BEFORE WE START

Assignment 2 deadline pushed to Tuesday noon

Error in Assignment 2:

Use “Pikachu”, “Charizard” and “Charmander”
(as opposed to “pikachu”, “charizard”, “charmander”)

Using NLTK n-grams is ok, but I think you could implement it on your own.

ANTONIS ANASTASOPOULOS

CS499 INTRODUCTION TO NLP

VECTOR SEMANTICS

https://cs.gmu.edu/~antonis/course/cs499-spring21/

With adapted slides by Graham Neubig

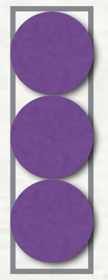
STRUCTURE OF THIS LECTURE

1. Why sentence representations?
2. Multi-task Learning
3. Training Sent. Representations
4. Contextualized Embeddings

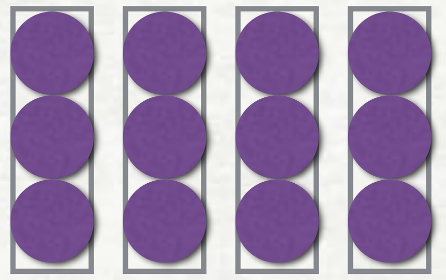
SENTENCE REPRESENTATIONS

We can create a vector or sequence of vectors from a sentence

this is an example →



this is an example →



<u>Obligatory Quote!</u>

> "You can't cram the meaning of a whole %&!$ing sentence into a single $&!*ing vector!"
> — Ray Mooney

GOAL FOR TODAY

Briefly Introduce **tasks**, **datasets** and **methods**

Introduce different **training objectives**

Talk about **multitask/transfer learning**

TASKS USING SENTENCE REPRESENTATIONS

WHERE WOULD WE NEED/USE SENTENCE REPRESENTATIONS?

Sentence Classification

Paraphrase Identification

Semantic Similarity

Entailment

Retrieval

SENTENCE CLASSIFICATION

Classify sentences according to various traits

Topic, sentiment, subjectivity/objectivity, etc.

I hate this movie

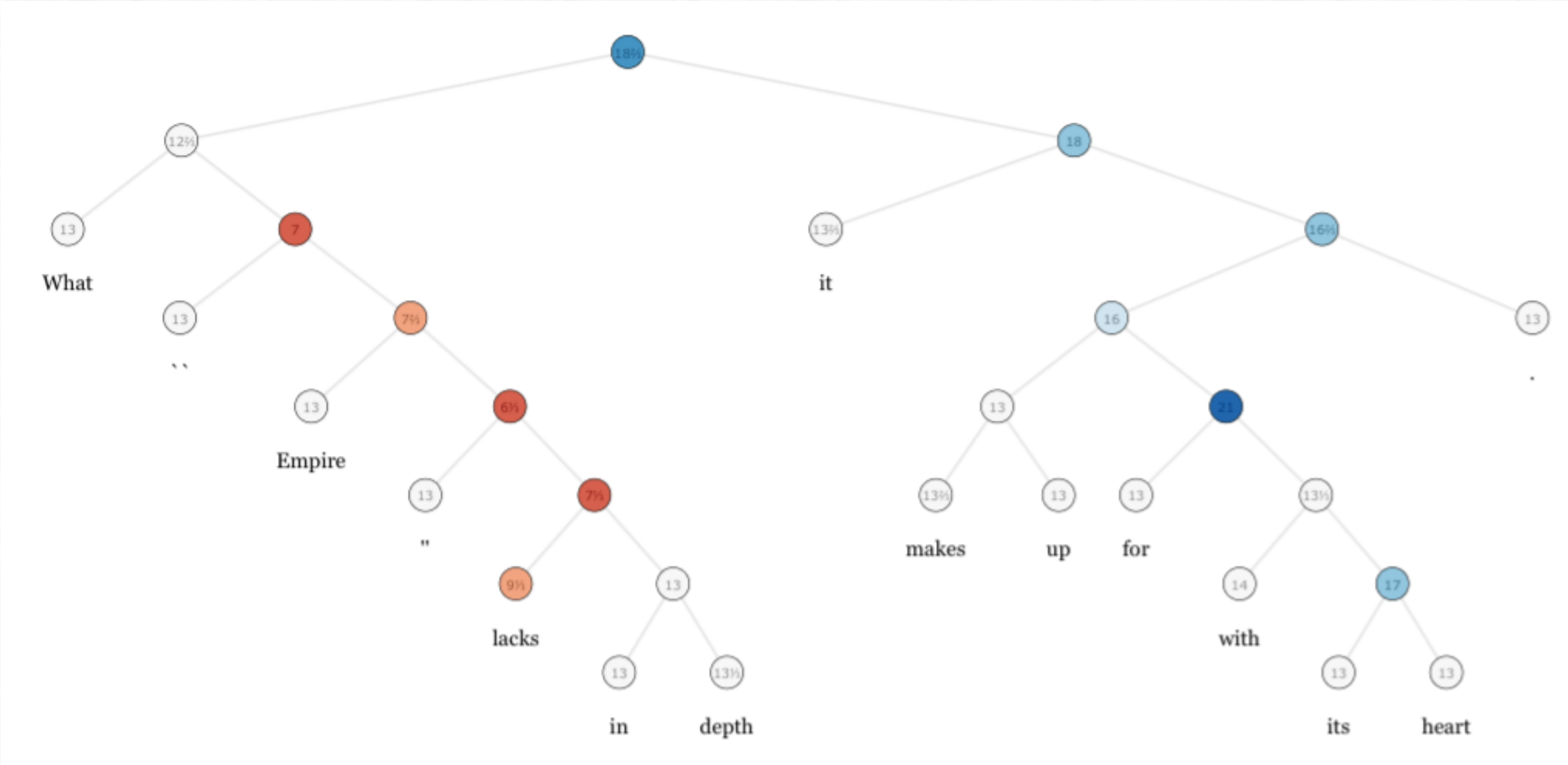
very good
good
neutral
bad
very bad

I love this movie

very good
good
neutral
bad
very bad

DATA EXAMPLE: STANFORD SENTIMENT TREEBANK (SOCHER ET AL. 2013)

In addition to standard tags, each constituent tagged with a sentiment value



PARAPHRASE IDENTIFICATION (DOLAN AND BROCKETT 2005)

Identify whether A and B mean the same thing

Charles O. Prince, 53, was named as Mr. Weill's successor.

↕

Mr. Weill's longtime confidant, Charles O. Prince, 53, was named as his successor.

- **Note:** *exactly* the same thing is too restrictive, so use a loose sense of similarity

SEMANTIC SIMILARITY/RELATEDNESS (MARELLI ET AL. 2014)

Do two sentences mean something similar?

Relatedness score	Example
1.6	A: "*A man is jumping into an empty pool*" B: "*There is no biker jumping in the air*"
2.9	A: "*Two children are lying in the snow and are making snow angels*" B: "*Two angels are making snow on the lying children*"
3.6	A: "*The young boys are playing outdoors and the man is smiling nearby*" B: "*There is no boy playing outdoors and there is no man smiling*"
4.9	A: "*A person in a black jacket is doing tricks on a motorbike*" B: "*A man in a black jacket is doing tricks on a motorbike*"

- Like paraphrase identification, but with shades of gray.

TEXTUAL ENTAILMENT (DAGAN ET AL. 2006, MARELLI ET AL. 2014)

Entailment: if A is true, then B is true (c.f. paraphrase, where opposite is also true)

The woman bought a sandwich for lunch
→ The woman bought lunch

Contradiction: if A is true, then B is not true

The woman bought a sandwich for lunch
→ The woman did not buy a sandwich

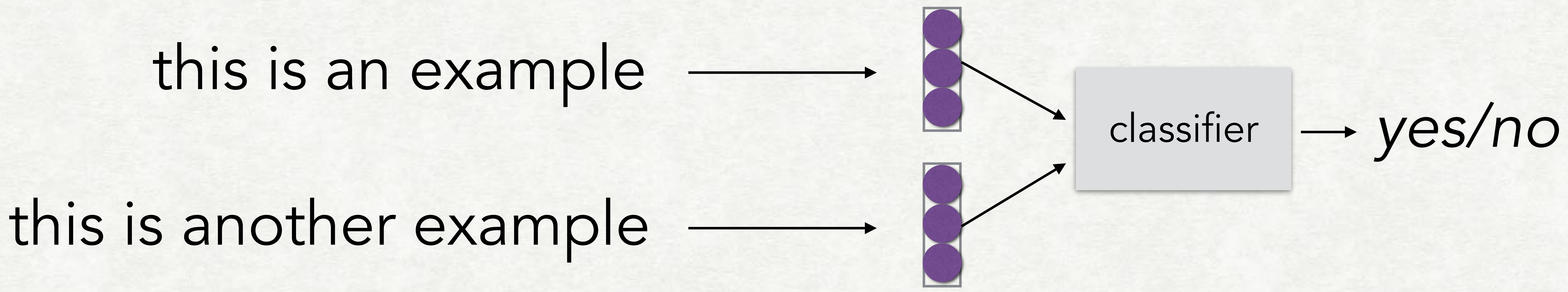
Neutral: cannot say either of the above

The woman bought a sandwich for lunch
→ The woman bought a sandwich for dinner

MODEL FOR SENTENCE PAIR PROCESSING

Calculate vector representation

Feed vector representation into classifier



How do we get such a representation?

MULTI-TASK LEARNING OVERVIEW

TYPES OF LEARNING

Multi-task learning is a general term for training on multiple tasks

Transfer learning is a type of multi-task learning where we only really care about one of the tasks

Domain adaptation is a type of transfer learning, where the output is the same, but we want to handle different topics or genres, etc.

PLETHORA OF TASKS IN NLP

In NLP, there are a plethora of tasks, each requiring different varieties of data

- **Only text:** e.g. language modeling
- **Naturally occurring data:** e.g. machine translation
- **Hand-labeled data:** e.g. most analysis tasks

And each in many languages, many domains!

RULE OF THUMB 1: MULTITASK TO INCREASE DATA

Perform multi-tasking when one of your two tasks has many fewer data

General domain → specific domain
(e.g. web text → medical text)

High-resourced language → low-resourced language
(e.g. English → Telugu)

Plain text → labeled text
(e.g. LM -> parser)

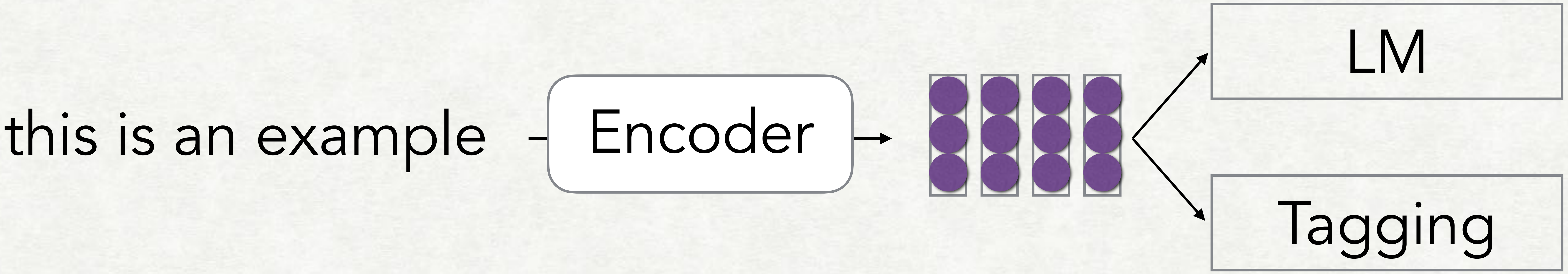
RULE OF THUMB 2: TASK RELATEDNESS

Perform multi-tasking when your **tasks are related**

e.g. predicting eye gaze and summarization (Klerke et al. 2016)

STANDARD MULTI-TASK LEARNING

Train representations to do well on multiple tasks at once

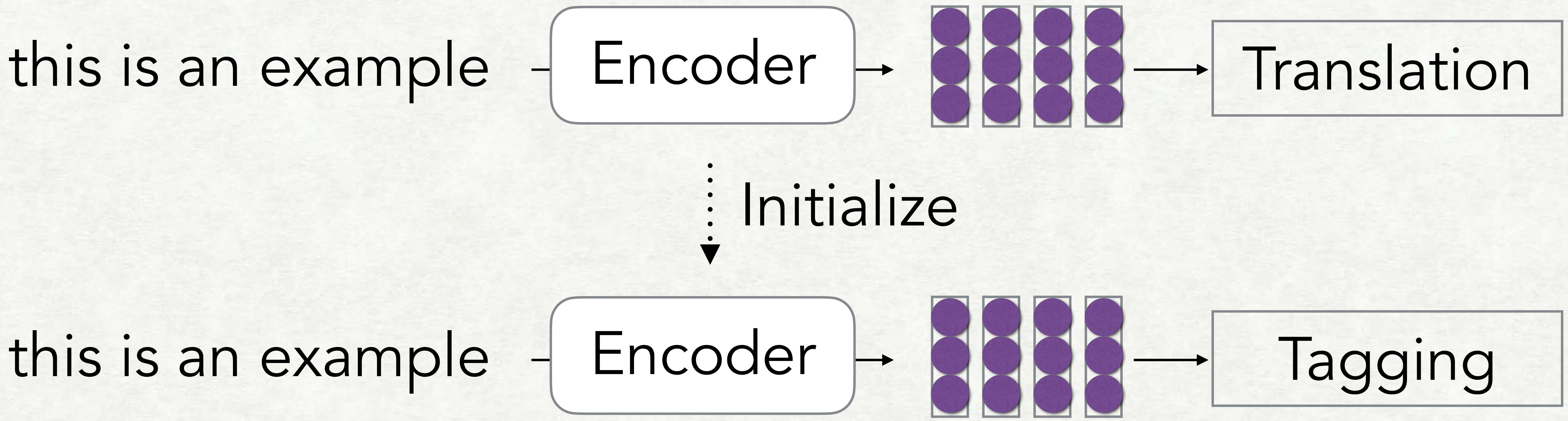


- In general, as simple as randomly choosing minibatch from one of multiple tasks

- Many many examples, starting with Collobert and Weston (2011)

PRE-TRAINING

First train on one task, then train on another



- Widely used in word embeddings (Turian et al. 2010)

- Also pre-training sentence encoders or contextualized word representations (Dai et al. 2015, Melamud et al. 2016)

THINKING ABOUT MULTI-TASKING, AND PRE-TRAINED REPRESENTATIONS

Many methods have names like SkipThought, ParaNMT, CoVe, ELMo, BERT along with pre-trained models

These often refer to a combination of

- **Model:** The underlying neural network architecture
- **Training Objective:** What objective is used to pre-train
- **Data:** What data the authors chose to use to train the model

Remember that these are often conflated (and don't need to be)!

END-TO-END VS. PRE-TRAINING

For any model, we can always use an end-to-end training objective

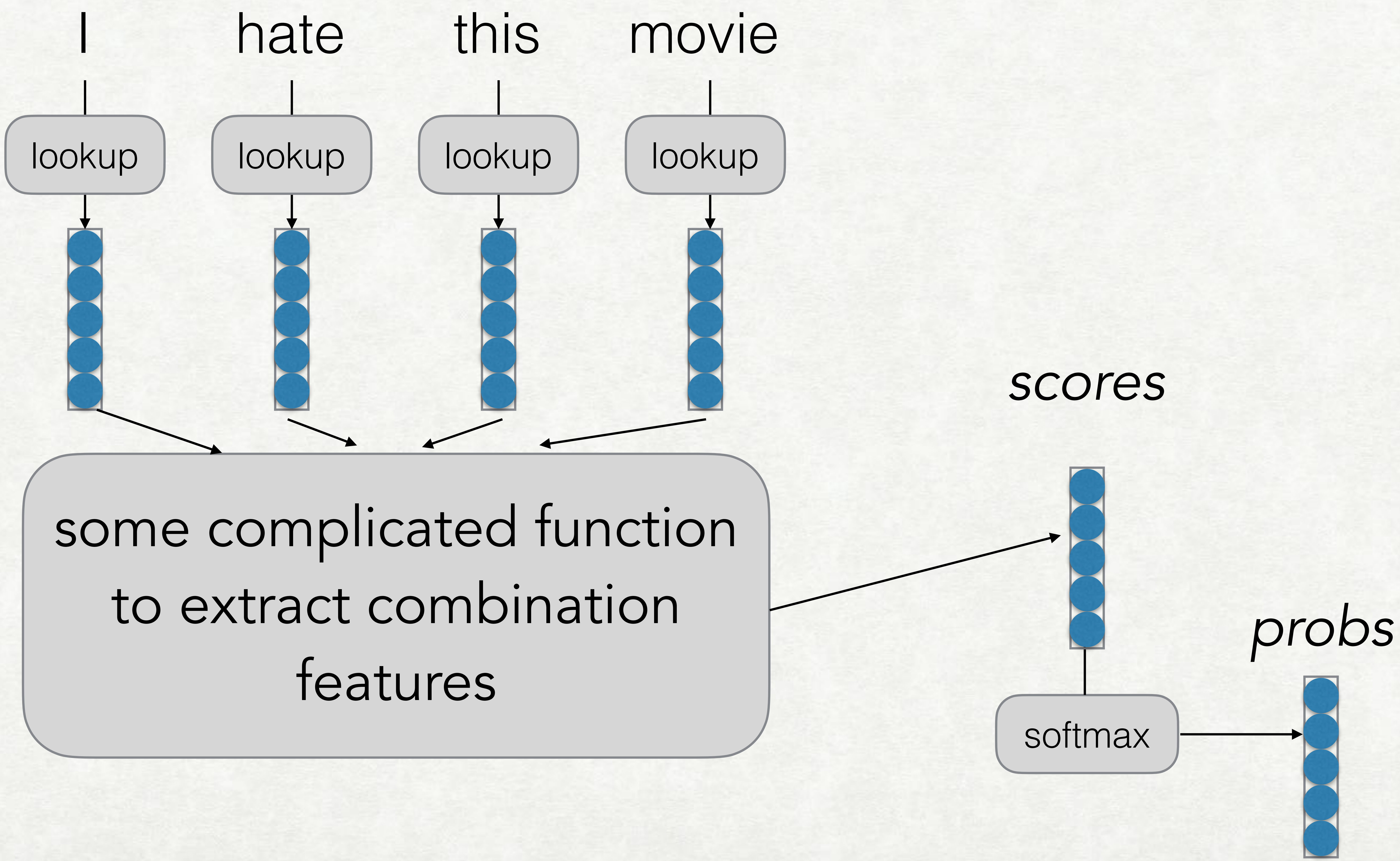
Problem: paucity of training data

Problem: weak feedback from end of sentence only for text classification, etc.

Often better to pre-train sentence embeddings on other task, then use or fine tune on target task

TRAINING SENTENCE REPRESENTATIONS

GENERAL MODEL OVERVIEW

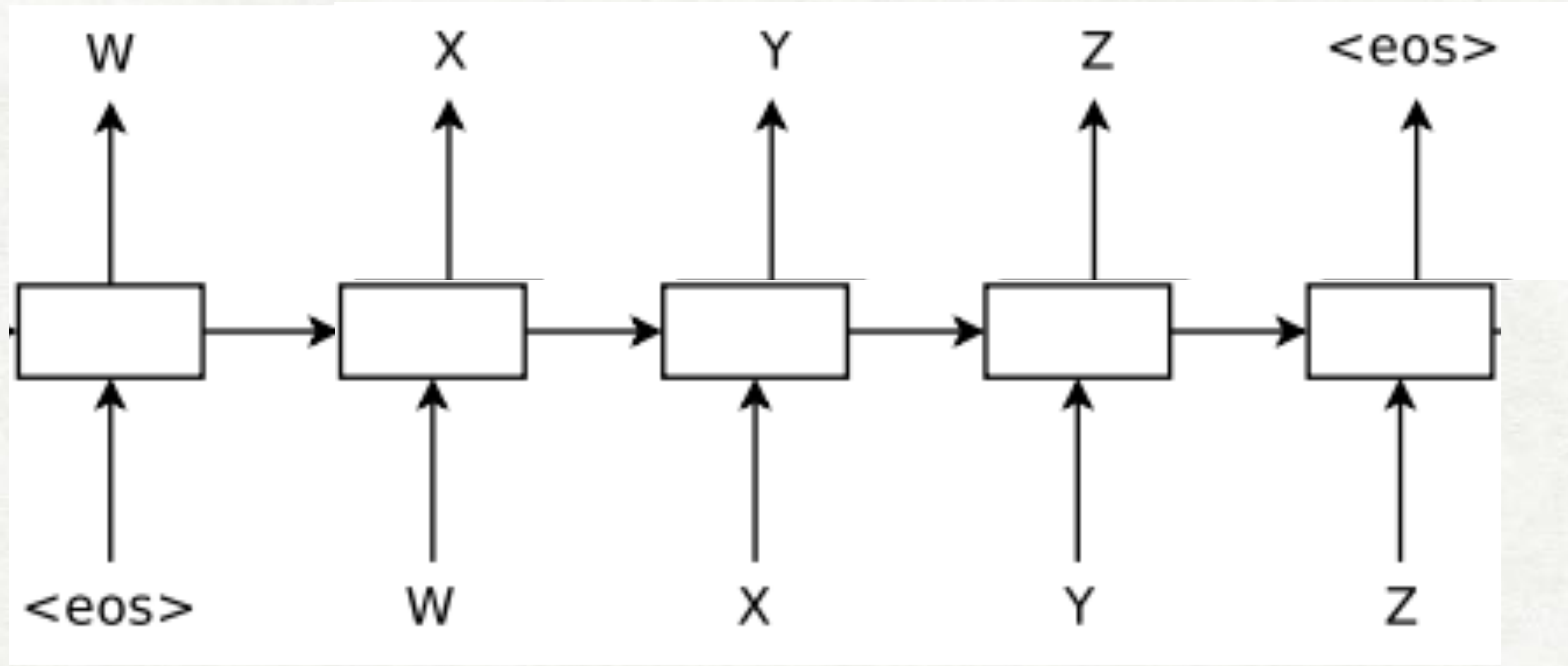


LANGUAGE MODEL TRANSFER (DAI AND LE 2015)

Model: LSTM

Objective: Language modeling objective

Data: Classification data itself, or Amazon reviews

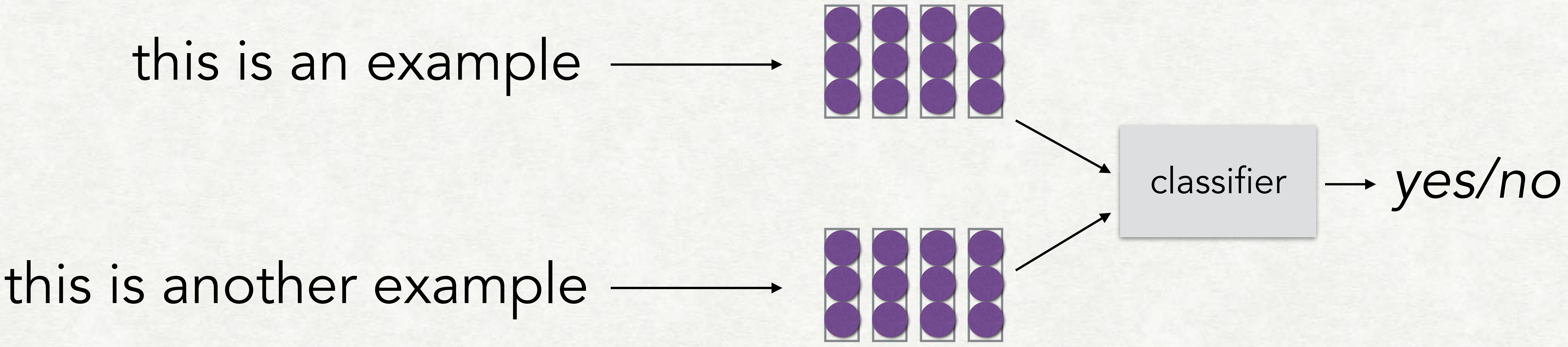


- **Downstream:** On text classification, initialize weights and continue training

CONTEXTUALIZED WORD REPRESENTATIONS

CONTEXTUALIZED WORD REPRESENTATIONS

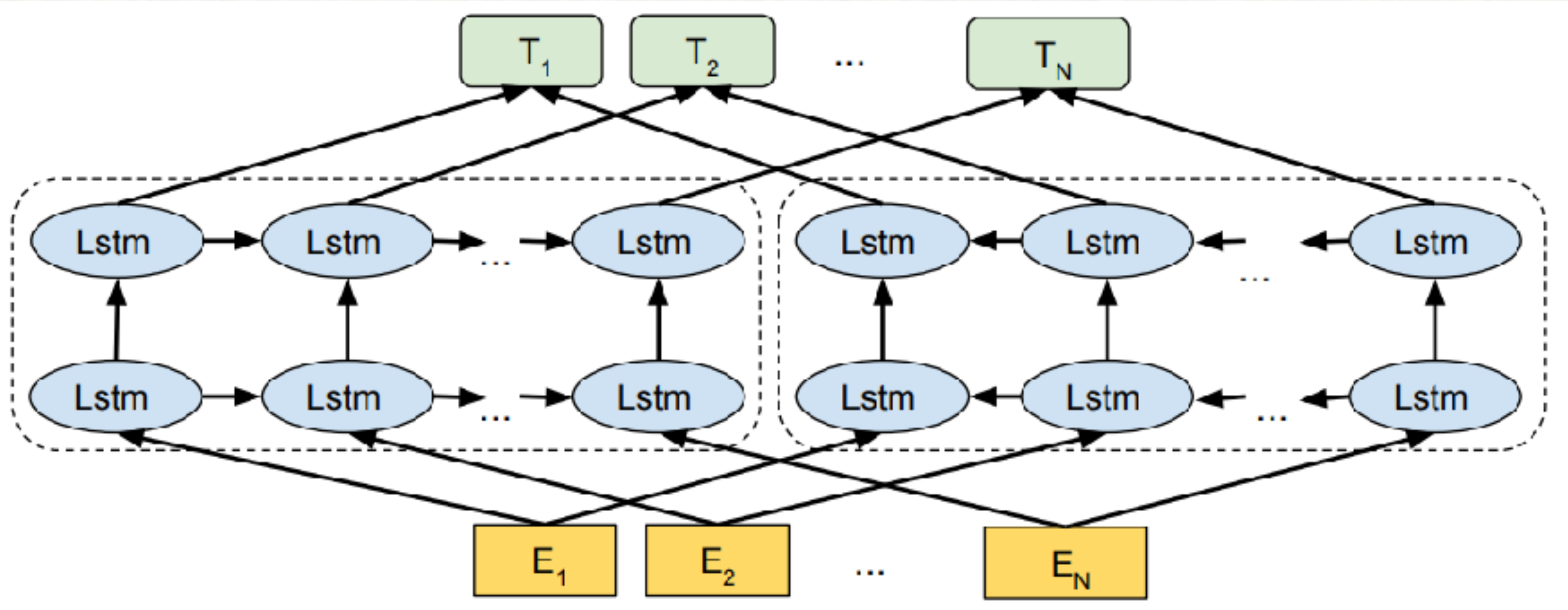
Instead of one vector per sentence, one vector per word!



How to train this representation?

BI-DIRECTIONAL LANGUAGE MODELING OBJECTIVE (ELMO; PETERS ET AL. 2018)

- **Model:** Multi-layer bi-directional LSTM
- **Objective:** Predict the next word left->right, next word right->left independently
- **Data:** 1B word benchmark LM dataset

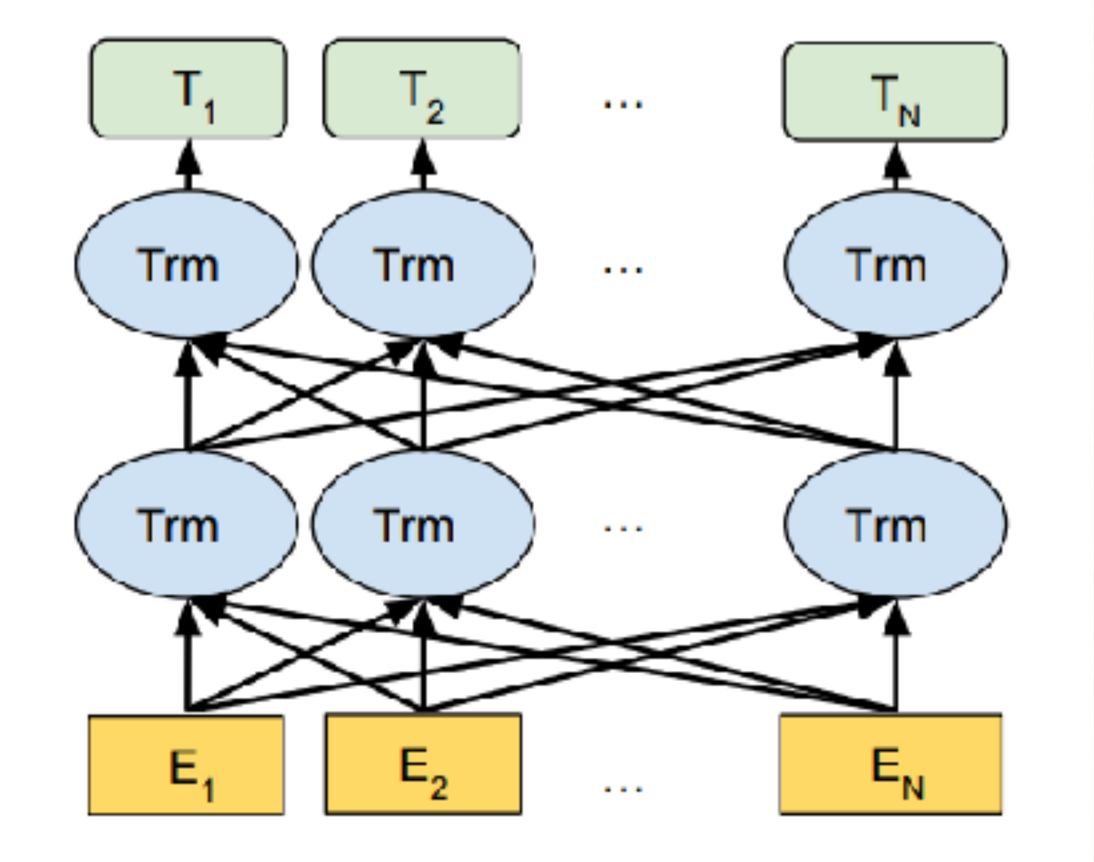


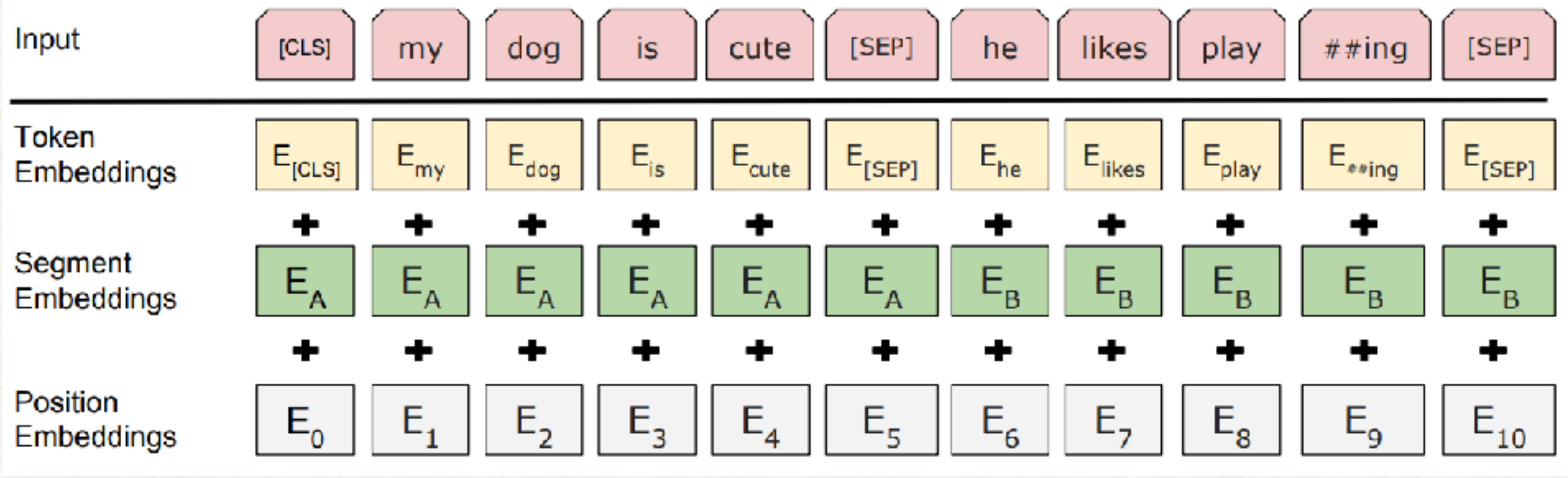
Downstream: Finetune the weights of the linear combination of layers on the downstream task

MASKED WORD PREDICTION (BERT; DEVLIN ET AL. 2018)

Like ELMo, uses bidirectional context, but with transformer model as base (+ tricks for efficient training)

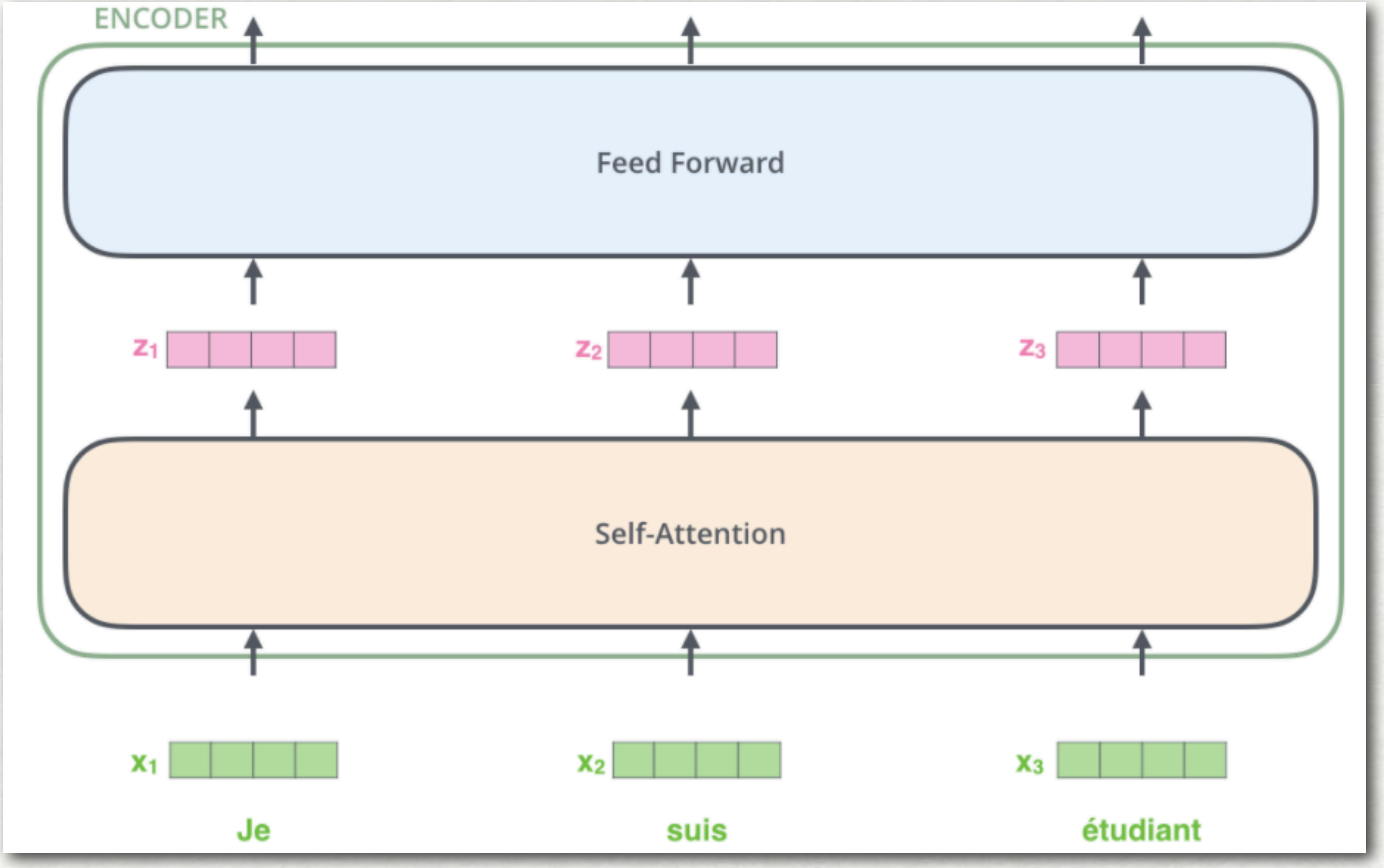
- **Model:** Multi-layer self-attention. Input sentence or pair, w/ [CLS] token, subword representation





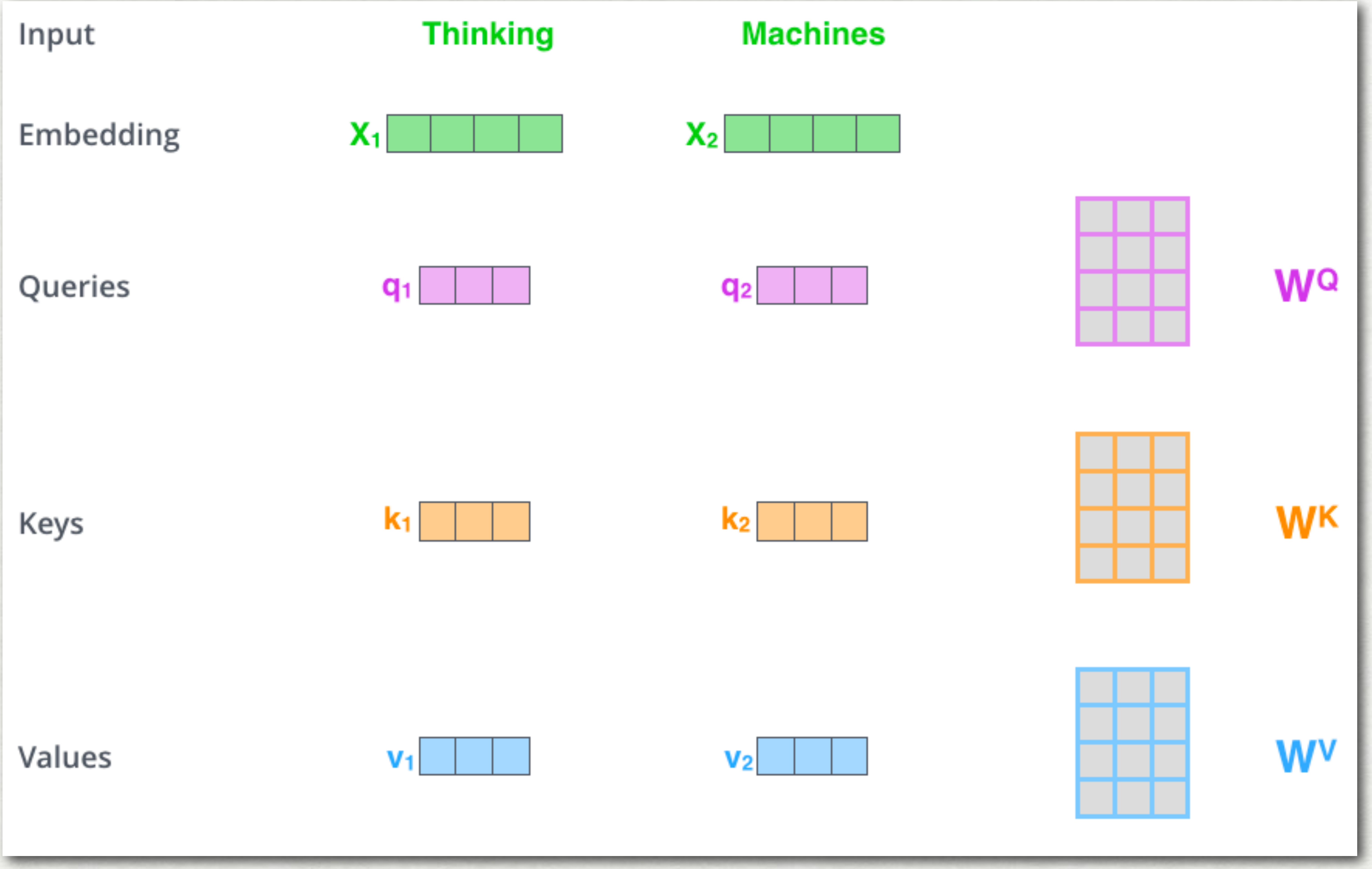
- **Objective:** Masked word prediction + next-sentence prediction
- **Data:** BooksCorpus + English Wikipedia

SELF-ATTENTION



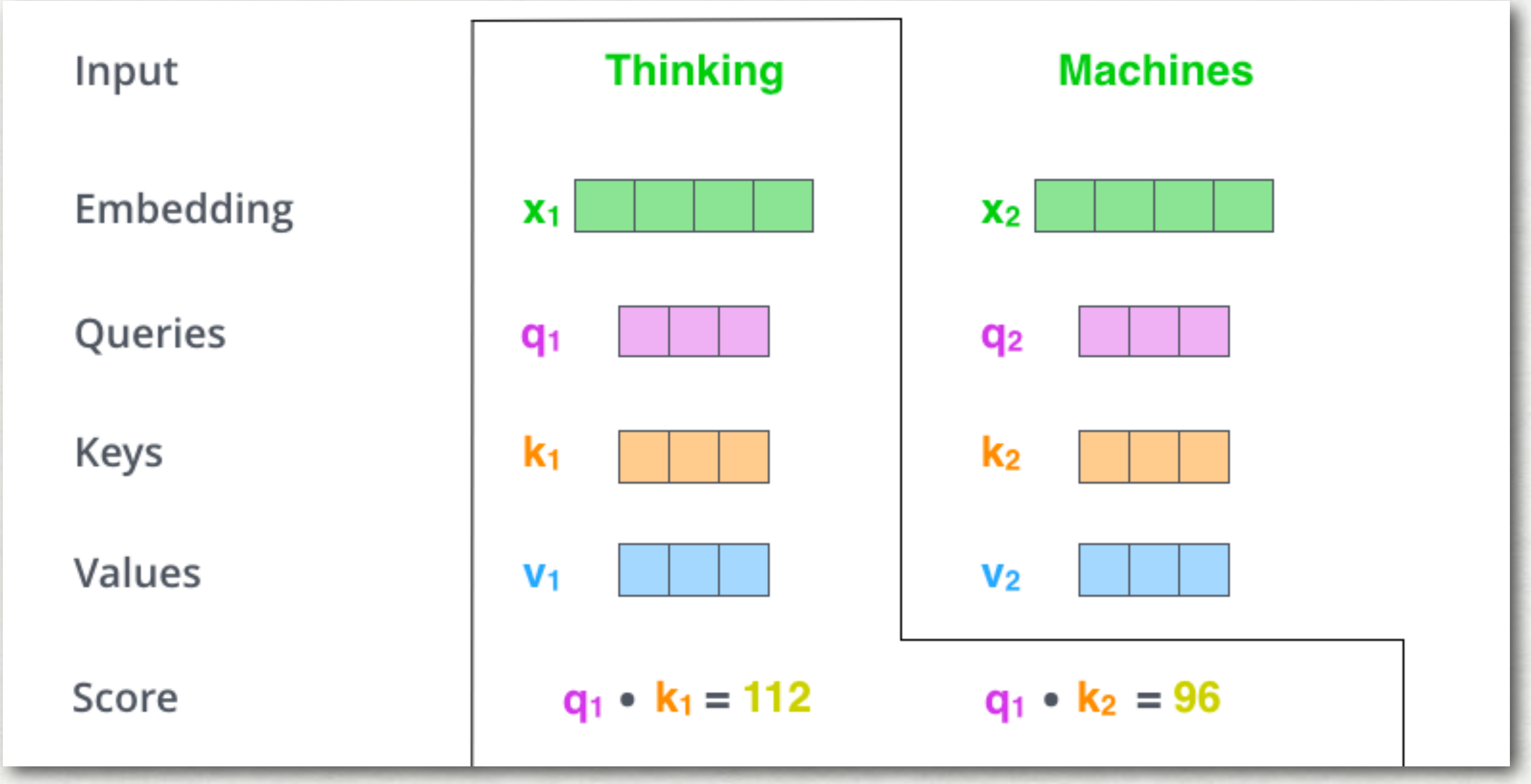
From http://jalammar.github.io/illustrated-transformer/

SELF-ATTENTION



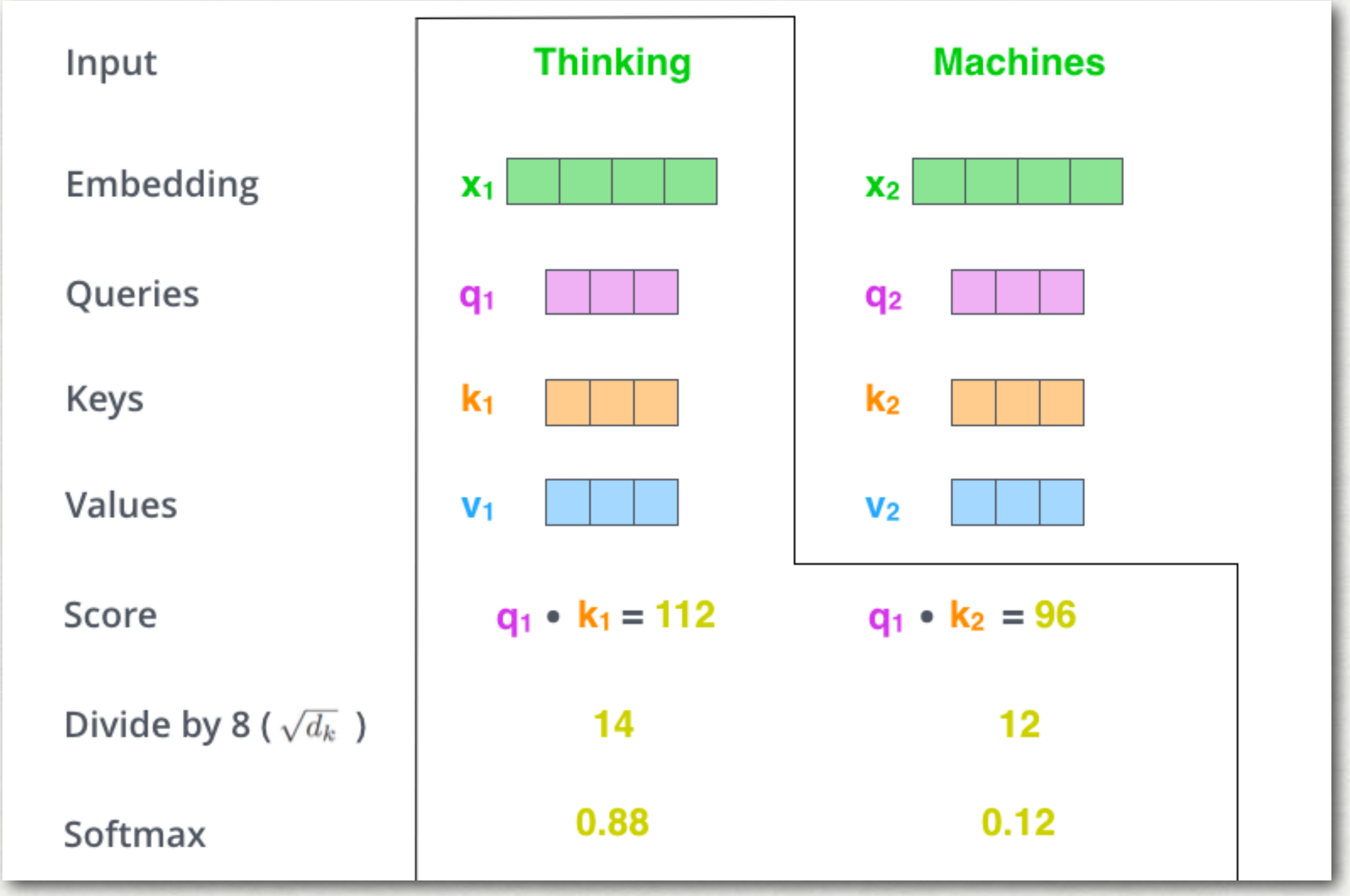
From http://jalammar.github.io/illustrated-transformer/

SELF-ATTENTION



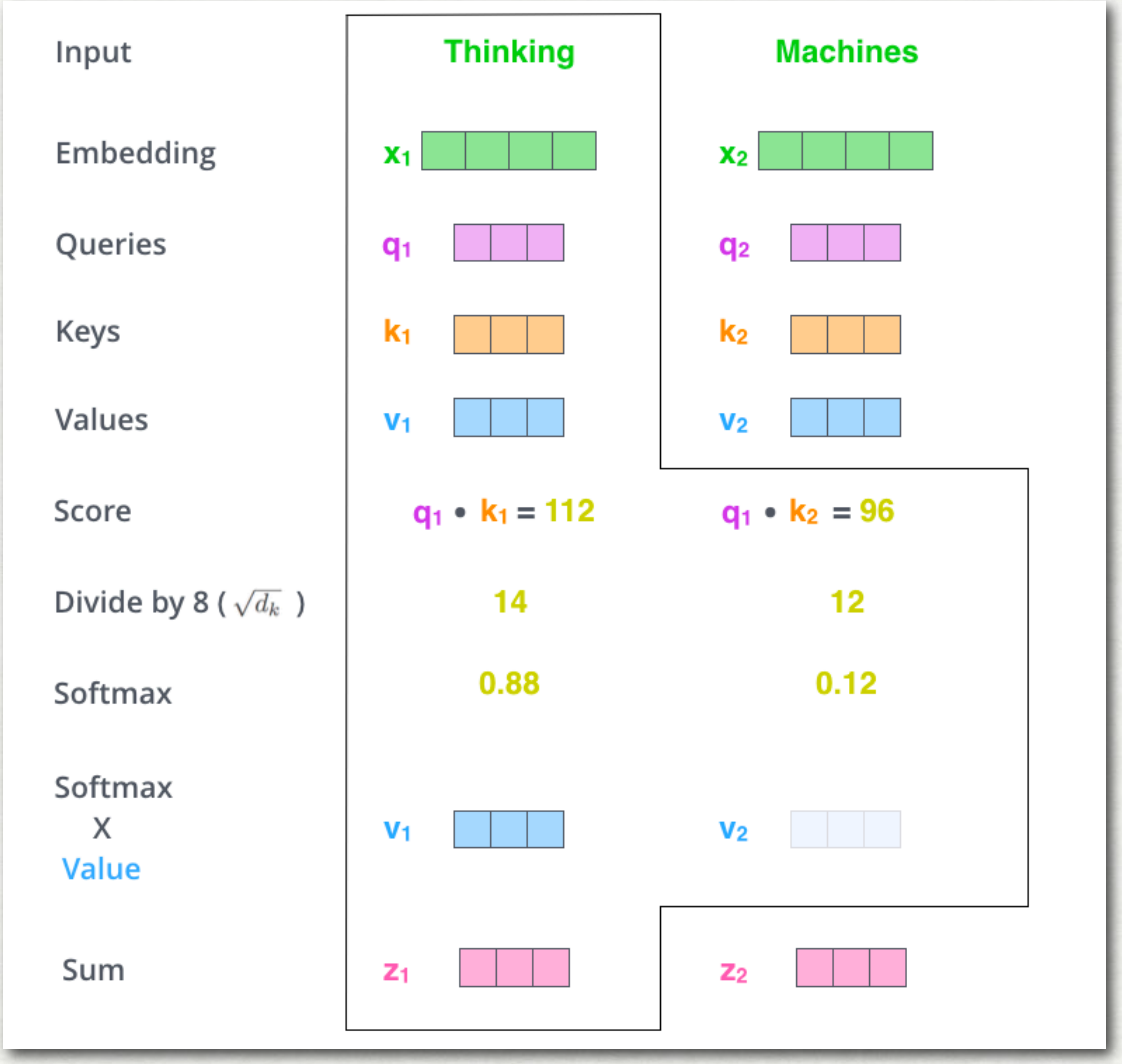
From http://jalammar.github.io/illustrated-transformer/

SELF-ATTENTION



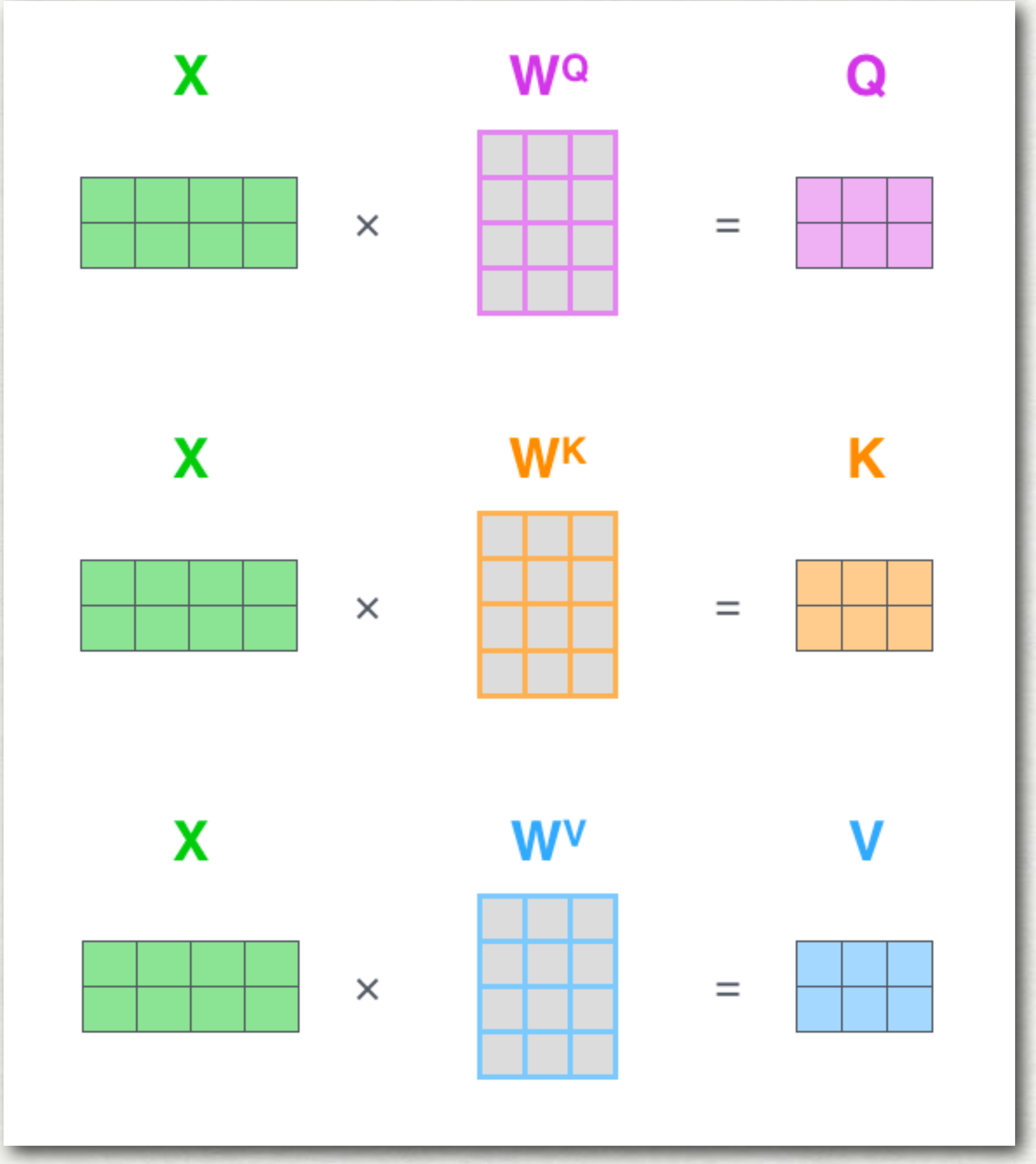
From http://jalammar.github.io/illustrated-transformer/

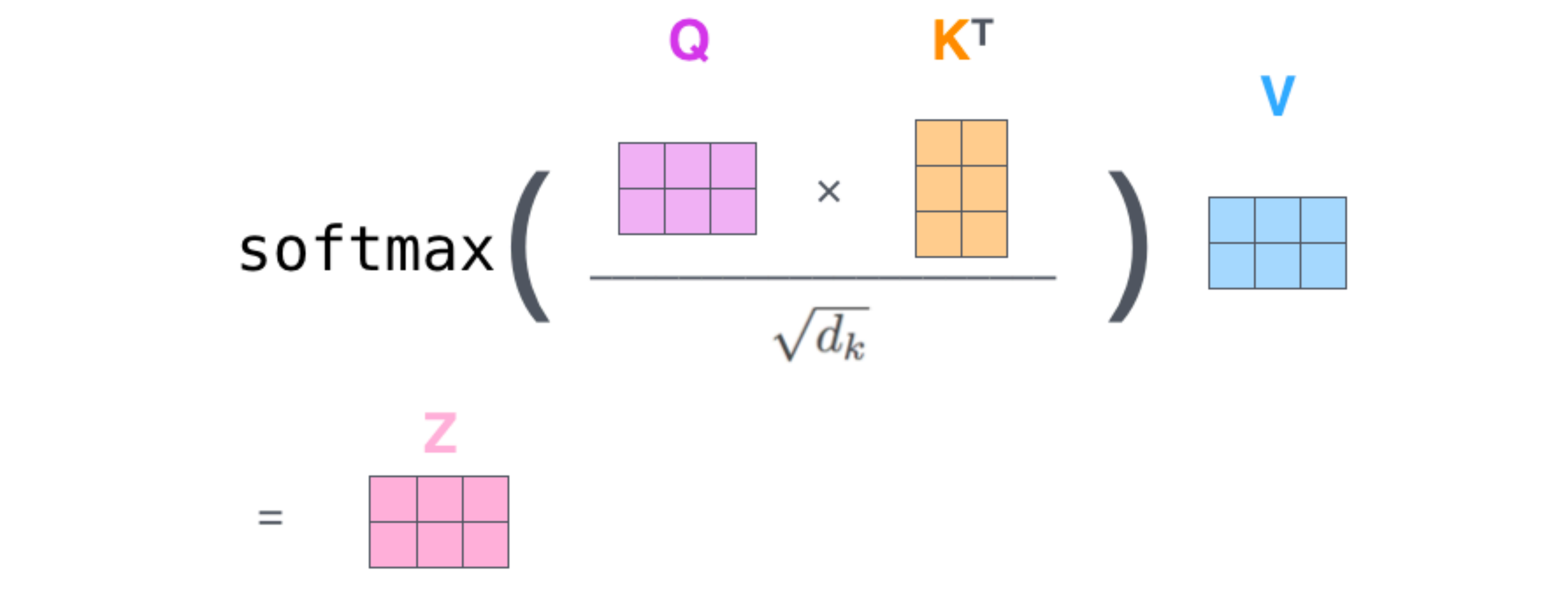
SELF-ATTENTION



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

SELF-ATTENTION





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

MASKED WORD PREDICTION (DEVLIN ET AL. 2018)

1. predict a masked word

80%: substitute input word with [MASK]

10%: substitute input word with random word

10%: no change

Like context2vec, but **better suited for multi-layer self attention**

CONSECUTIVE SENTENCE PREDICTION (DEVLIN ET AL. 2018)

1. classify two sentences as consecutive or not:

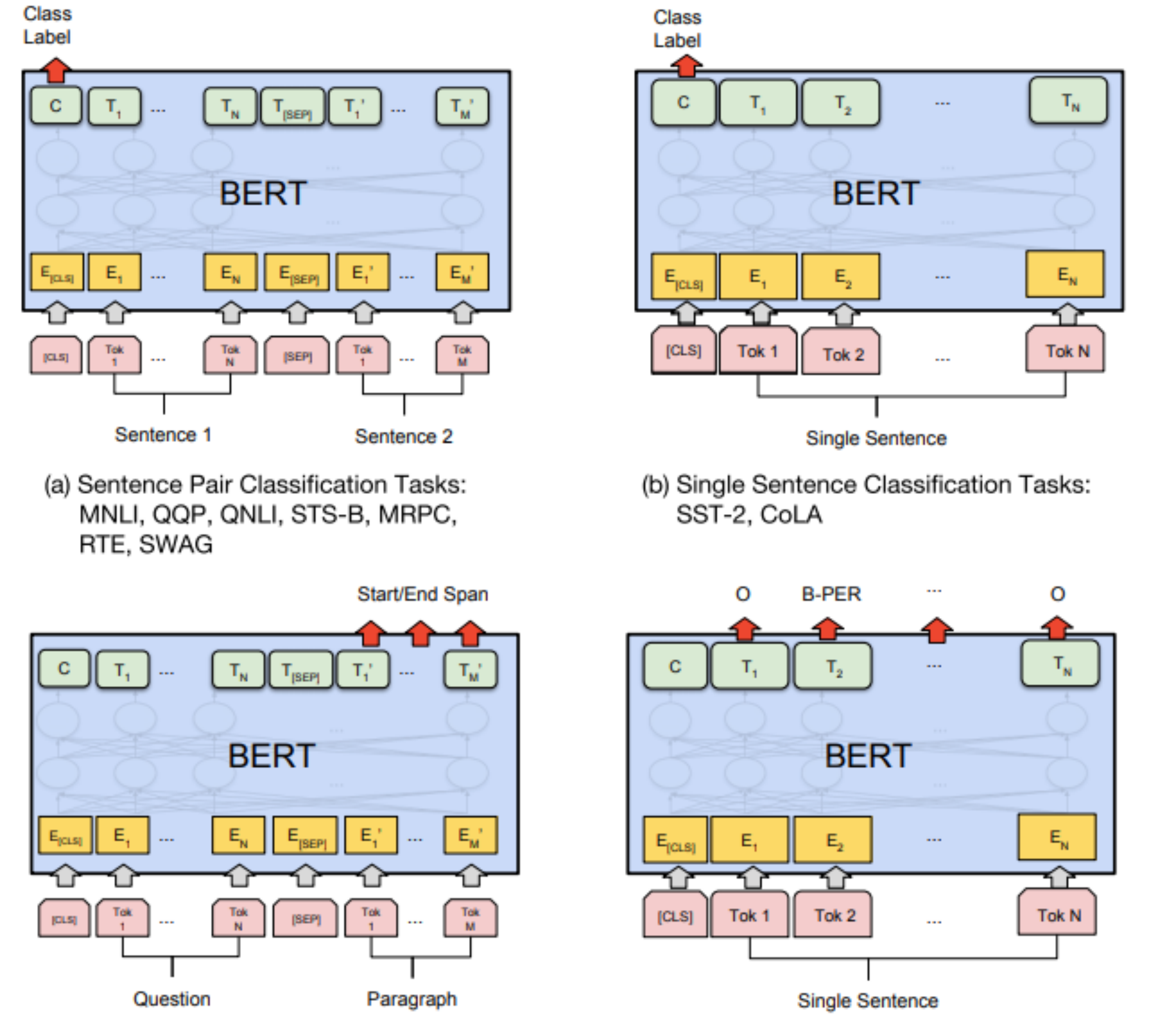
50% of training data (from OpenBooks) is "consecutive"

Input = [CLS] the man [MASK] to the store [SEP]
penguin [MASK] are flight ##less birds [SEP]
Label = NotNext

Input = [CLS] the man went to [MASK] store [SEP]
he bought a gallon [MASK] milk [SEP]
Label = IsNext

USING BERT WITH PRE-TRAINING/FINETUNING

Use the pre-trained model as the first “layer” of the final model, then train on the desired task



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

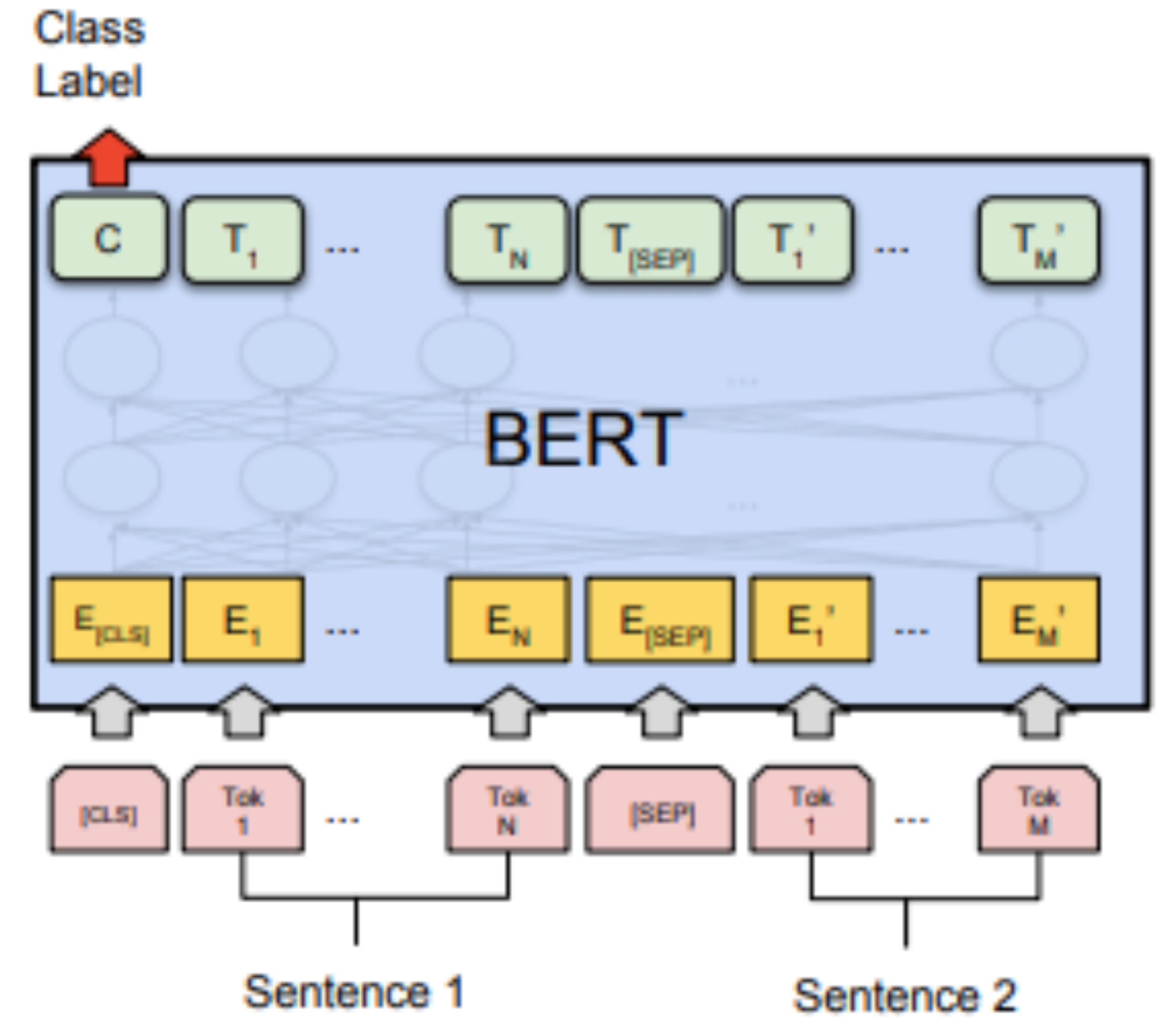
(b) Single Sentence Classification Tasks: SST-2, CoLA

(c) Question Answering Tasks: SQuAD v1.1

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

USING BERT WITH PRE-TRAINING/FINETUNING

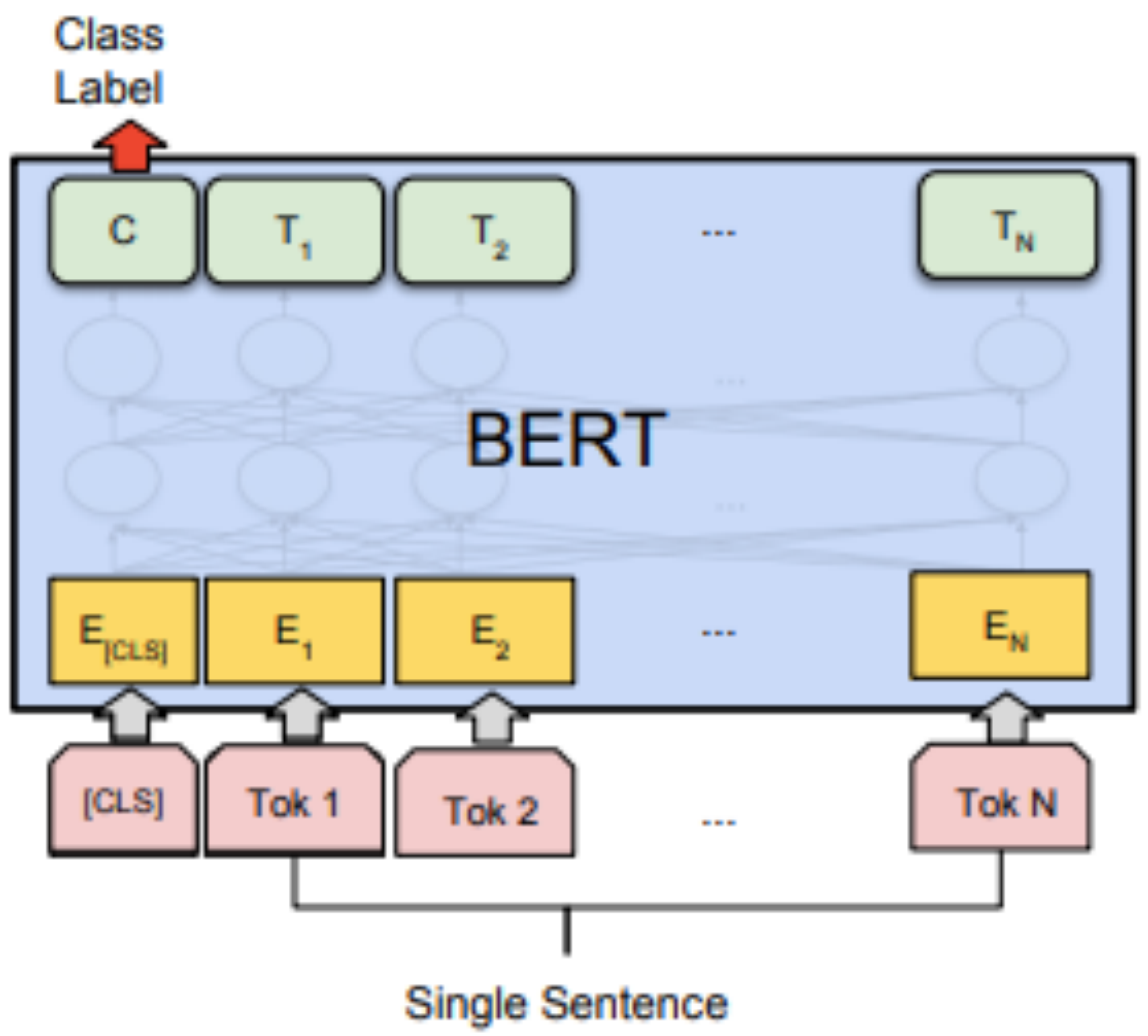
Use the pre-trained model as the first “layer” of the final model, then train on the desired task



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

USING BERT WITH PRE-TRAINING/FINETUNING

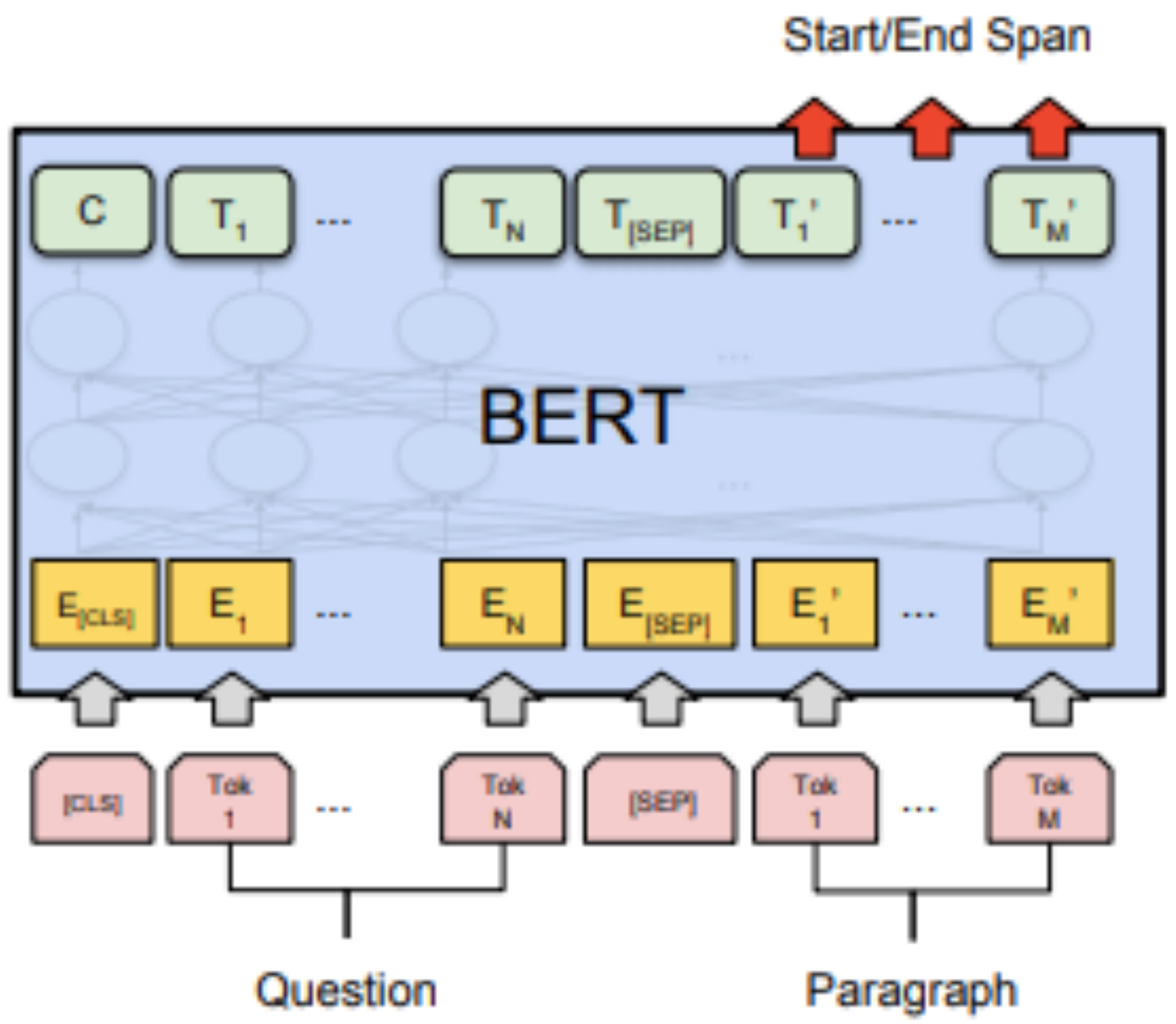
Use the pre-trained model as the first "layer" of the final model, then train on the desired task



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

USING BERT WITH PRE-TRAINING/FINETUNING

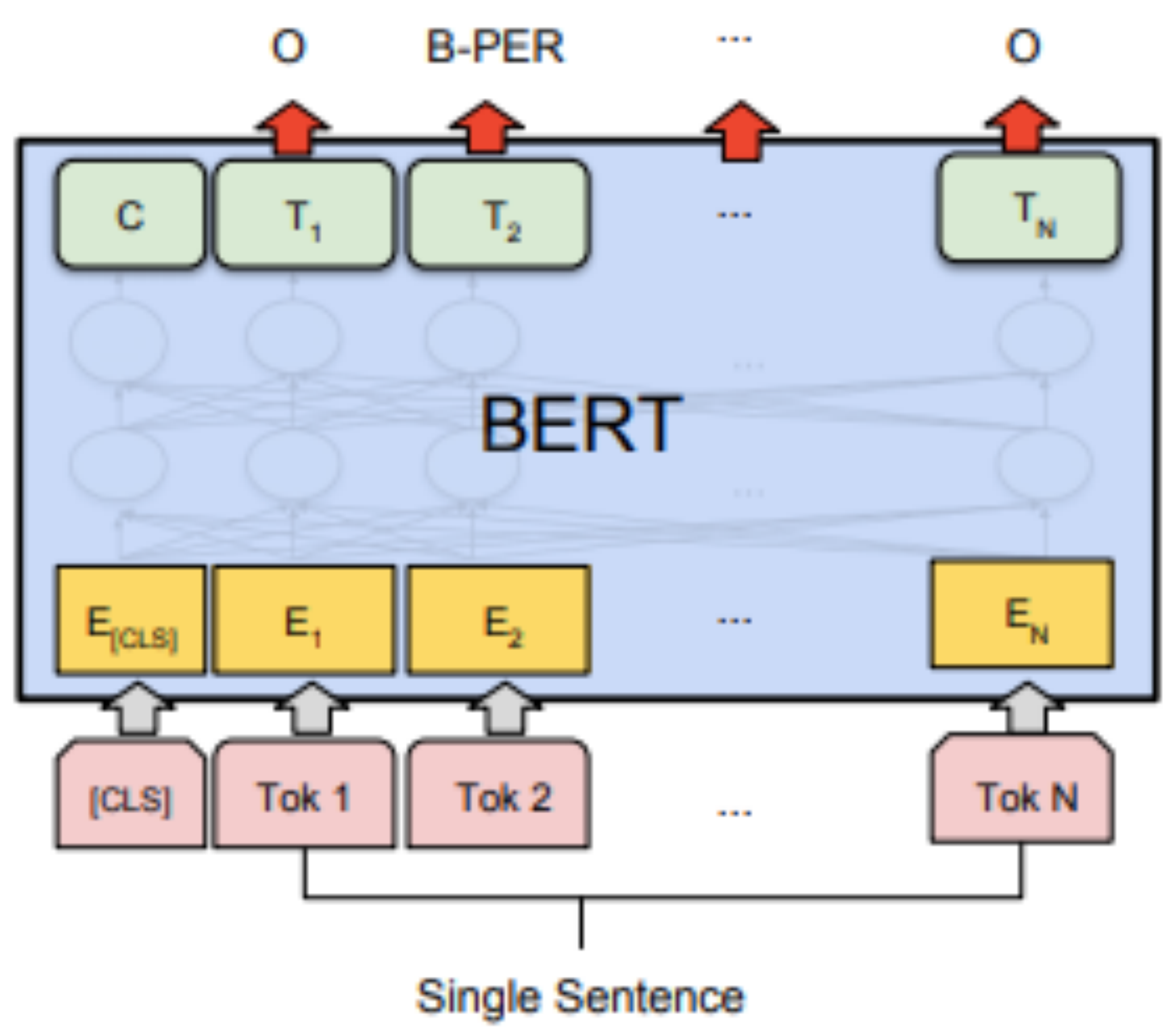
Use the pre-trained model as the first “layer” of the final model, then train on the desired task



(c) Question Answering Tasks:
SQuAD v1.1

USING BERT WITH PRE-TRAINING/FINETUNING

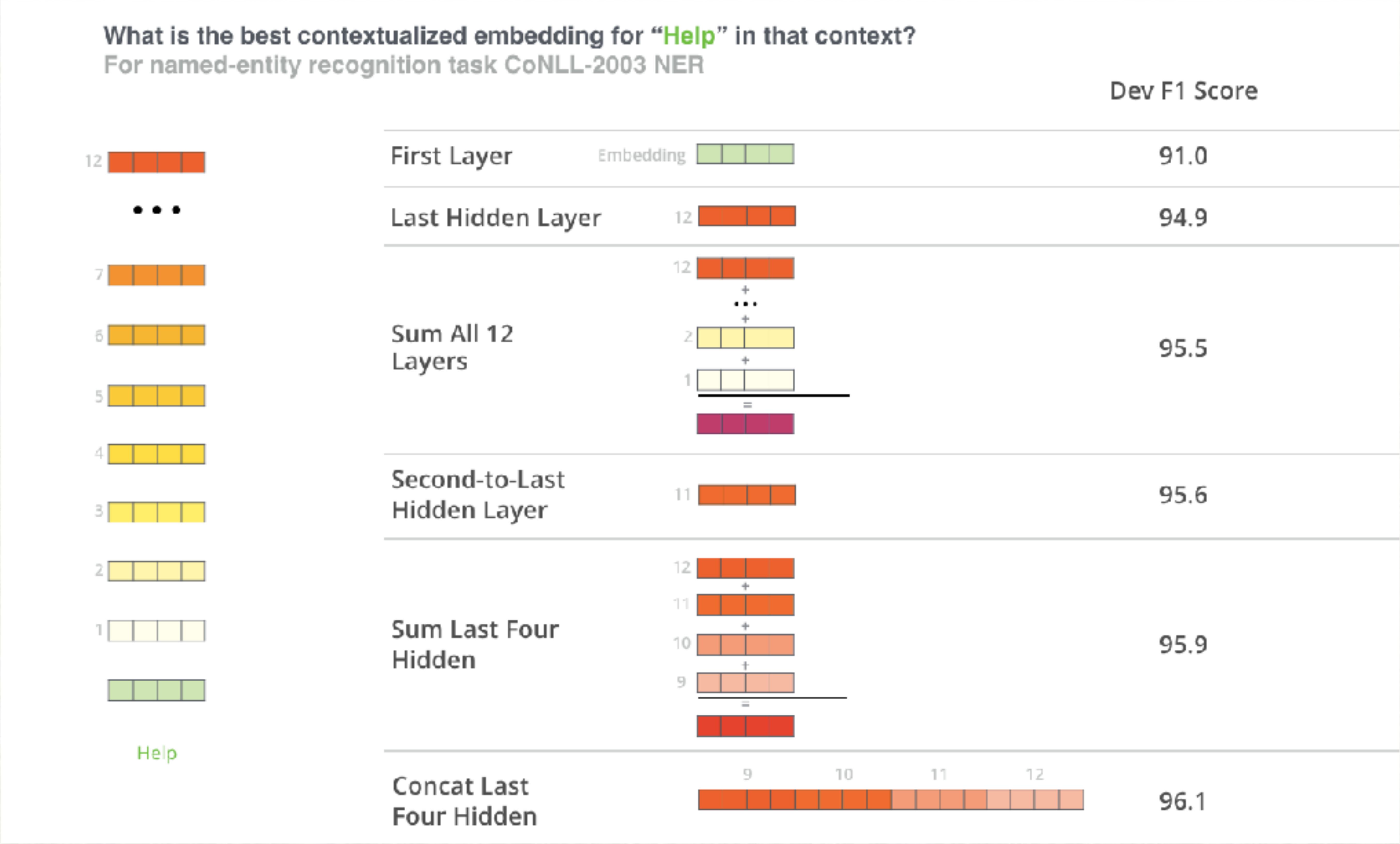
Use the pre-trained model as the first "layer" of the final model, then train on the desired task



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

USING BERT FOR REPRESENTATIONS

Use the pre-trained model to obtain contextualized word representations for the input



[visualization from The Illustrated BERT: https://jalammar.github.io/illustrated-bert/]

WHICH METHOD IS BETTER?

WHICH MODEL?

Not very extensive comparison…

Wieting et al. (2015) find that simple word averaging is more robust out-of-domain

Devlin et al. (2018) compare unidirectional and bi-directional transformer, but no comparison to LSTM like ELMo (for performance reasons?)

WHICH TRAINING OBJECTIVE?

Not very extensive comparison…

Zhang and Bowman (2018) control for training data, and find that bi-directional LM seems better than MT encoder

Devlin et al. (2018) find next-sentence prediction objective good compliment to LM objective

WHICH DATA?

Not very extensive comparison…

Zhang and Bowman (2018) find that more data is probably better, but results preliminary.

Data with context is probably essential.

SOME RECENT IMPROVEMENTS

VARIOUS MONOLINGUAL BERTS

French: FlauBERT, CamemBERT

BERTje, ALBERTO, BETO, KoBERT, FinBERT, Bangla-BERT, German, Chinese, Russian, Japanese, etc

web-scale scraped corpora: https://oscar-corpus.com/

MBERT

BERT trained on more than 100 languages

Really good starting point, but also issues for low-resource languages, e.g. over-segmentation

ROBERTA

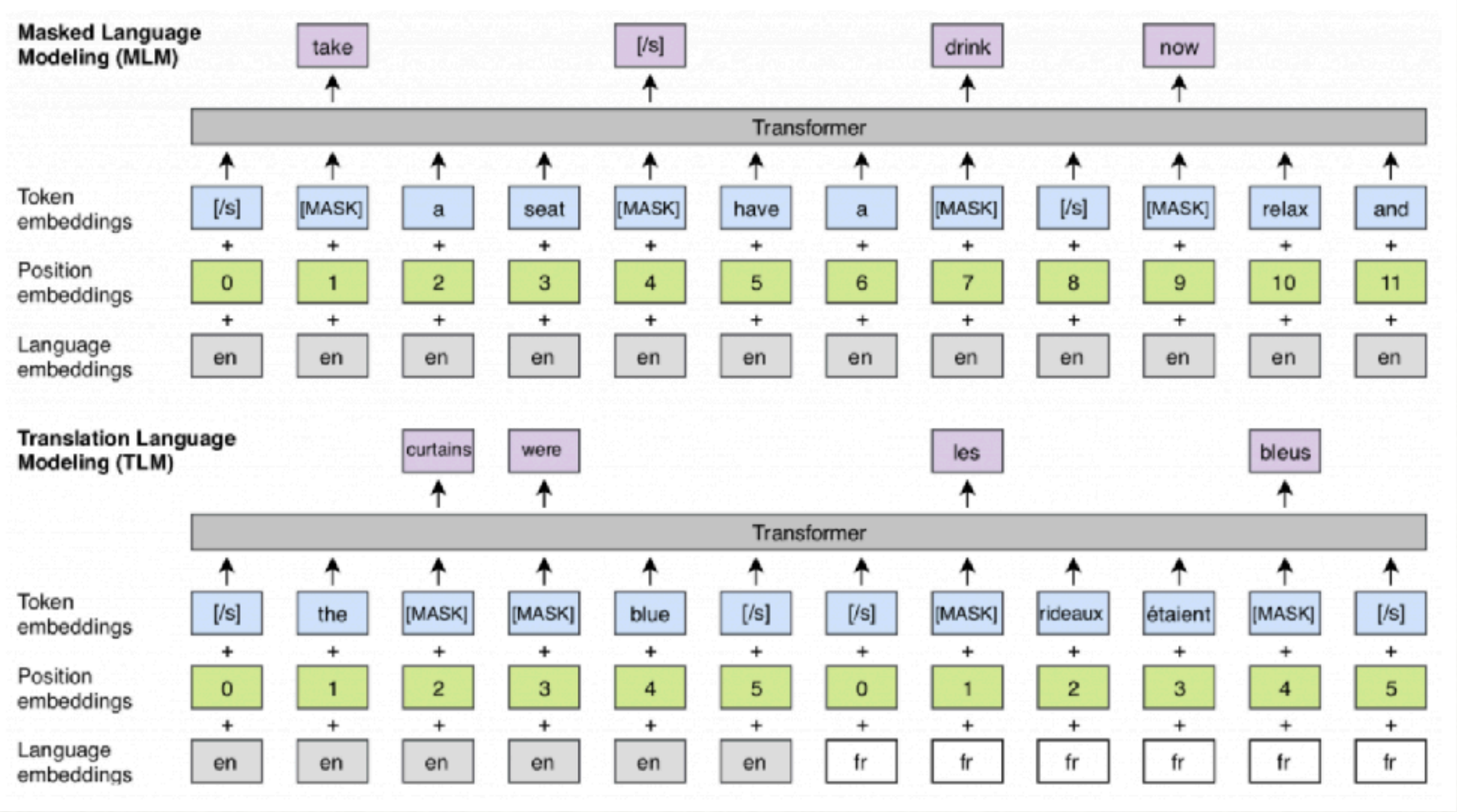
Original BERT model was under-trained

Better trained, more data, and more robust model

XLM AND XLM-R

BERT problem: each sample in a single language

Combine MLM with Translation LM



NEXT CLASS PREVIEW

Part-of-speech and Part-of-speech tagging