LLM Workshop session 2:

Working with Textual Data

Mahdi Samiei

21 Feb 2024

What we will discuss …

- Text, NLP and ML
- Simple ML for NLP and Tokenization
- Simple DL for NLP
- Deep Architectures and Attention
- Language Models
- Contrastive Learning for NLP

What is Text

Just Another Modal

Difference to Image:

- The discreteness and meaninglessness of the components
- variable sentence length

Text Tasks

Classification

Generation

Representation Learning

💡 *The proposed solution can be independent of the problem: for example, try to solve classification with generation.*

Rule Based Viewpoint

Let's classify sports and economic texts.

The first thing that occurs in our intuition as a rule is the presence of certain words.

We can’t write all rules of the world!

Now let's try to do this automatically and approximately with machine learning …

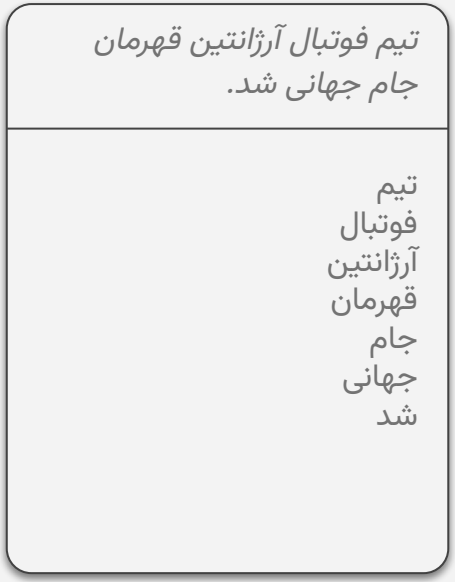
```
sports_words = ['شنا' ,'هندبال' ,'بسکتبال' ,'فوتبال']

economic_words = ['بورس' ,'تورم' ,'طلا' ,'مالیات']

for word in sports:

    if word in string:

        return 'ورزشی'

for word in economic_words:

    if word in string:

        return 'اقتصادی'
```

ML: Feature Engineering from Text

In machine learning, we must first be able to extract features from the data (textual data here).

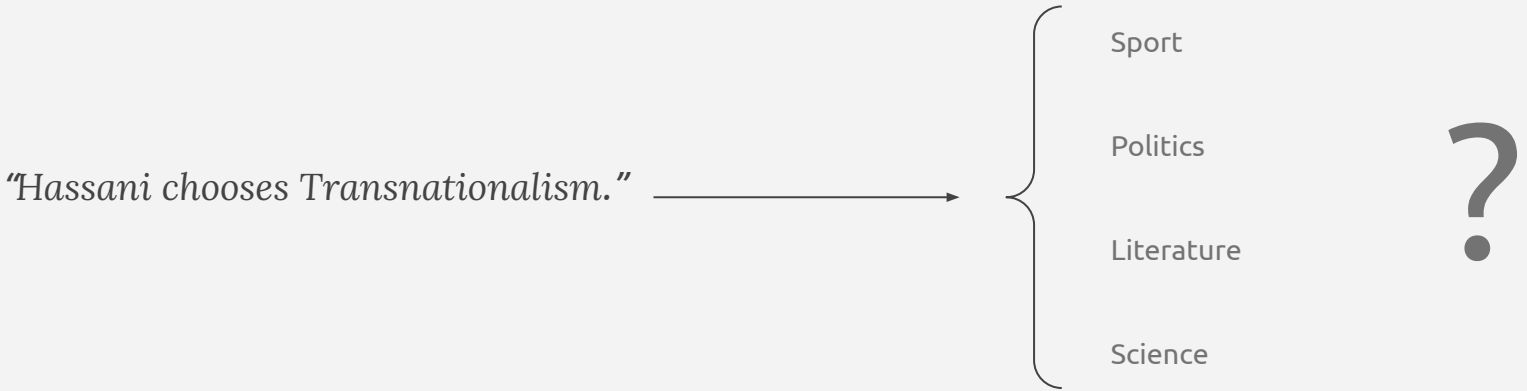
Feature extraction: Extraction of Semantic Units: Tokenization

Then we apply machine learning models such as naive-bayes or logistic regression or neural network on these features



Feature Engineering from Text
Presence, Naive Bayes

Feature Engineering from Text Presence, Why We need Subword?



Feature Engineering from Text
Presence, Why We need embedding?

Question: Given this two classes predict the class for given sample:

Class 1: Men usually cry alone.

Class 2: Women are meticulous in details.

Sample: Every day, the queen asks herself why the situation is like this.

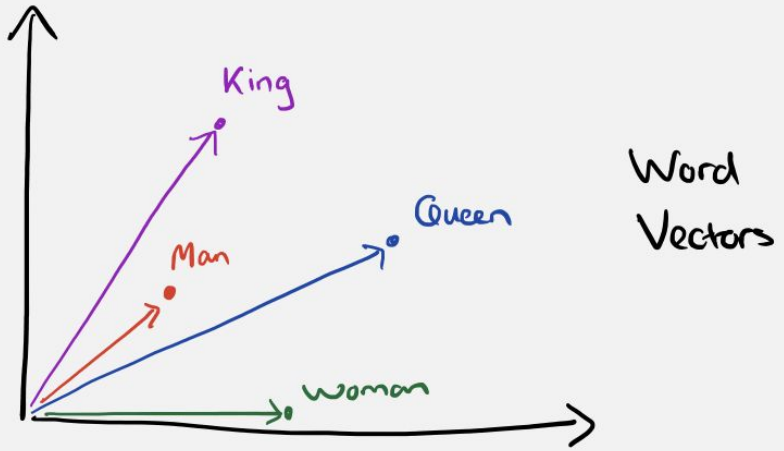
Class 1 or Class 2?

Feature Engineering from Text Presence, Why We need embedding?

So far we have proceeded by assuming the presence or absence of each word in the sentence as a feature.

If we have 30,000 words in the language, then we have to keep a vector with size of 30000 element for each word.

This is while some words are similar and related in meaning.



Feature Engineering from Text
Why We need embedding in context?

Non-Contextualized word embeddings:
Embedding(شیر) = f(شیر)

آن یکی شیر است که آدم می خورد

آن یکی شیر است که آدم می درد

Feature Engineering from Text
Why We need embedding in context?

Non-Contextualized word embeddings: Embedding(شیر) = f(شیر)

آن یکی شیر است که آدم می خورد

آن یکی شیر است که آدم می درد

Contextualized word embeddings: Embedding(شیر) = f(شیر | آن یکی شیر است که آدم می‌خورد)

Contextualized word embeddings: Embedding(شیر) = f(شیر | آن یکی شیر است که آدم می‌درد)

Before RNNs:

Embedding (شیر است که آدم را می‌خورد) =

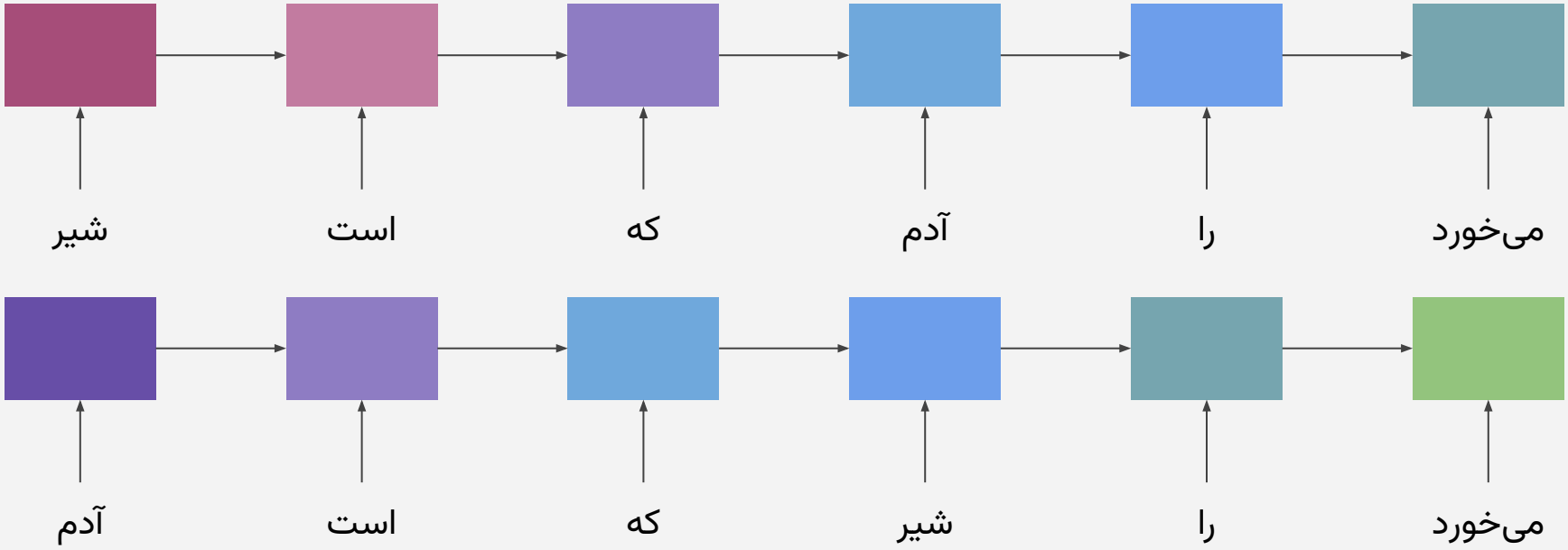
Average(e(شیر), e(است), e(که), e(آدم), e(را), e(می‌خورد))

Embedding (آدم است که شیر را می‌خورد) =

Average(e(آدم), e(است), e(که), e(شیر), e(را), e(می‌خورد))

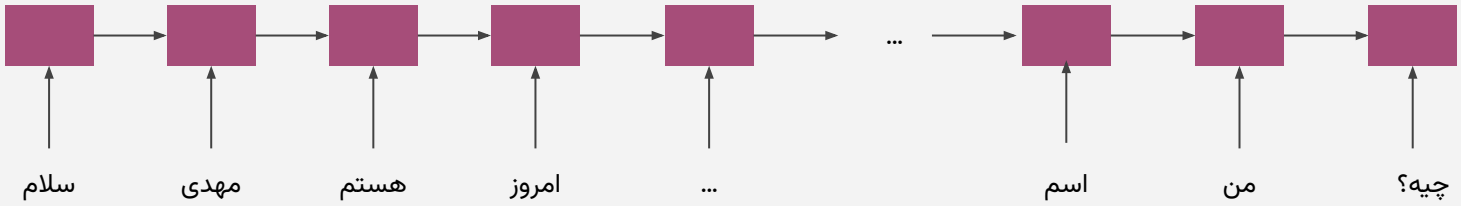
Deep Models: RNNS: The Role of State

Embedding is function of context and current word



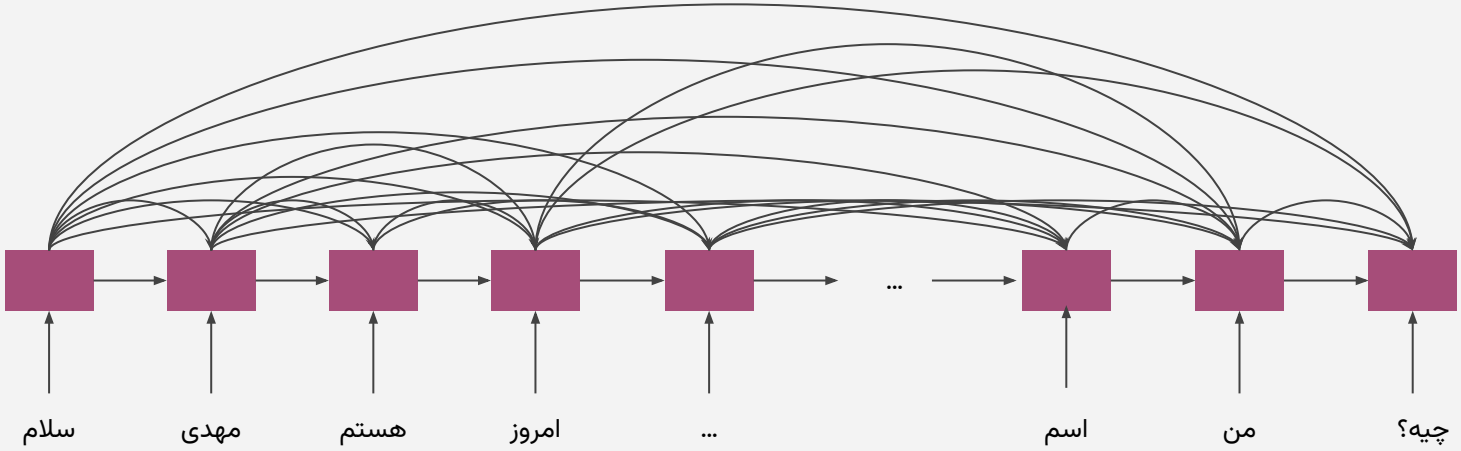
Why RNN Fails?

When the sentence becomes long, the words at the end of the sentence forget the impact of the words at the beginning.

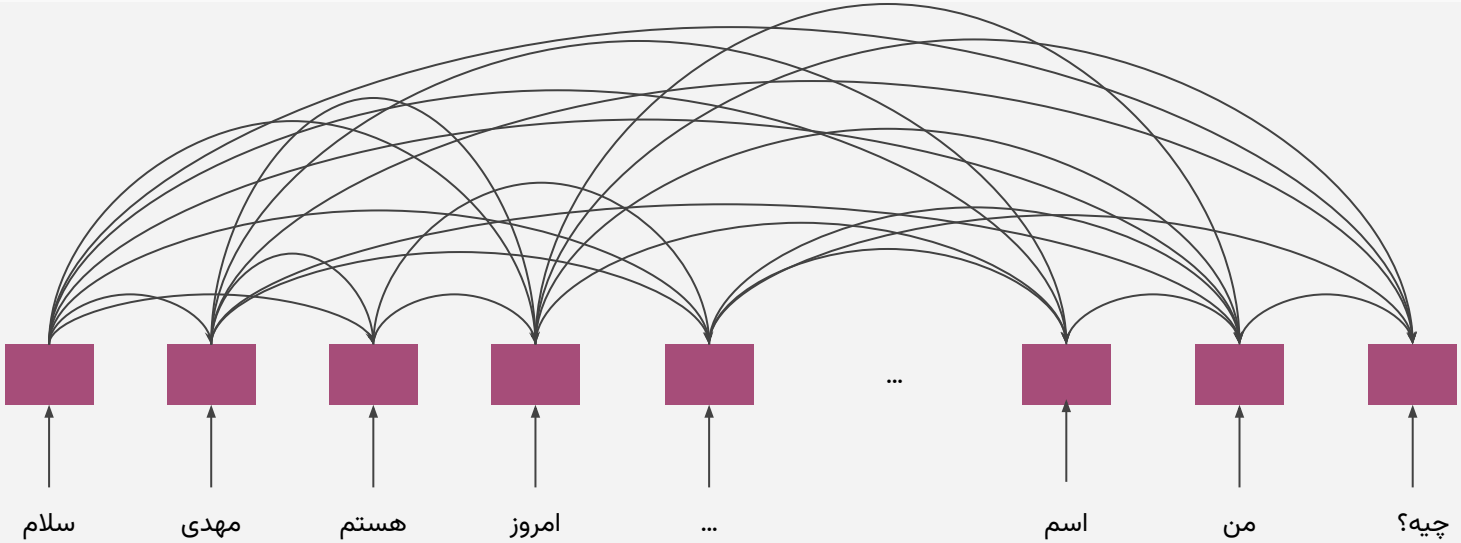


Why RNN Fails? Use Attention

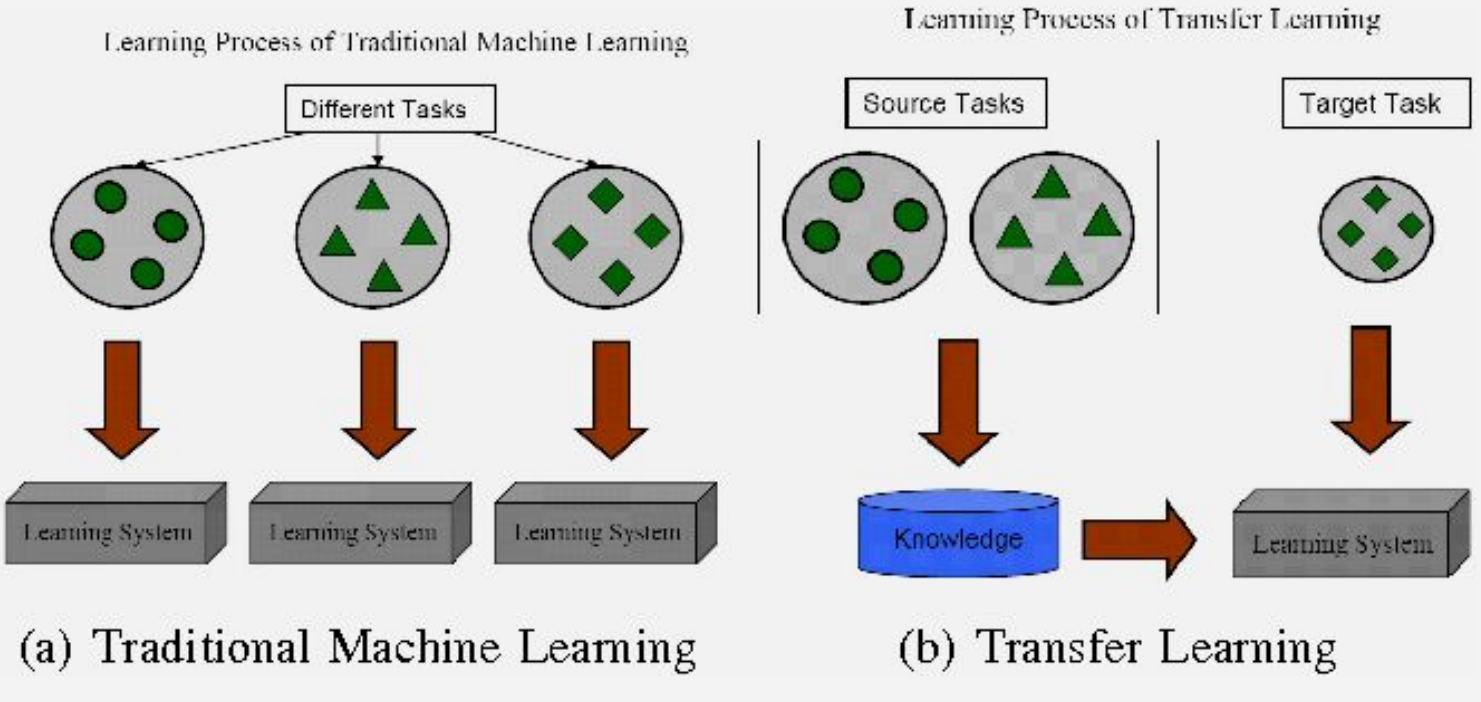
In addition to being a function of the state and the present word, embedding is also a function of paying attention to the previous words.



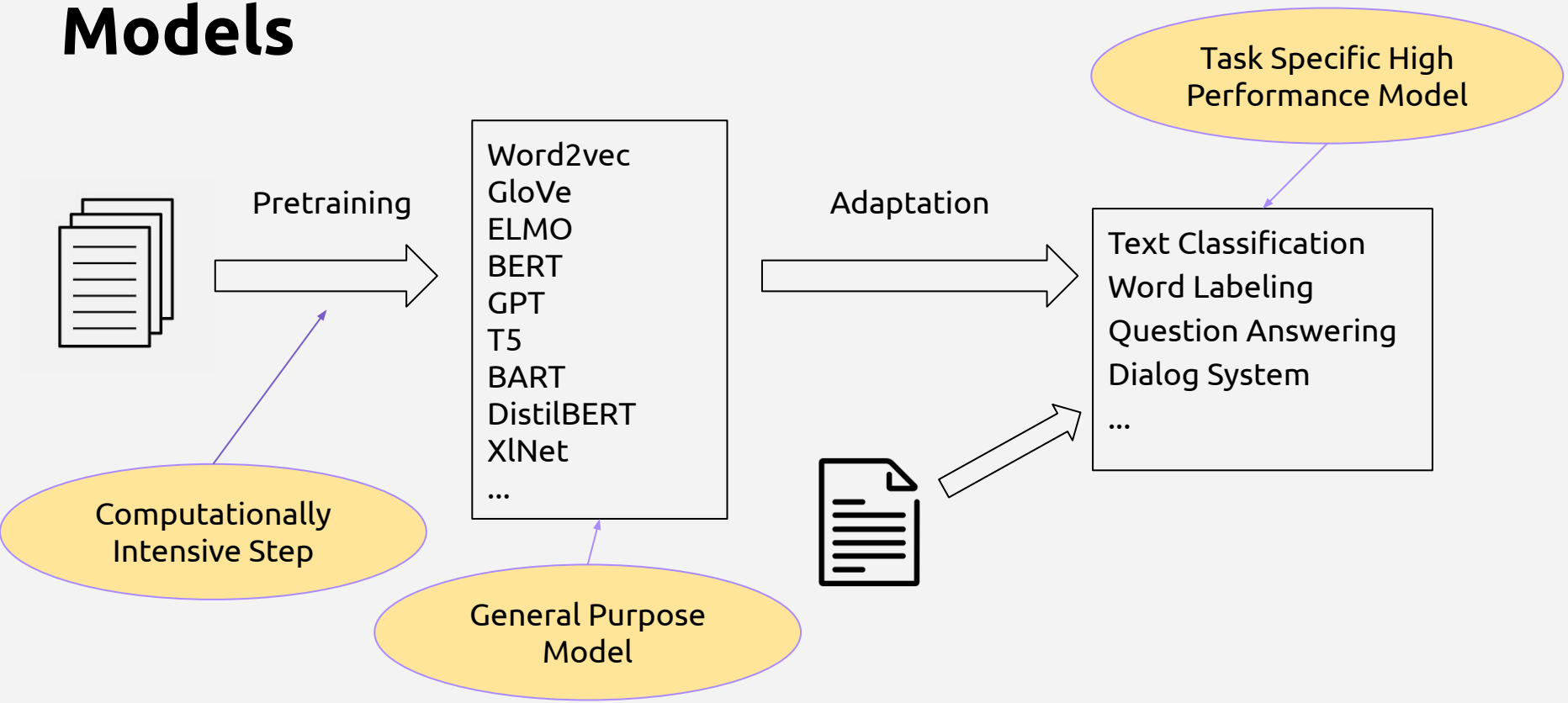
Why RNN Fails? Use Just Attention: Transformer



Pretrained Models



Transfer Learning in NLP: Language Models



Let’s see two pretraining task

Masker Language Model

In the [MASK] of God.

In the name of God.

Next Sentence Prediction

- The man went to the store.
- He bought a gallon of milk.

Label: Is Next

- The man went to the store.
- Penguins are flightless birds

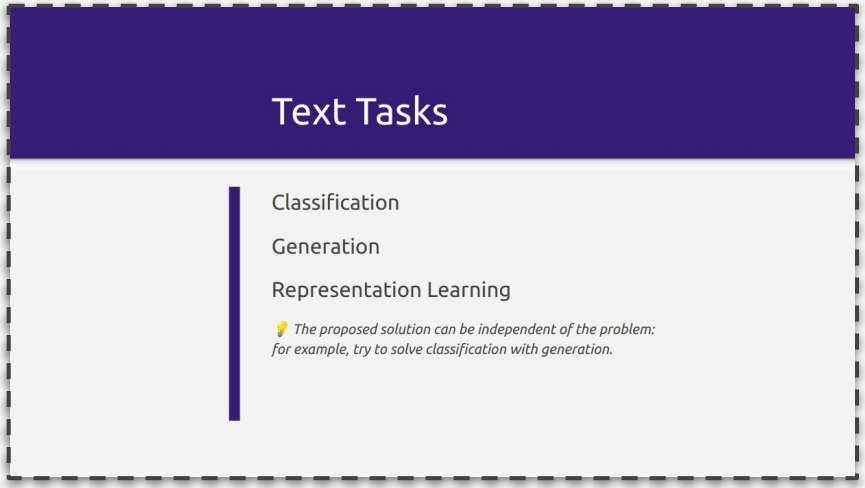
Label: Not Next

LLM Trailer!

A model without parameter tuning

Next Sentence Prediction is not Enough for Retrieval (just use [cls] is bad)

- One of the important open problems is similarity measurement for two sentences.
- For This we have two options:
  - Let's use Bert's nsp classification number. (massive pairs number)
  - Let's use the comparison of CLS representations. (low quality)



Better Representation using Contrastive Learning

Use Contrastive Learning for Represenation Learning:
Hard Negative Problem: NLI Dataset

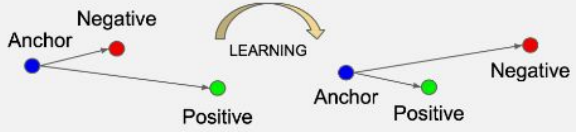
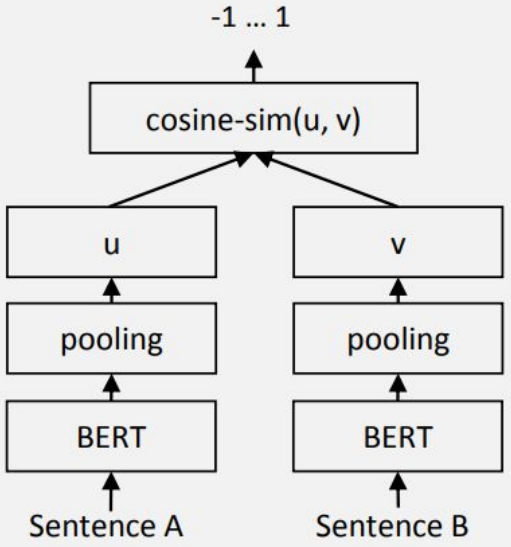


Fig. 1. Illustration of triplet loss given one positive and one negative per anchor.
(Image source: Schroff et al. 2015)



Lets see Set fit, fewshot contrastive classification

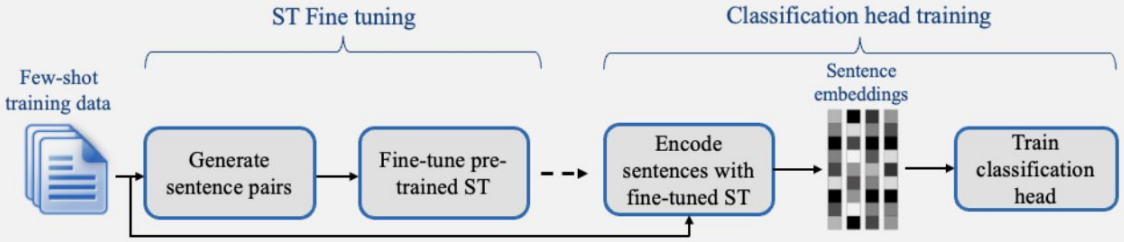


Figure 2: SETFIT 's fine-tuning and training block diagram.

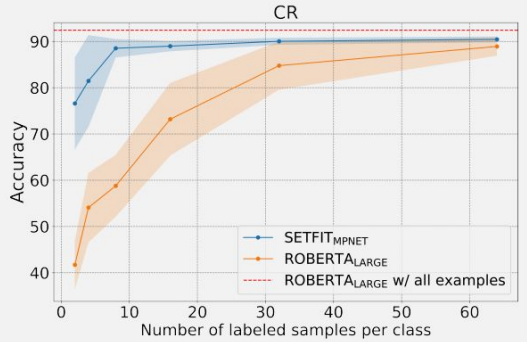


Figure 1: Compared to standard fine-tuning, SETFIT is more sample efficient and exhibits less variability when trained on a small number of labeled examples.

Thank You!