Thursday • October 28, 2021

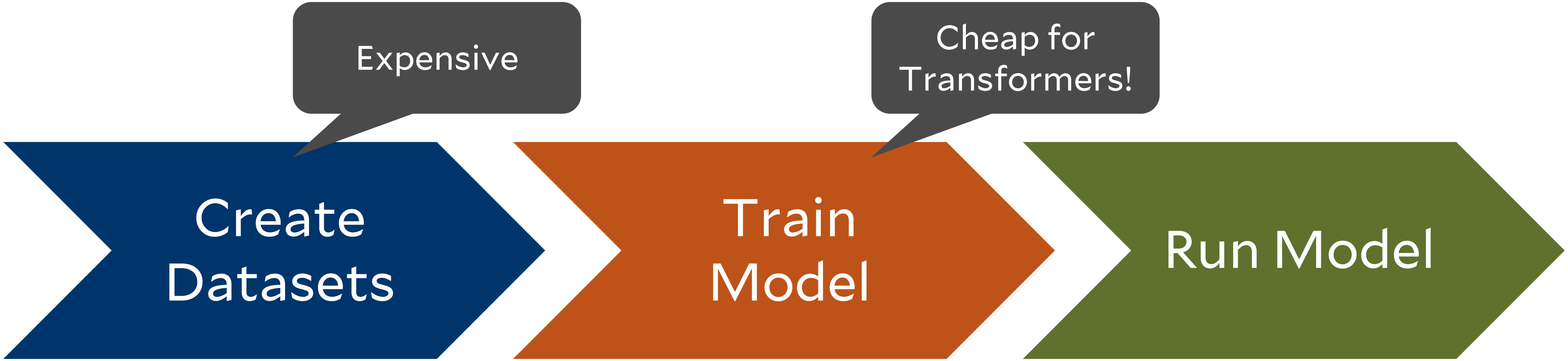
Large-Scale Transfer Learning

Yale

LING 380/780

Neural Network Models of Linguistic Structure

NLP Model Development Process



Expensive

Cheap for Transformers!

Create Datasets

Train Model

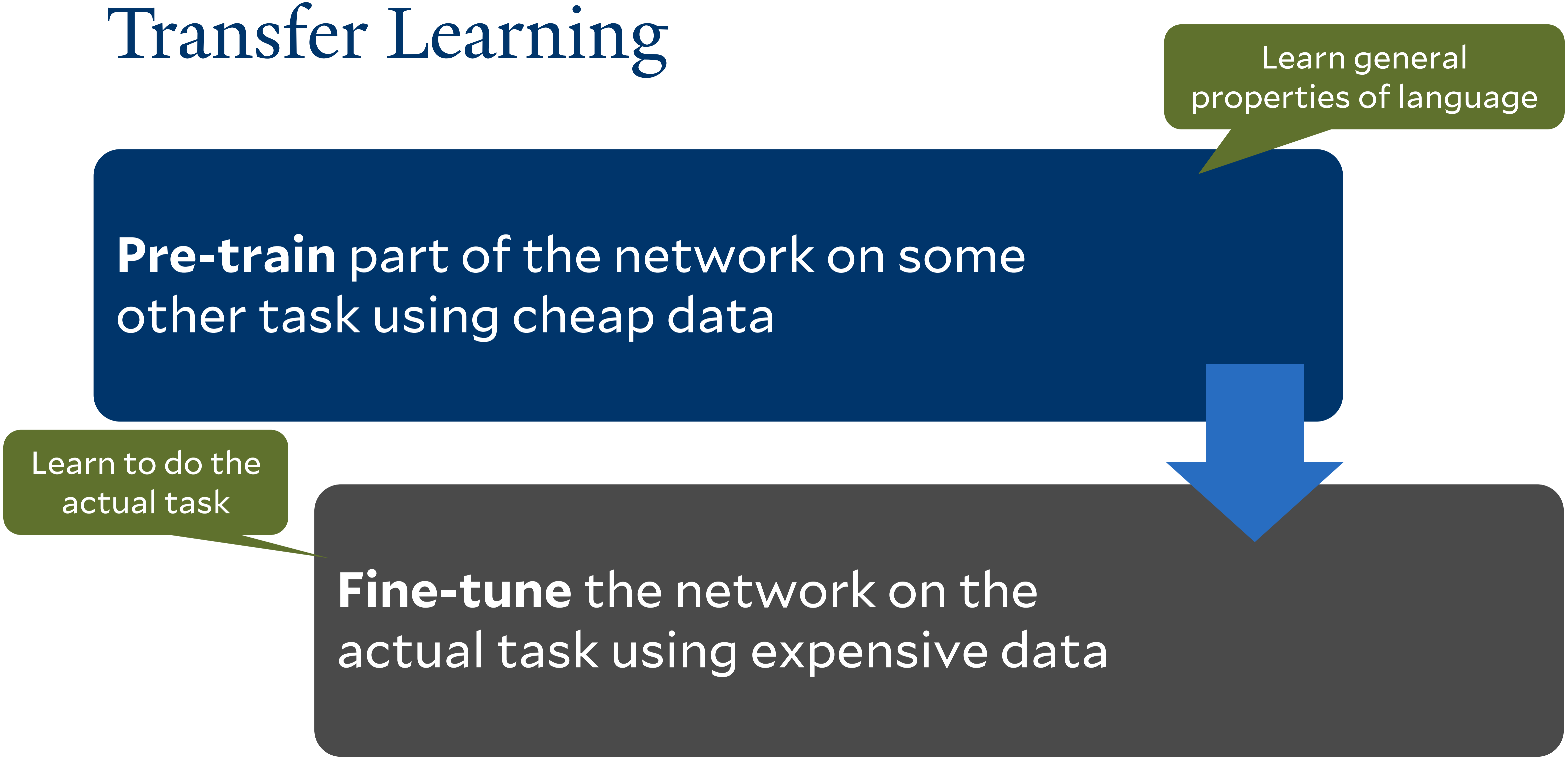
Run Model

- **Cost:** Human Labor (hrs)
- **Cost:** Computation Time (PFLOPs)
- **Cost:** User Computation Time (ms/example)

Training on Large Datasets

- Training a Transformer is very cheap thanks to GPUs and parallel computation.
- In theory, Transformers can be trained on very large datasets.
- But creating datasets is still very labor-intensive.
- We need a cheap way of getting large datasets to train Transformers with.

Transfer Learning



Learn general properties of language

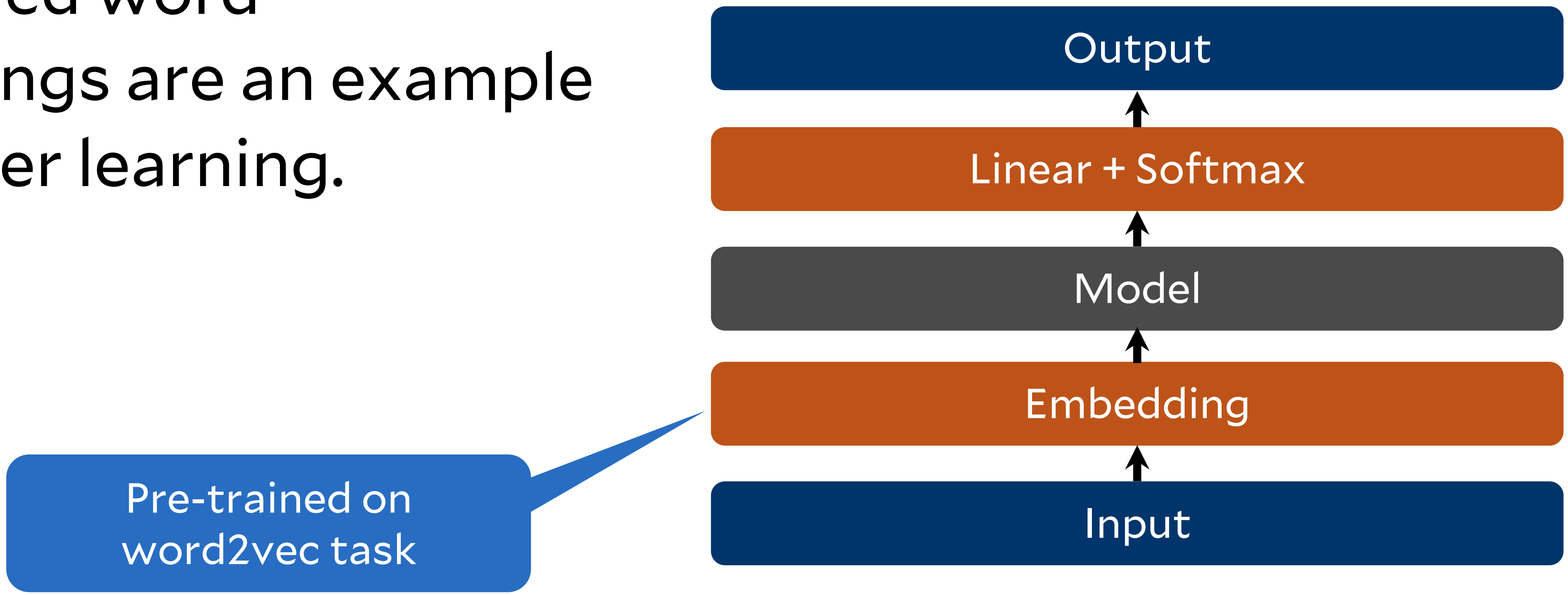
Pre-train part of the network on some other task using cheap data

Learn to do the actual task

Fine-tune the network on the actual task using expensive data

Example: Word Embeddings

- Pre-trained word embeddings are an example of transfer learning.



Output

Linear + Softmax

Model

Embedding

Input

Pre-trained on word2vec task

Two Methods for Transfer Learning

- Pre-trained Transformers
- In-context adaptation (a.k.a. “few-shot transfer”)

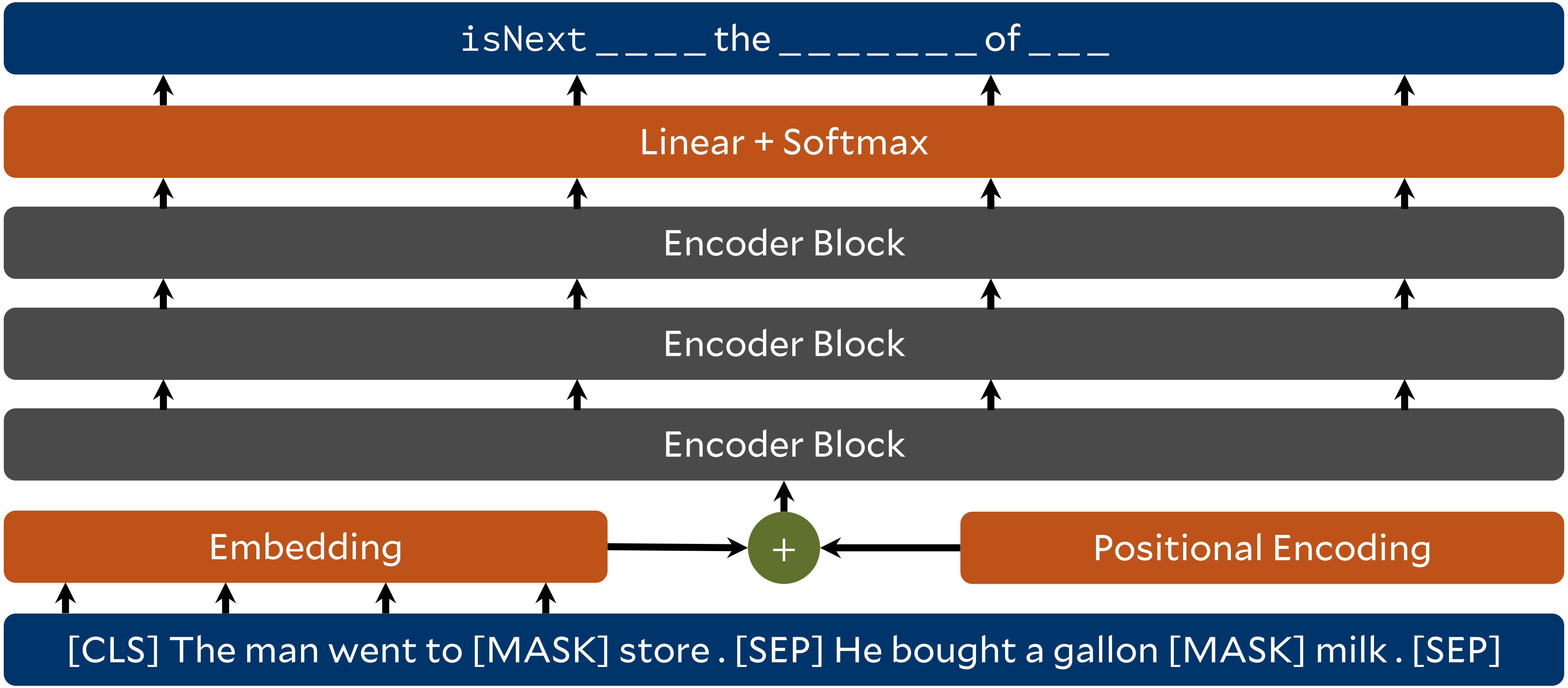
Pre-Trained Transformers

Transfer Learning Method #1

Pre-Training an Entire Transformer

- The **Bidirectional Encoder Representations from Transformer** (BERT) model is a pre-trained Transformer.
- Jointly trained on two tasks.
  - **Masked language modeling** (MLM)**:** Take a sentence with blanks (represented by the [MASK] token), and fill in the blanks.
  - **Next sentence prediction:** Take two sentences and predict whether the second one comes immediately after the first in the corpus.

BERT Model



isNext _ _ _ _ the _ _ _ _ _ _ _ _ of _ _ _

Linear + Softmax

Encoder Block

Encoder Block

Encoder Block

Embedding

+

Positional Encoding

[CLS] The man went to [MASK] store . [SEP] He bought a gallon [MASK] milk . [SEP]

BERT Embeddings and Classification

- For input sequence $w_1 w_2 \dots w_n$, the outputs $\boldsymbol{h}^{(1)}, \boldsymbol{h}^{(2)}, \dots, \boldsymbol{h}^{(n)}$ from the last encoder block are known as **contextual embeddings** (or **BERT embeddings**).
- When BERT fine-tuned on a classification task, the contextual embedding for the [CLS] token (i.e., $\boldsymbol{h}^{(1)}$) is used by the linear decoder to predict the output.
- During pre-training, $\boldsymbol{h}^{(1)}$ is used for next sentence prediction.

Masked Language Modeling

- For MLM, 15% of words in a sentence are “masked out.” BERT needs to predict what these words are.
- 80% of the time, the masked out words are replaced by [MASK].
- 10% of the time, they are replaced by a random word.
- 10% of the time, the original masked-out word is kept.
- Loss is only evaluated for [CLS] and masked-out words.

BERT Model Specs

Model	Encoder Blocks	Hidden Size	Attn. Heads	Total Params
BERT Base Cased BERT Base Uncased	12	768	12	110 million
BERT Large Cased BERT Large Uncased	24	1024	16	340 million

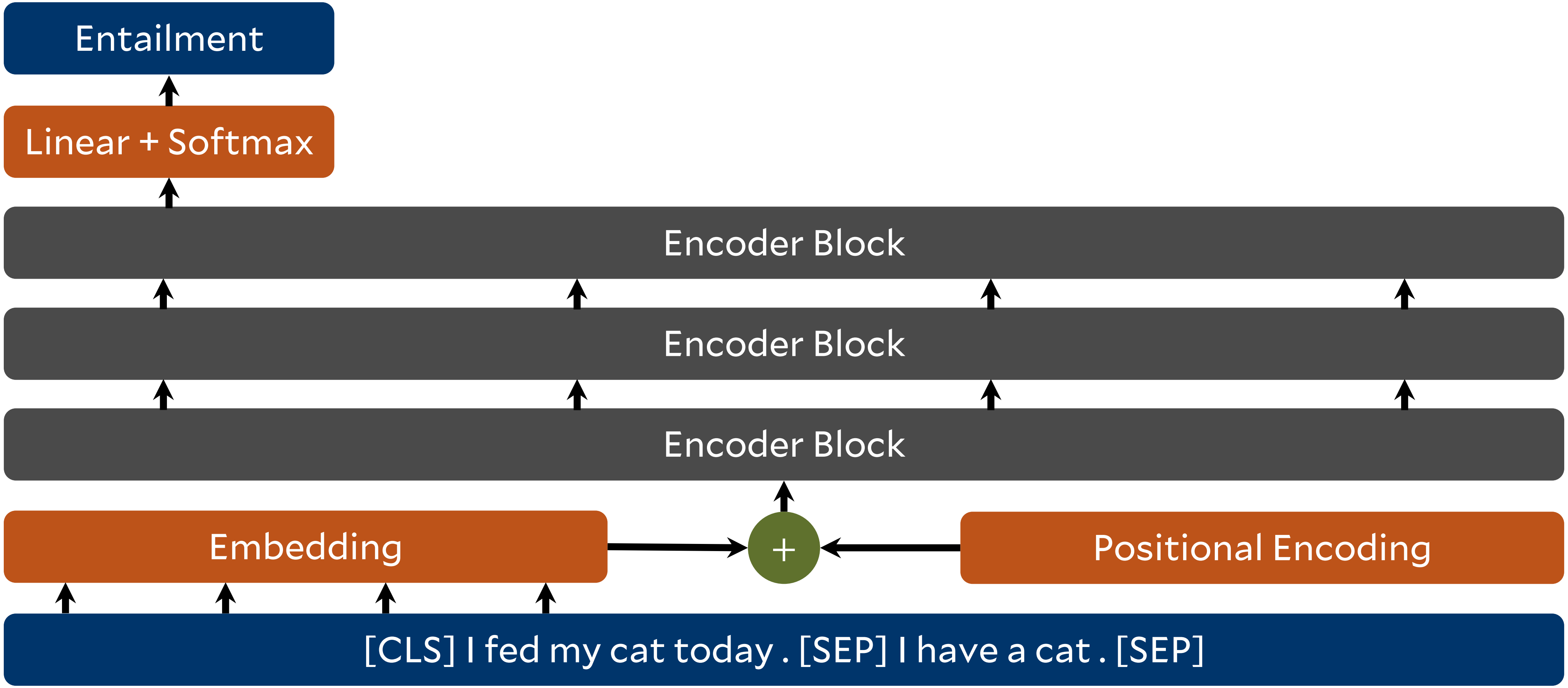
BERT Training Data

Corpus	**Size** (Millions of Words)
BooksCorpus (Zhu et al., 2015)	800
English Wikipedia	2,500
Total	3,300

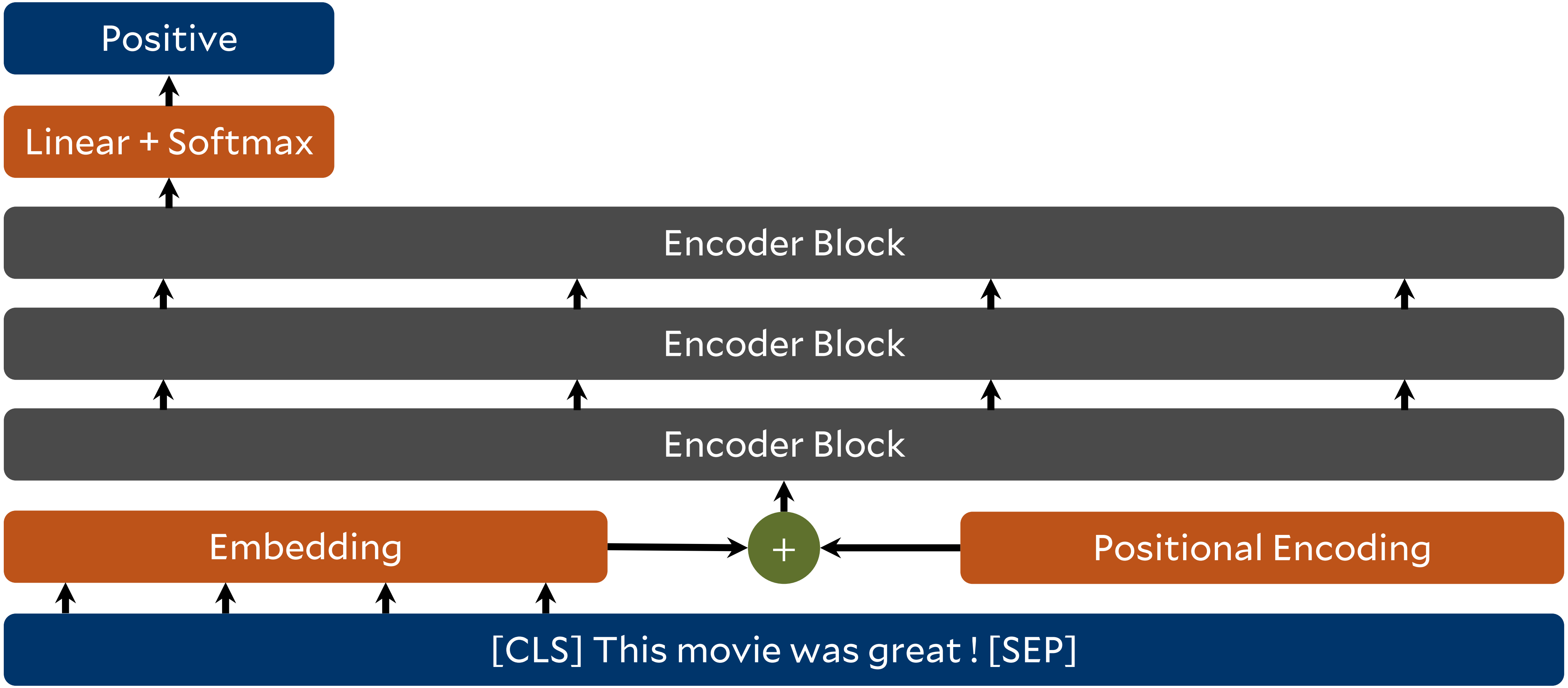
Fine-Tuning BERT

- To fine-tune BERT, just download a pre-trained BERT and start training it on some task.
- Instead of [BOS] and [EOS], use [CLS] and [SEP].
- For two-sentence inputs, put a [SEP] between the two sentences.

BERT for Two-Sentence Classification



BERT for One-Sentence Classification



Positive

Linear + Softmax

Encoder Block

Encoder Block

Encoder Block

Embedding

+

Positional Encoding

[CLS] This movie was great ! [SEP]

Classification Tasks

- **Natural Language Inference:** Do two sentences have a relation of entailment/contradiction/neither?
- **Sentiment Analysis:** Does a sentence have positive/negative/neutral sentiment?
- **Linguistic Acceptability:** Is a sentence grammatical?
- **Paraphrase Classification:** do two sentences mean the same thing?

BERT Classification Accuracy (%)

Task	SOTA	BERT Base Uncased	BERT Large Uncased
Natural Language Inference	80.6	84.6	86.7
Sentiment Analysis	93.2	93.5	94.9
Linguistic Acceptability	35.0	52.1	60.5
Paraphrase Classification	86.0	88.9	89.3

Other Versions of BERT

- **RoBERTa:** Facebook’s (improved) version of BERT
- **DistilBERT, ALBERT:** Smaller versions of BERT
- **CamemBERT, FlauBERT:** BERT in French
- **PhoBERT, herBERT:** BERT in Vietnamese and Polish, resp.
- **mBERT**: BERT in 104 languages
- **SpanBERT:** BERT for phrase-level tasks (e.g., named entity recognition, coreference resolution, etc.)