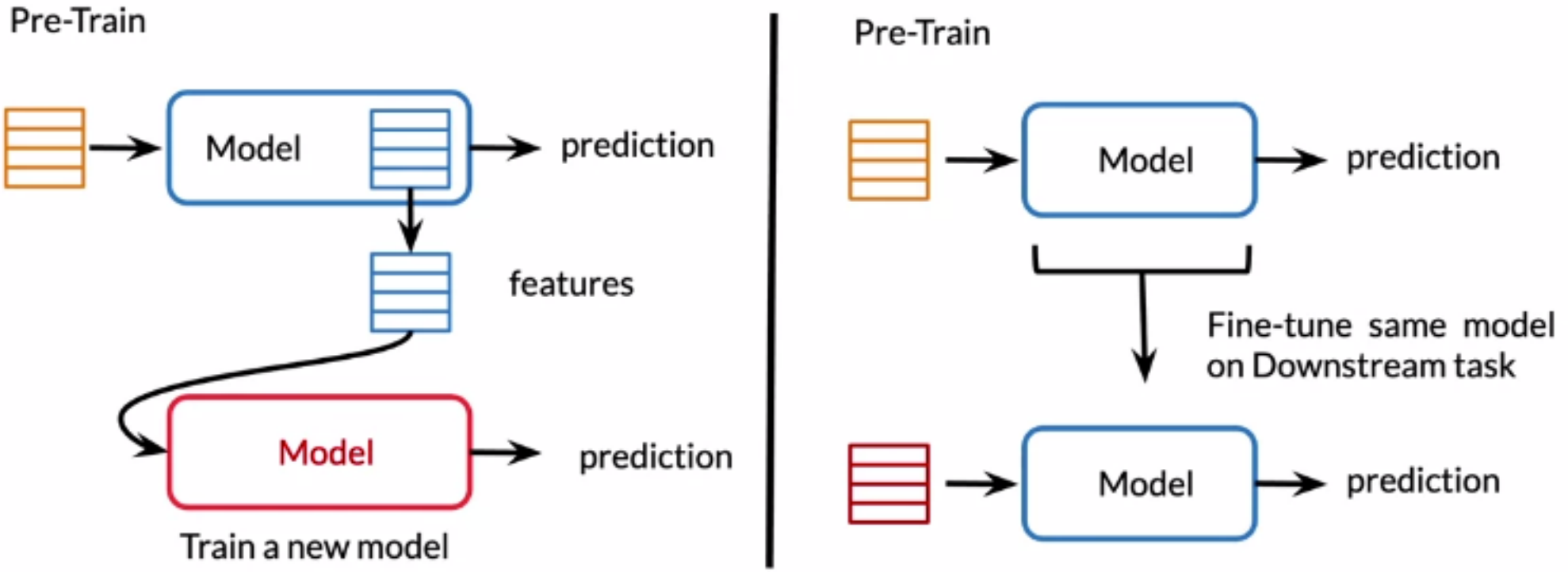




Feature-based vs. Fine-Tuning



Feature-based vs. Fine-Tuning

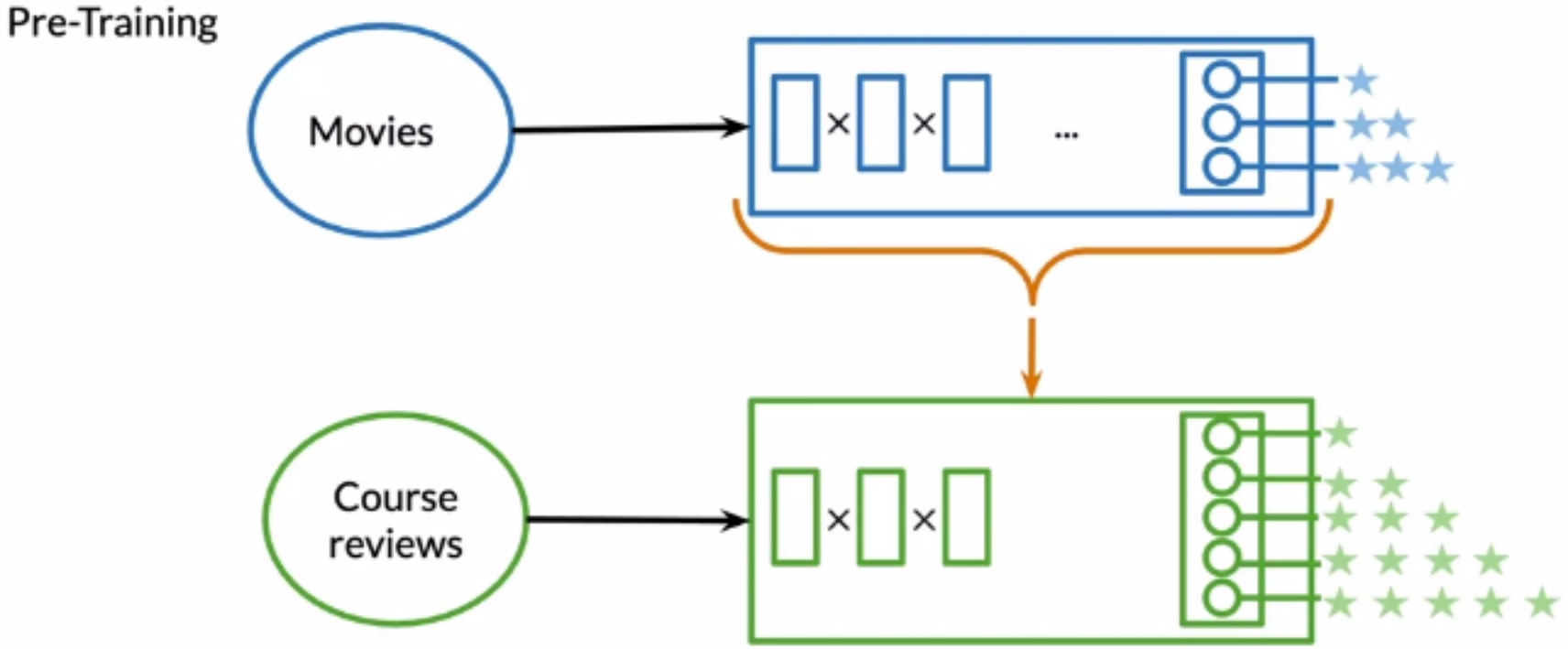


Fine-tune: adding a layer

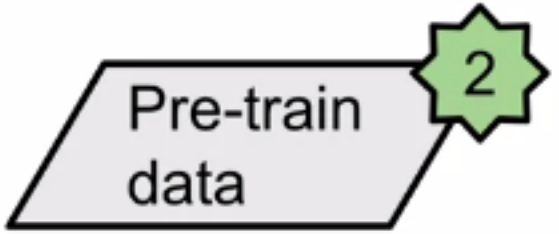
Pre-Training

Fine-tune: adding a layer

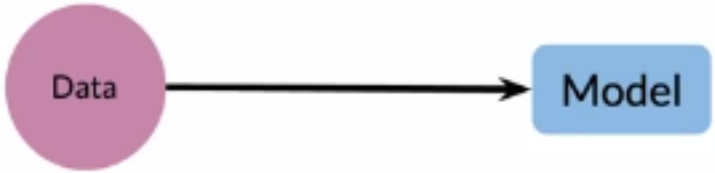
Transfer 1

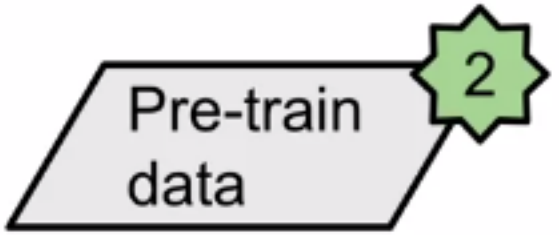


Data and performance

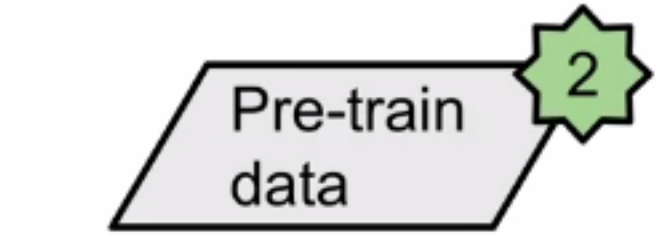


Data and performance

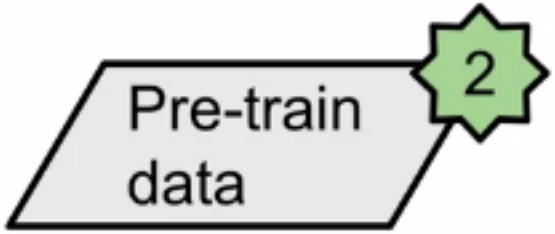




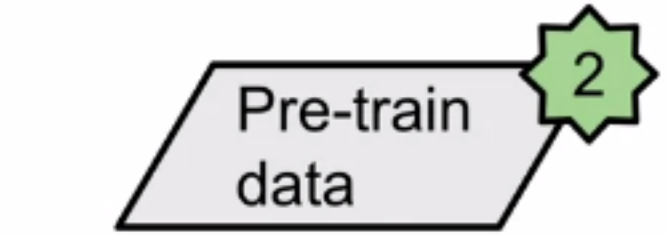
Data and performance

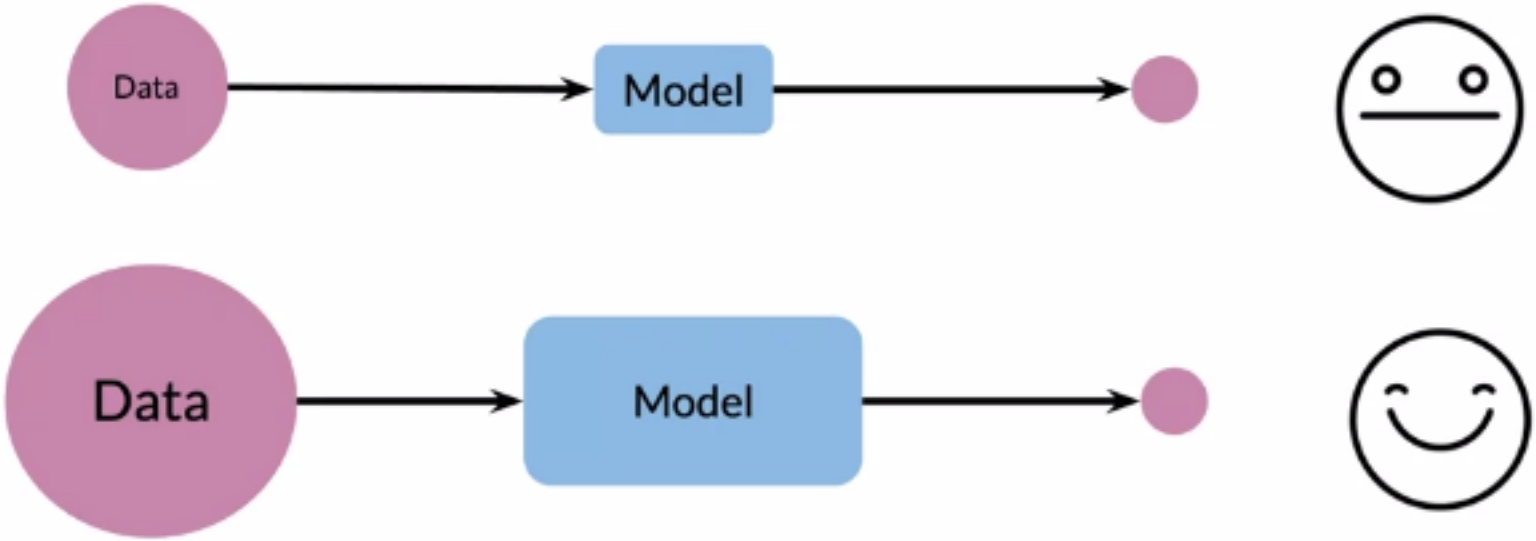


Data and performance

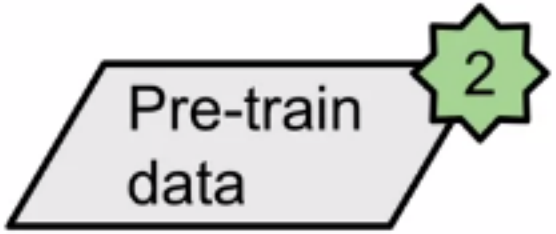


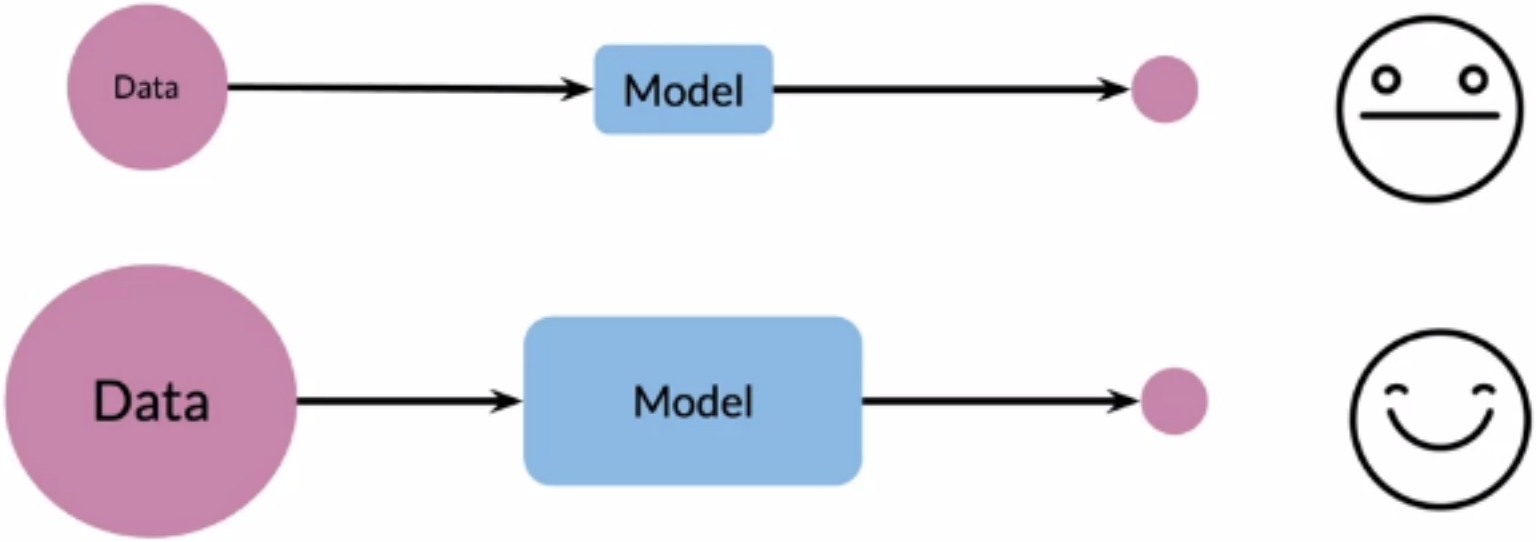
Data and performance



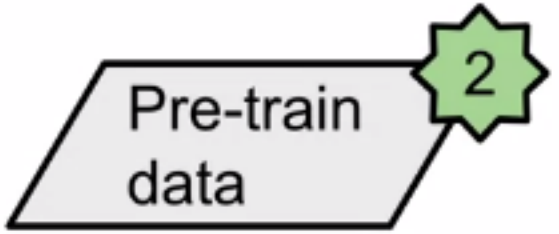


Data and performance

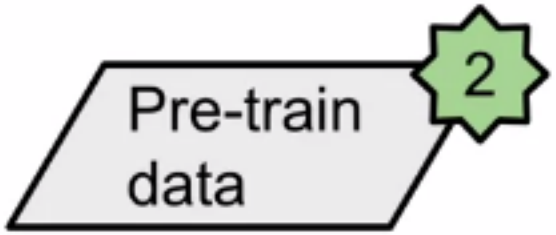




Labeled vs Unlabeled Data

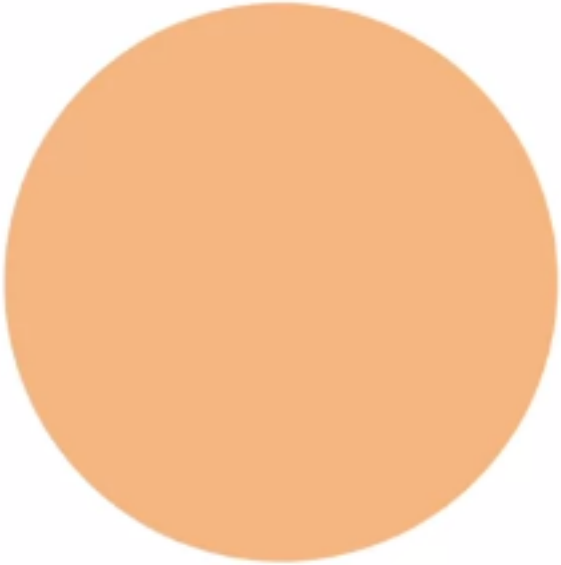


Labeled vs Unlabeled Data

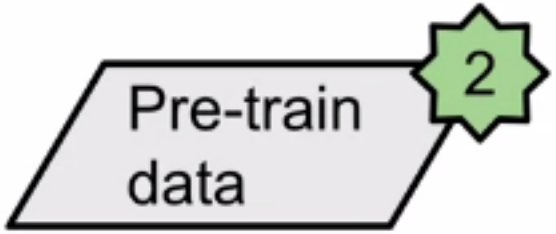


Labeled text data

Unlabeled text data

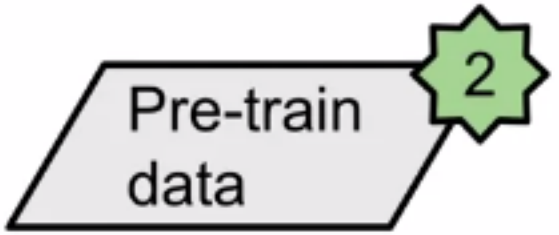


Transfer learning with unlabeled data



Pre-Training

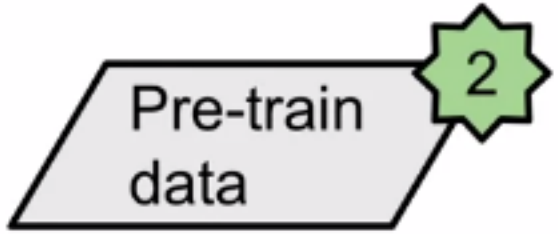
Transfer learning with unlabeled data

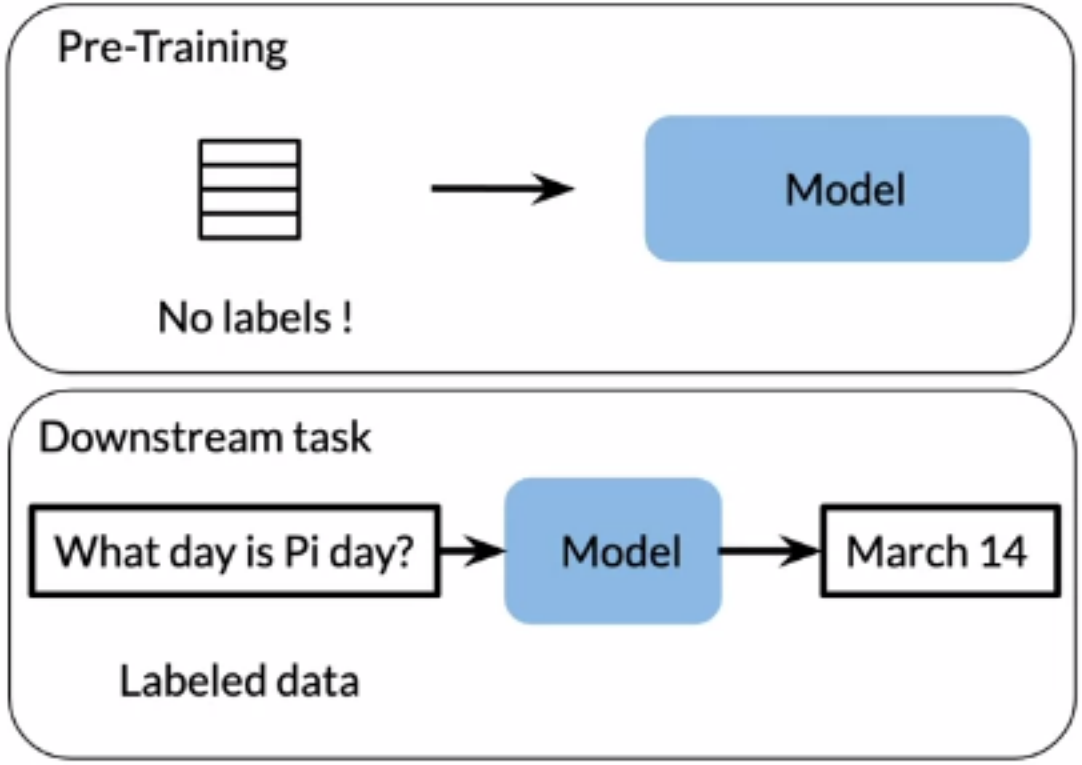


Pre-Training

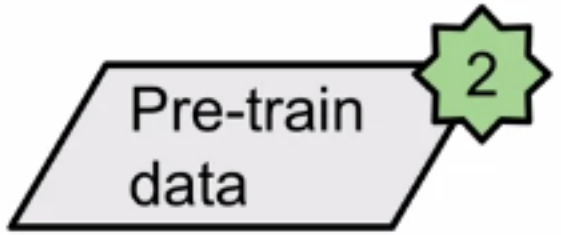
No labels !

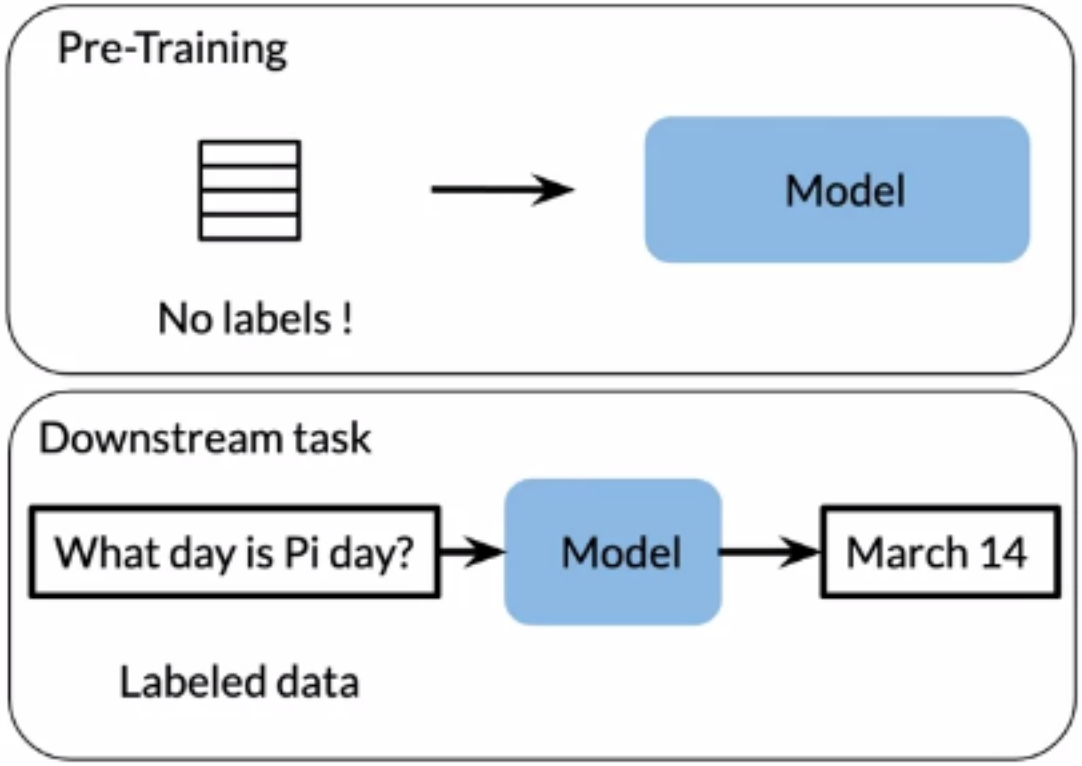
Transfer learning with unlabeled data





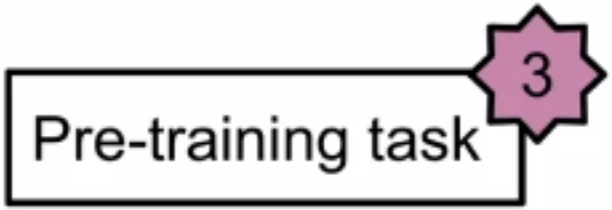
Transfer learning with unlabeled data





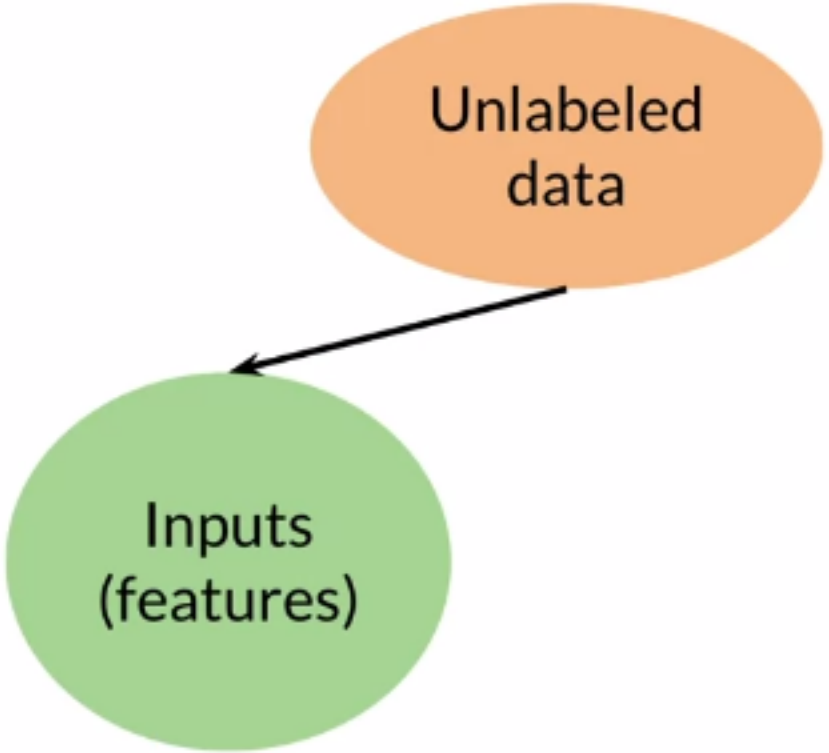
Which tasks work with **unlabeled** data?

Self-supervised task



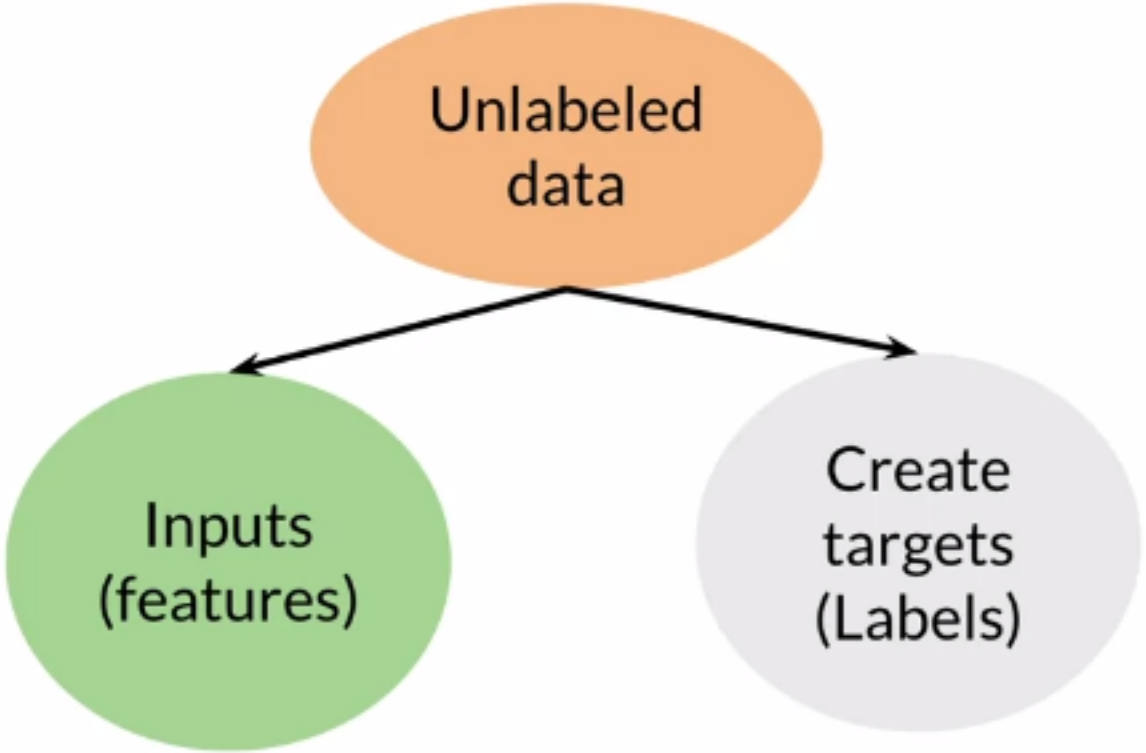
Self-supervised task

Pre-training task 3

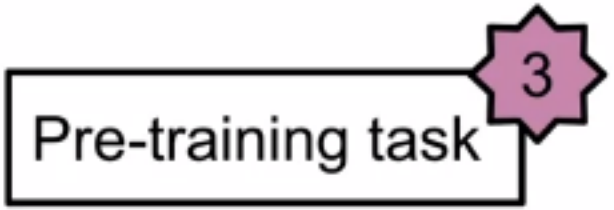


Self-supervised task

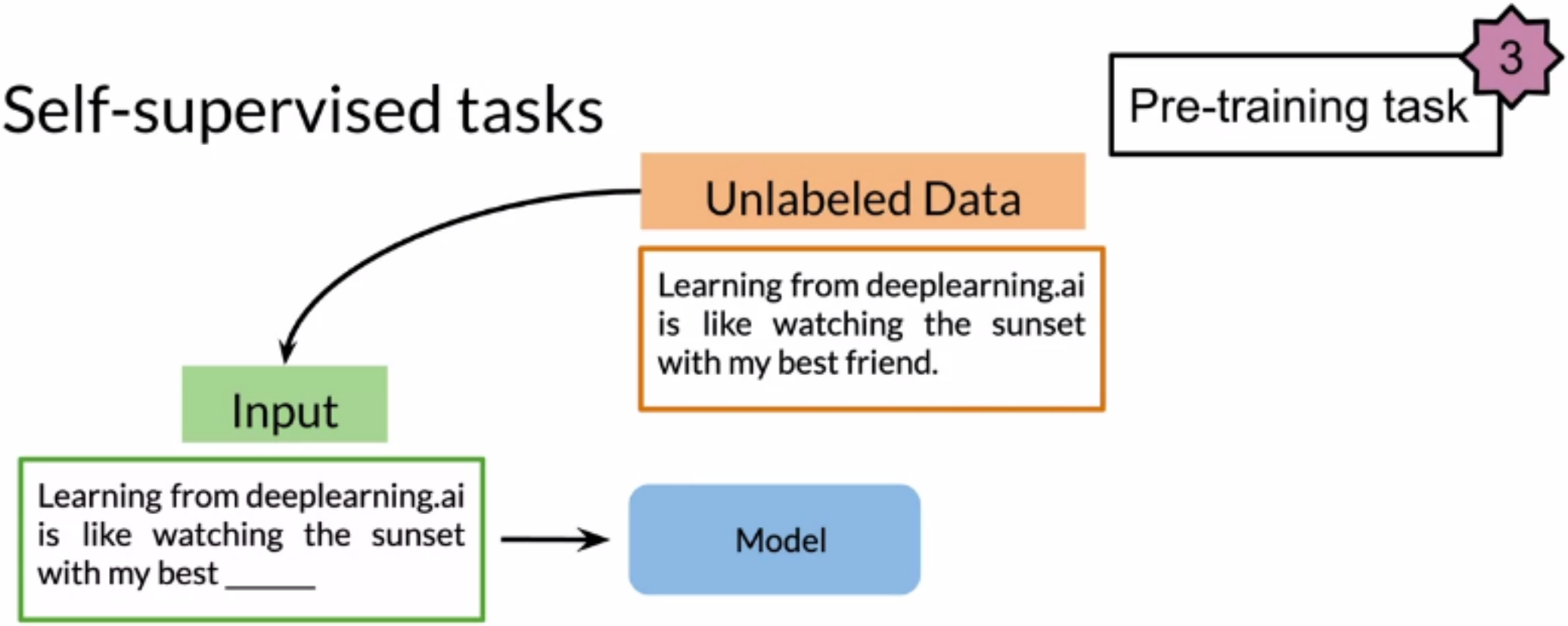
Pre-training task 3



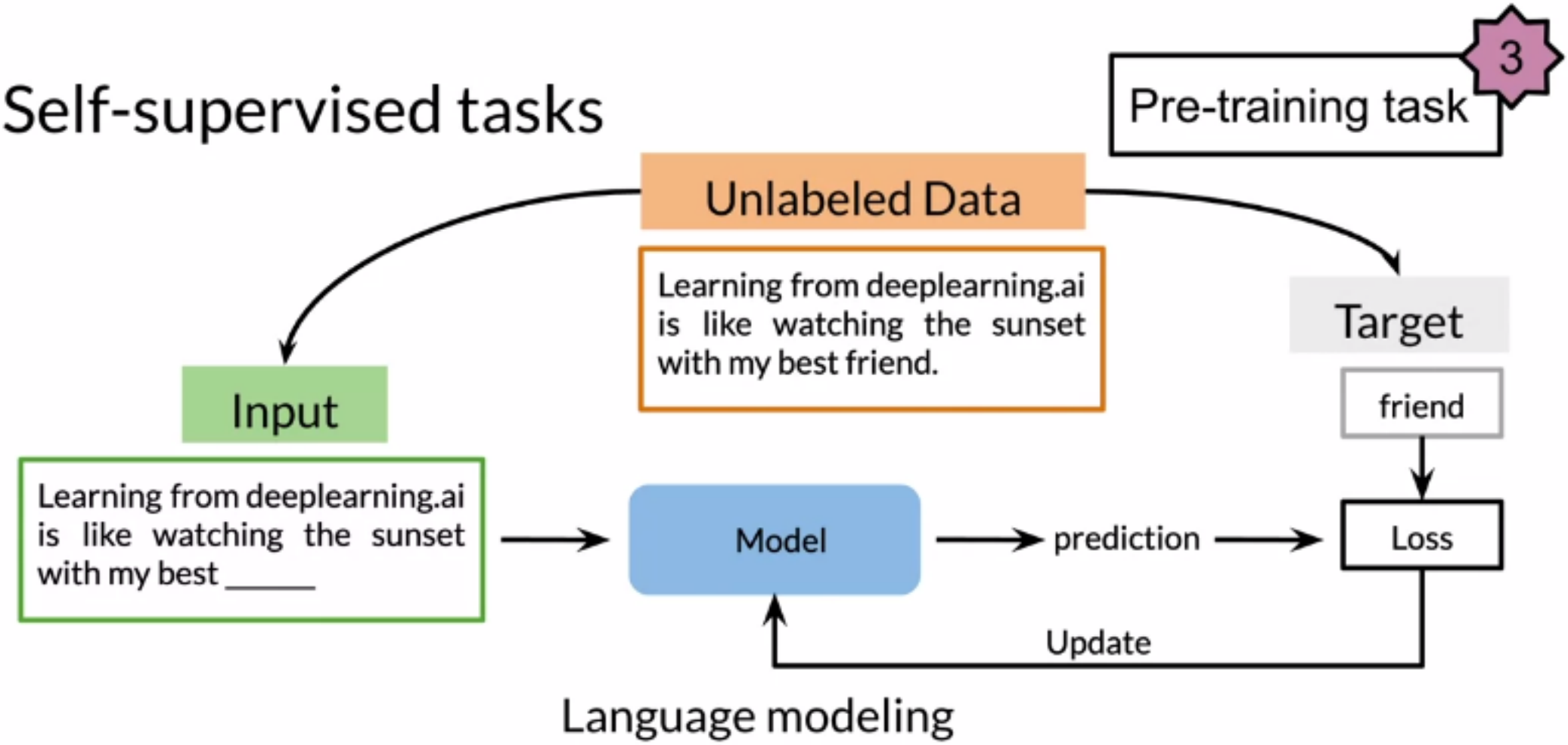
Self-supervised tasks



Self-supervised tasks



Self-supervised tasks

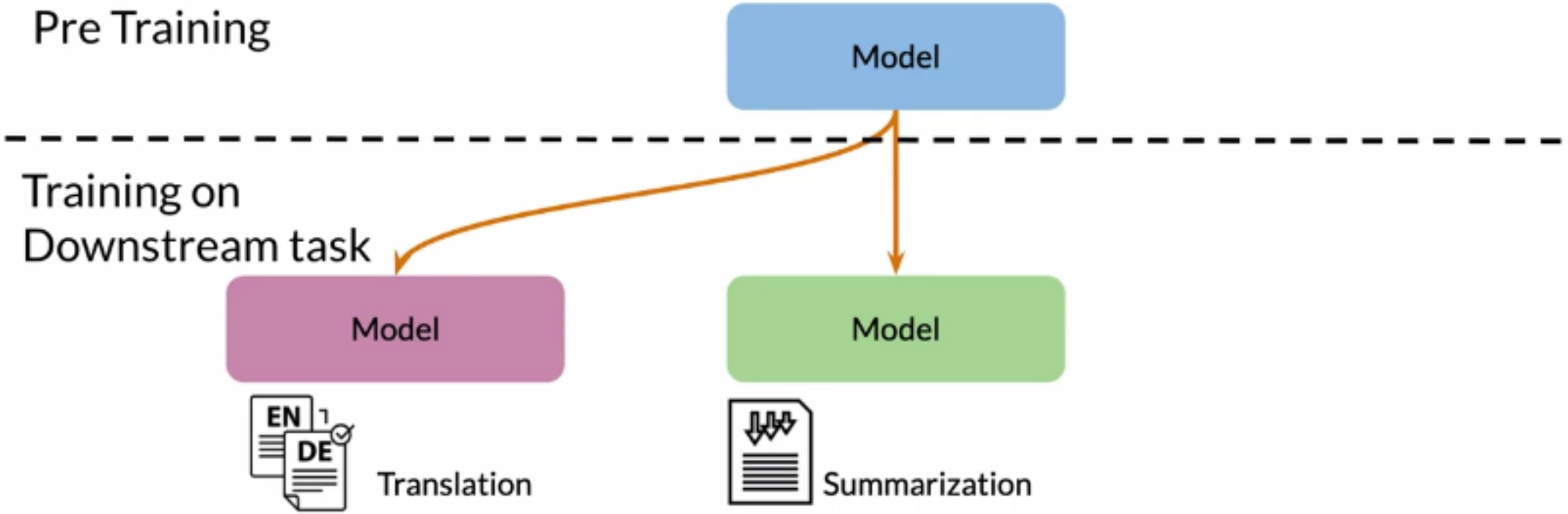


Fine-tune a model for each downstream task

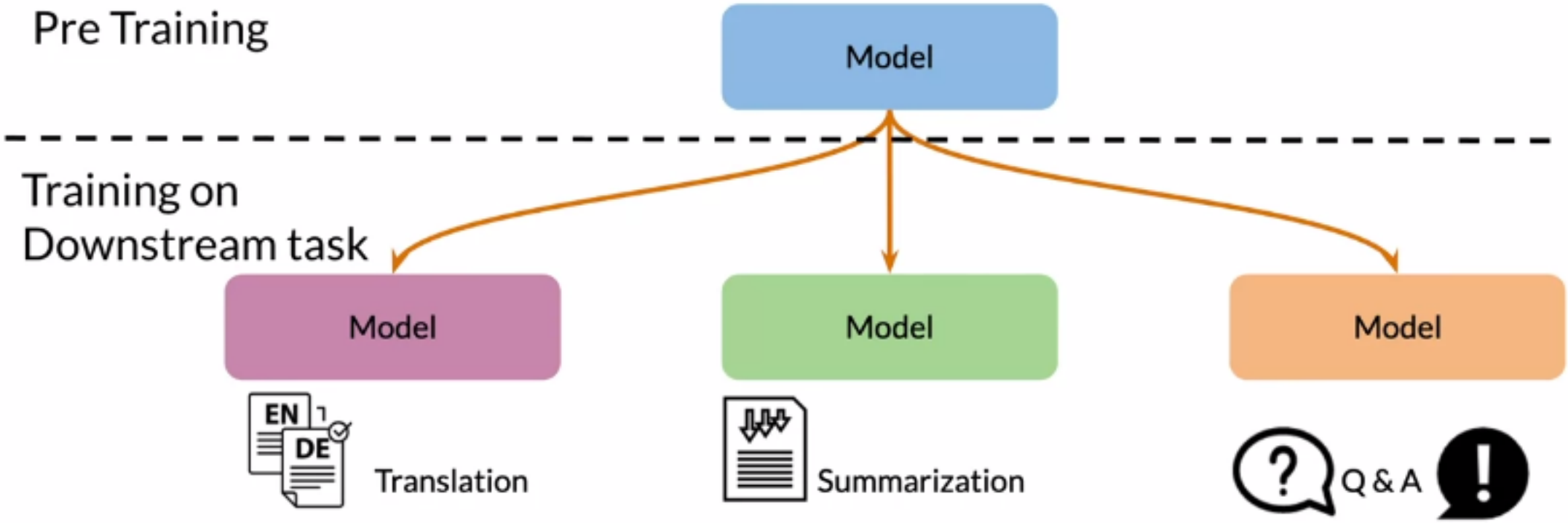
Pre Training

Training on
Downstream task

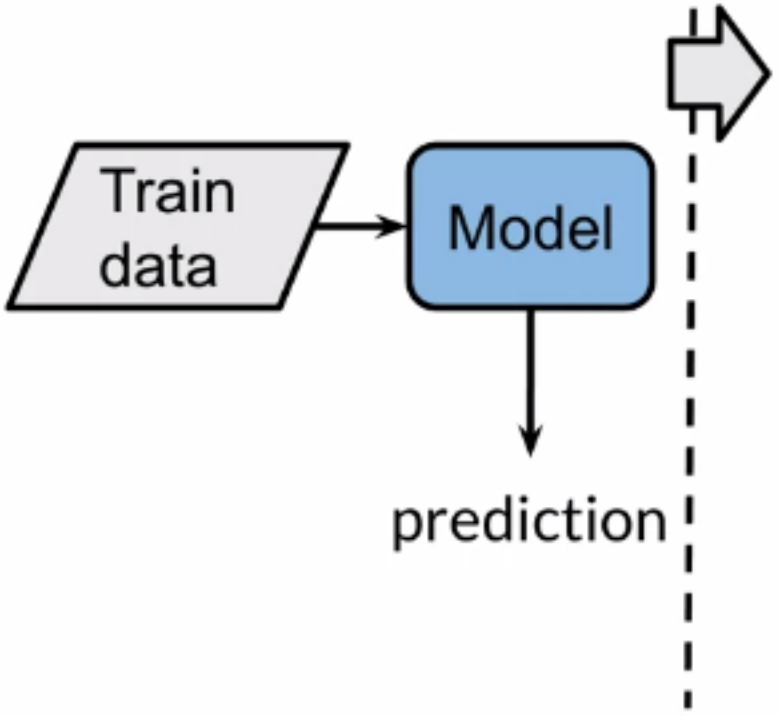
Fine-tune a model for each downstream task



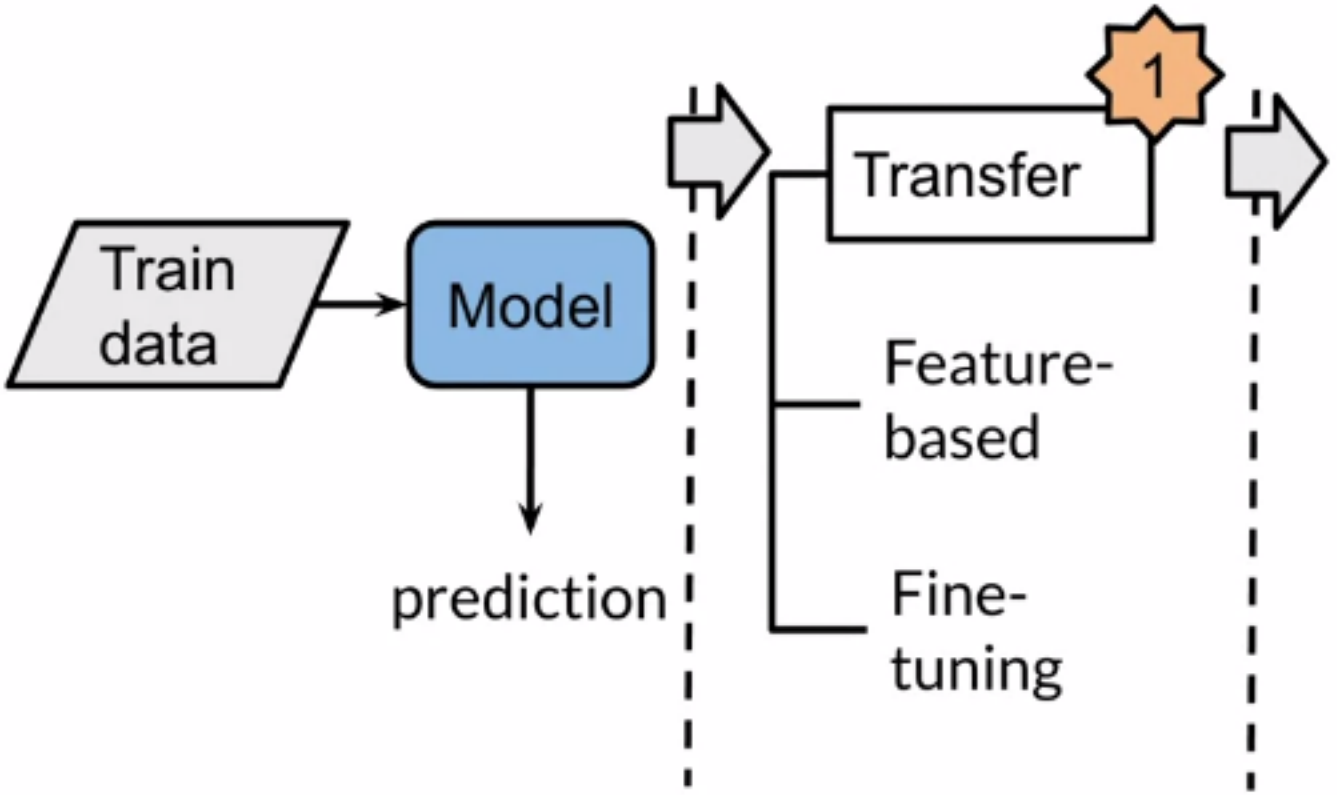
Fine-tune a model for each downstream task



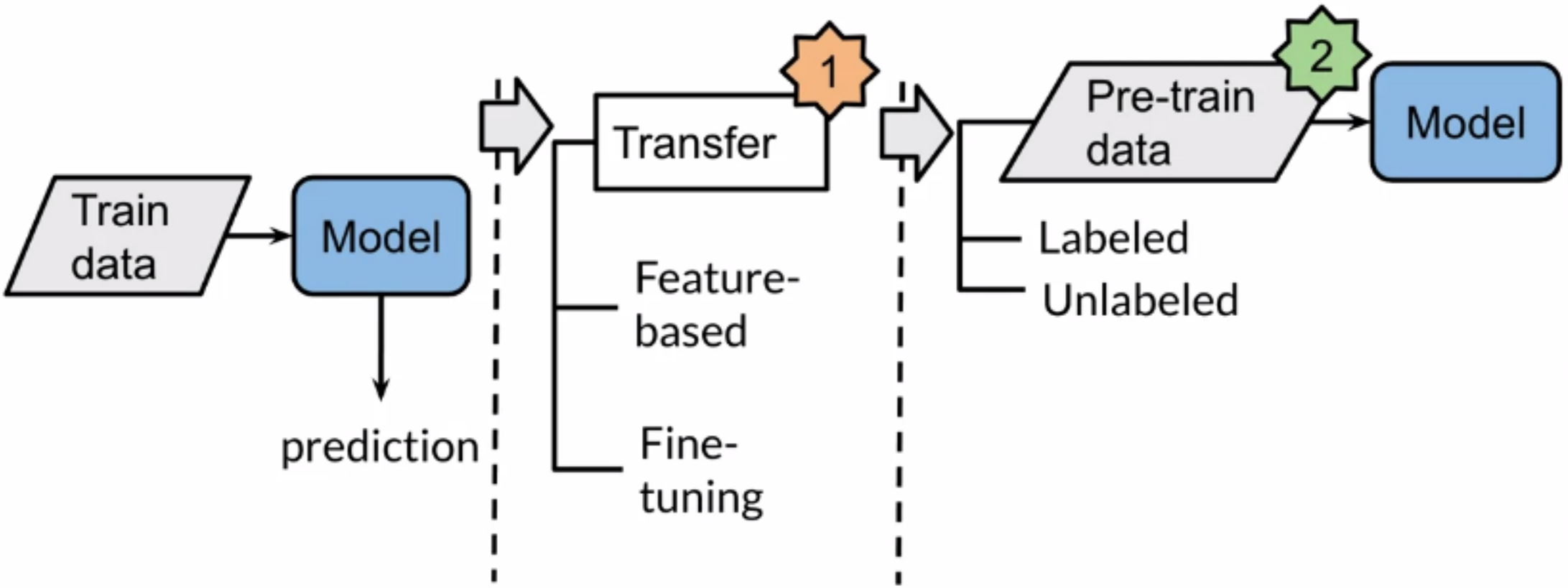
Summary



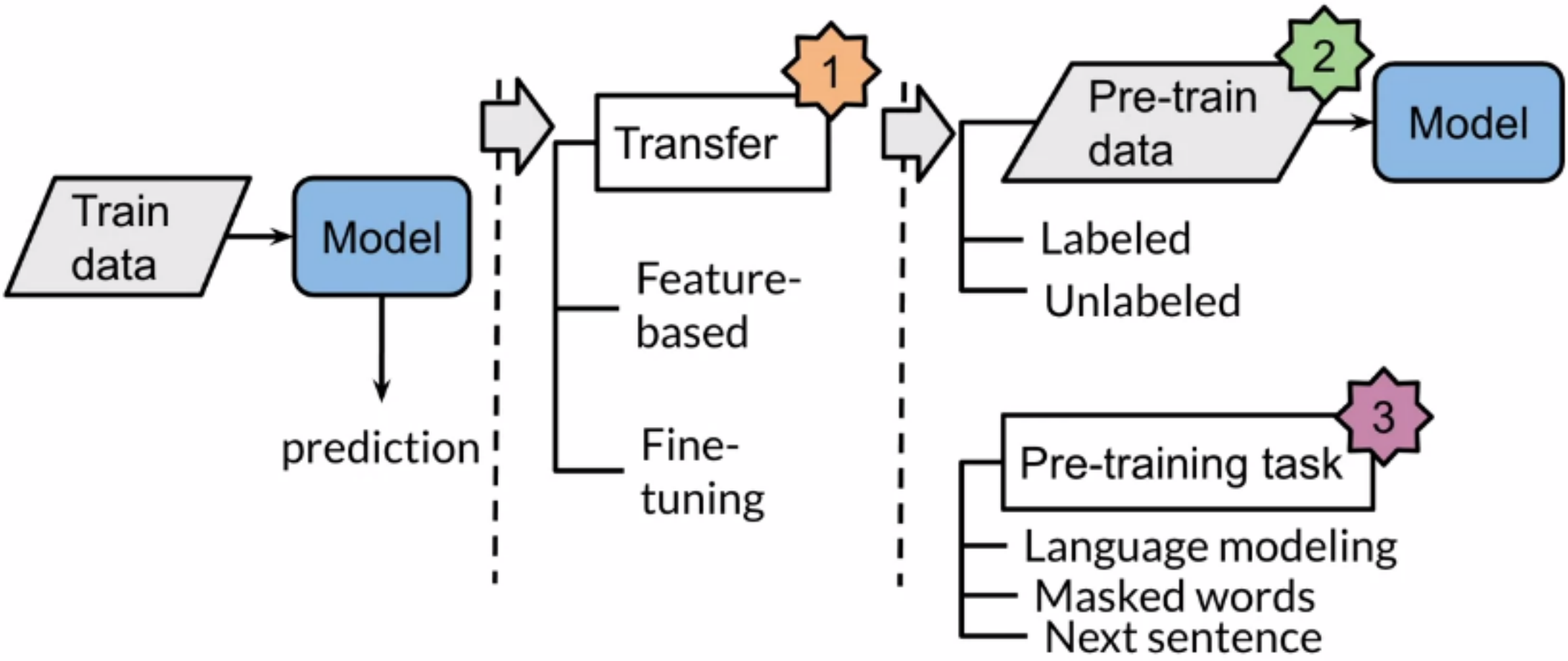
Summary


1
Transfer
Train data
Model
prediction
Feature-based
Fine-tuning

Summary


Train data
Model
prediction
1
Transfer
Feature-based
Fine-tuning
2
Pre-train data
Model
Labeled
Unlabeled

Summary


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

deeplearning.ai

ELMo, GPT, BERT, T5

Outline

Context

... right ...

Context

… right …

… they were on the right …

… they were on the right side of the street

Continuous Bag of Words

… they were on the right side of the street

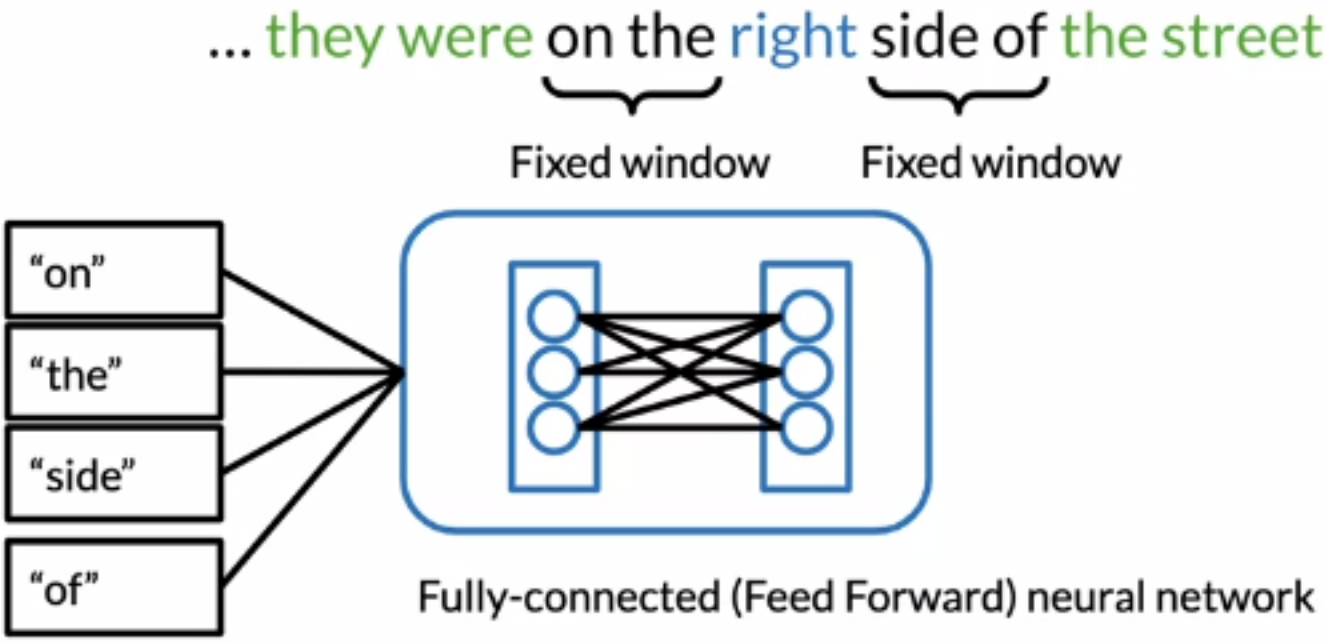
Fixed window Fixed window

Continuous Bag of Words

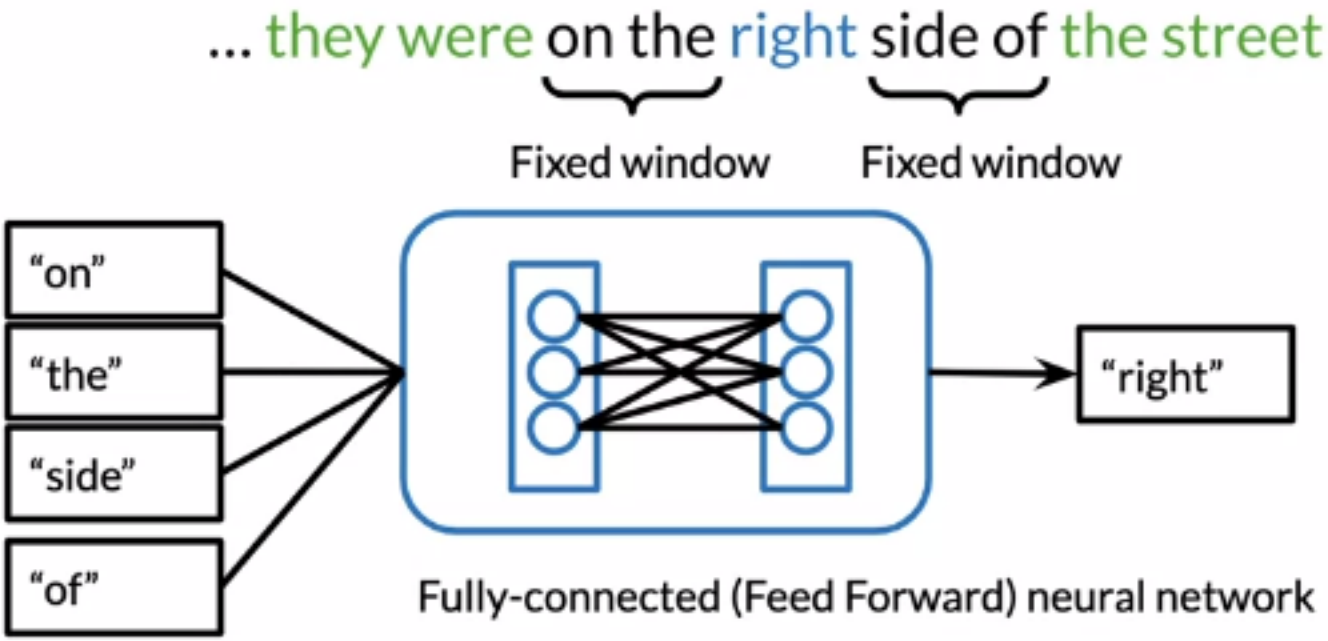
… they were on the right side of the street

Fixed window Fixed window

Continuous Bag of Words



Continuous Bag of Words



Need more context?

… they were on the right side of the street.

Fixed window Fixed window

Need more context?

… they were on the right side of the street.

Fixed window Fixed window

… 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.

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

ELMo: Full context using RNN

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

ELMo: Full context using RNN

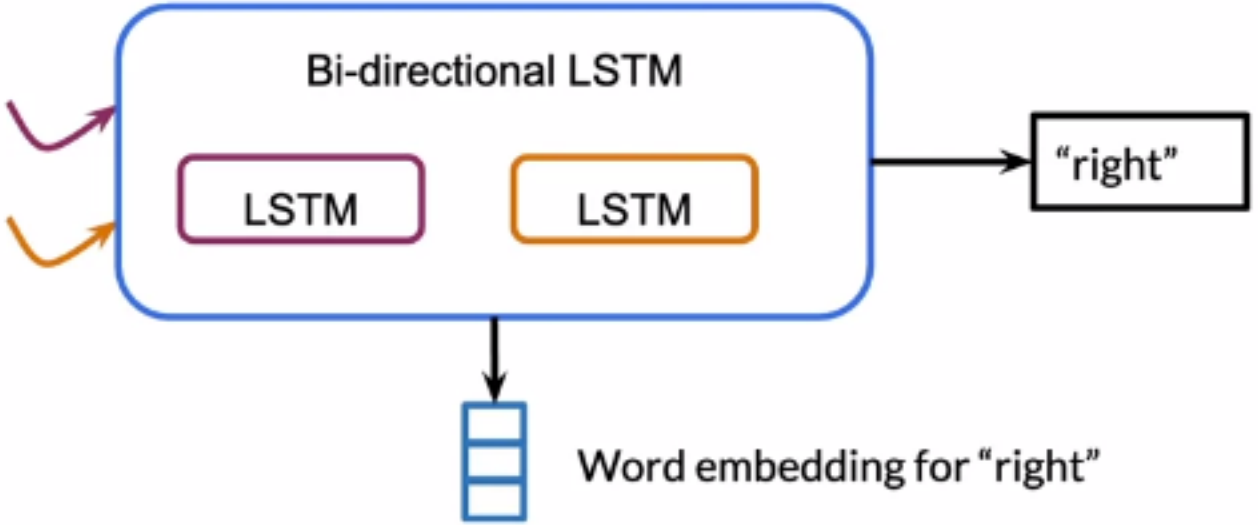
The legislators believed that they were on the ____ side of history so they changed the law.

ELMo: Full context using RNN

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

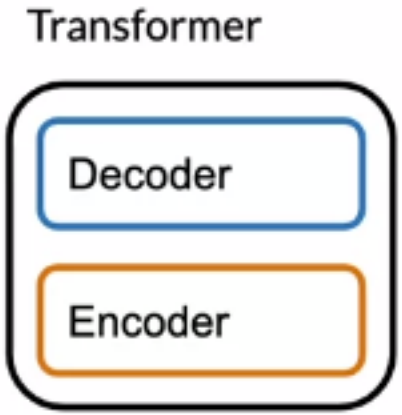
ELMo: Full context using RNN

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

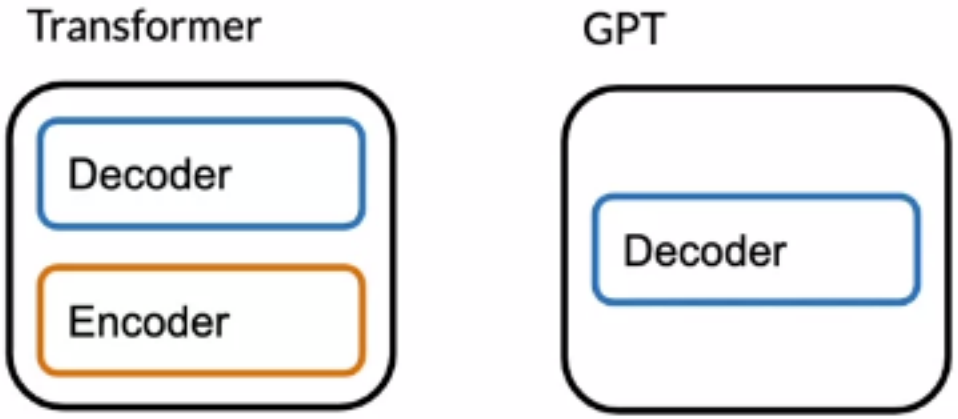


Open AI GPT

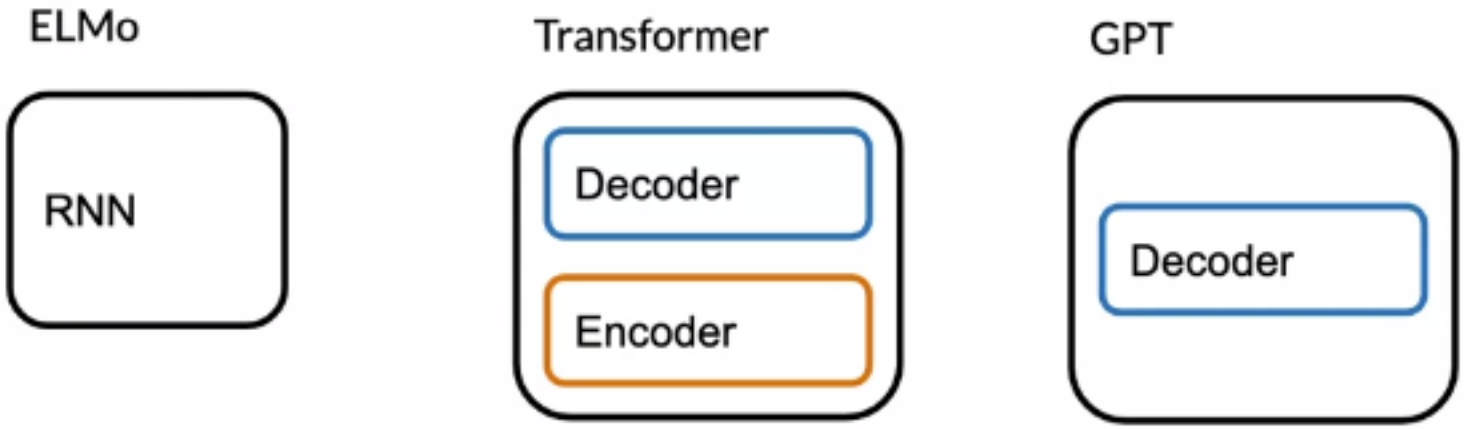
Open AI GPT



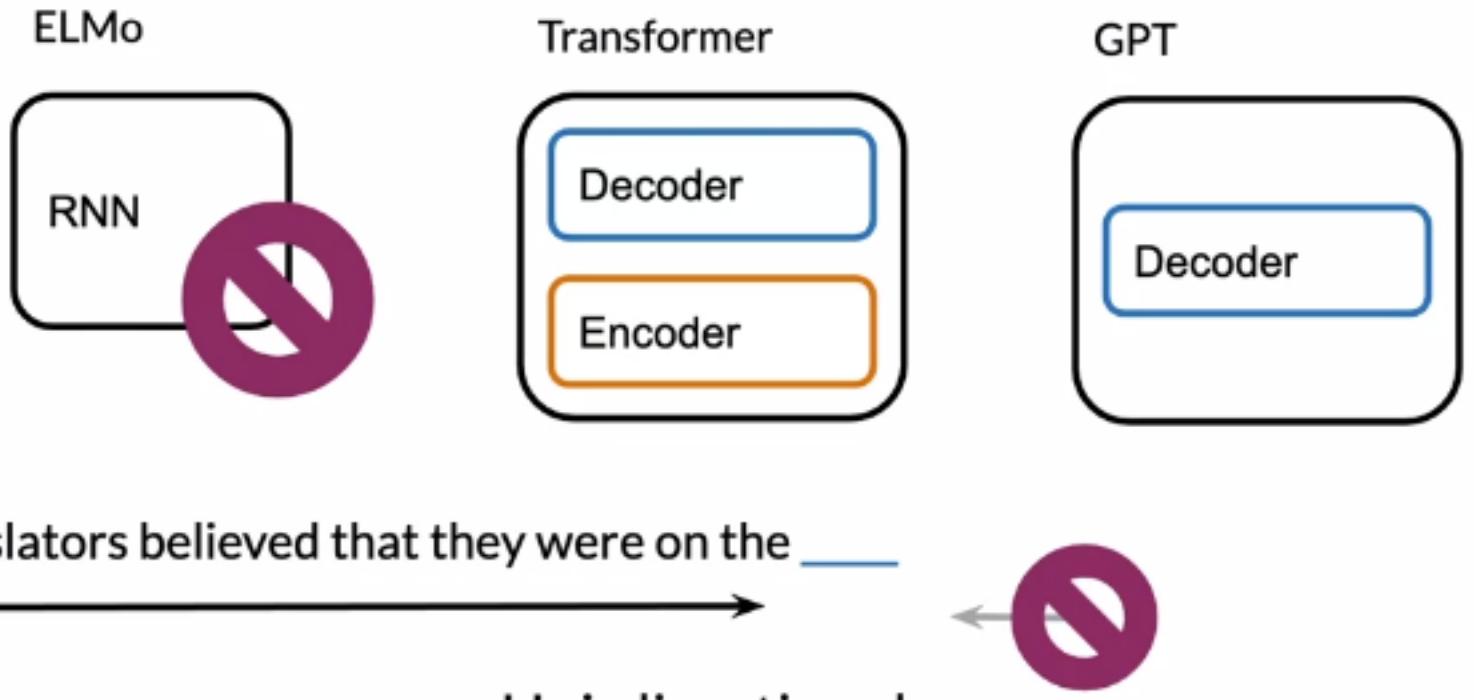
Open AI GPT



Open AI GPT

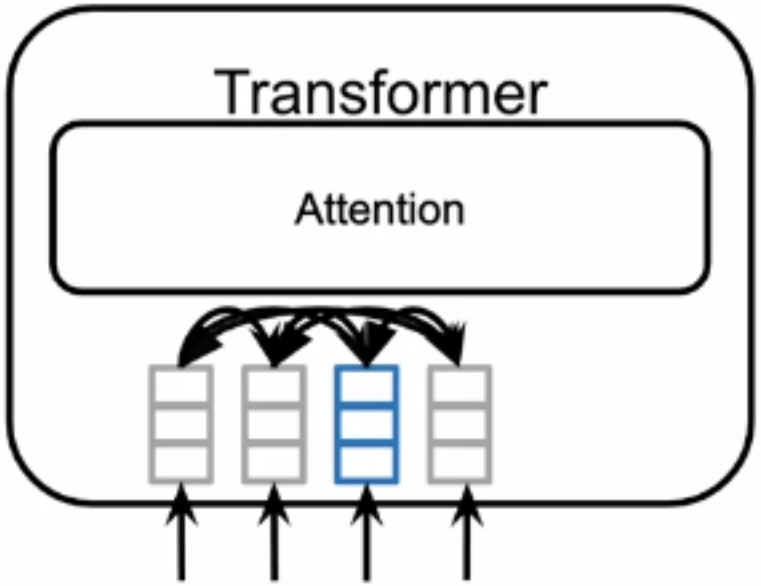


Open AI GPT

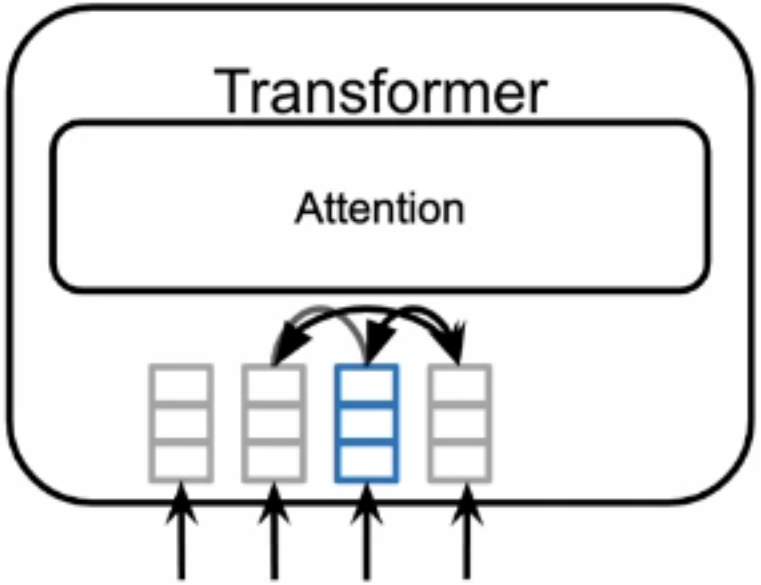

ELMo
RNN
Transformer
Decoder
Encoder
GPT
Decoder
The legislators believed that they were on the ____
Uni-directional

Why not bi-directional?

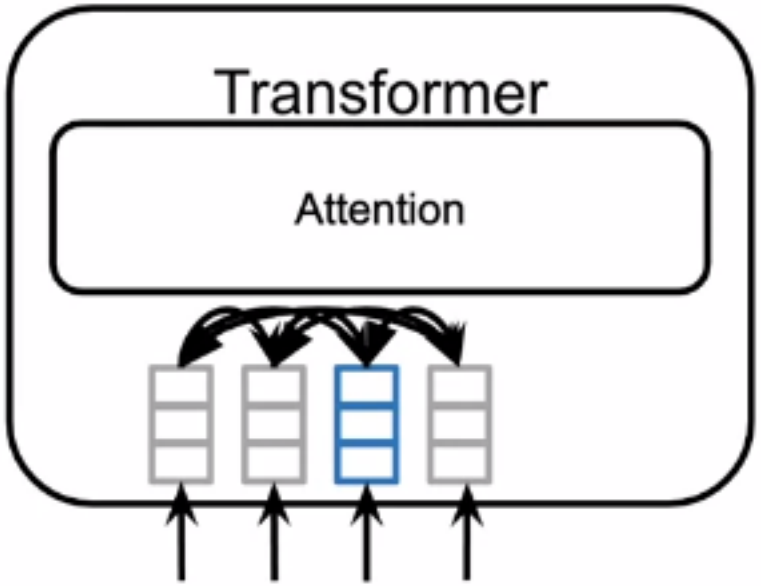
Why not bi-directional?



Why not bi-directional?



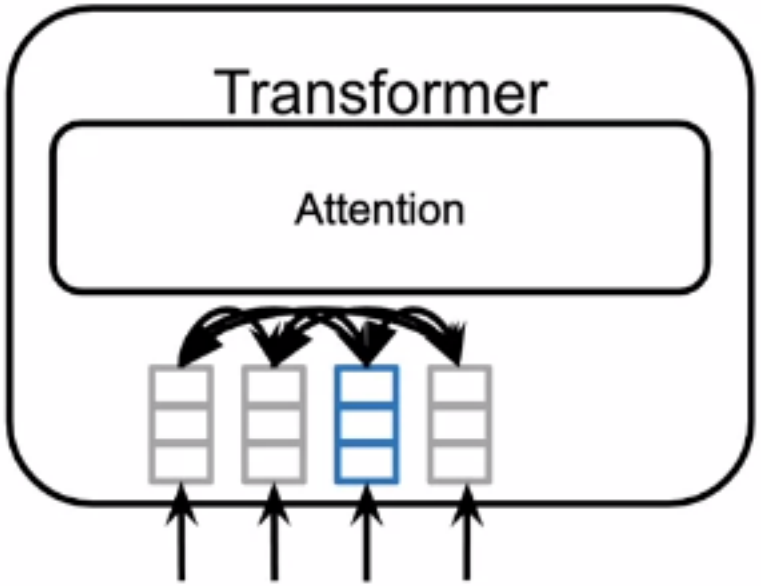
Why not bi-directional?



... on the right side...

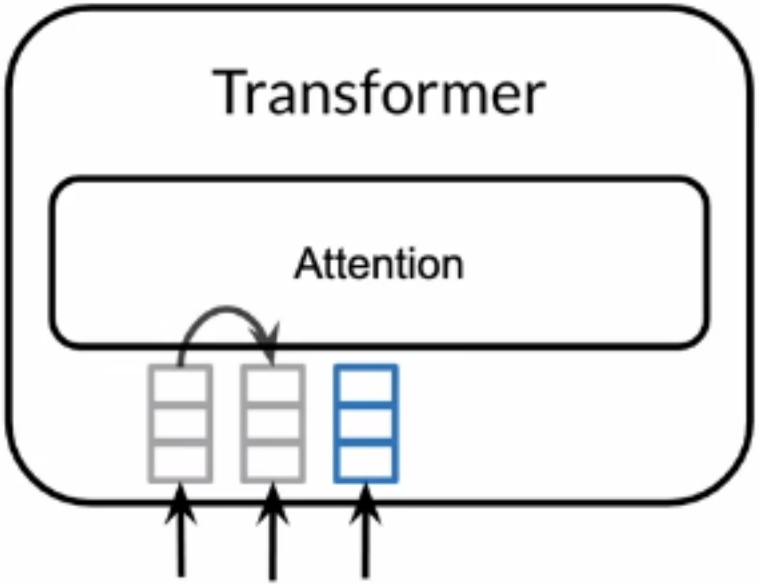
Each word can peek at itself!

GPT: Uni-directional

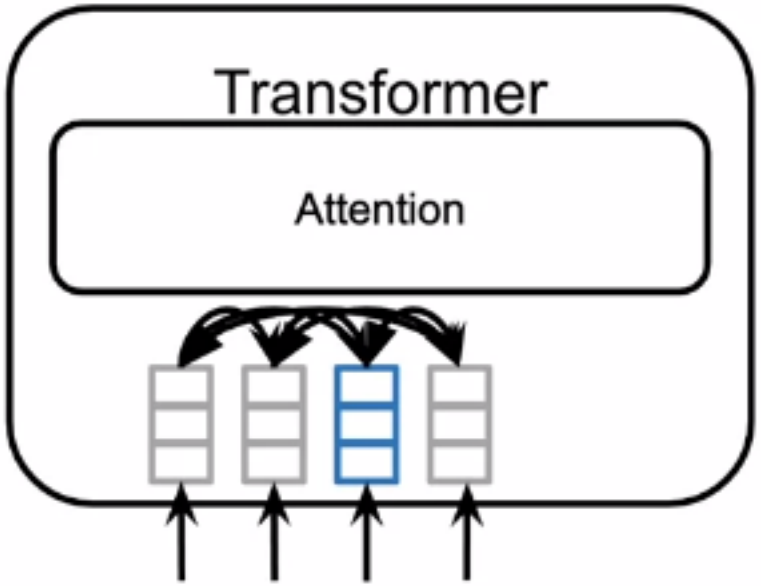


... on the right side...

Each word can peek at itself!

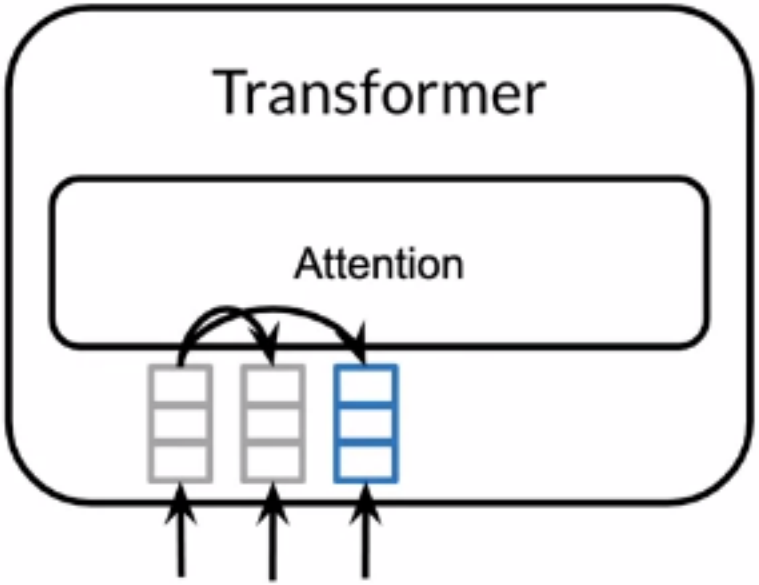


GPT: Uni-directional



... on the right side...

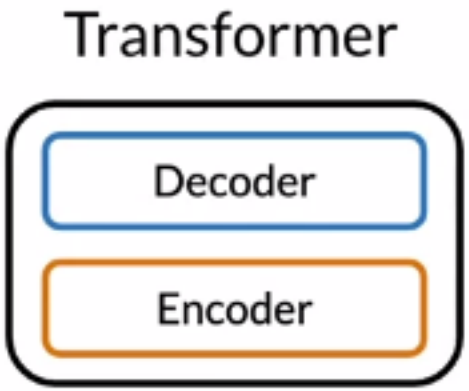
Each word can peek at itself!



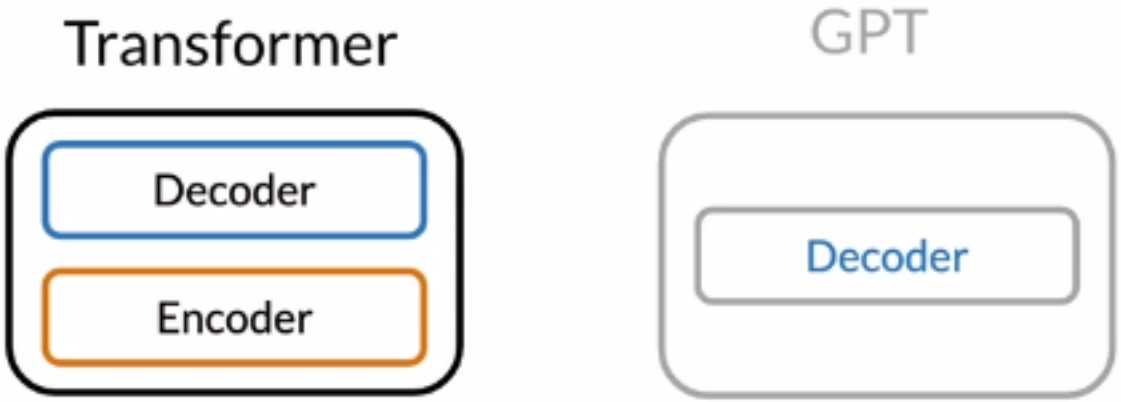
... on the right

No peeking!

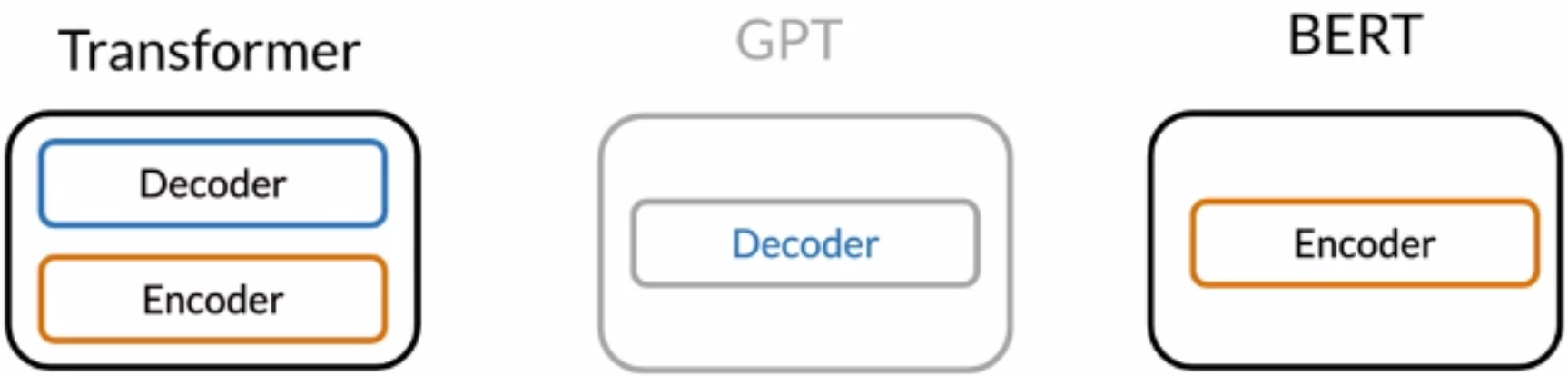
BERT



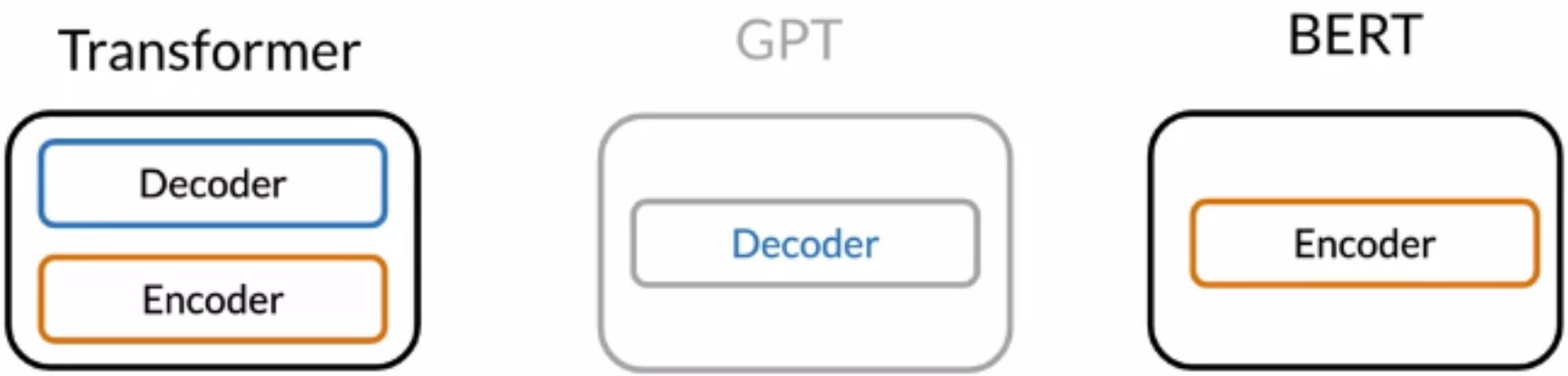
BERT



BERT

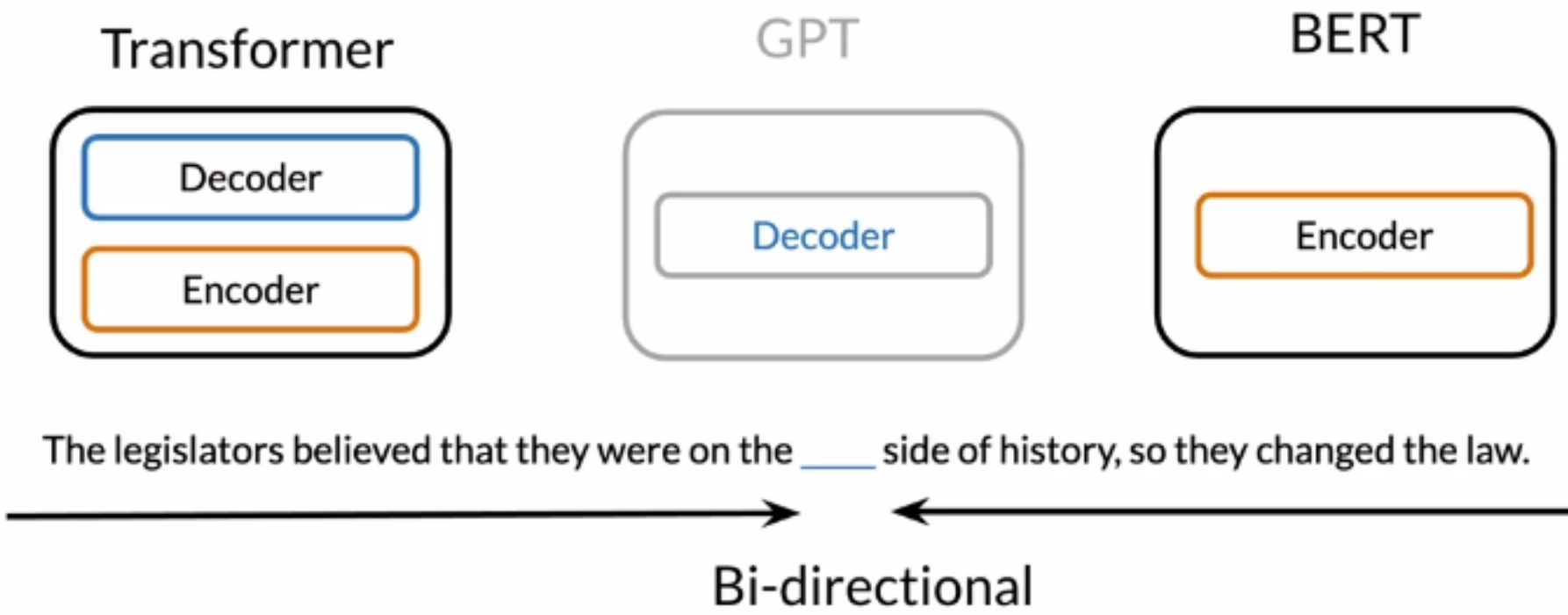


BERT



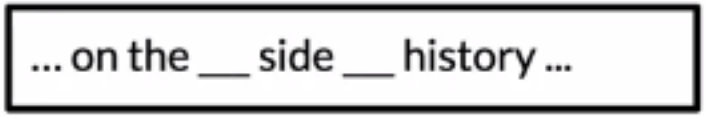
The legislators believed that they were on the ____ side of history, so they changed the law.

BERT

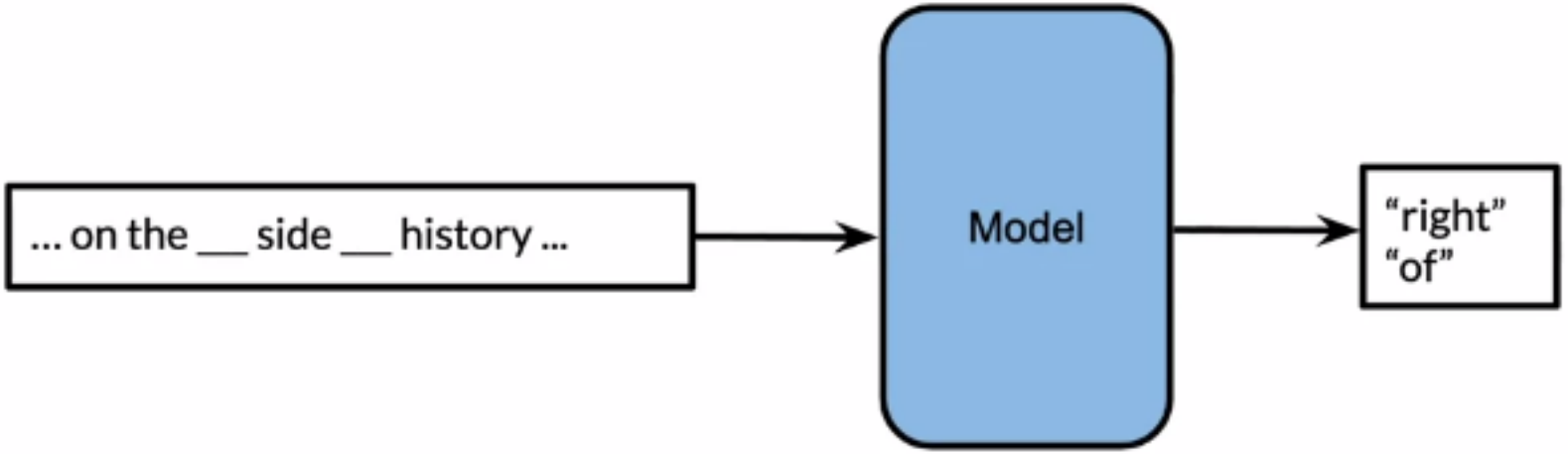

Transformer
Decoder
Encoder
GPT
Decoder
BERT
Encoder
The legislators believed that they were on the ____ side of history, so they changed the law.
Bi-directional

Transformer + Bi-directional Context

Transformer + Bi-directional Context

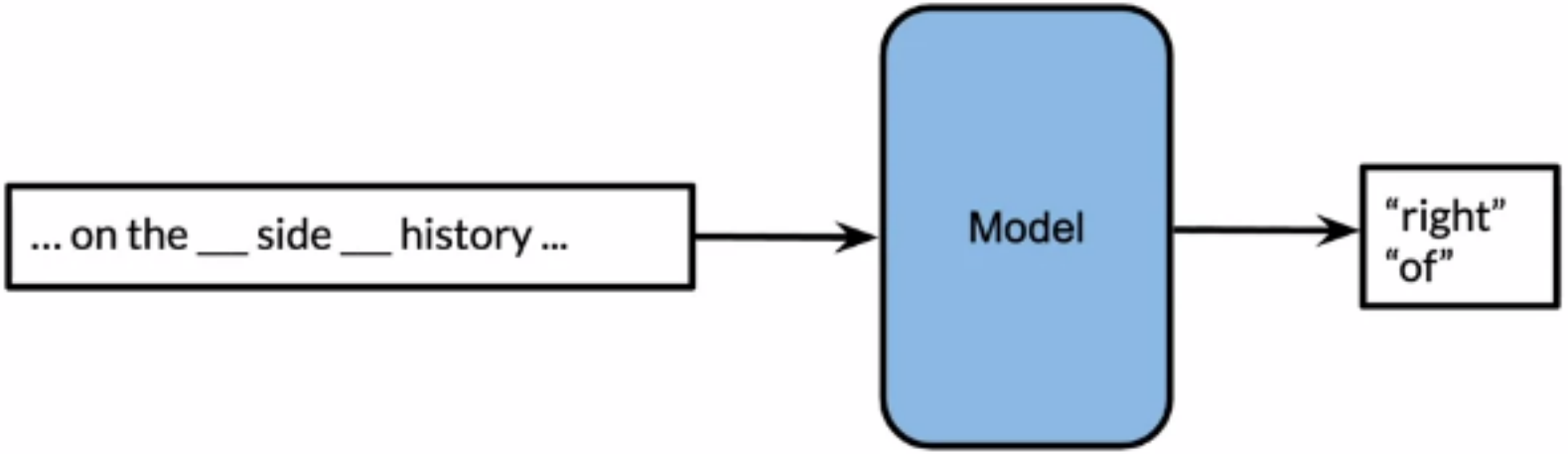


Transformer + Bi-directional Context



Multi-Mask Language Modeling

Transformer + Bi-directional Context



Multi-Mask Language Modeling

BERT: Words to Sentences

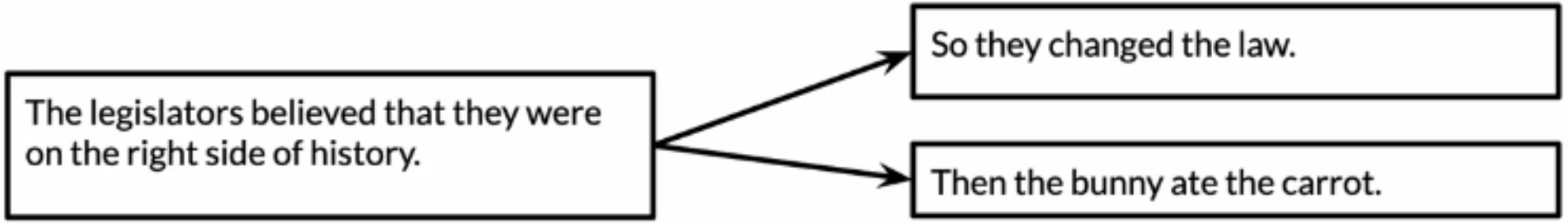
BERT: Words to Sentences

The legislators believed that they were on the right side of history.

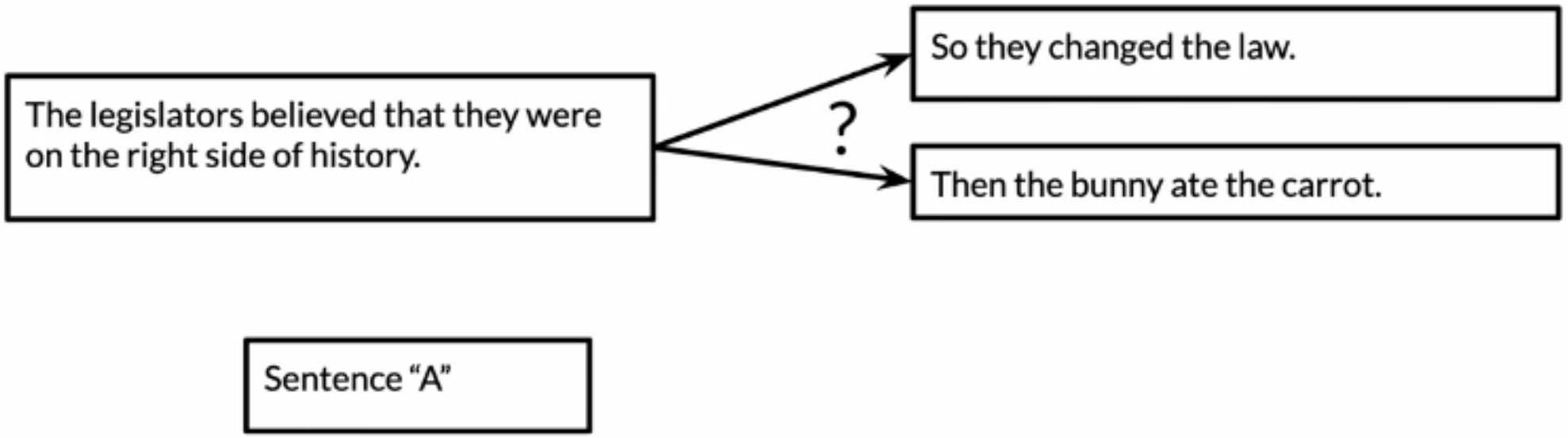
BERT: Words to Sentences

The legislators believed that they were on the right side of history.

BERT: Words to Sentences

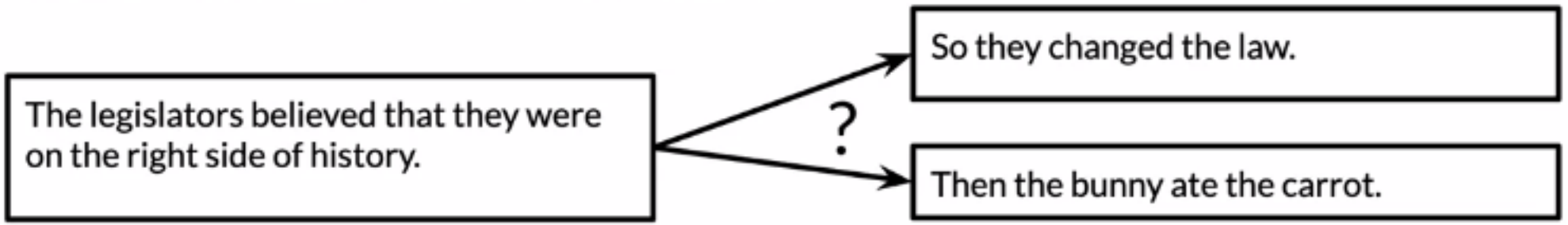


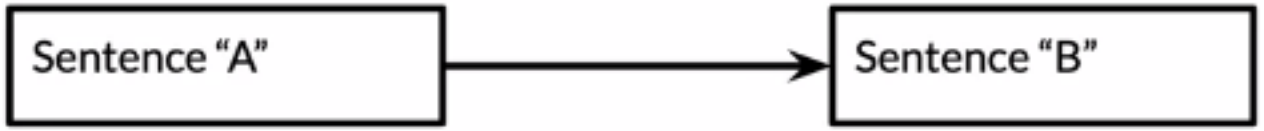
BERT: Words to Sentences



Next Sentence Prediction

BERT: Words to Sentences





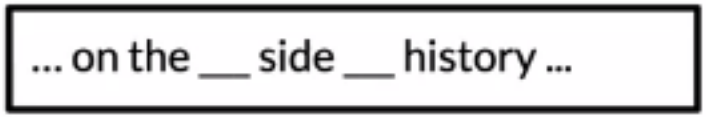
Next Sentence Prediction

BERT Pre-training Tasks

Multi-Mask Language Modeling

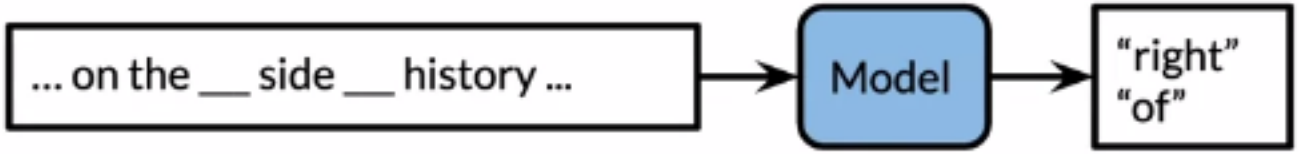
BERT Pre-training Tasks

Multi-Mask Language Modeling



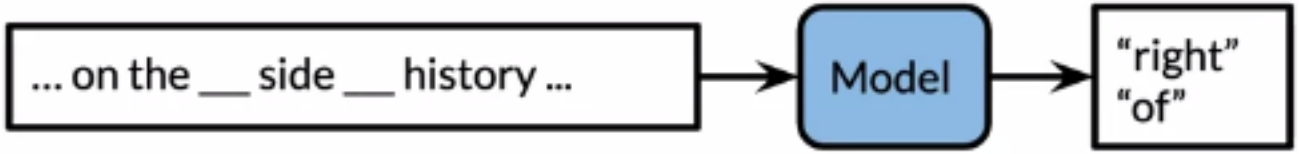
BERT Pre-training Tasks

Multi-Mask Language Modeling



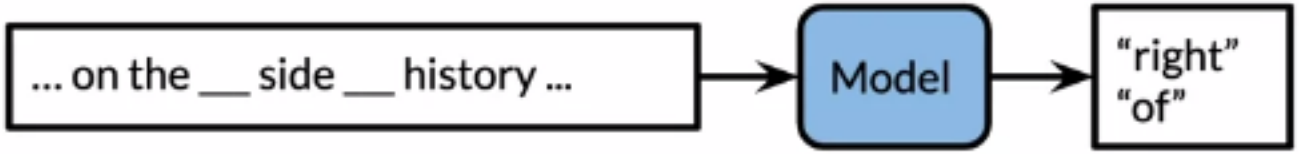
BERT Pre-training Tasks

Multi-Mask Language Modeling



BERT Pre-training Tasks

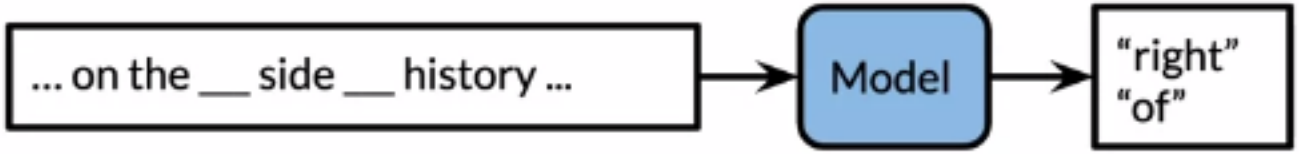
Multi-Mask Language Modeling



Next Sentence Prediction

BERT Pre-training Tasks

Multi-Mask Language Modeling

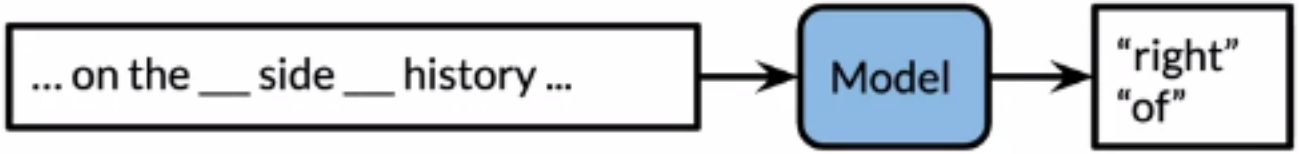


Next Sentence Prediction

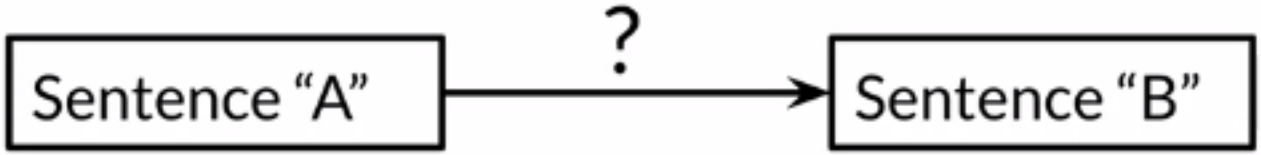
Sentence “A”

BERT Pre-training Tasks

Multi-Mask Language Modeling

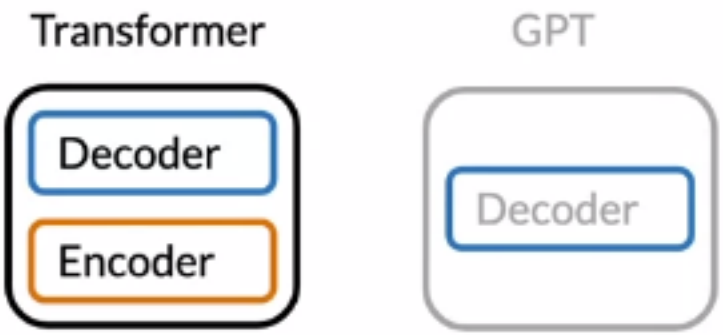


Next Sentence Prediction



T5: Encoder vs. Encoder-Decoder

T5: Encoder vs. Encoder-Decoder

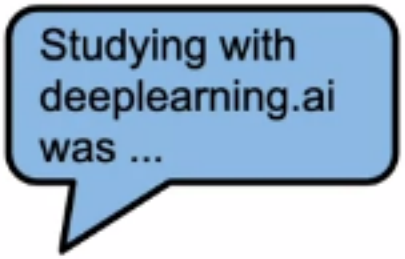


T5: Encoder vs. Encoder-Decoder

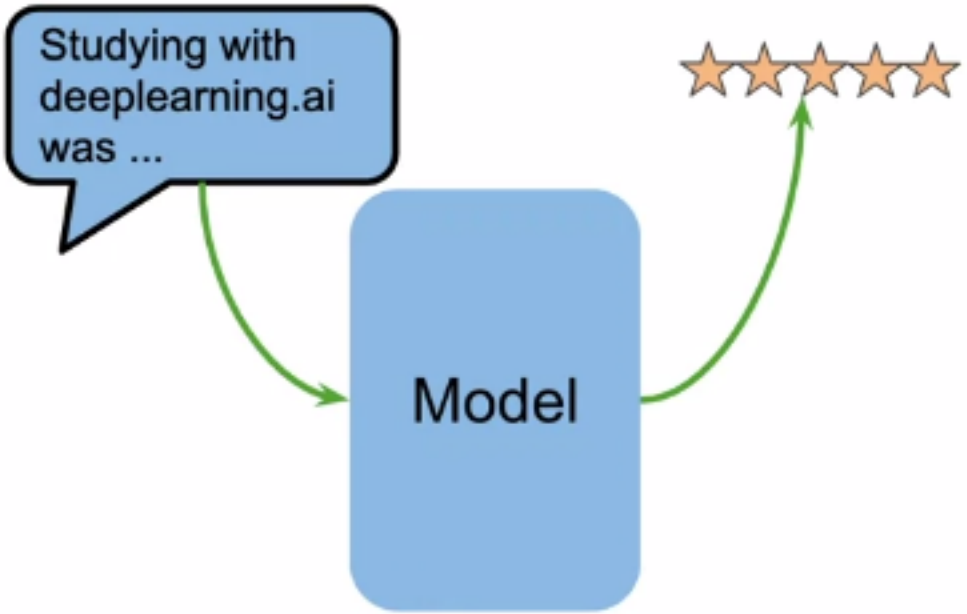
T5: Encoder vs. Encoder-Decoder

T5: Multi-task

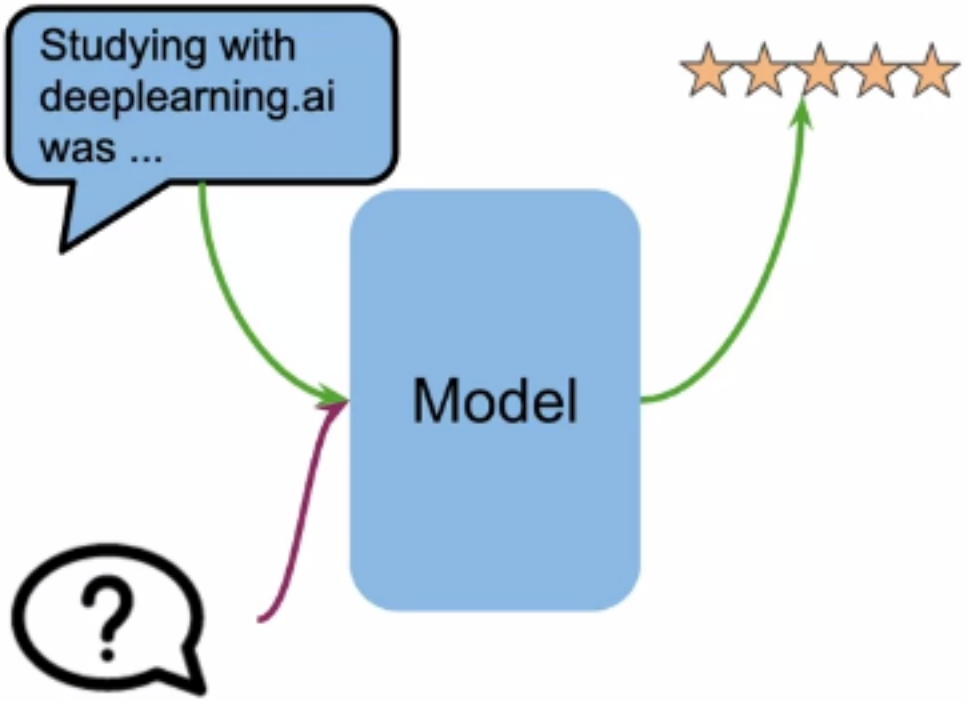
T5: Multi-task



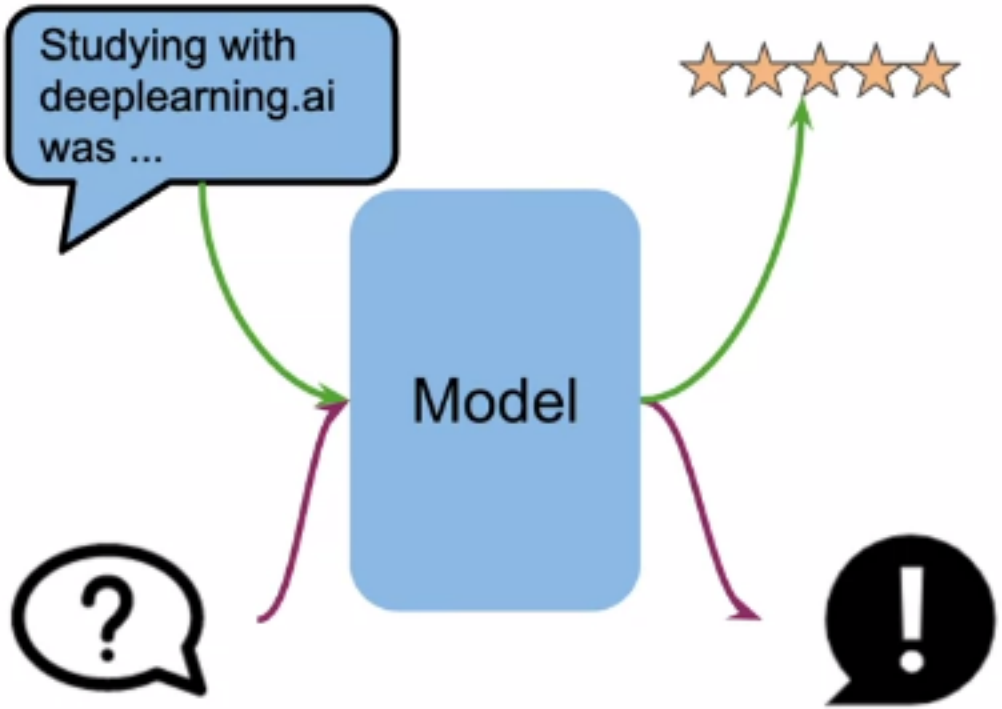
T5: Multi-task



T5: Multi-task

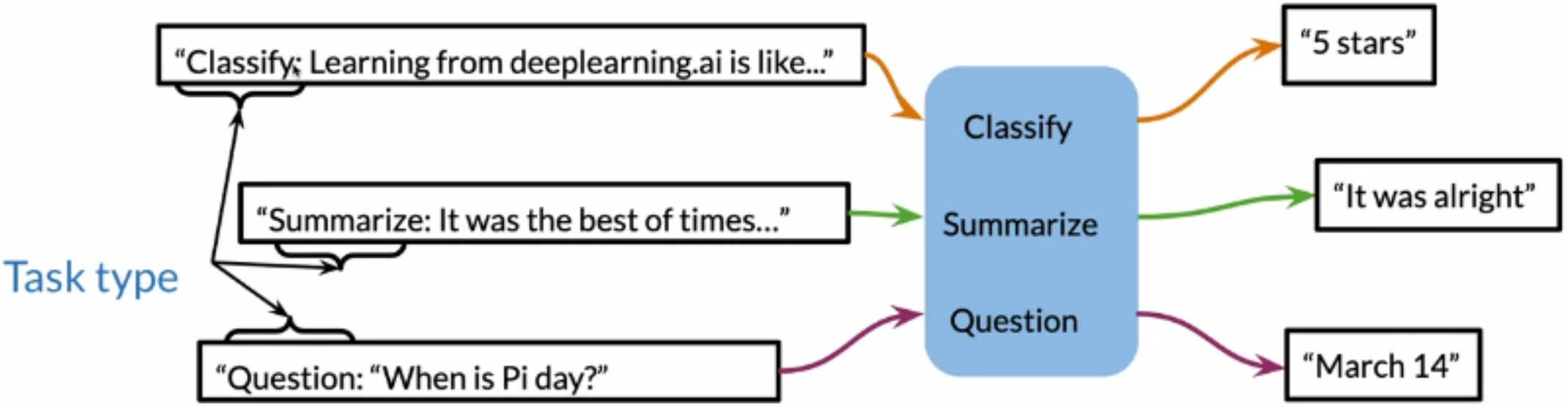


T5: Multi-task

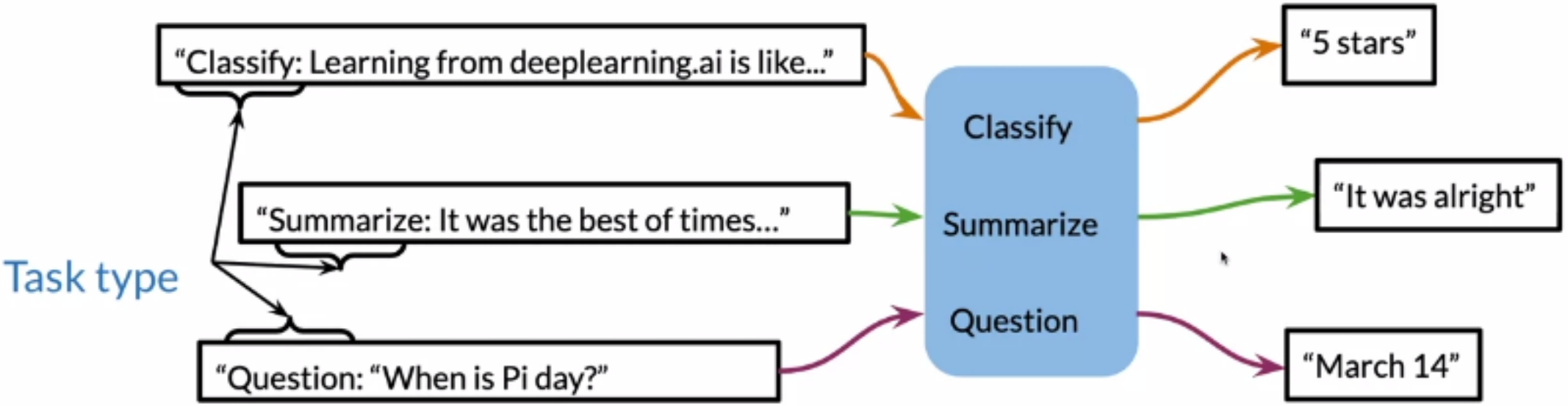


How?

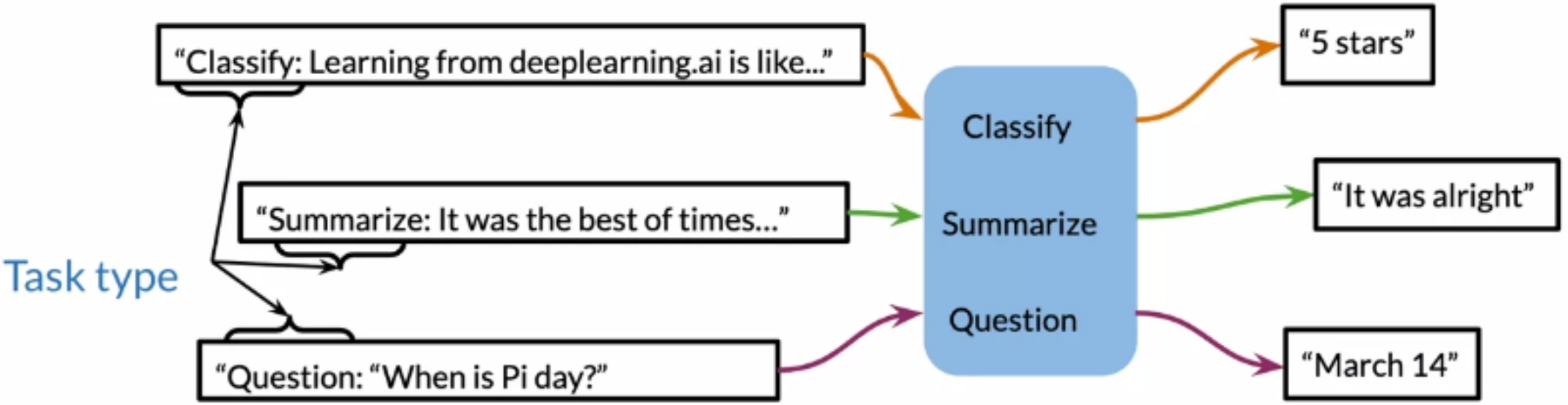
T5: Text-to-Text



T5: Text-to-Text



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	

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	

Summary

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	

Summary


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

Outline

- Learn about the BERT architecture
- Understand how BERT pre-training works

BERT

- Makes use of transfer learning/pre-training:

BERT

- Makes use of transfer learning/pre-training:

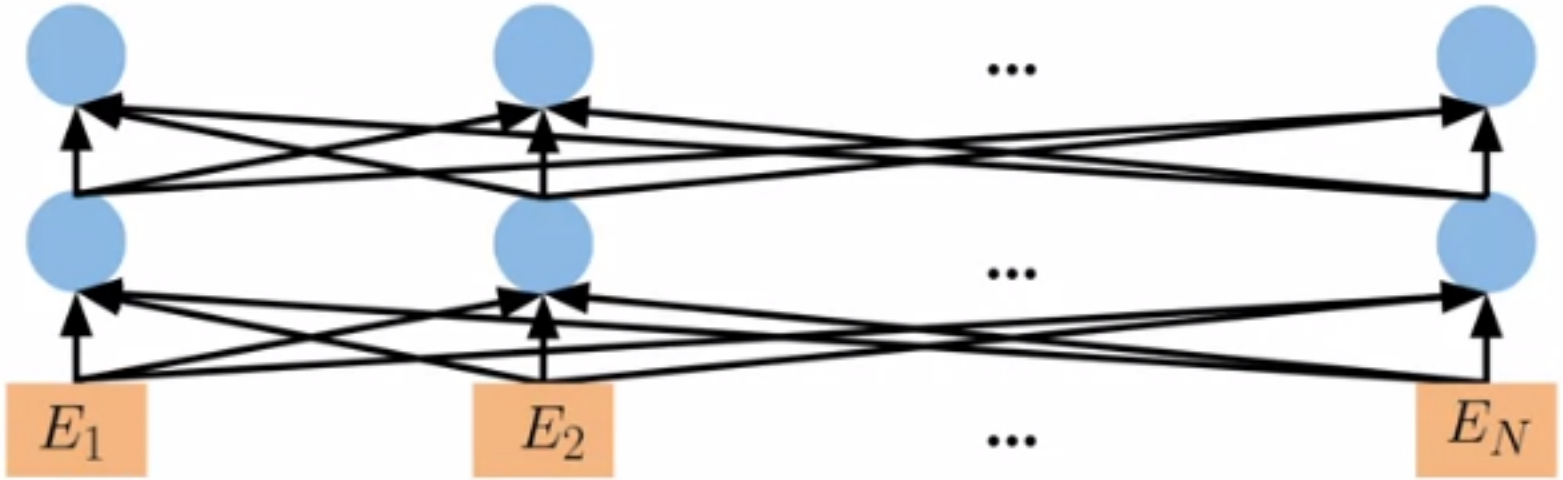
E_1 E_2 ... E_N

BERT

- Makes use of transfer learning/pre-training:

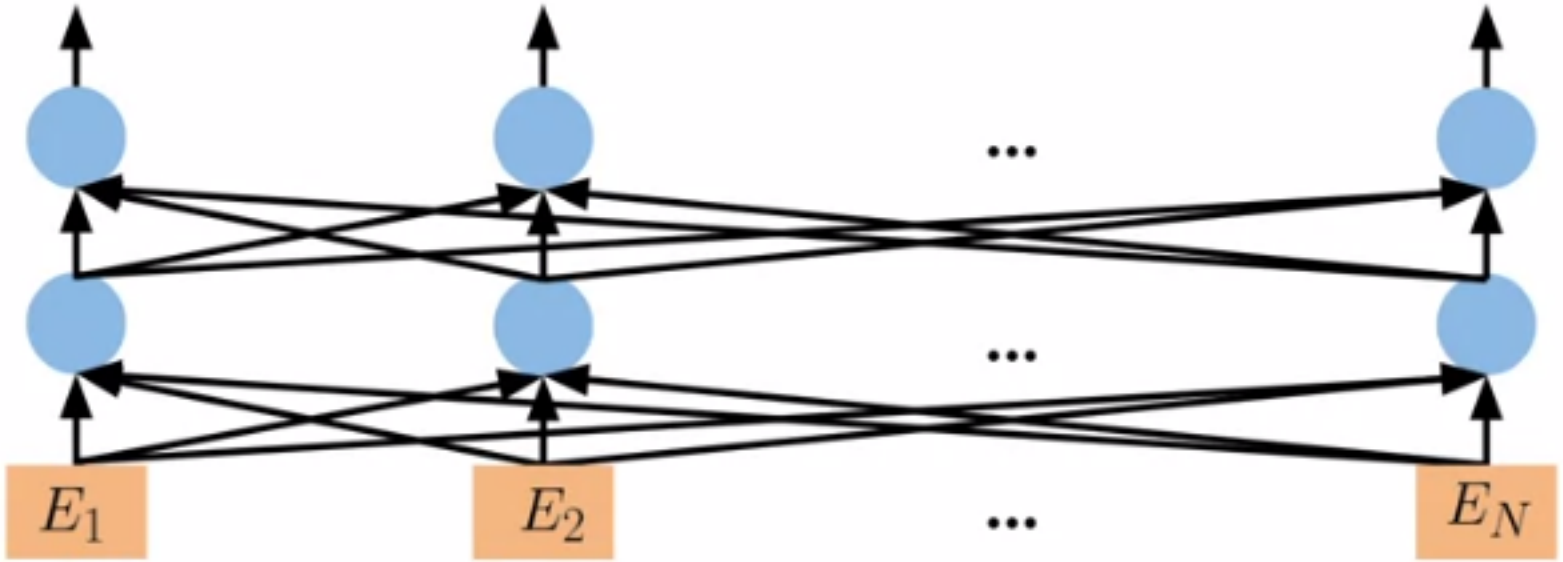
BERT

- Makes use of transfer learning/pre-training:



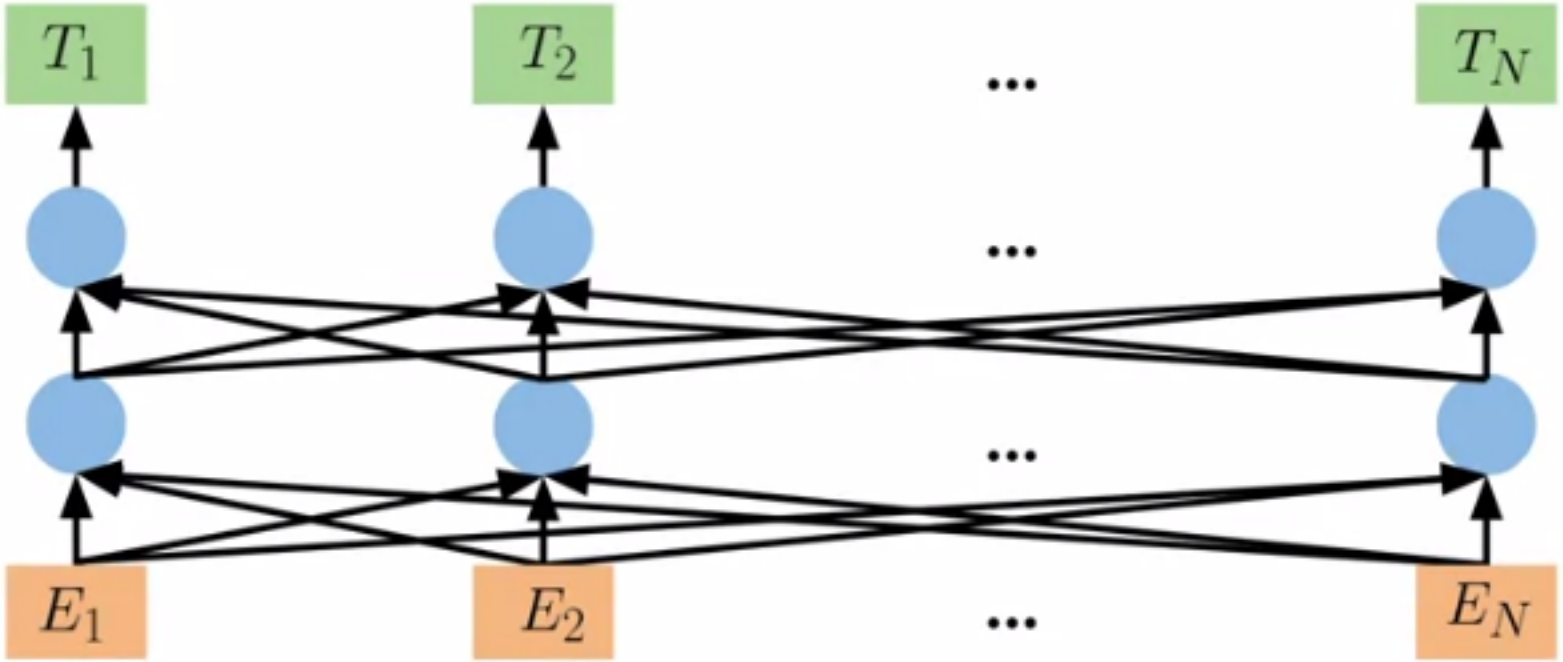
BERT

- Makes use of transfer learning/pre-training:



BERT

- Makes use of transfer learning/pre-training:



BERT

- A multi layer bidirectional transformer

BERT

- A multi layer bidirectional transformer
- Positional embeddings

BERT

- A multi layer bidirectional transformer
- Positional embeddings
- BERT_base:
    - 12 layers (12 transformer blocks)
    - 12 attentions heads
    - 110 million parameters

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.

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.

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.

Outline

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

Formalizing the input

Formalizing the input

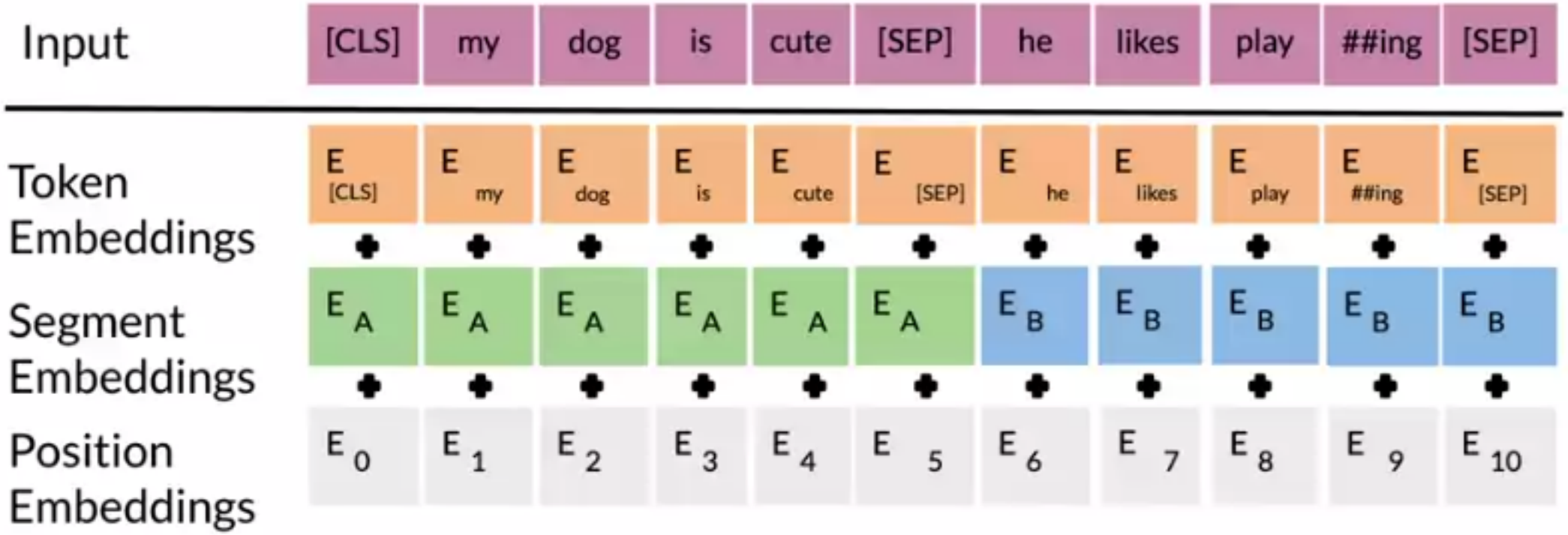
Position Embeddings

Formalizing the input

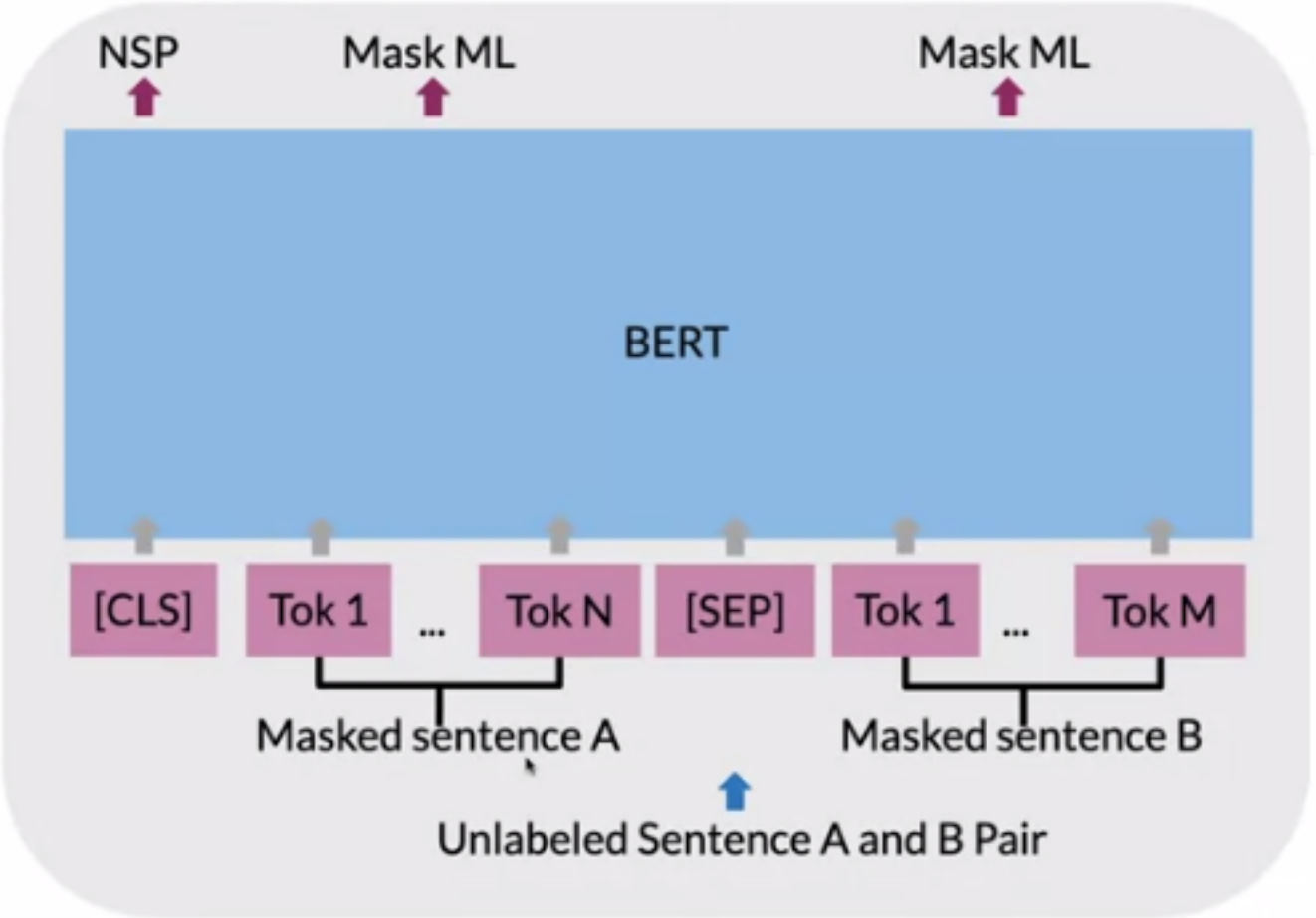
Formalizing the input

Formalizing the input

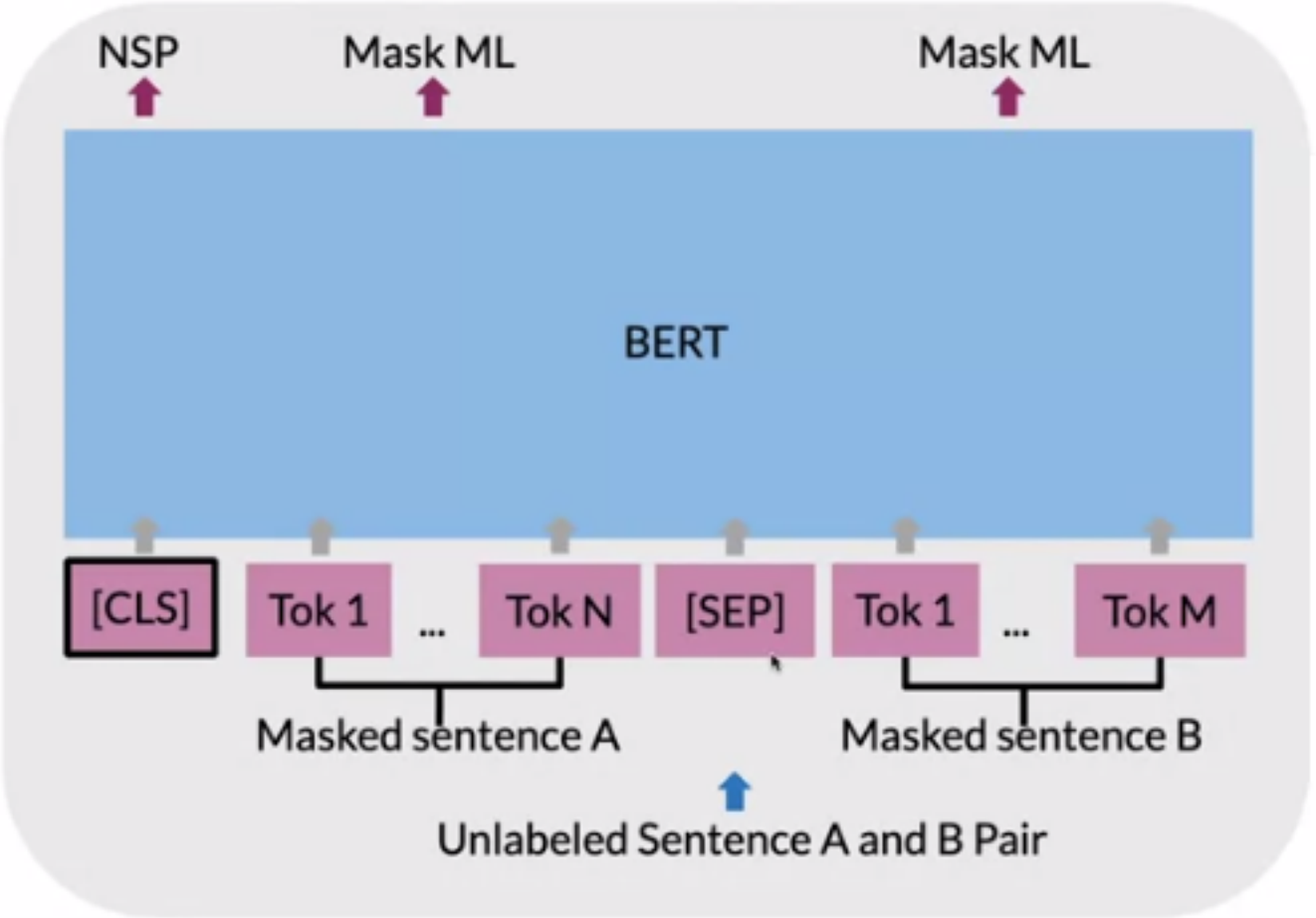
Formalizing the input



Visualizing the output

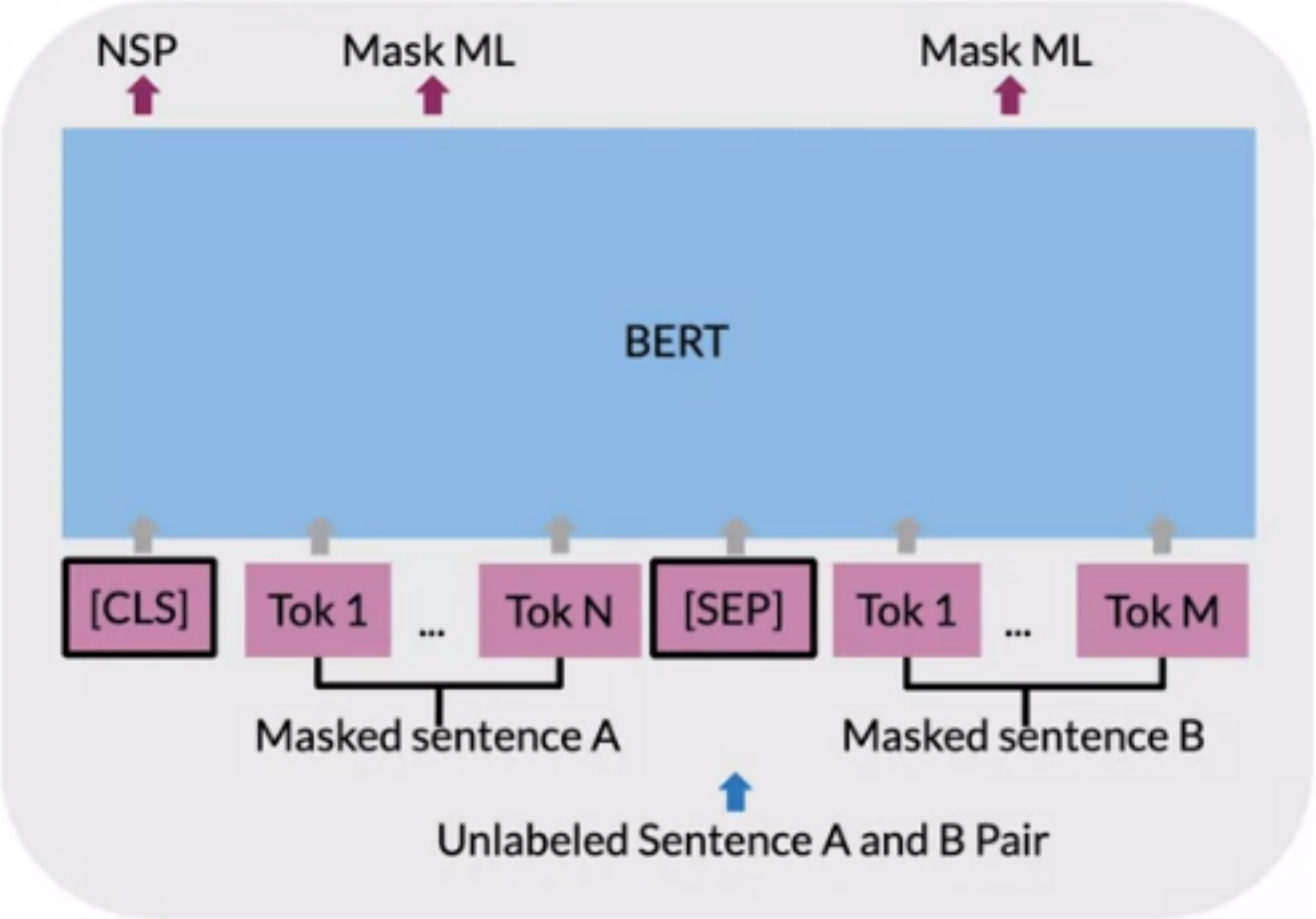


Visualizing the output



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

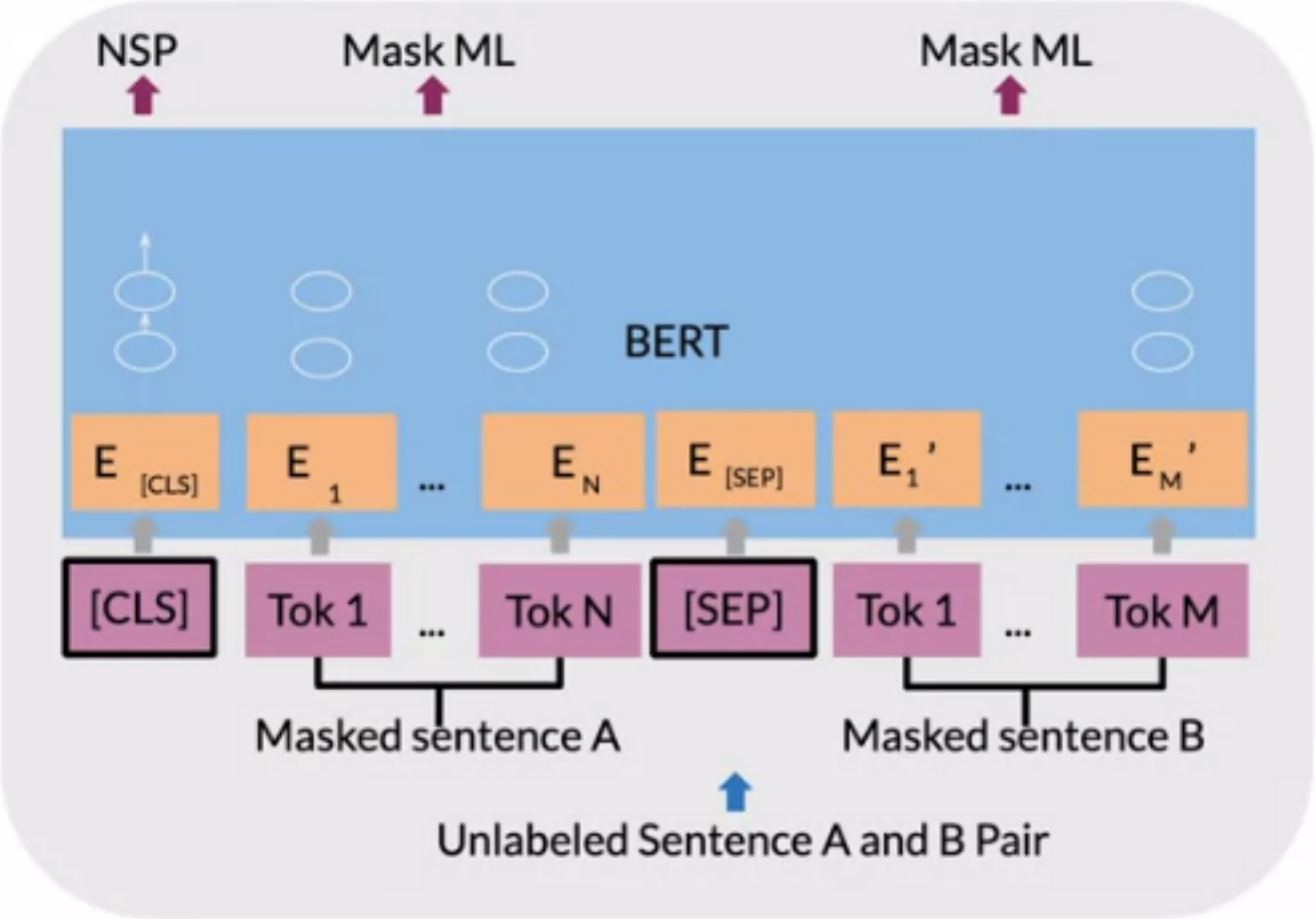
Visualizing the output



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

- [SEP]: a special separator token

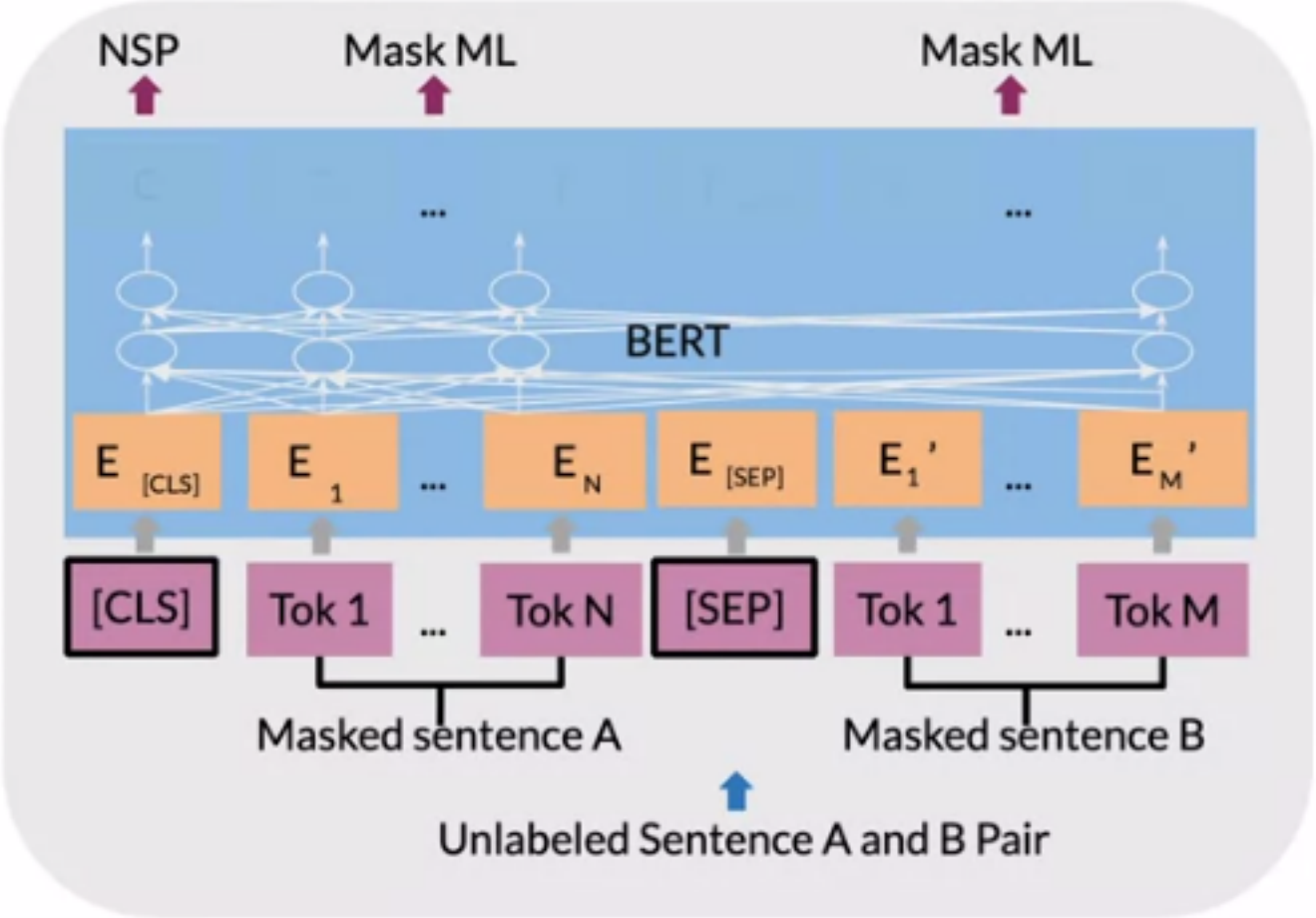
Visualizing the output



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

- [SEP]: a special separator token

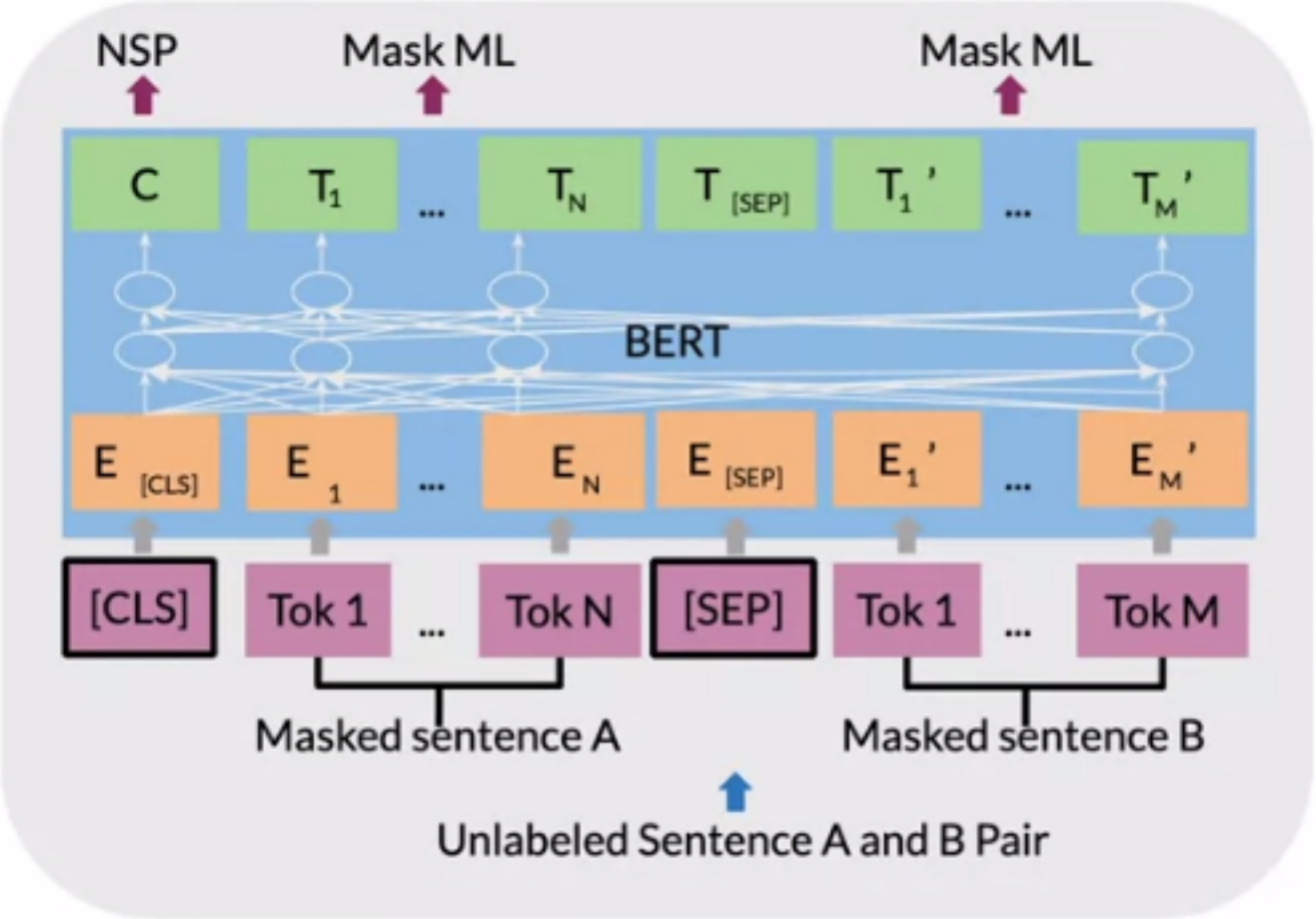
Visualizing the output



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

- [SEP]: a special separator token

Visualizing the output



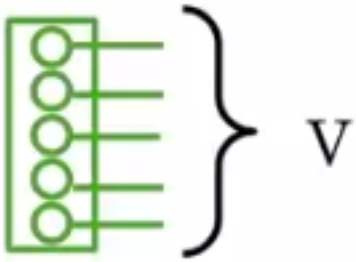
- [CLS]: a special classification symbol added in front of every input

- [SEP]: a special separator token

BERT Objective

Objective 1:
Multi-Mask LM

Loss: Cross Entropy Loss



BERT Objective

Objective 1:
Multi-Mask LM

Loss: Cross Entropy Loss

Objective 2:
Next Sentence Prediction

Loss: Binary Loss

BERT Objective

Objective 1:
Multi-Mask LM

Loss: Cross Entropy Loss

Objective 2:
Next Sentence Prediction

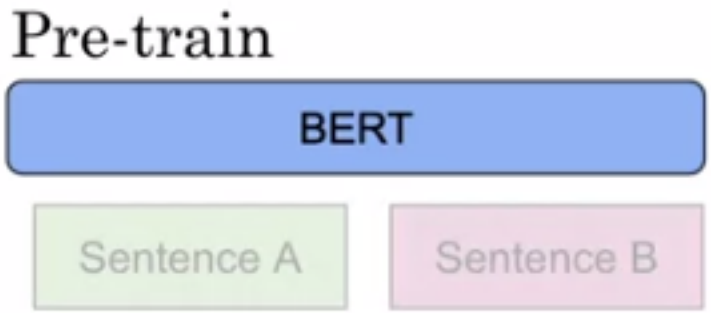
Loss: Binary Loss

Summary

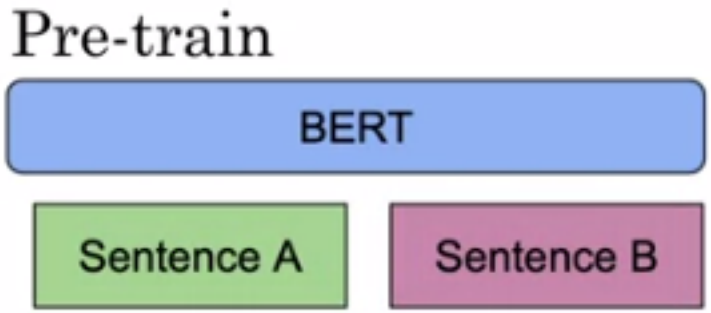
- BERT objective
- Model inputs/outputs

Fine-tuning BERT: Outline

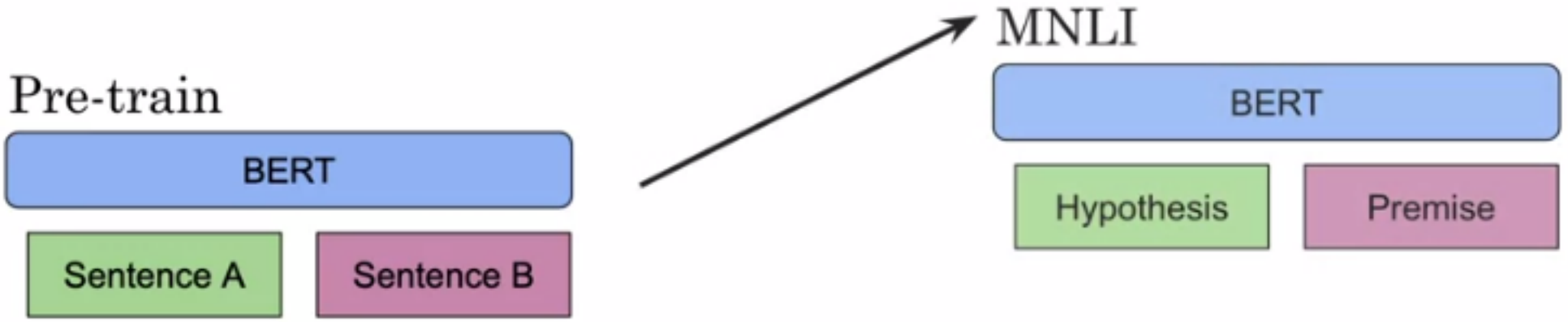
Fine-tuning BERT: Outline



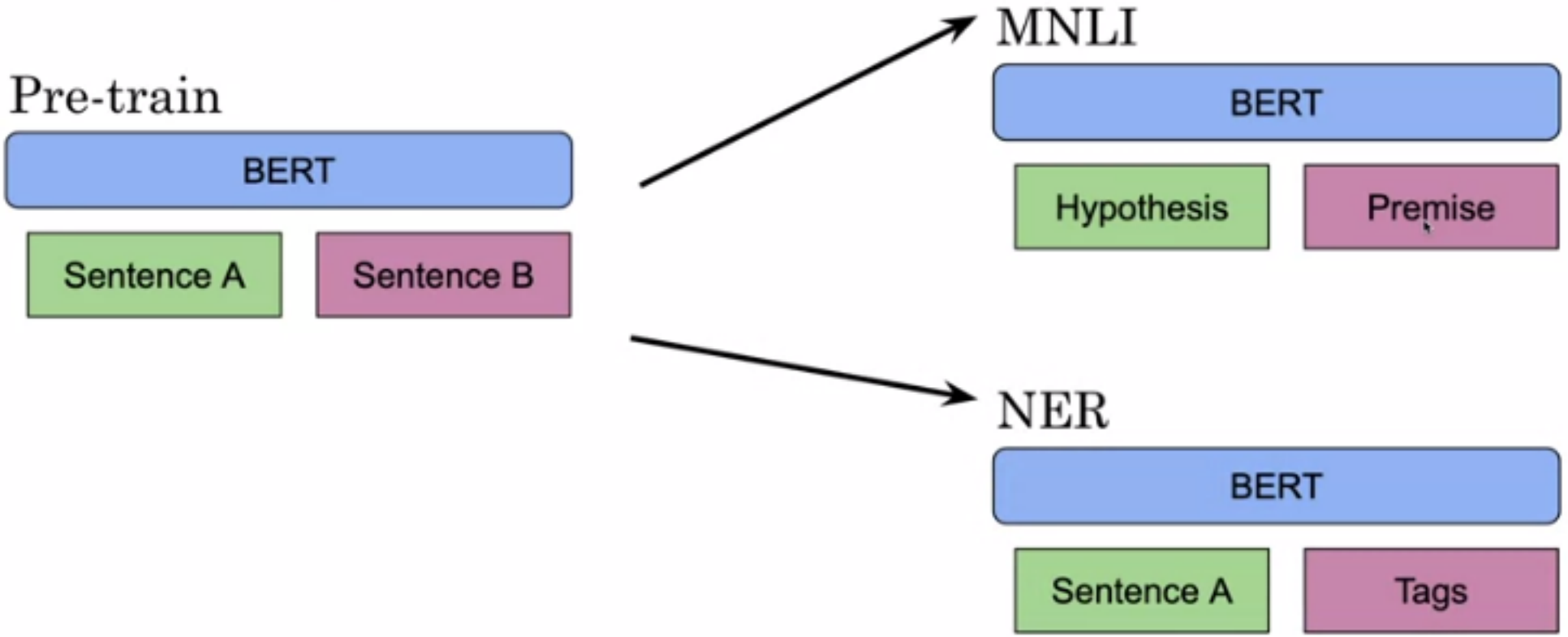
Fine-tuning BERT: Outline



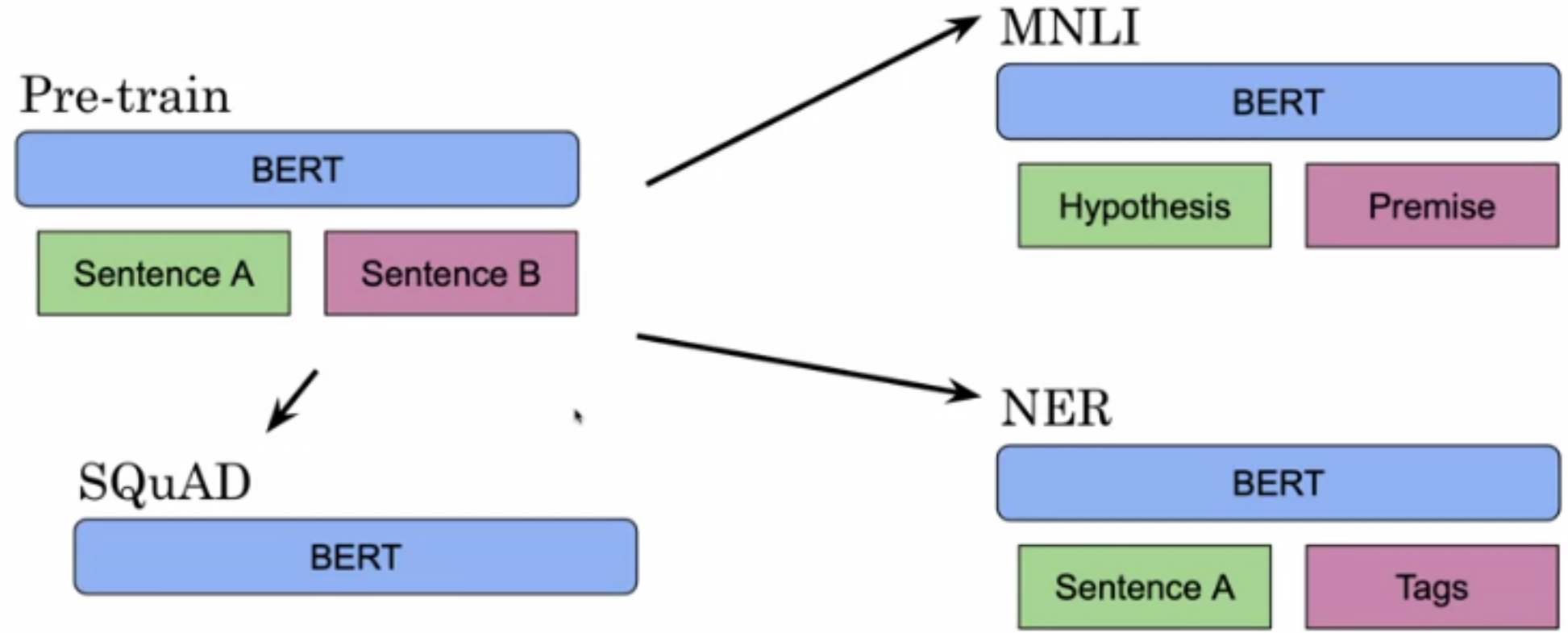
Fine-tuning BERT: Outline



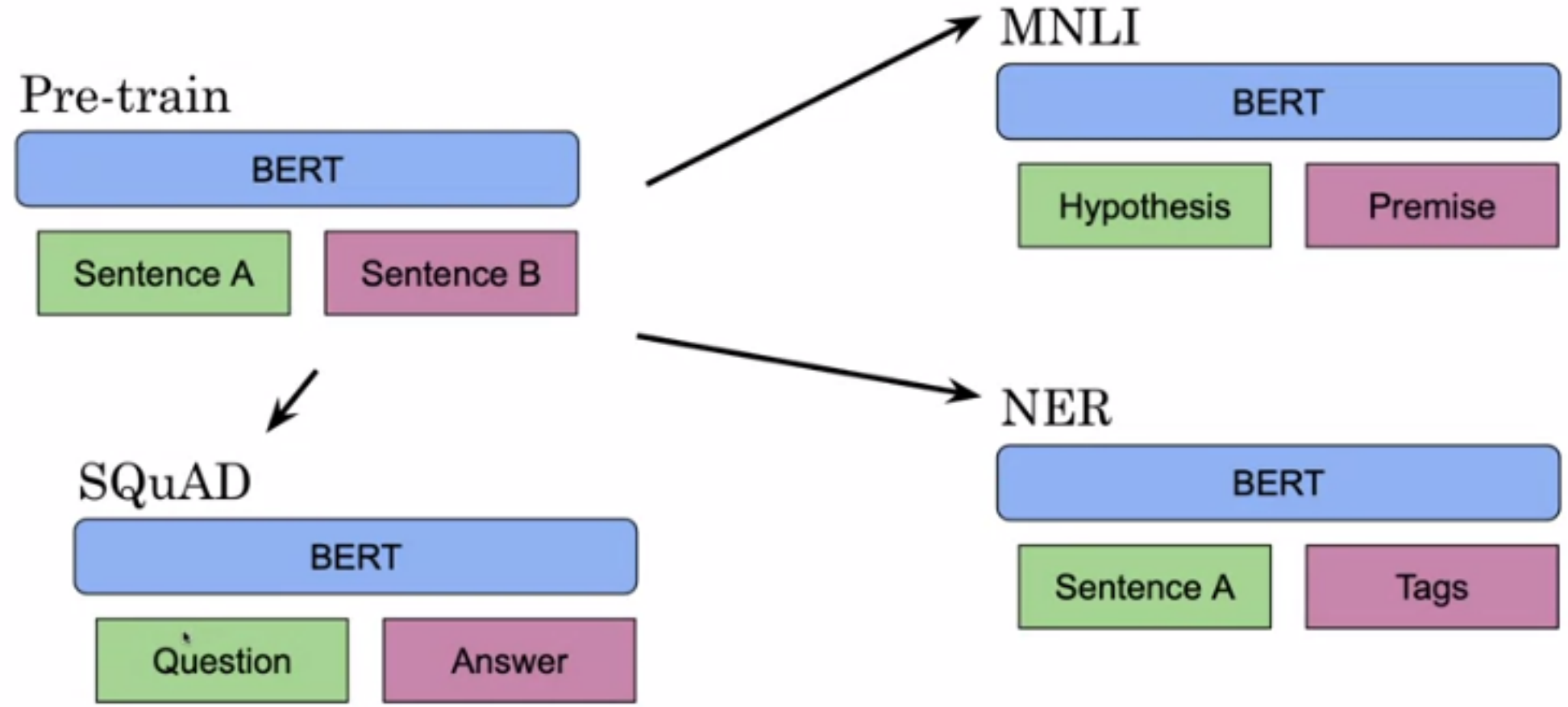
Fine-tuning BERT: Outline



Fine-tuning BERT: Outline

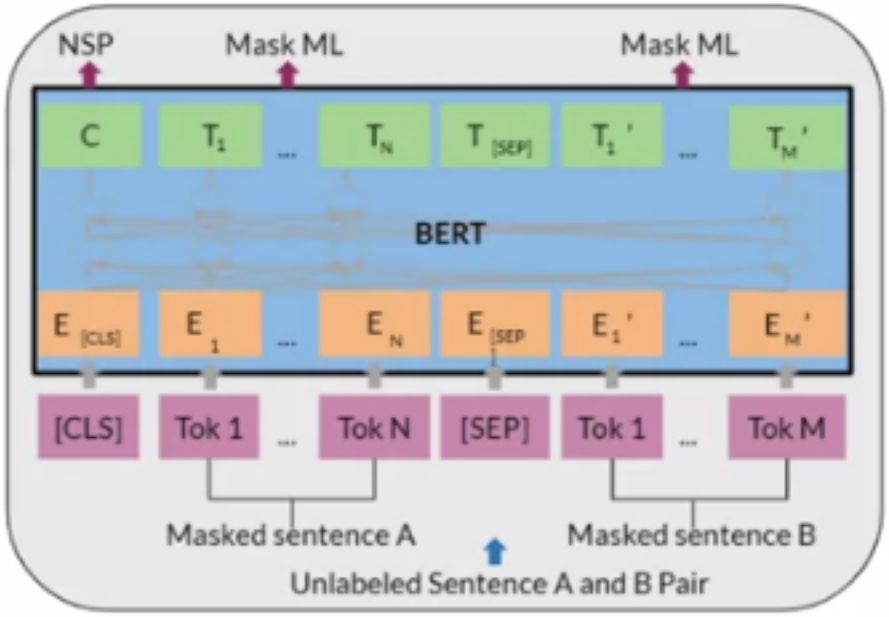


Fine-tuning BERT: Outline

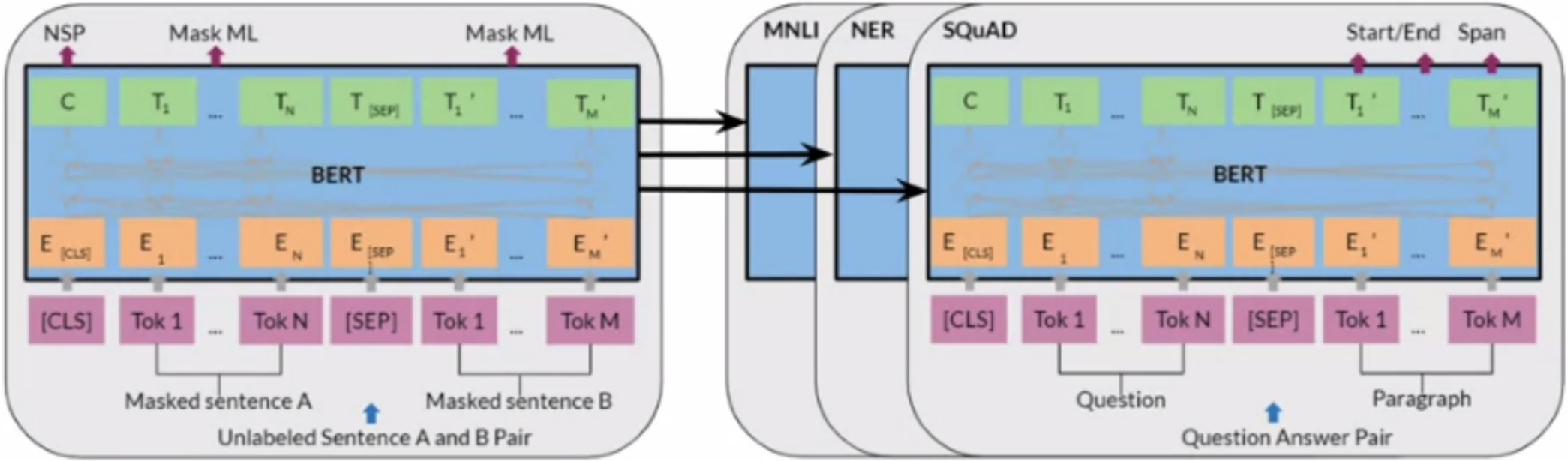


Inputs

Inputs

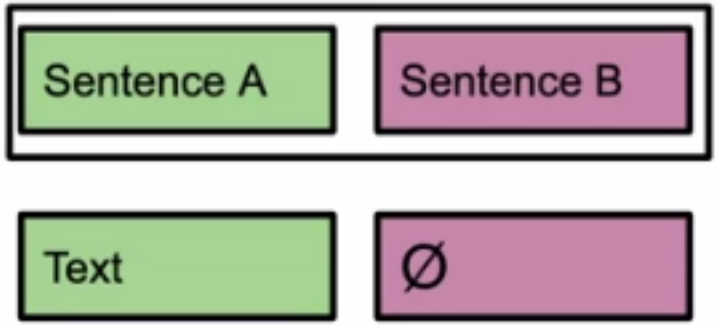

NSP
Mask ML
Mask ML
C
T1
...
TN
T [SEP]
T1'
...
TM'
BERT
E [CLS]
E1
...
EN
E [SEP]
E1'
...
EM'
[CLS]
Tok 1
...
Tok N
[SEP]
Tok 1
...
Tok M
Masked sentence A
Masked sentence B
Unlabeled Sentence A and B Pair

Inputs

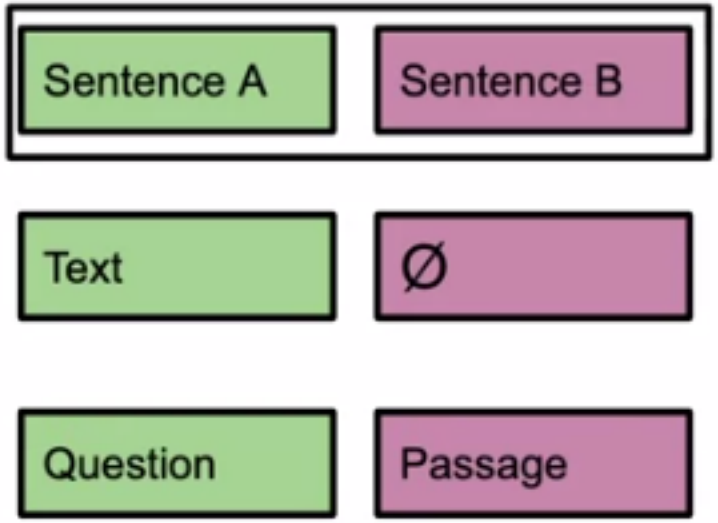

NSP
Mask ML
Mask ML
C
T1
...
TN
T[SEP]
T1'
...
TM'
BERT
E[CLS]
E1
...
EN
E[SEP]
E1'
...
EM'
[CLS]
Tok 1
...
Tok N
[SEP]
Tok 1
...
Tok M
Masked sentence A
Masked sentence B
Unlabeled Sentence A and B Pair
MNLI
NER
SQuAD
Start/End Span
Question
Paragraph
Question Answer Pair

Summary

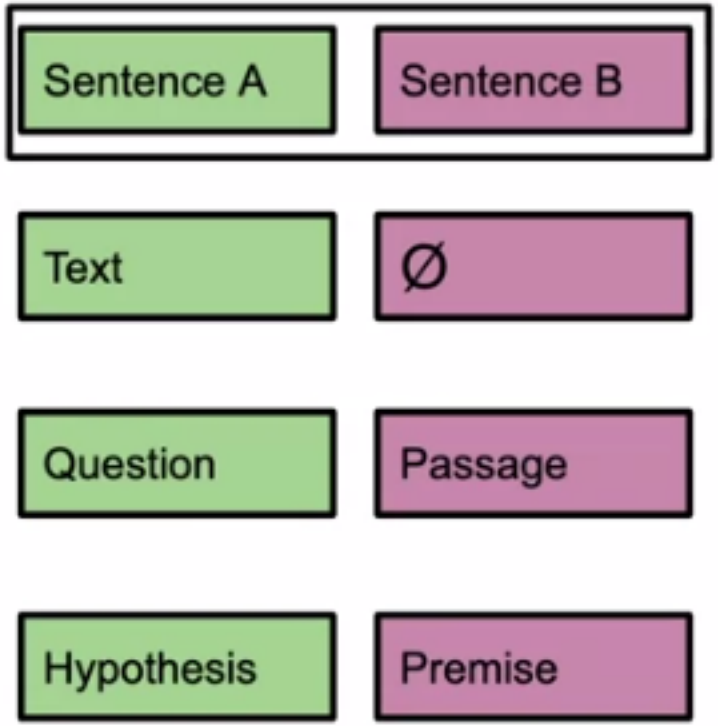
Summary



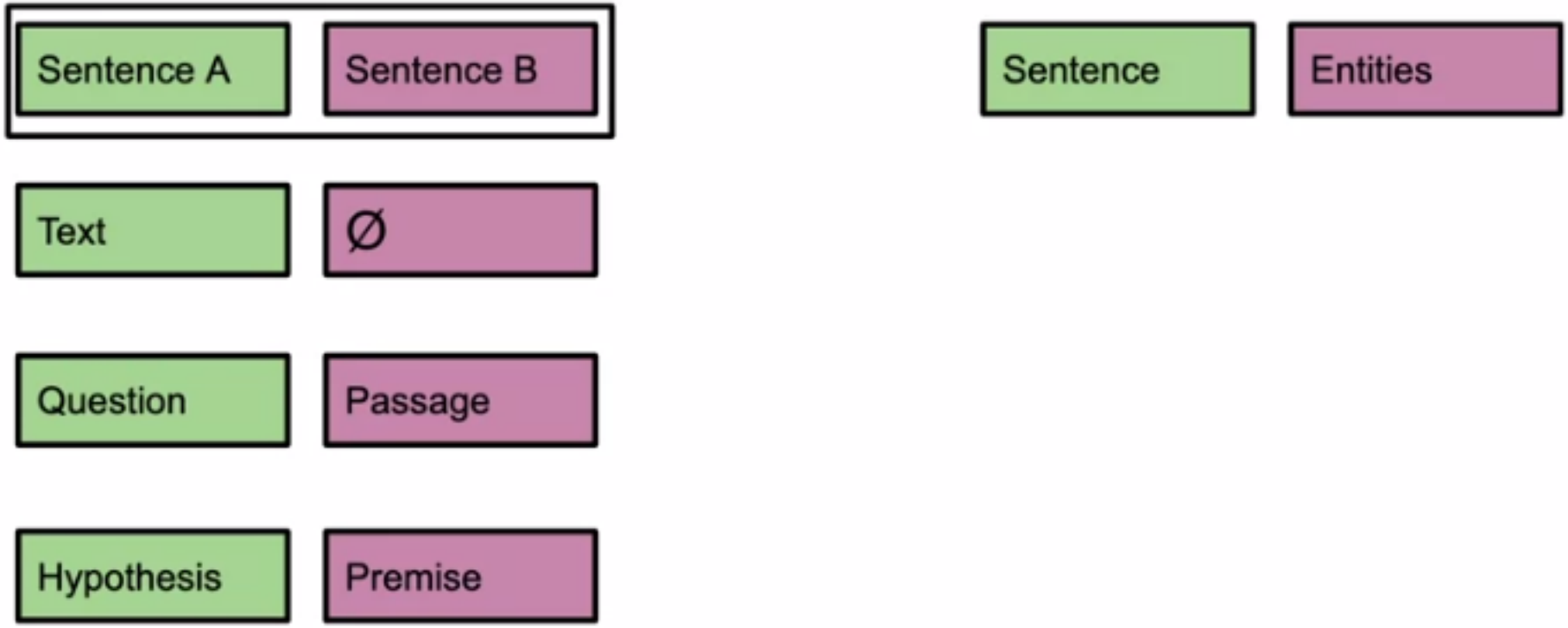
Summary



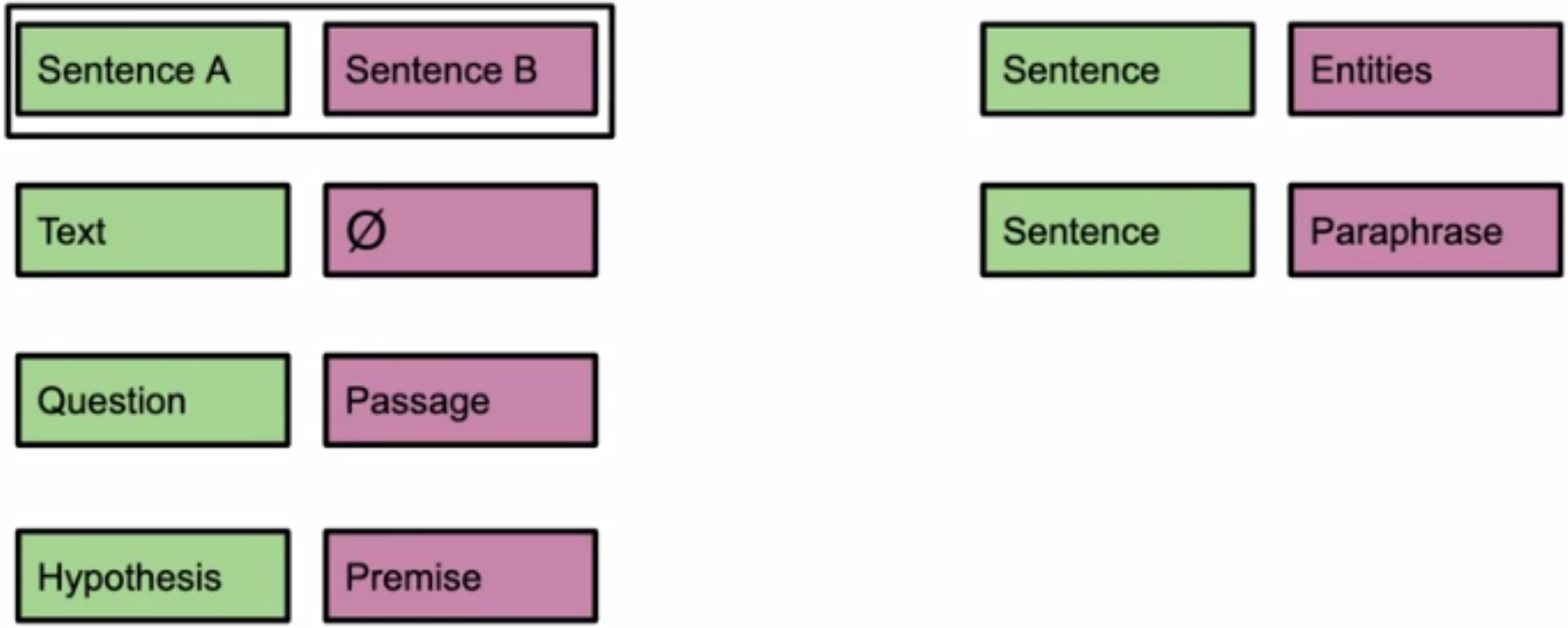
Summary


Sentence A
Sentence B
Text
Ø
Question
Passage
Hypothesis
Premise

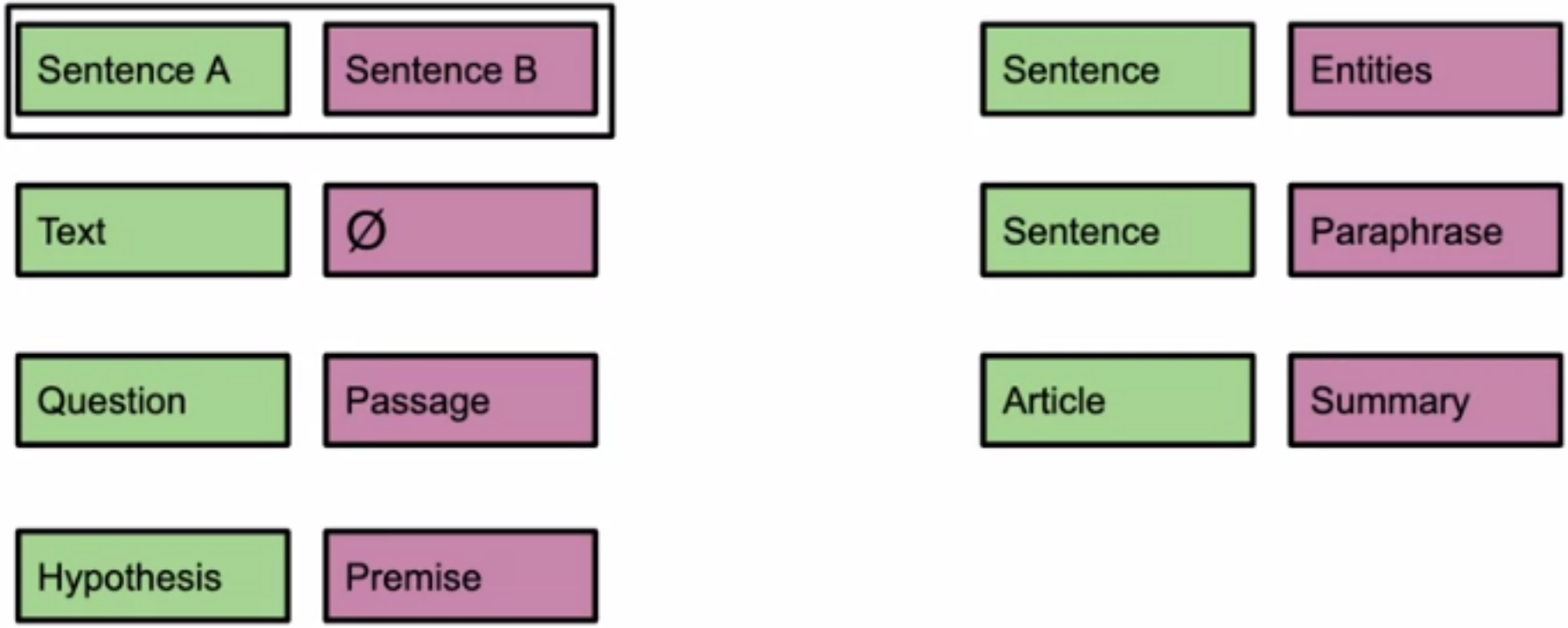
Summary


Sentence A
Sentence B
Text
Ø
Question
Passage
Hypothesis
Premise
Sentence
Entities

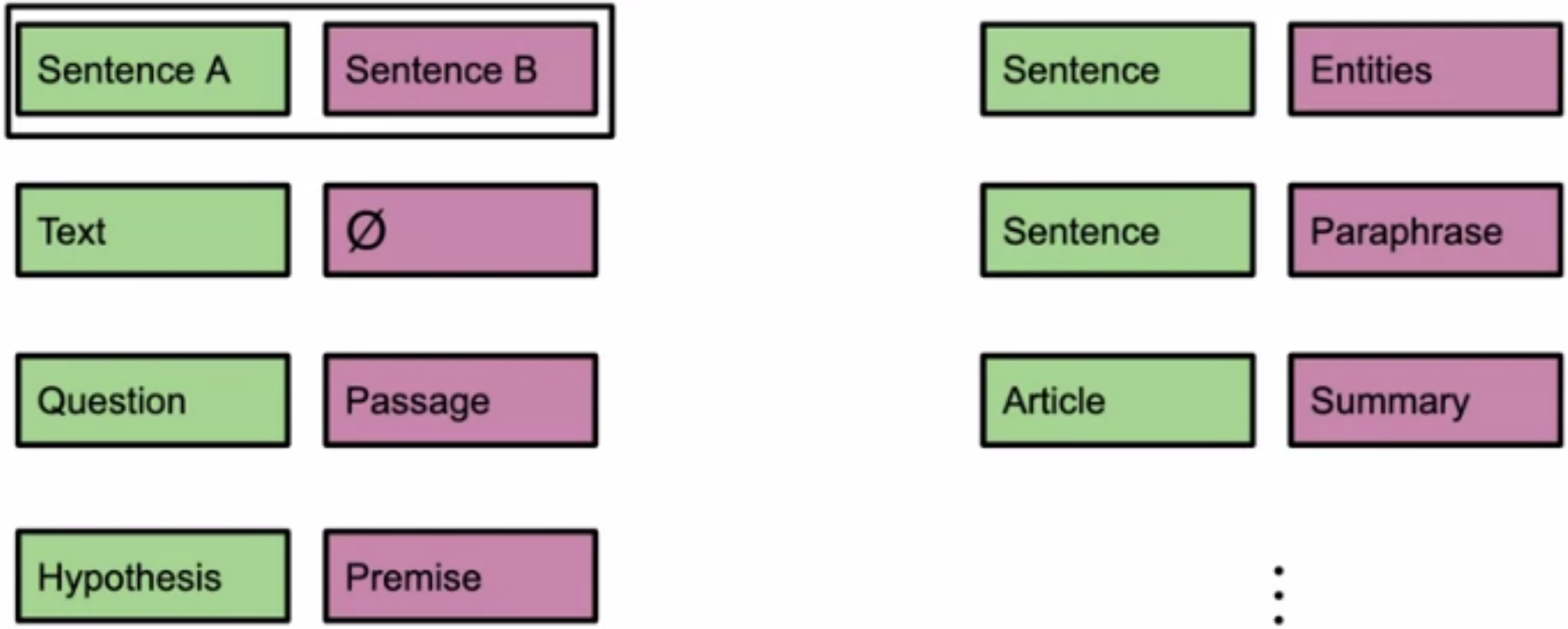
Summary


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

Summary


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

Summary


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

Outline

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

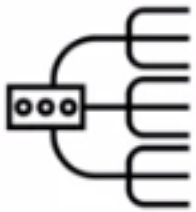
Transformer - T5 Model

Text to Text

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

Transformer - T5 Model

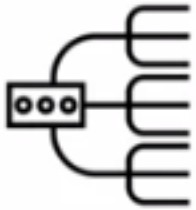
Text to Text



Classification

Transformer - T5 Model

Text to Text



Classification

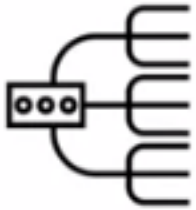
Question Answering (Q&A)

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

Transformer - T5 Model

Text to Text

Machine Translation

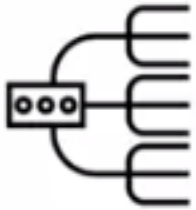


Classification

Question Answering (Q&A)

Transformer - T5 Model

Text to Text



Classification

Question Answering (Q&A)

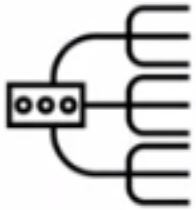
Machine Translation

Summarization

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

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.

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

Transformer - T5 Model

Original text

Thank you for inviting me to your party last week.

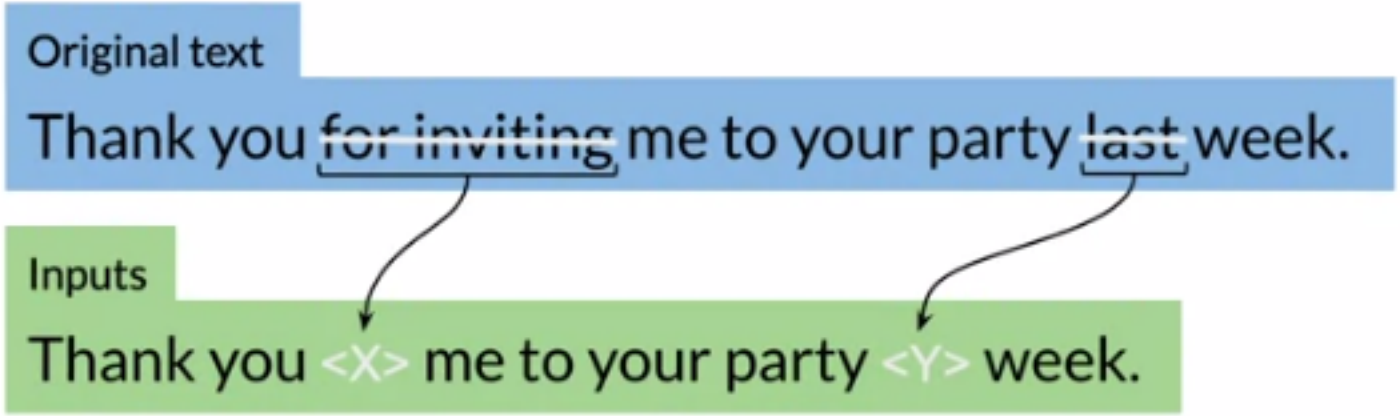
©Exploring the Limits of Transfer learning with a unified text to Text Transformer. Raffel et. al. 2020

Transformer - T5 Model

Original text

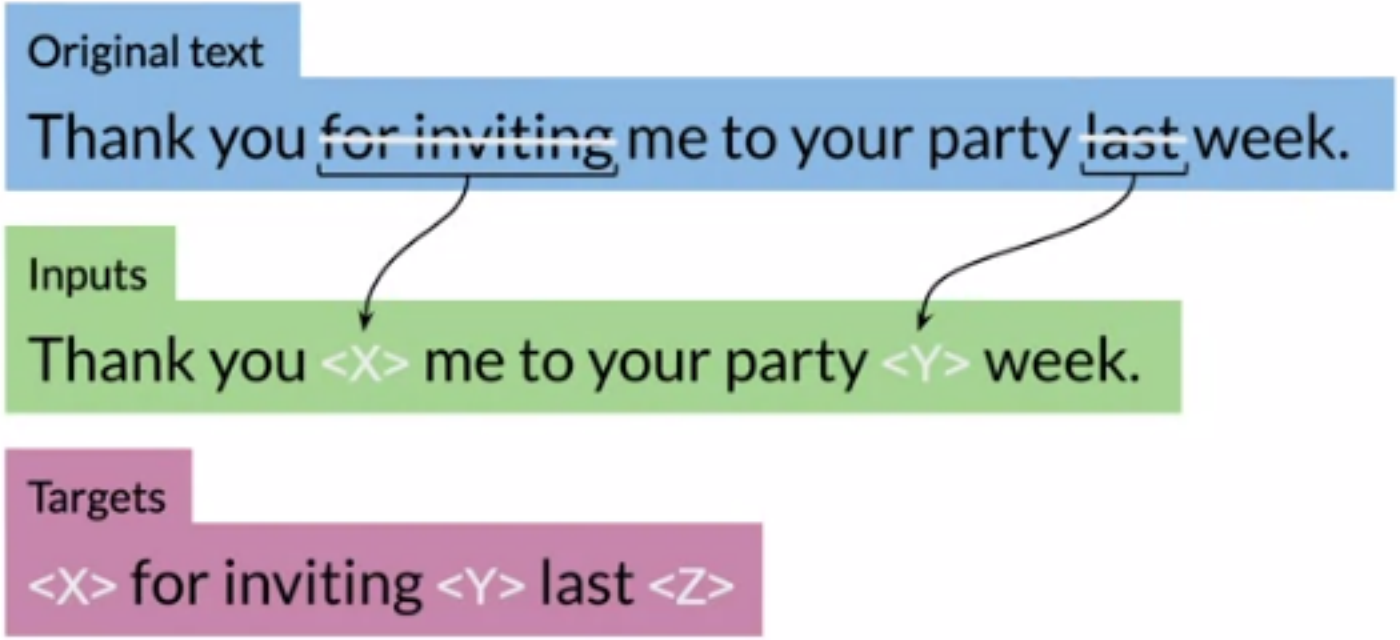
Thank you ~~for inviting~~ me to your party ~~last~~ week.

Transformer - T5 Model



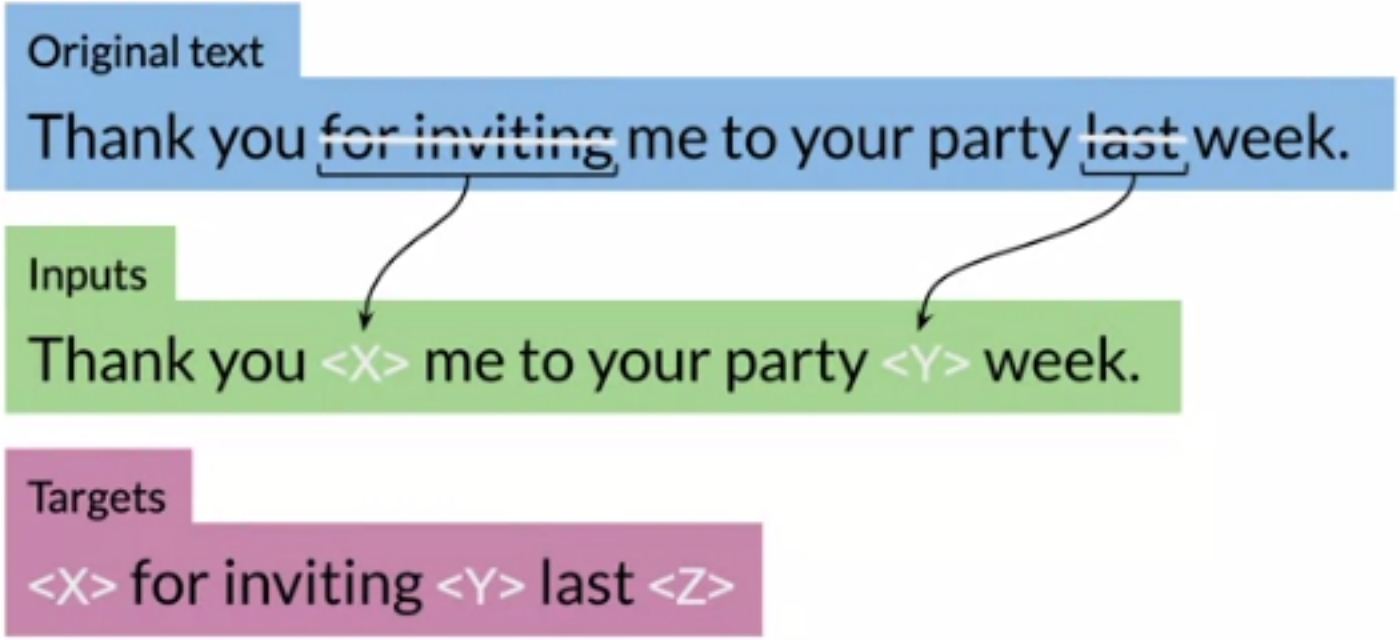
©Exploring the Limits of Transfer learning with a unified text to Text Transformer. Raffel et. al. 2020

Transformer - T5 Model



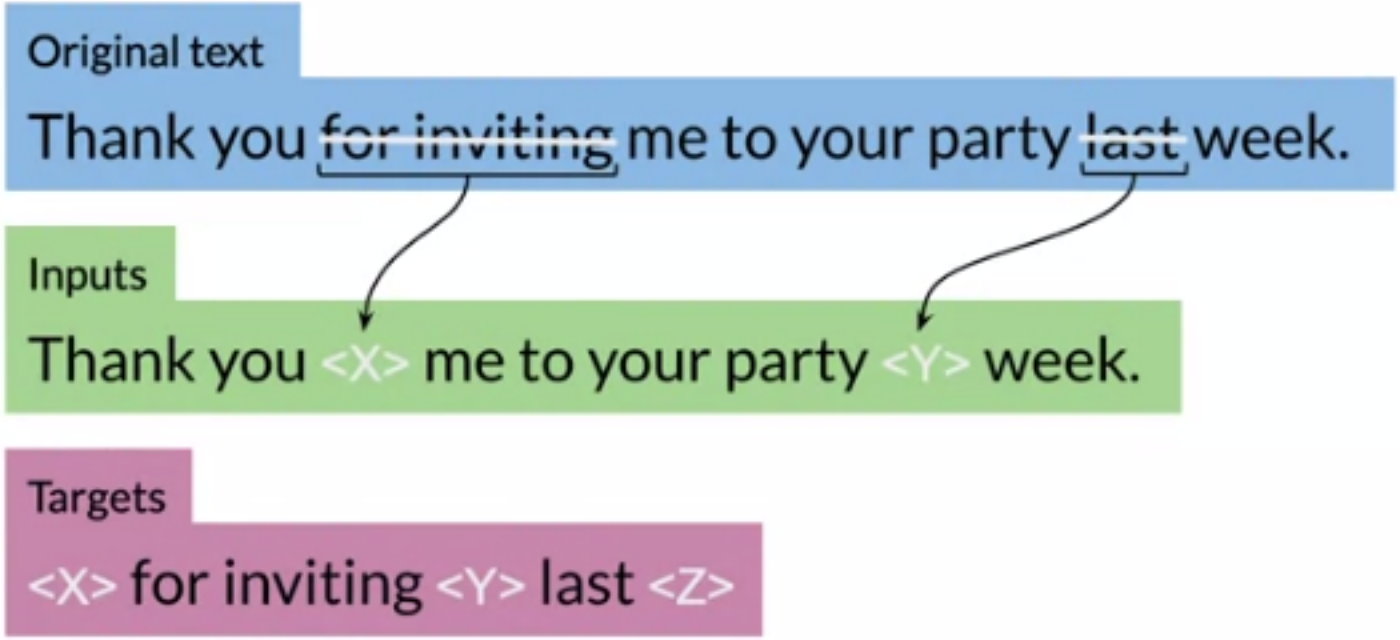
©Exploring the Limits of Transfer learning with a unified text to Text Transformer. Raffel et. al. 2020

Transformer - T5 Model



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

Transformer - T5 Model



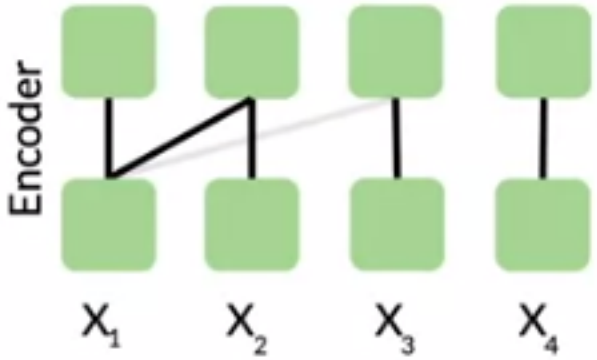
©Exploring the Limits of Transfer learning with a unified text to Text Transformer. Raffel et. al. 2020

Model Architecture

X_1 X_2 X_3 X_4

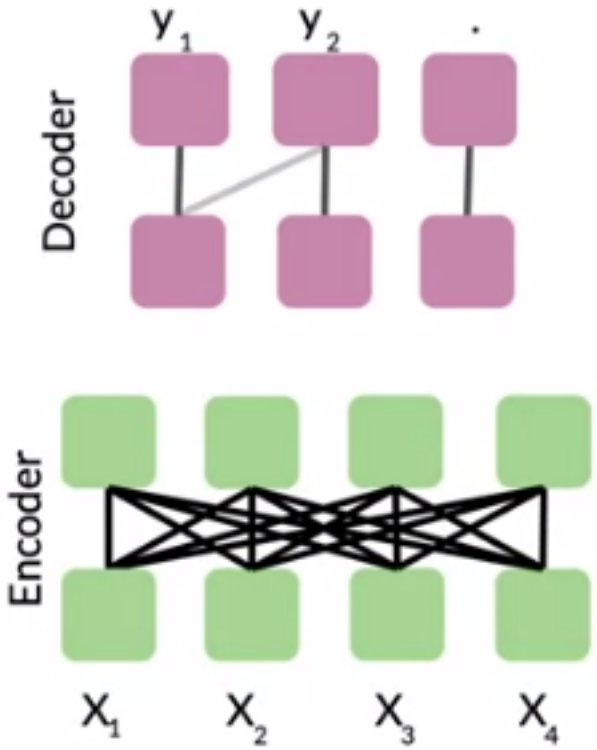
©Exploring the Limits of Transfer learning with a unified text to Text Transformer. Raffel et. al. 2020

Model Architecture



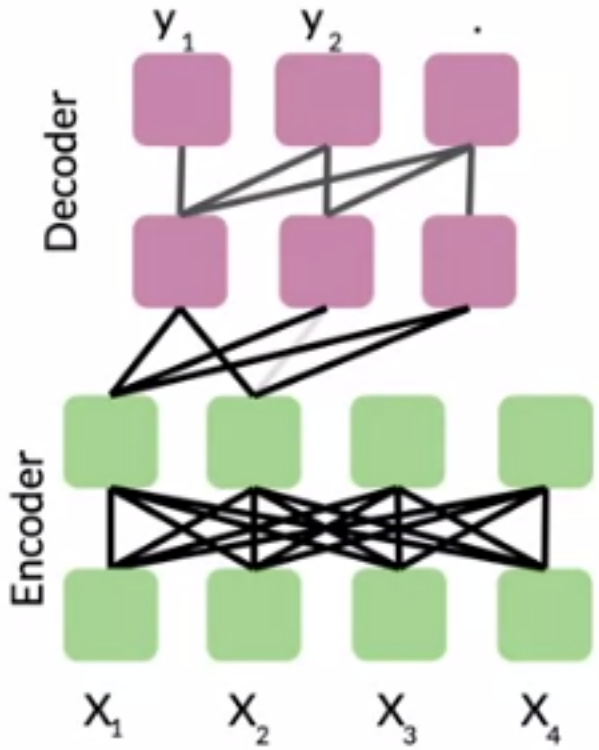
©Exploring the Limits of Transfer learning with a unified text to Text Transformer. Raffel et. al. 2020

Model Architecture



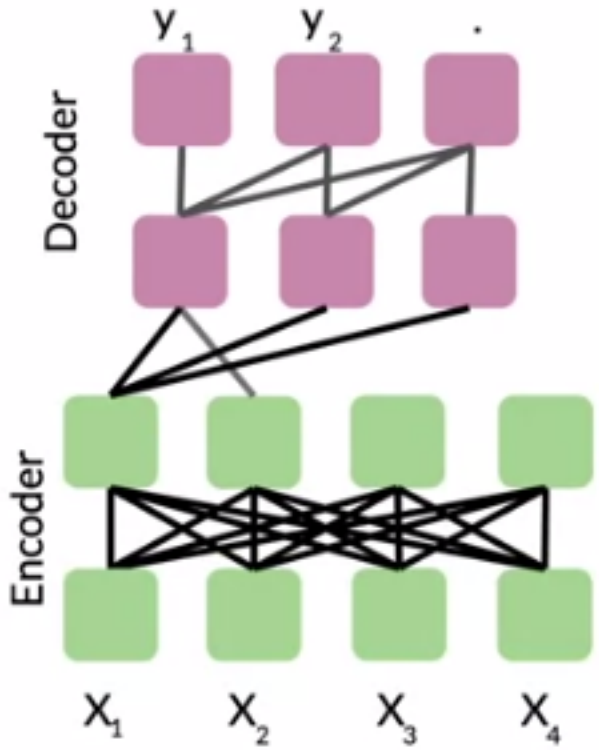
©Exploring the Limits of Transfer learning with a unified text to Text Transformer. Raffel et. al. 2020

Model Architecture



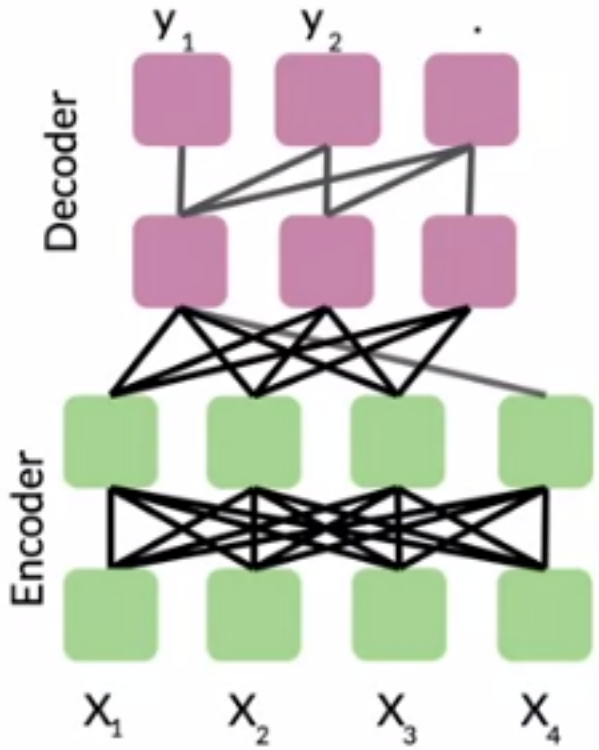
©Exploring the Limits of Transfer learning with a unified text to Text Transformer. Raffel et. al. 2020

Model Architecture



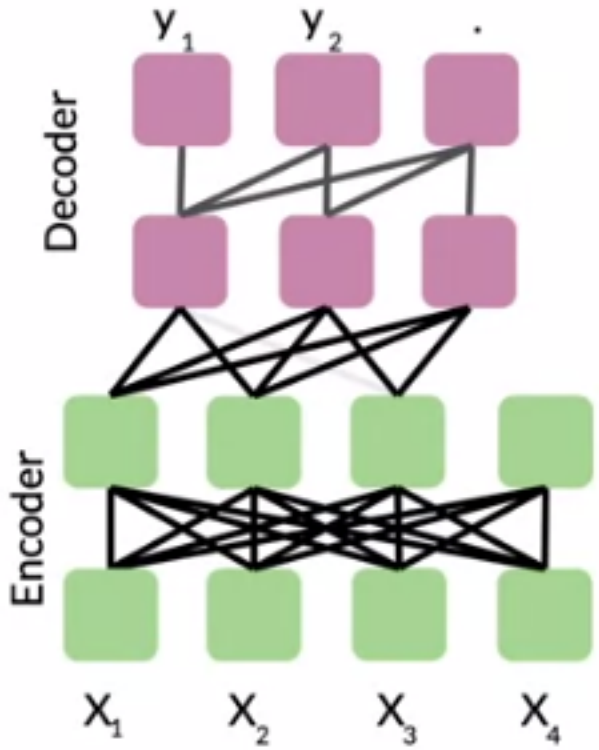
©Exploring the Limits of Transfer learning with a unified text to Text Transformer. Raffel et. al. 2020

Model Architecture



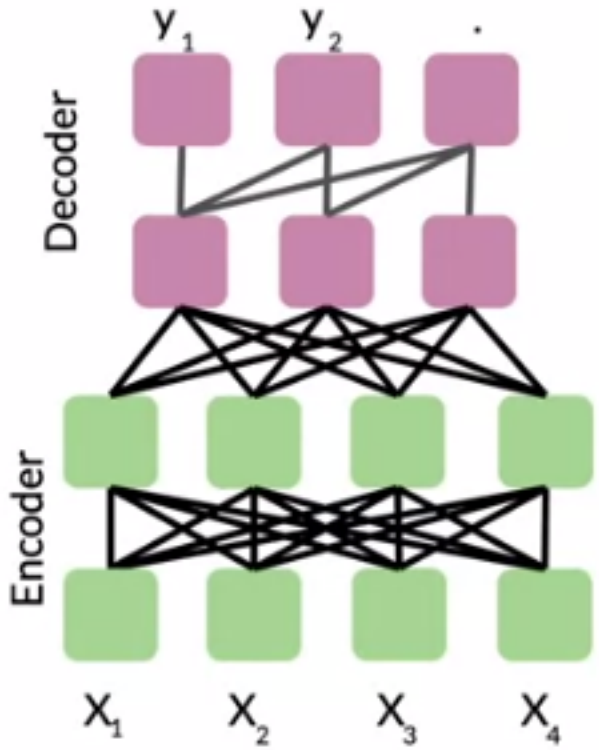
©Exploring the Limits of Transfer learning with a unified text to Text Transformer. Raffel et. al. 2020

Model Architecture



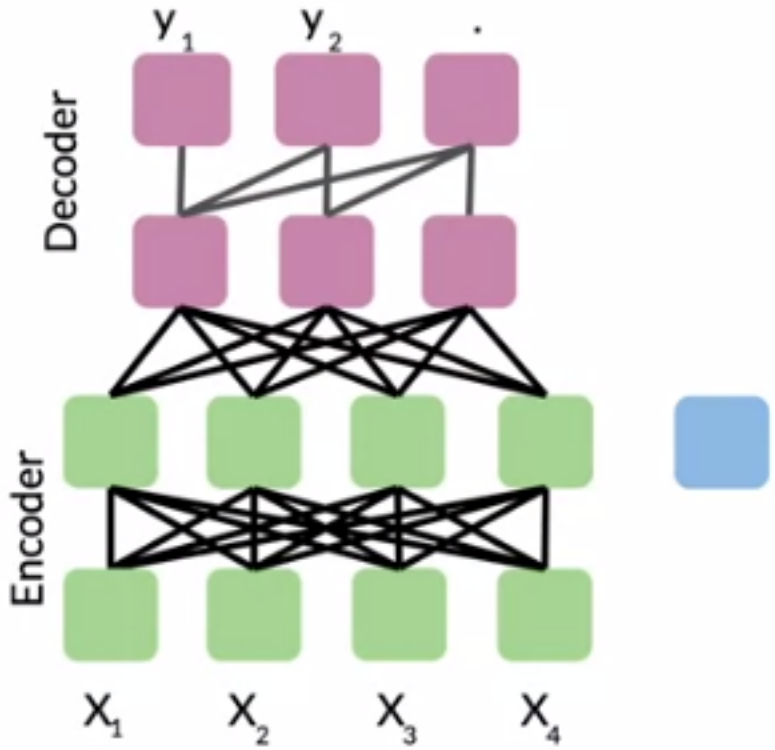
©Exploring the Limits of Transfer learning with a unified text to Text Transformer. Raffel et. al. 2020

Model Architecture



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

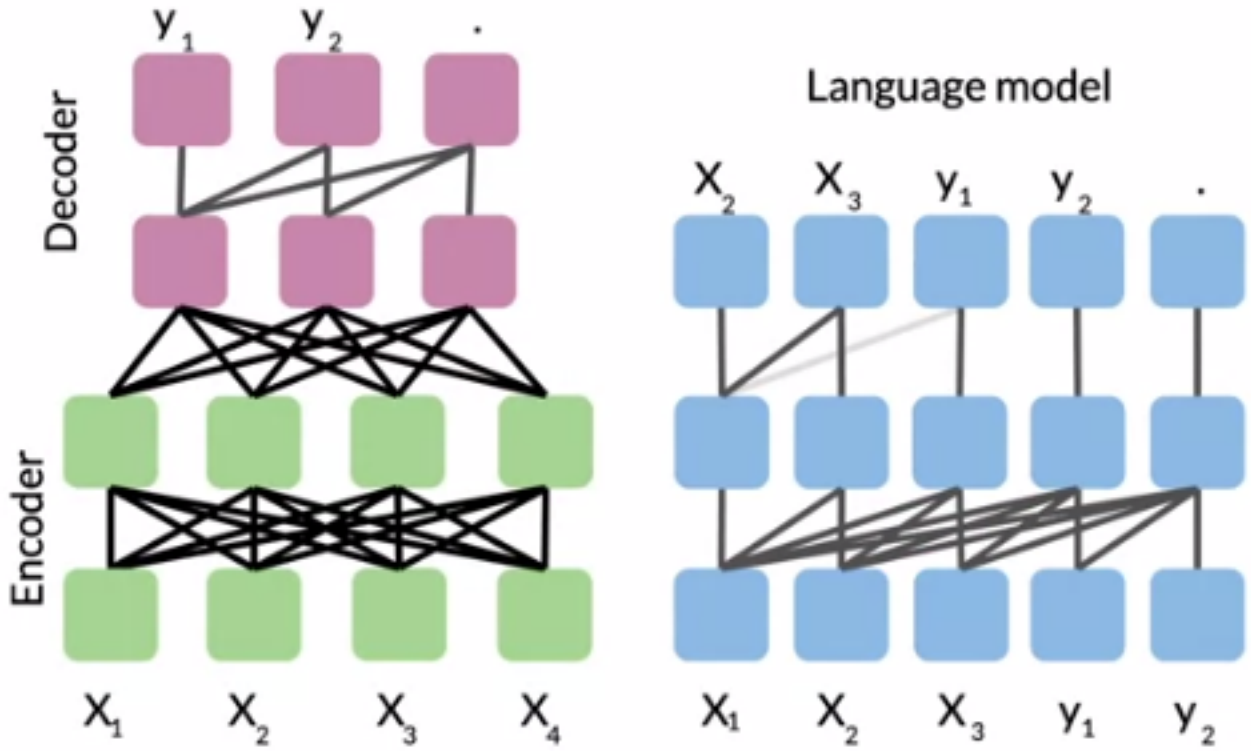
Model Architecture



Language model

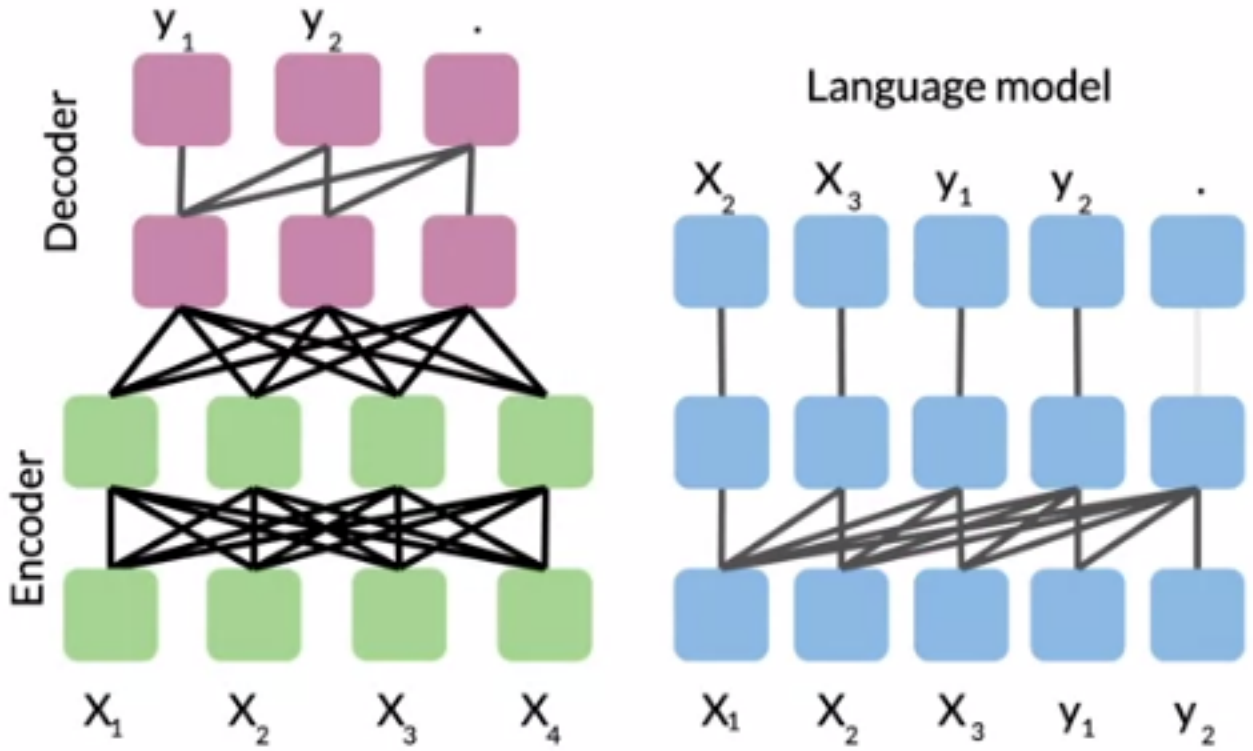
©Exploring the Limits of Transfer learning with a unified text to Text Transformer. Raffel et. al. 2020

Model Architecture



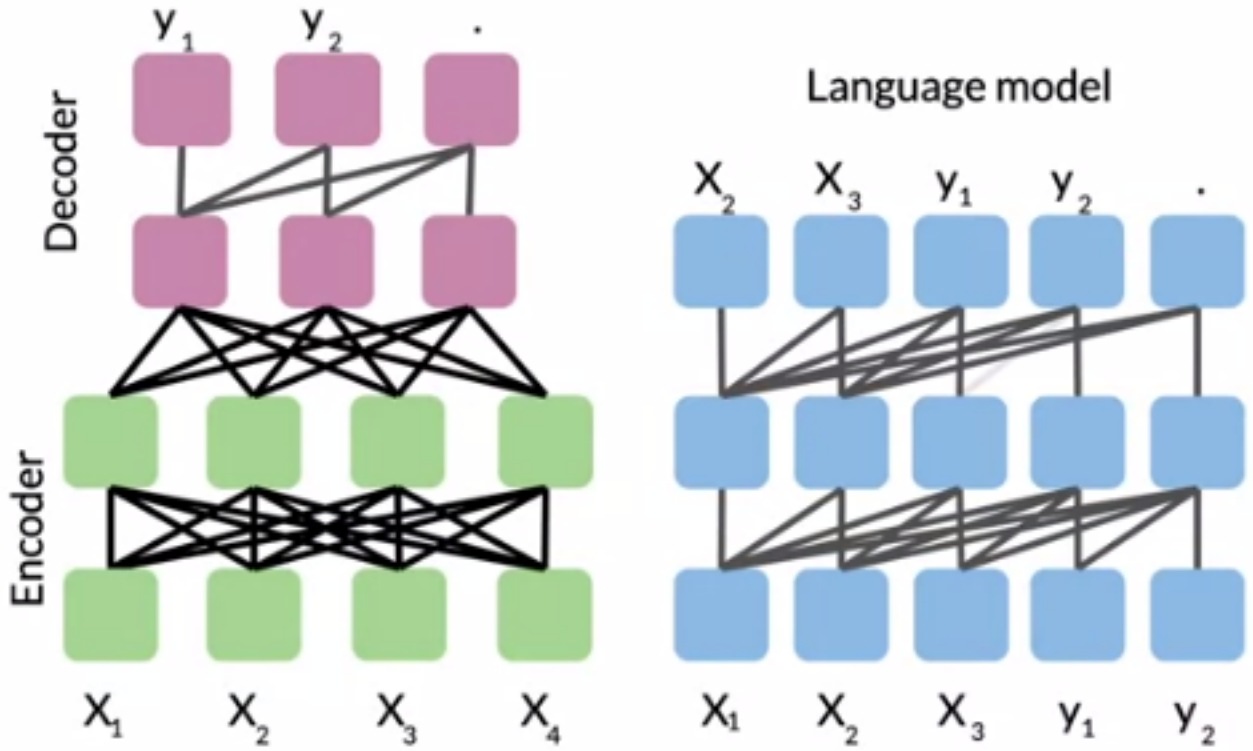
©Exploring the Limits of Transfer learning with a unified text to Text Transformer. Raffel et. al. 2020

Model Architecture



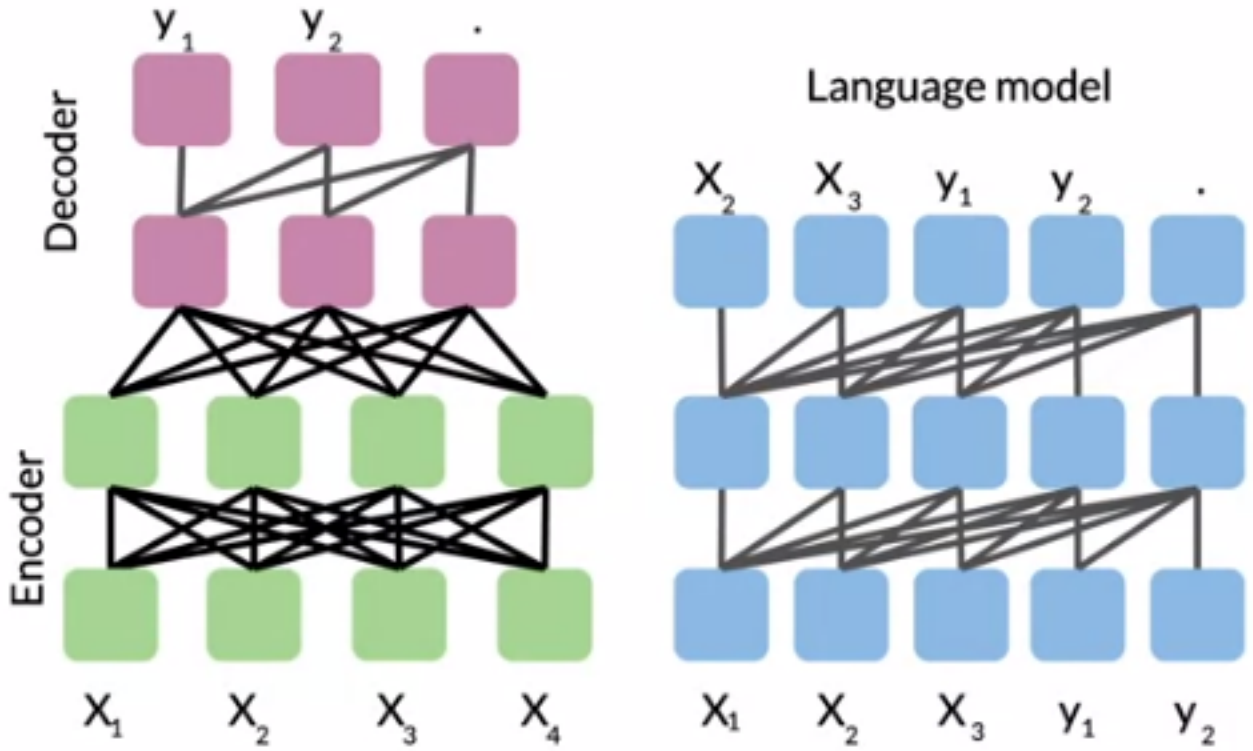
©Exploring the Limits of Transfer learning with a unified text to Text Transformer. Raffel et. al. 2020

Model Architecture



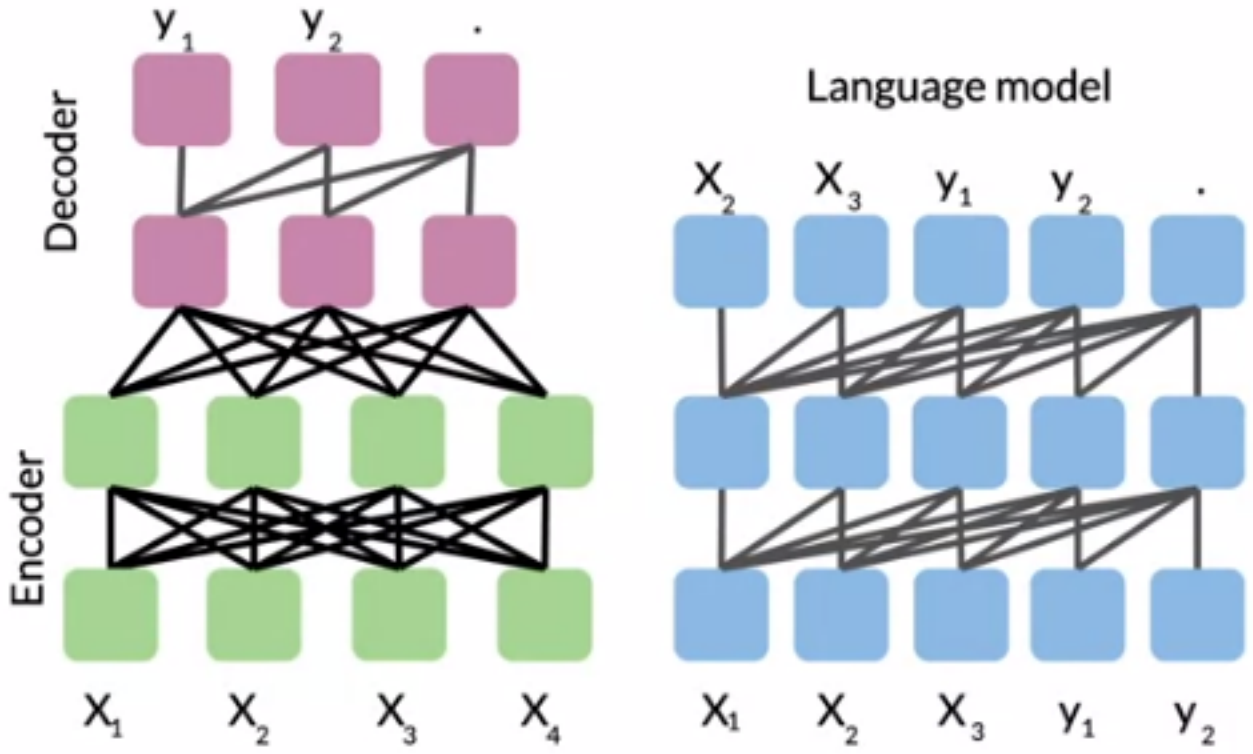
©Exploring the Limits of Transfer learning with a unified text to Text Transformer. Raffel et. al. 2020

Model Architecture



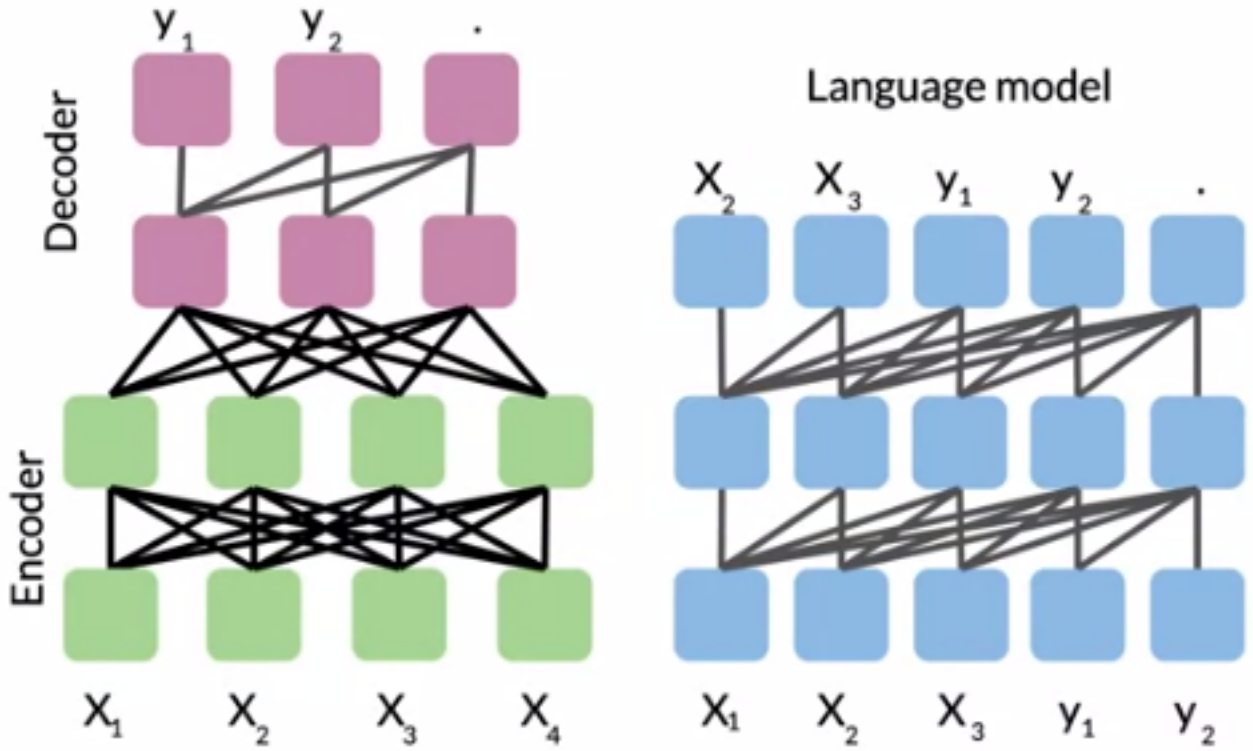
©Exploring the Limits of Transfer learning with a unified text to Text Transformer. Raffel et. al. 2020

Model Architecture



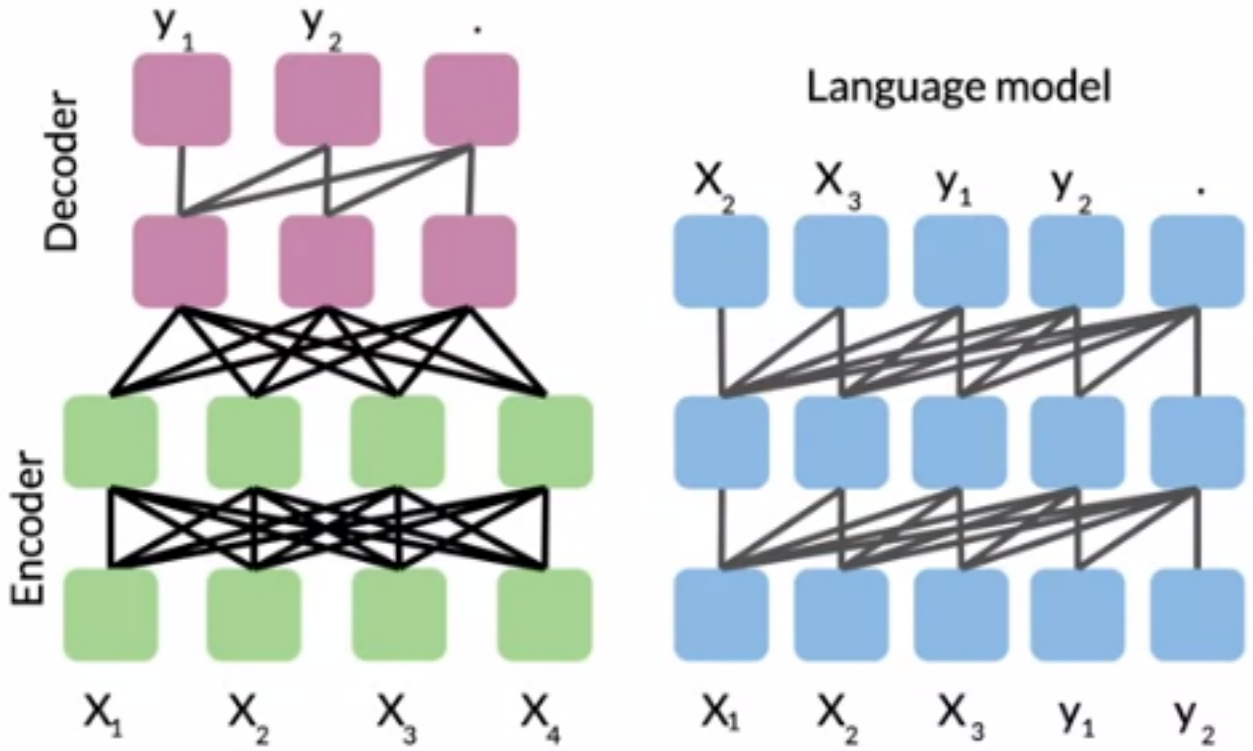
©Exploring the Limits of Transfer learning with a unified text to Text Transformer. Raffel et. al. 2020

Model Architecture



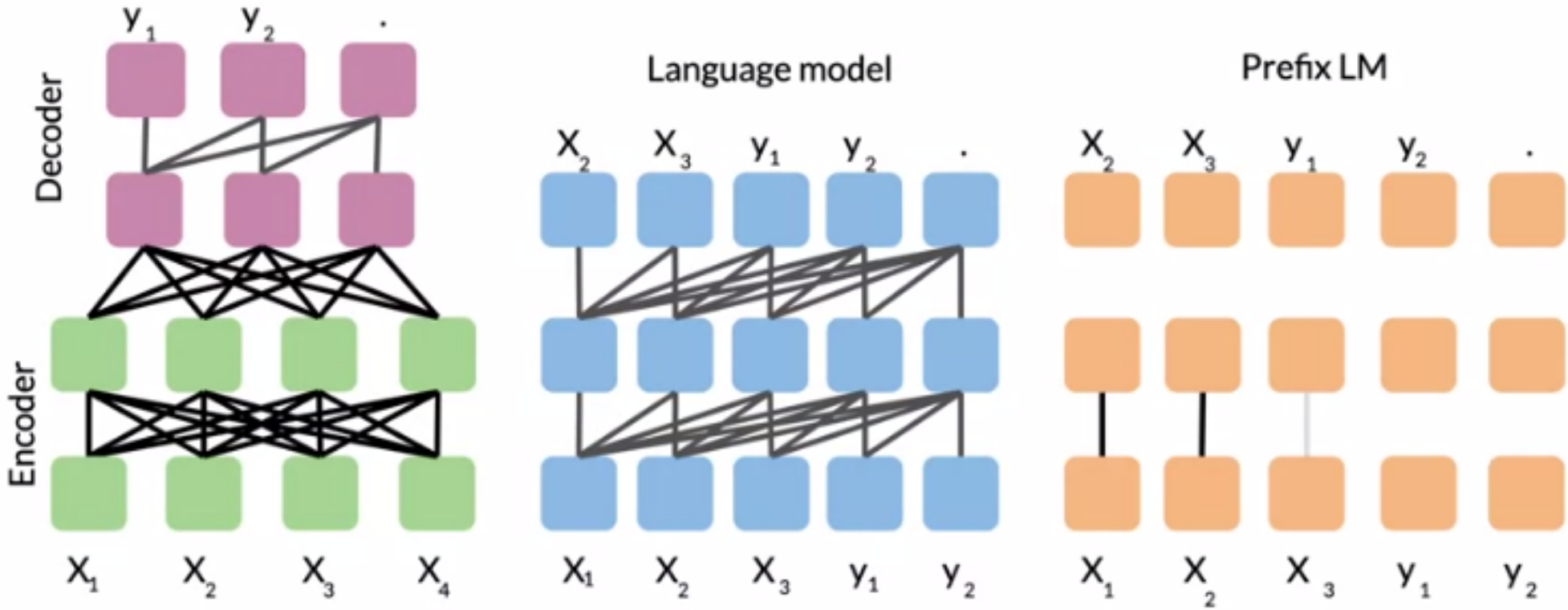
©Exploring the Limits of Transfer learning with a unified text to Text Transformer. Raffel et. al. 2020

Model Architecture



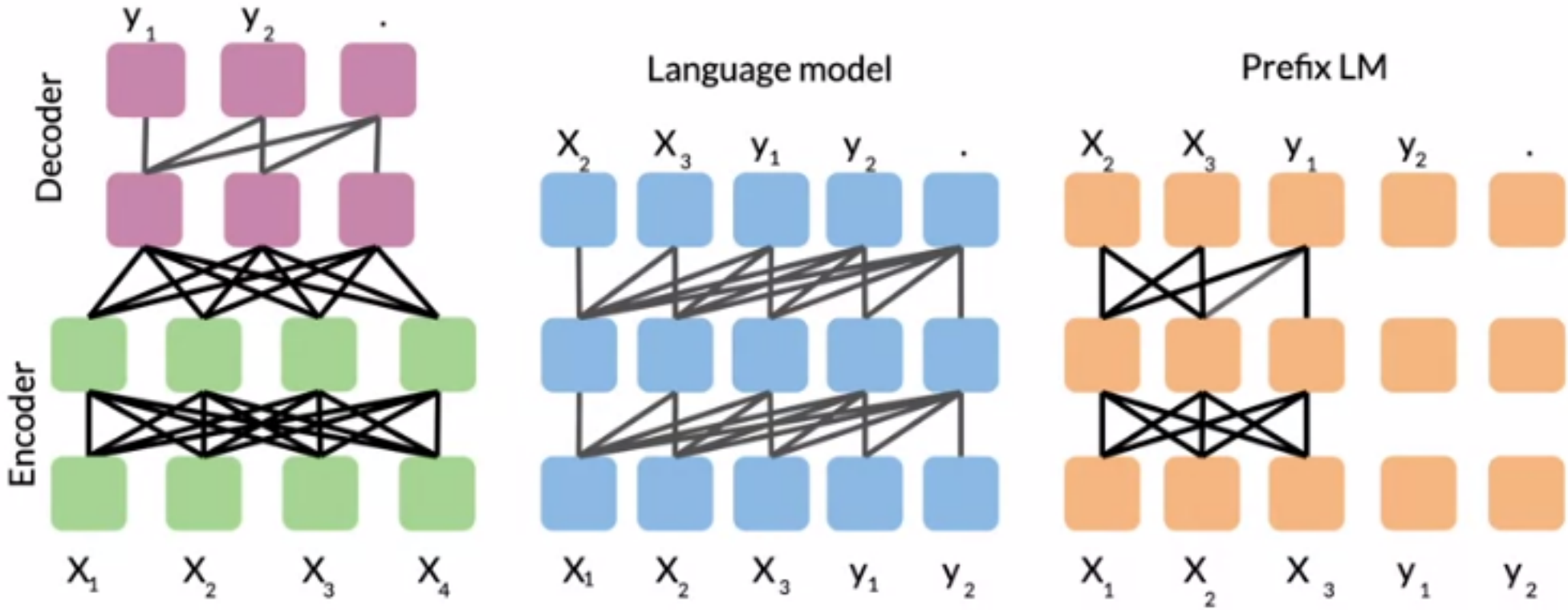
©Exploring the Limits of Transfer learning with a unified text to Text Transformer. Raffel et. al. 2020

Model Architecture



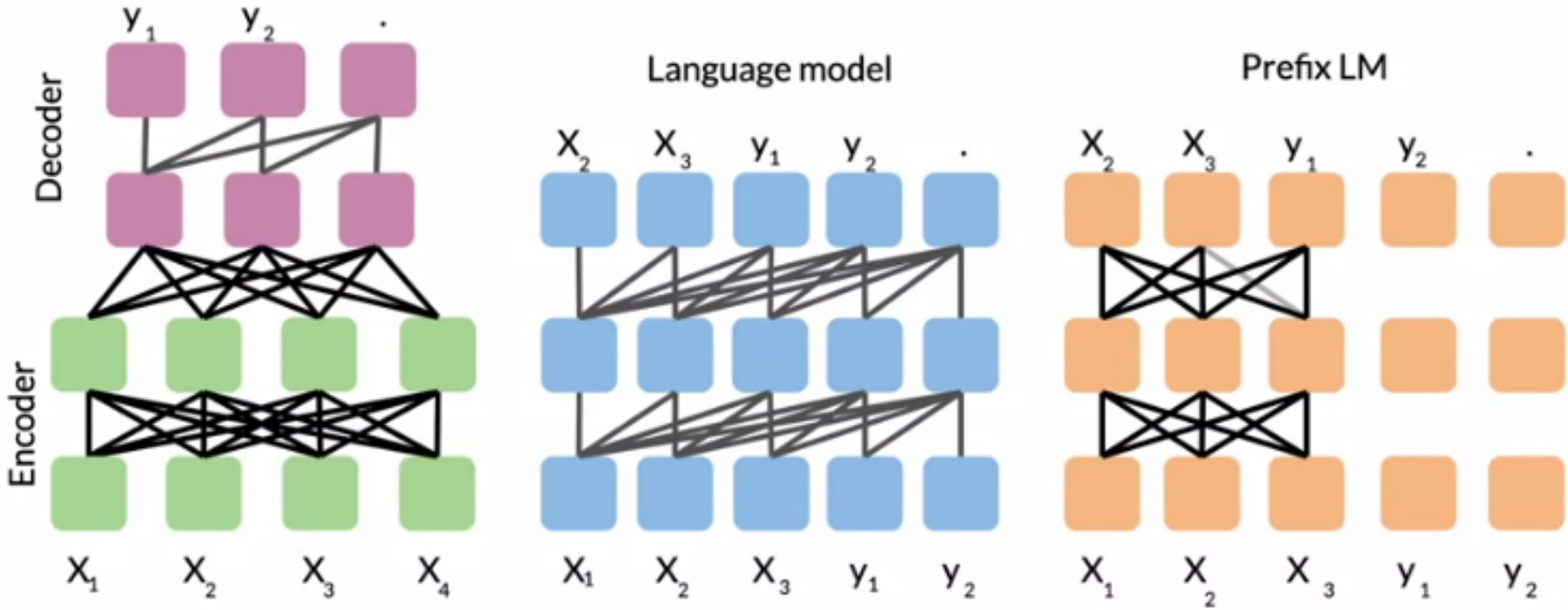
©Exploring the Limits of Transfer learning with a unified text to Text Transformer. Raffel et. al. 2020

Model Architecture



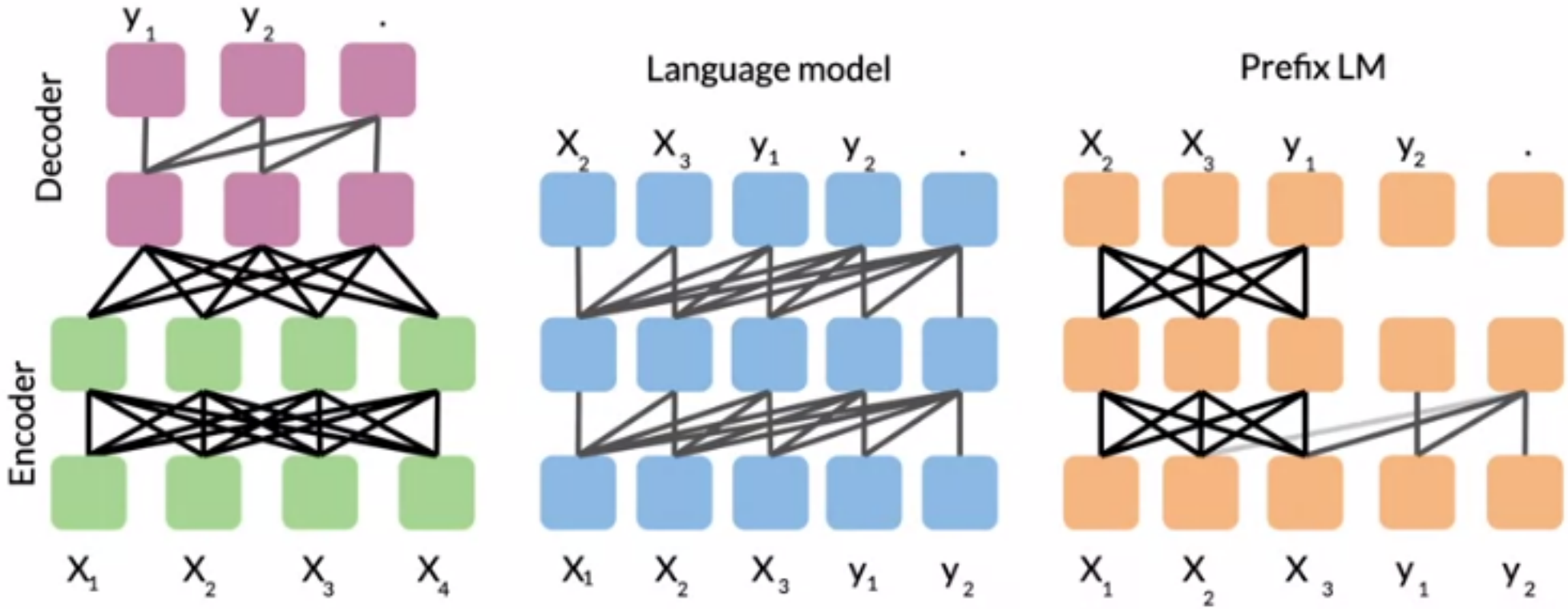
©Exploring the Limits of Transfer learning with a unified text to Text Transformer. Raffel et. al. 2020

Model Architecture



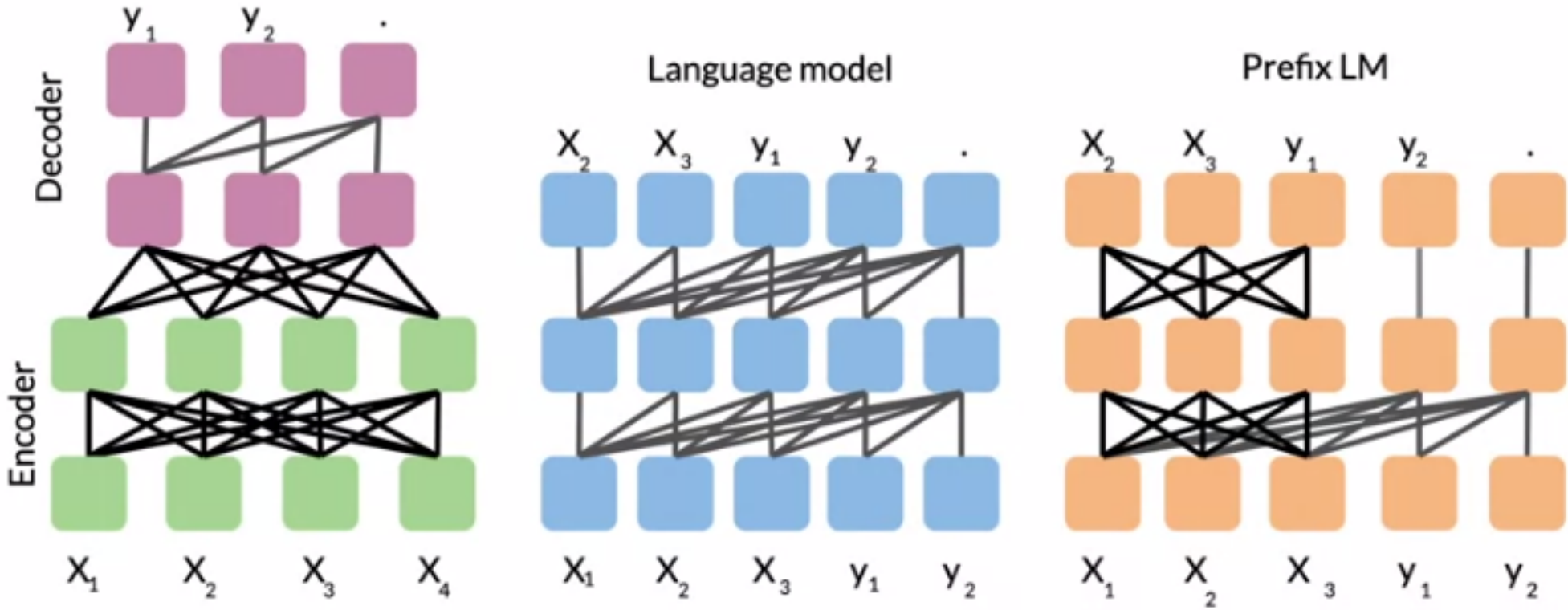
©Exploring the Limits of Transfer learning with a unified text to Text Transformer. Raffel et. al. 2020

Model Architecture



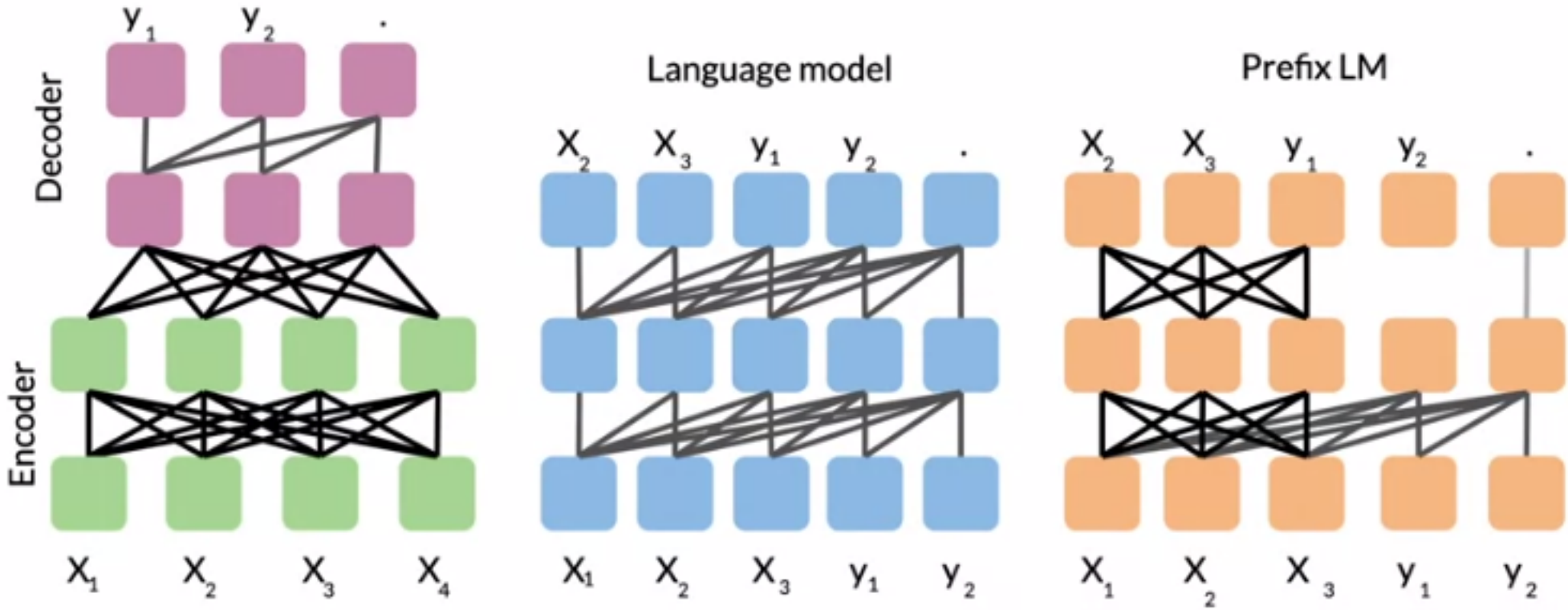
©Exploring the Limits of Transfer learning with a unified text to Text Transformer. Raffel et. al. 2020

Model Architecture



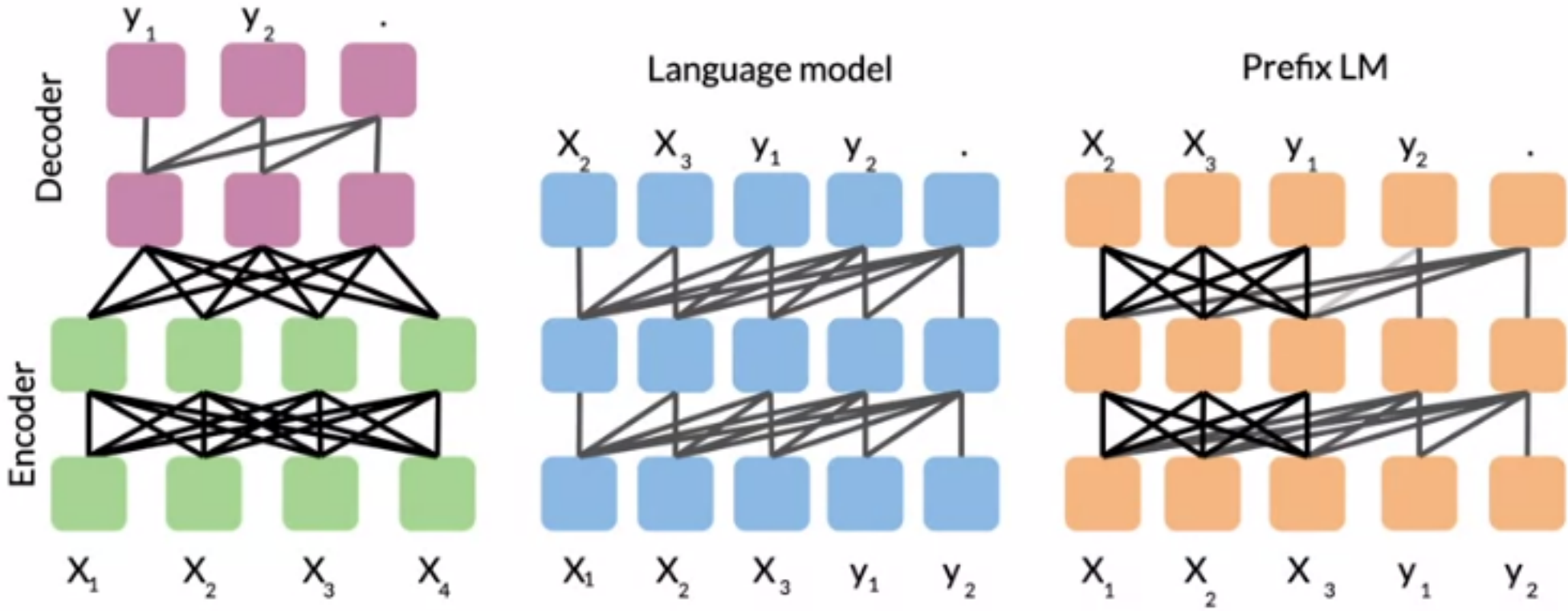
©Exploring the Limits of Transfer learning with a unified text to Text Transformer. Raffel et. al. 2020

Model Architecture



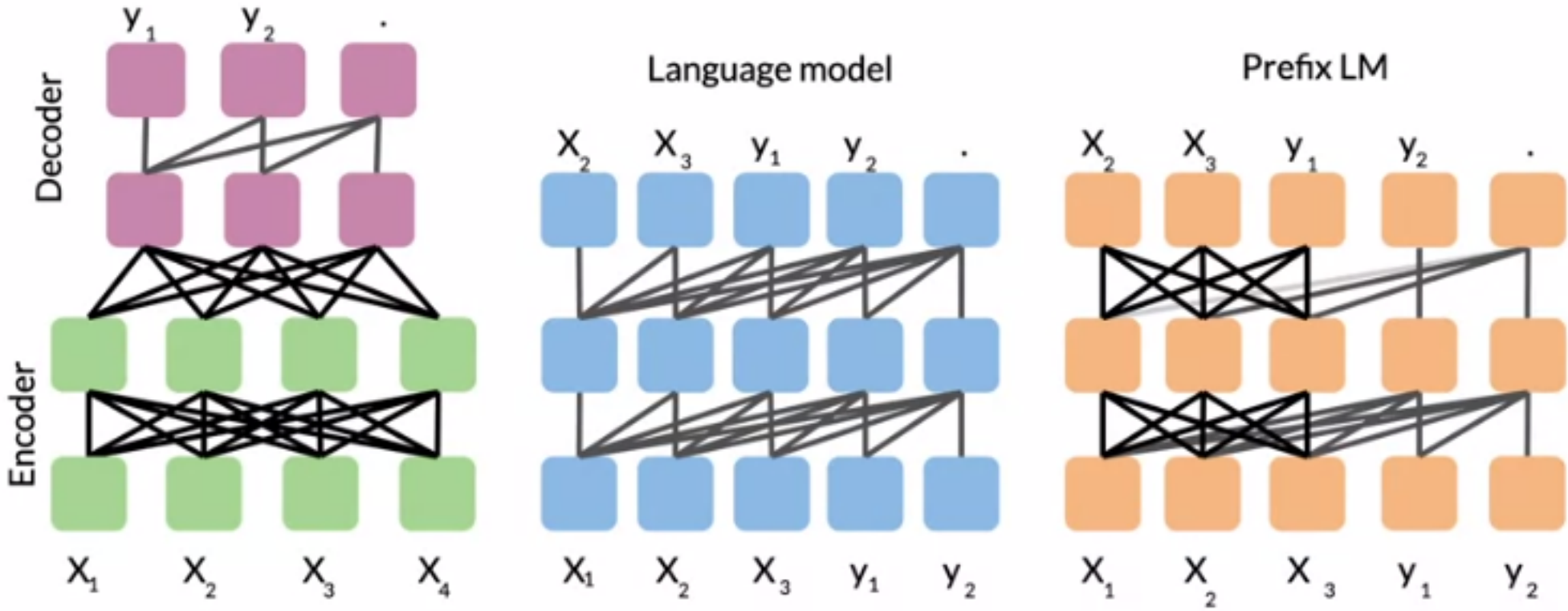
©Exploring the Limits of Transfer learning with a unified text to Text Transformer. Raffel et. al. 2020

Model Architecture



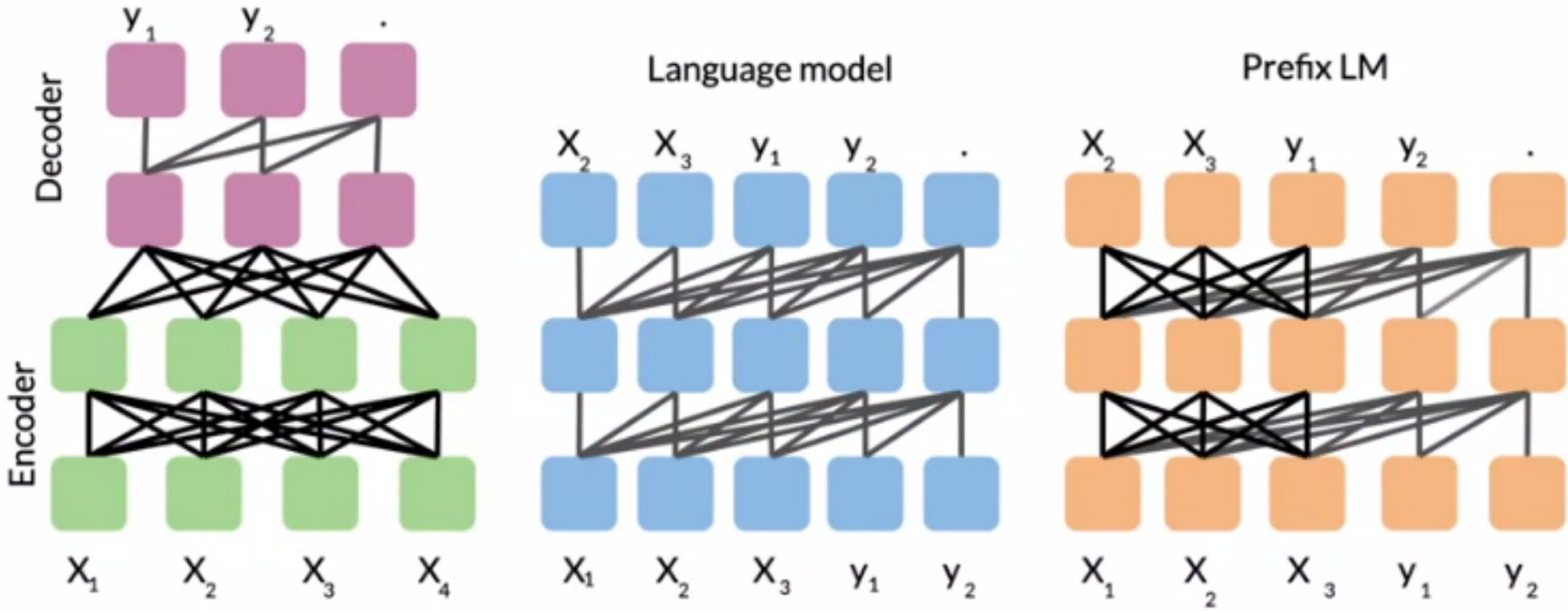
©Exploring the Limits of Transfer learning with a unified text to Text Transformer. Raffel et. al. 2020

Model Architecture



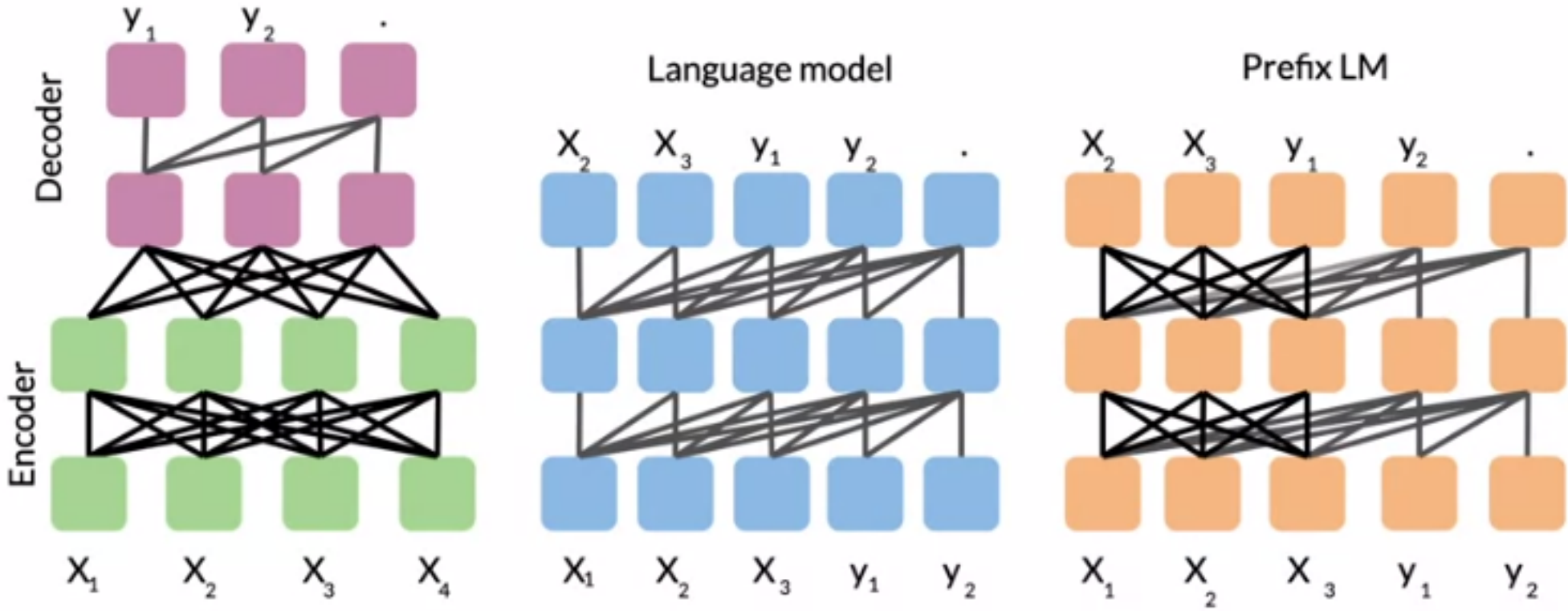
©Exploring the Limits of Transfer learning with a unified text to Text Transformer. Raffel et. al. 2020

Model Architecture



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

Model Architecture



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

Model Architecture

- Encoder/decoder

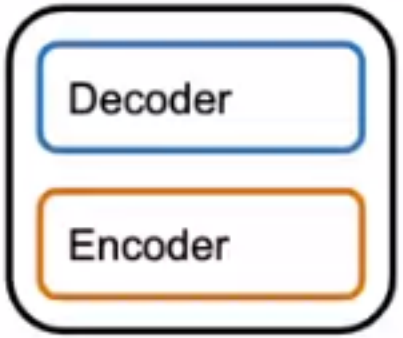
Model Architecture

- Encoder/decoder

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

Model Architecture

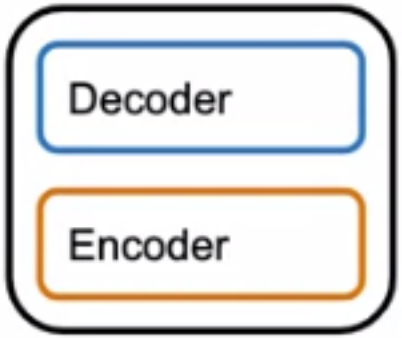
- Encoder/decoder



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

Model Architecture

- Encoder/decoder

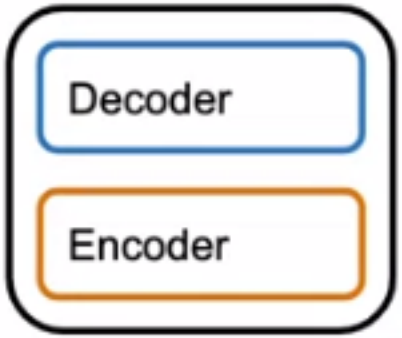


- 12 transformer blocks each

©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

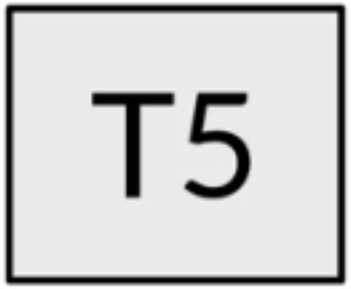
- 220 million parameters

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

Summary

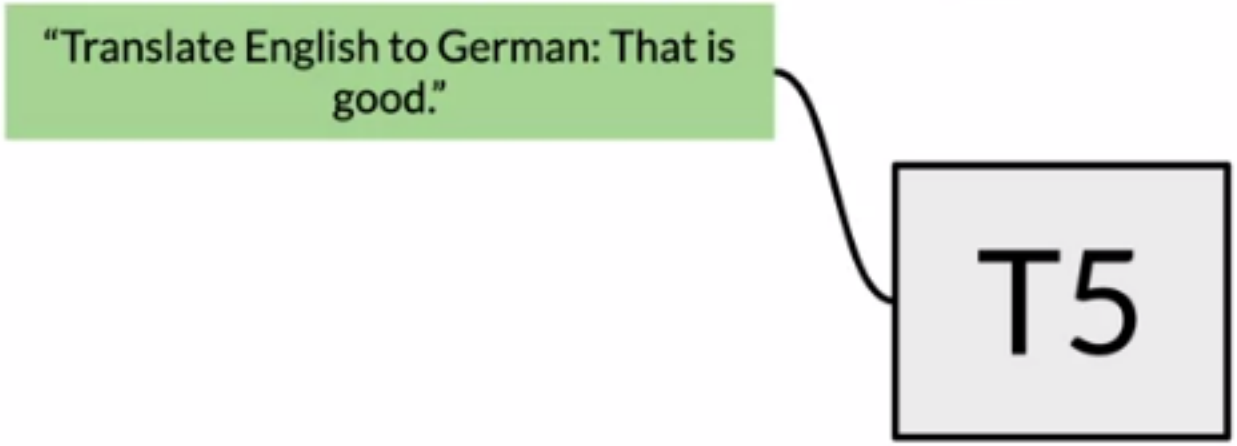
- Prefix LM attention
- Model architecture
- Pre-training T5 (MLM)

Multi-task training strategy



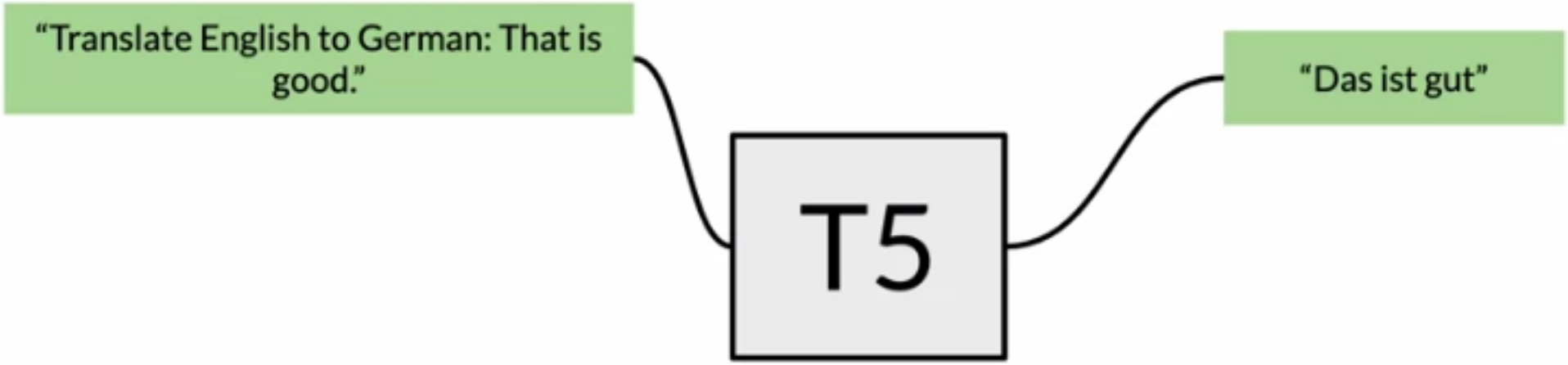
©Exploring the Limits of Transfer learning with a unified text to Text Transformer. Raffel et. al. 2020

Multi-task training strategy



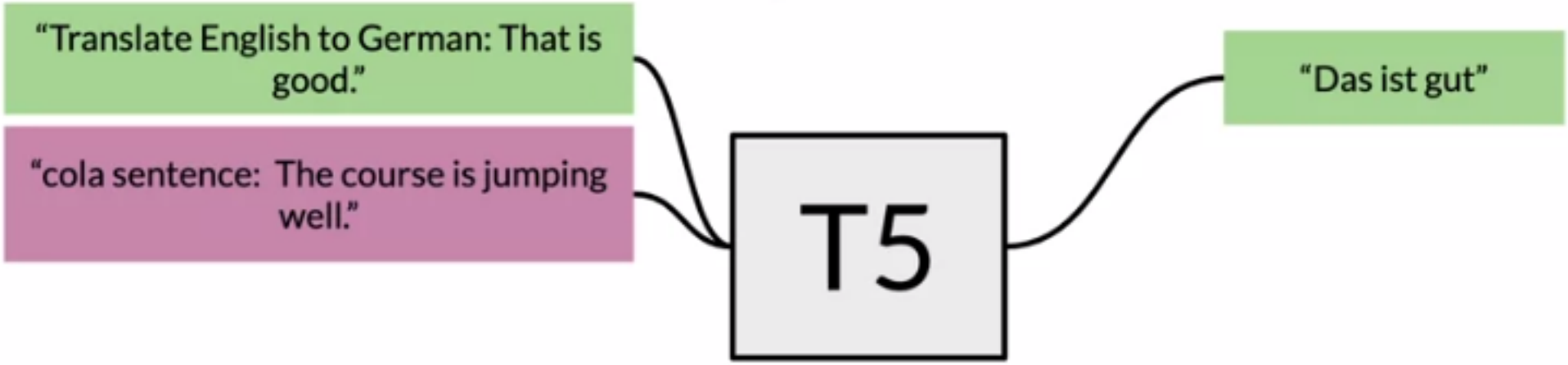
©Exploring the Limits of Transfer learning with a unified text to Text Transformer. Raffel et. al. 2020

Multi-task training strategy



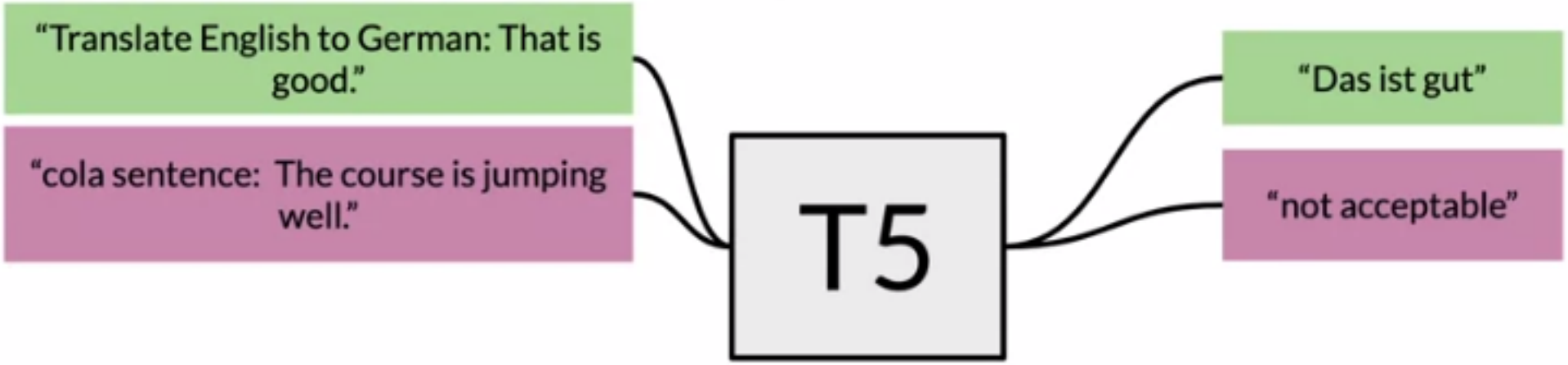
©Exploring the Limits of Transfer learning with a unified text to Text Transformer. Raffel et. al. 2020

Multi-task training strategy



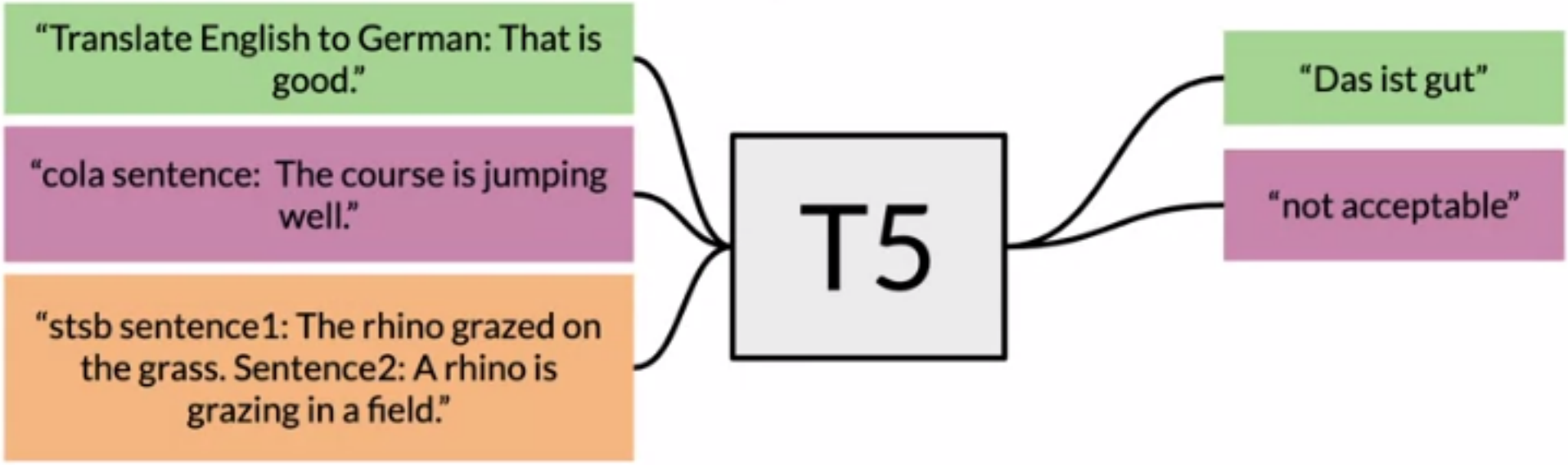
©Exploring the Limits of Transfer learning with a unified text to Text Transformer. Raffel et. al. 2020

Multi-task training strategy



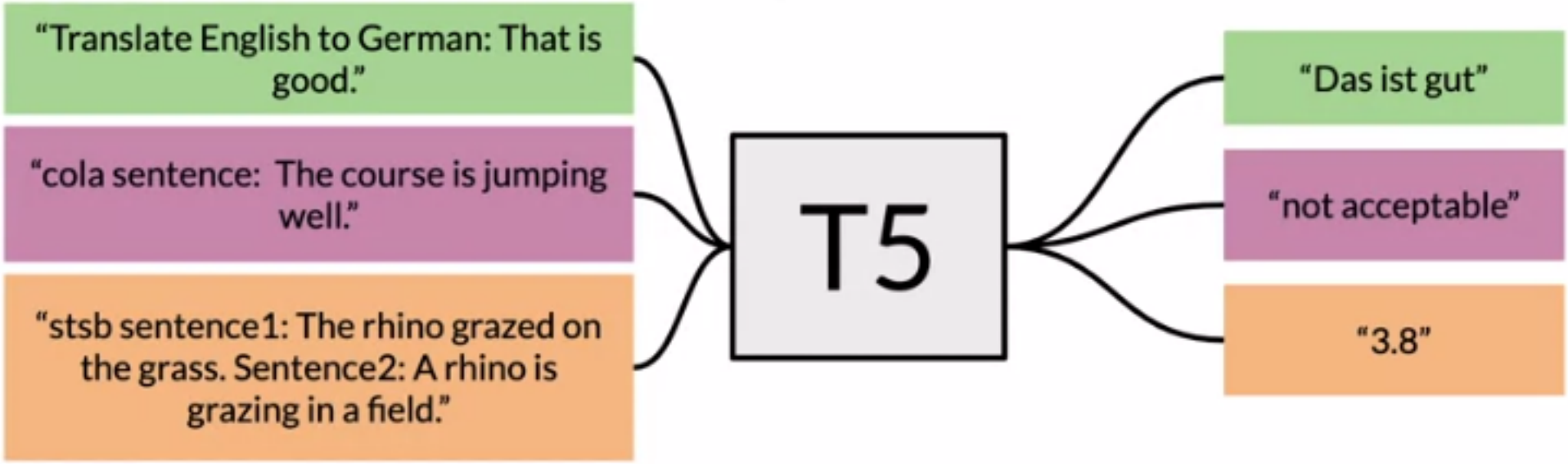
©Exploring the Limits of Transfer learning with a unified text to Text Transformer. Raffel et. al. 2020

Multi-task training strategy



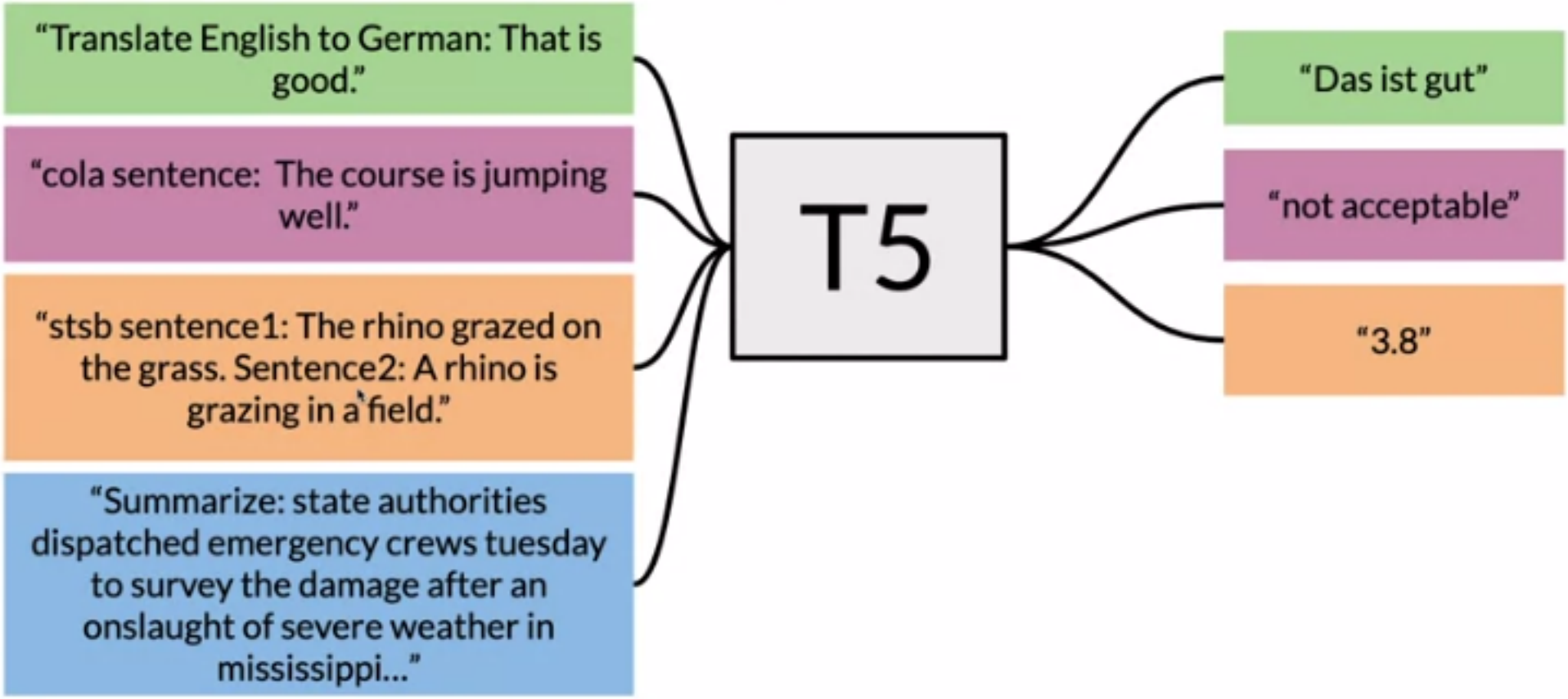
©Exploring the Limits of Transfer learning with a unified text to Text Transformer. Raffel et. al. 2020

Multi-task training strategy



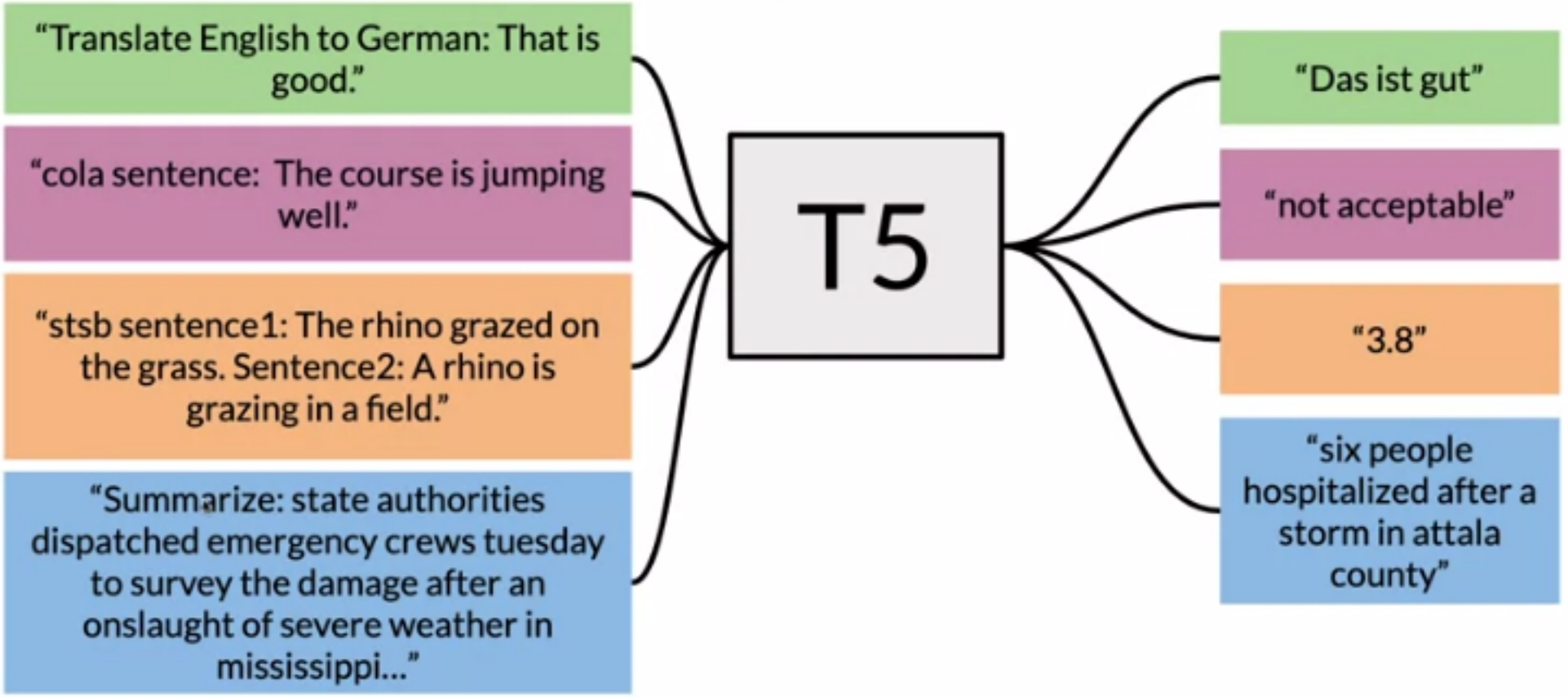
©Exploring the Limits of Transfer learning with a unified text to Text Transformer. Raffel et. al. 2020

Multi-task training strategy



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

Multi-task training strategy



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

Input and Output Format

Machine translation:

- translate English to German: That is good.

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

Input and Output Format

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

Input and Output Format

Machine translation:

- translate English to German: That is good.

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

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

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

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

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

Multi-task Training Strategy

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

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

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

Data Training Strategies

Examples-proportional mixing

Equal mixing

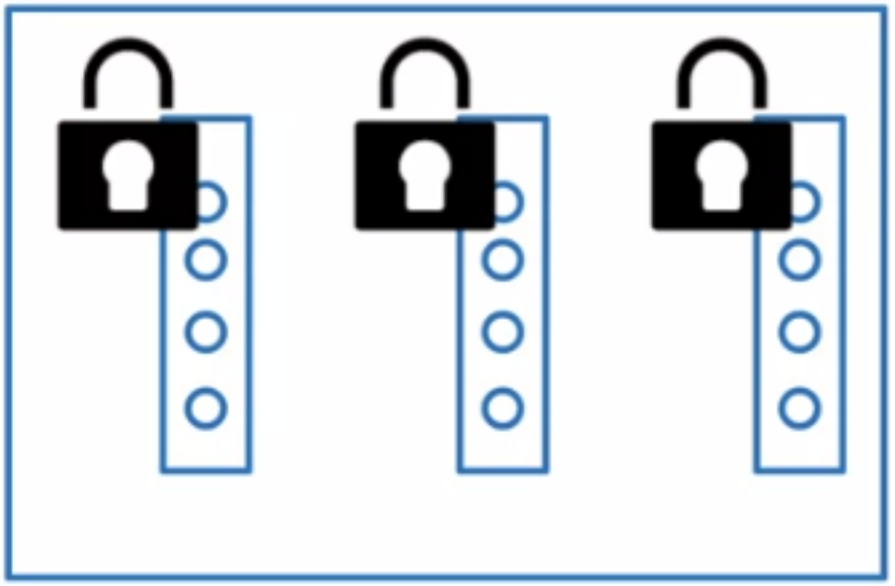
Data Training Strategies

Examples-proportional mixing

Equal mixing

Temperature-scaled mixing

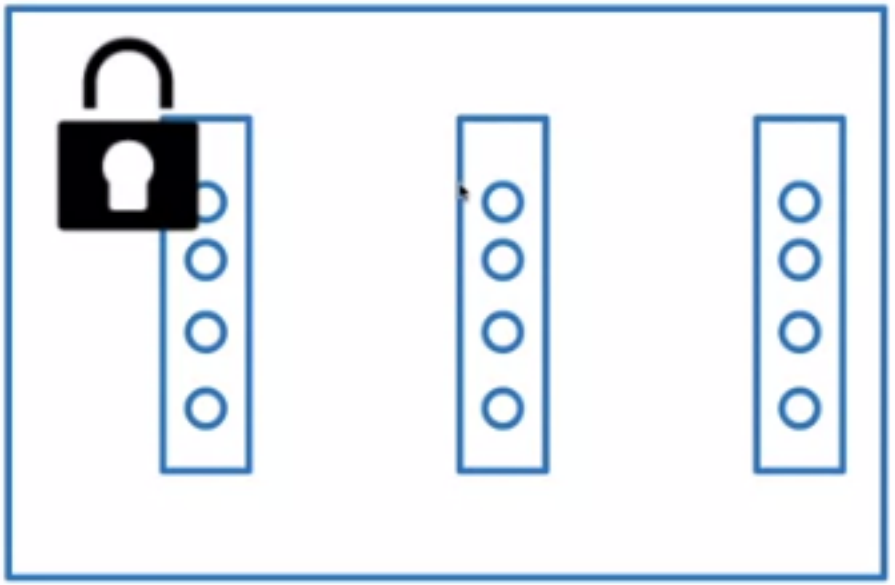
Gradual unfreezing vs. Adapter layers



Gradual unfreezing

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

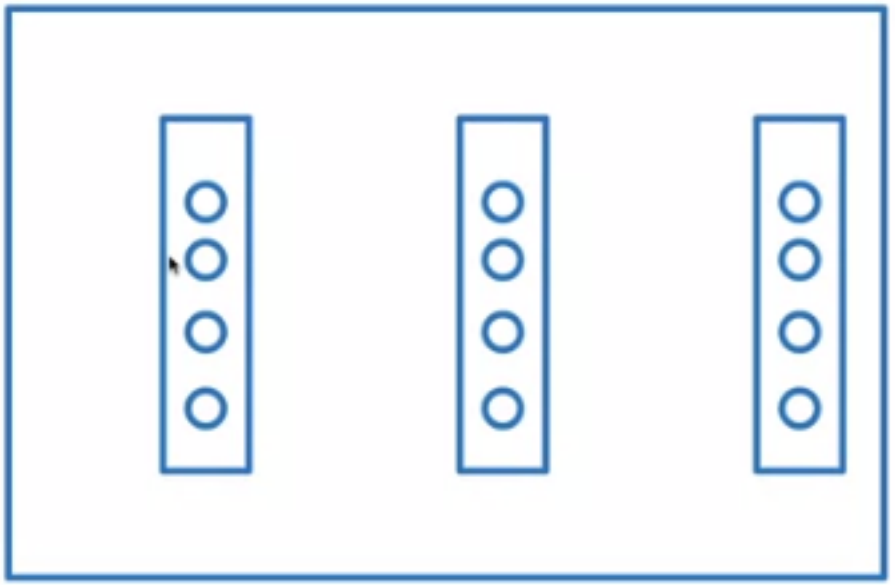
Gradual unfreezing vs. Adapter layers



Gradual unfreezing

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

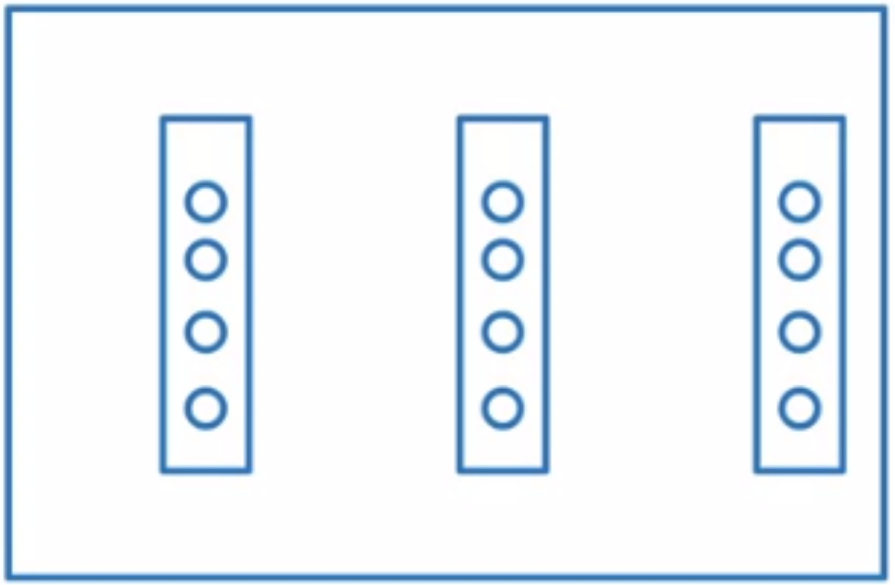
Gradual unfreezing vs. Adapter layers



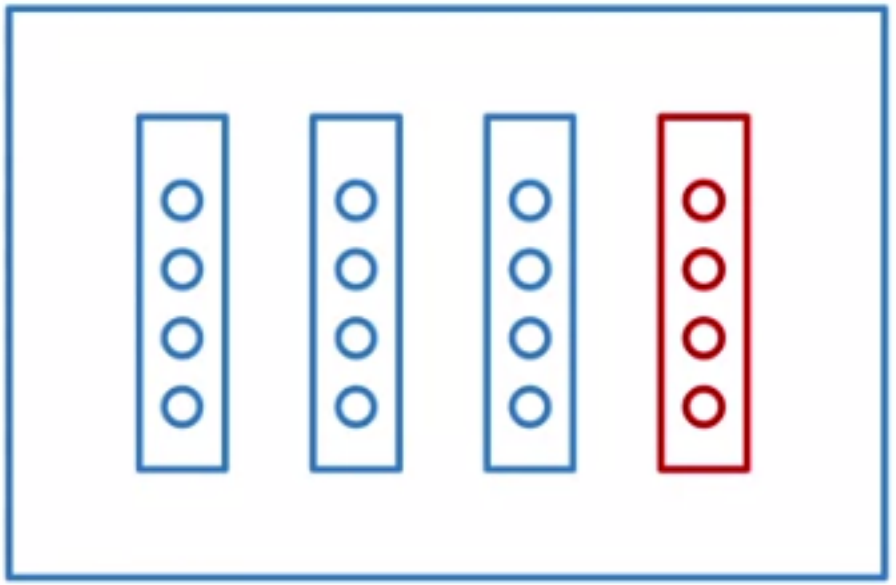
Gradual unfreezing

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

Gradual unfreezing vs. Adapter layers



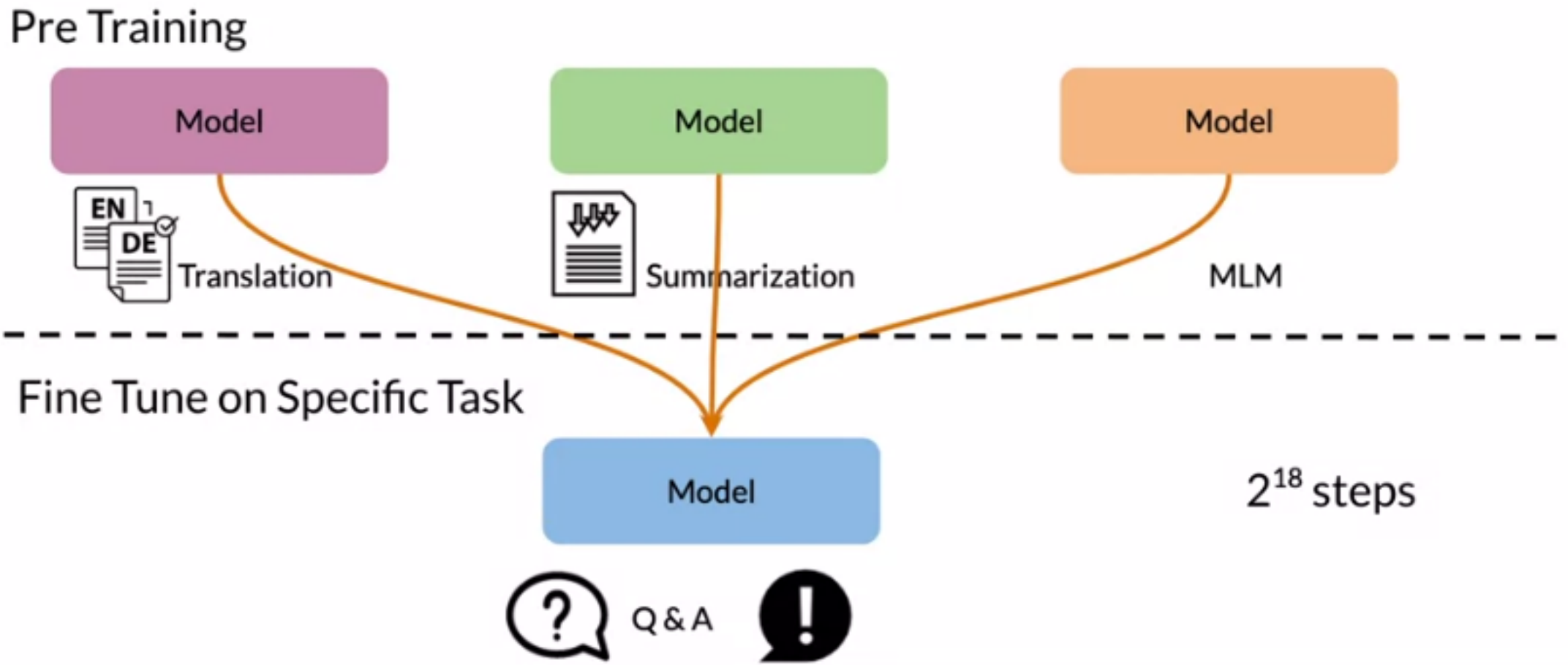
Gradual unfreezing



Adapter layers

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

Fine-tuning


Pre Training
Model
Model
Model
EN
DE
Translation
Summarization
MLM
Fine Tune on Specific Task
Model
2^{18} steps
Q&A

General Language Understanding Evaluation

General Language Understanding Evaluation

- A collection used to train, evaluate, analyze natural language understanding systems

General Language Understanding Evaluation

- A collection used to train, evaluate, analyze natural language understanding systems

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

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

Tasks Evaluated on

- Sentence grammatical or not?

Tasks Evaluated on

- Sentence grammatical or not?
- Sentiment

Tasks Evaluated on

- Sentence grammatical or not?
- Sentiment
- Paraphrase

Tasks Evaluated on

- Sentence grammatical or not?
- Sentiment
- Paraphrase

Tasks Evaluated on

- Sentence grammatical or not?
- Sentiment
- Paraphrase
- Similarity

Tasks Evaluated on

- Sentence grammatical or not?
- Sentiment
- Paraphrase
- Similarity
- Questions duplicates

Tasks Evaluated on

- Sentence grammatical or not?
- Sentiment
- Paraphrase
- Similarity
- Questions duplicates
- Answerable

Tasks Evaluated on

- Sentence grammatical or not?
- Sentiment
- Paraphrase
- Similarity
- Questions duplicates
- Answerable
- Contradiction

Tasks Evaluated on

- Sentence grammatical or not?
- Sentiment
- Paraphrase
- Similarity
- Questions duplicates
- Answerable
- Contradiction
- Entailment

Tasks Evaluated on

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

General Language Understanding Evaluation

- Drive research

General Language Understanding Evaluation

- Drive research
- Model agnostic

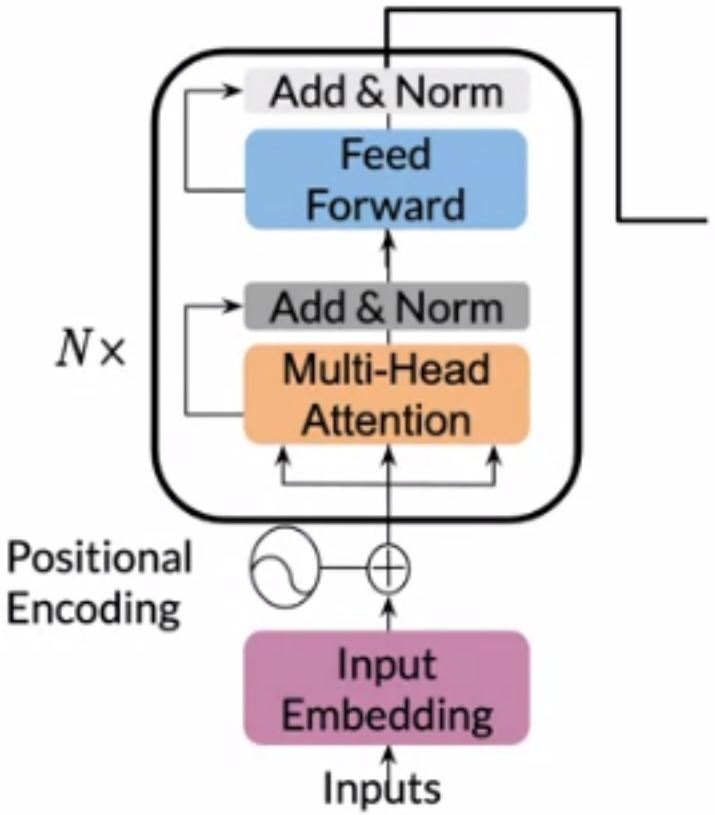
General Language Understanding Evaluation

- Drive research

- Model agnostic

- Makes use of transfer learning

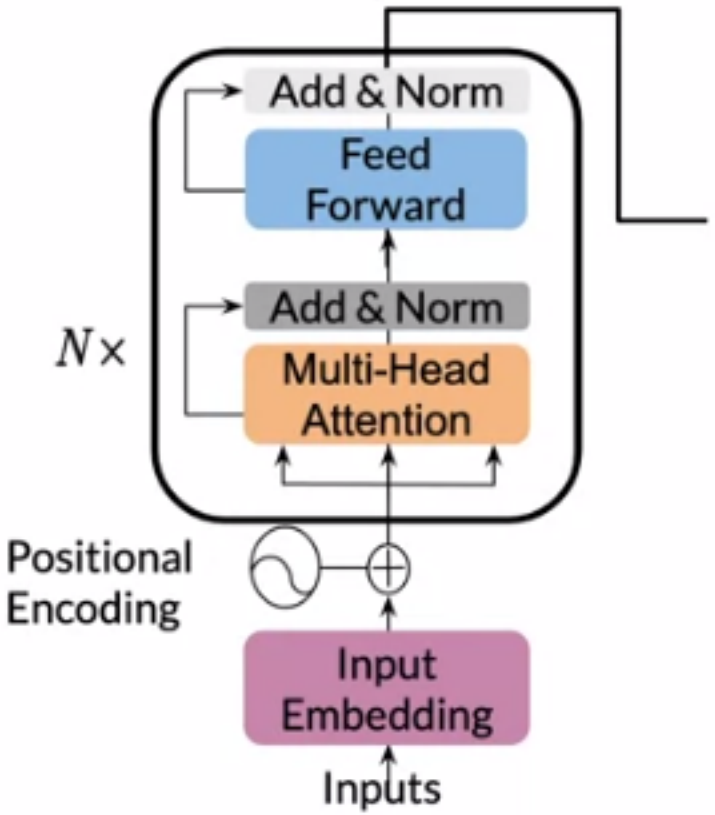
Transformer encoder



Feedforward:

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

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

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