



# TRENDS: PRE-TRAINED MODELS, MULTI-TASK LEARNING, COMPRESSION TECHNIQUES AND APPLICATIONS

Luísa Coheur

# OVERVIEW

- Learning objectives
- Topics
  - Pre-trained models
  - How to use pre-trained models
    - Direct Use
    - Transfer Learning
      - Feature-based Transfer Learning
      - Fine-tuning
  - Parameter-efficient fine-tuning methods
  - Multi-task learning
  - Compressing Techniques
  - Applications
- key takeaways
- Suggested readings

# LEARNING OBJECTIVES

## LEARNING OBJECTIVES

- After this class, students should be able to define pre-trained models and explain different ways to leverage them, from simple Prompting to Transfer Learning.
  - In particular, students should be able to explain and apply Feature-based Transfer Learning and Fine-tuning.
- Additionally, students should be able to explain the concept of Multi-task Learning and how it can be conducted.
- Furthermore, students should be able to describe well-known compression techniques and NLP tasks.

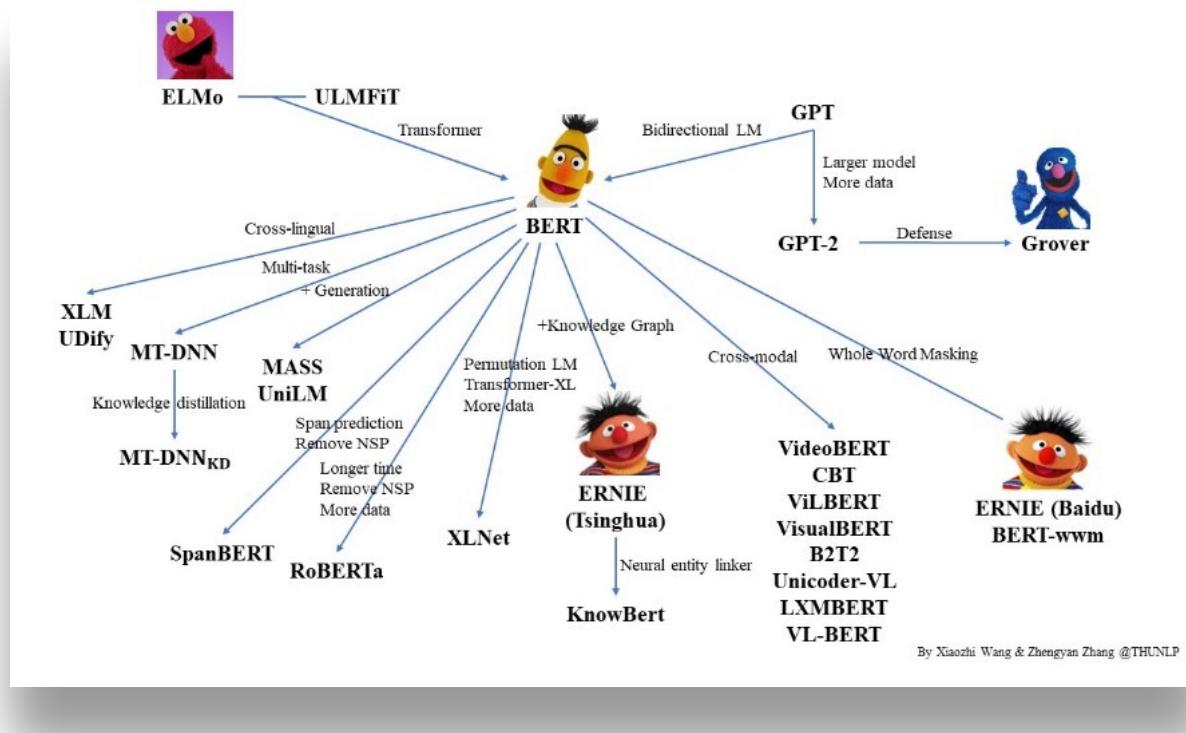
# TOPICS

# OVERVIEW

- Learning objectives
- Topics
  - Pre-trained models
  - How to use pre-trained models
    - Direct Use
    - Transfer Learning
      - Feature-based Transfer Learning
      - Fine-tuning
  - Parameter-efficient fine-tuning methods
  - Multi-task learning
  - Compressing Techniques
  - Applications
- key takeaways
- Suggested readings

# PRE-TRAINED MODELS

- Pre-trained models are machine learning models that have been previously trained on large datasets and saved for future use
  - Examples:
    - BERT and GPT (Generative Pre-trained Transformer)



## PRE-TRAINED MODELS

- Pre-trained models are then used in different specific tasks
  - Idea:
    - You learn your native language. It take years, but finally, you level is excellent (pre-trained model). Future use: you find a job as a salesman in the pharmaceutical industry. You need to acquire new, specific vocabulary ← REMEMBER?

From your colleagues (2024):

<https://www.youtube.com/watch?v=spwSbuSG6c0>

## BY THE WAY... ARE LLMs PRE-TRAINED MODELS?

- YES – an LLM is a type of pre-trained model
  - trained on massive amounts of text data with the objective of predicting the next token
- And NO – LLMs are more than just pre-trained models
  - They are also fine-tuned with techniques like Instruction Tuning or Reinforcement Learning from Human Feedback (RLHF), and they are ready to perform many tasks
  - They are usually used via prompting rather than retraining

# OVERVIEW

- Learning objectives
- Topics
  - Pre-trained models
  - How to use pre-trained models
    - Direct Use
    - Transfer Learning
      - Feature-based Transfer Learning
      - Fine-tuning
    - Parameter-efficient fine-tuning methods
    - Multi-task learning
    - Compressing Techniques
    - Applications
  - key takeaways
  - Suggested readings

# HOW TO USE PRE-TRAINED MODELS (DIRECT USE)

- The direct use is usually via **inference** – the internal process the model uses to generate an output from a given input
  - For instance, using BERT to extract representations
- Notice that:
  - Prompts – the external act of providing a Natural language input (the “prompt”) to the model – is usually not applied to simple pre-trained models (just to LLMs)

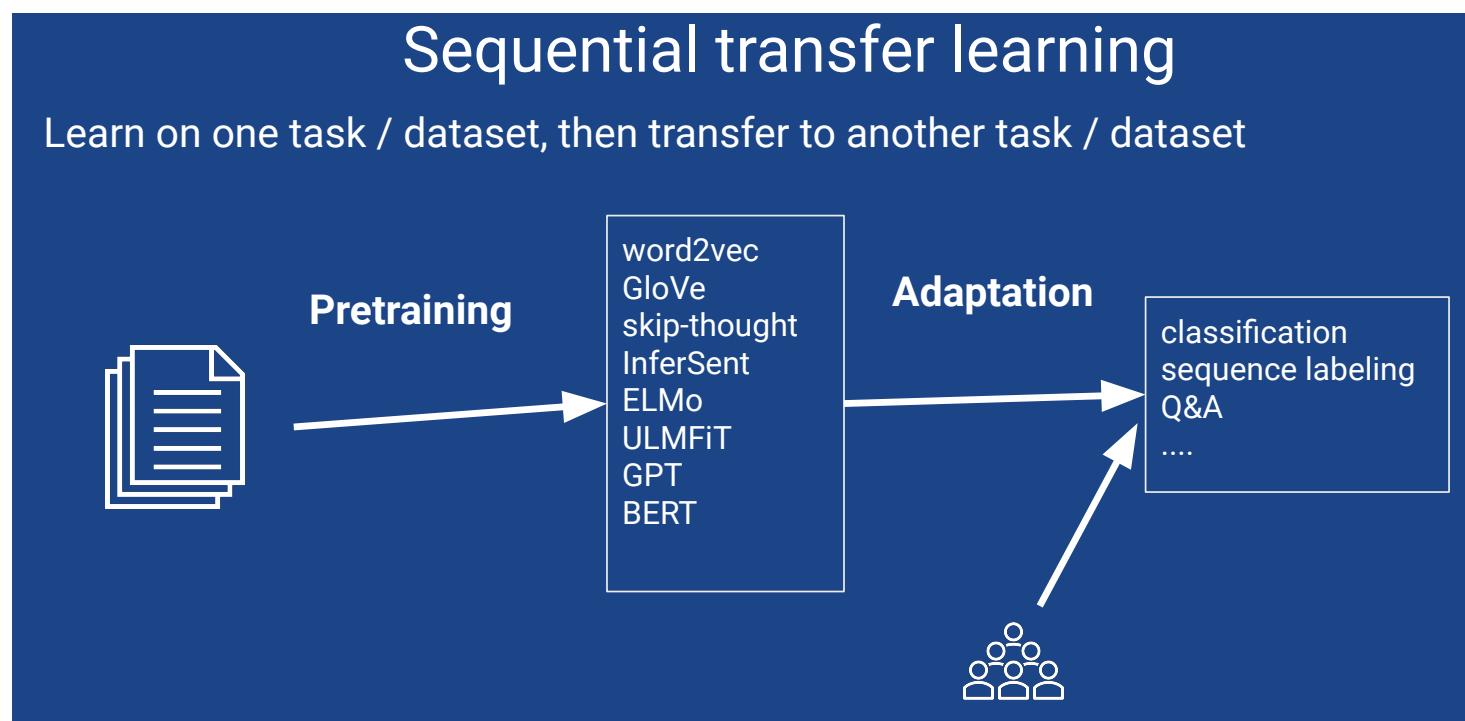
**We will talk about prompt engineering in the next class**

# OVERVIEW

- Learning objectives
- Topics
  - Pre-trained models
  - How to use pre-trained models
    - Direct Use
    - Transfer Learning
      - Feature-based Transfer Learning
      - Fine-tuning
  - Parameter-efficient fine-tuning methods
  - Multi-task learning
  - Compressing Techniques
  - Applications
- key takeaways
- Suggested readings

# TRANSFER LEARNING

- Transfer learning is a machine learning technique in which a pre-trained model, developed for a particular task, is reused as the starting point for a model on a second task



# TRANSFER LEARNING

- Some types of Transfer Learning
  - Feature-Based Transfer Learning: the features learned from the source domain/task are used as a starting point for the target domain/task
  - Fine-tuning: the pre-trained model's weights are adjusted (the model is re-trained on a variety of concrete examples) to better fit the target domain/task data

# OVERVIEW

- Learning objectives
- Topics
  - Pre-trained models
  - How to use pre-trained models
    - Direct Use
    - Transfer Learning
      - Feature-based Transfer Learning
      - Fine-tuning
  - Parameter-efficient fine-tuning methods
  - Multi-task learning
  - Compressing Techniques
  - Applications
- key takeaways
- Suggested readings

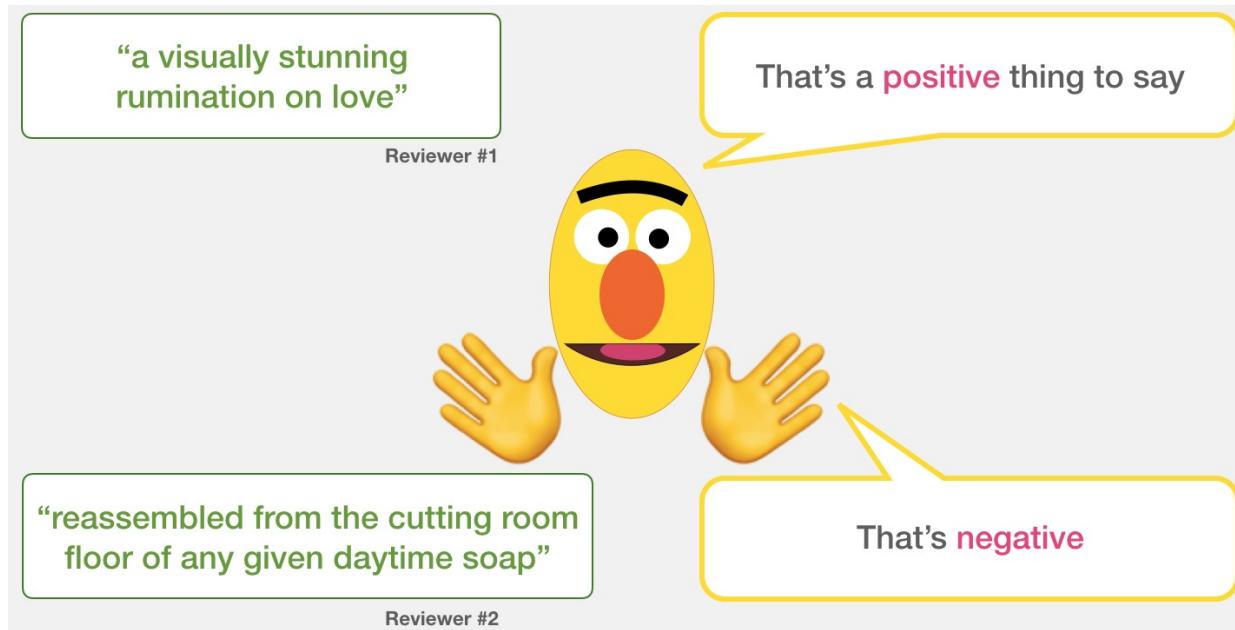
## FEATURE-BASED TRANSFER LEARNING

- The next slides are from from “A Visual Guide to USING BERT for the First Time” (by Jay Alammar)



# FEATURE-BASED TRANSFER LEARNING

Sentences (data from the “Stanford Sentiment Treebank”)	label
a stirring , funny and finally transporting re imagining of beauty and the beast and 1930s horror films	1
apparently reassembled from the cutting room floor of any given daytime soap	0
they presume their audience won't sit still for a sociology lesson	0
this is a visually stunning rumination on love , memory , history and the war between art and commerce	1
jonathan parker 's bartleby should have been the be all end all of the modern office anomie films	1



**Sentiment  
Analysis is an  
NLP TASK! (we  
already saw this)**

# FEATURE-BASED TRANSFER LEARNING

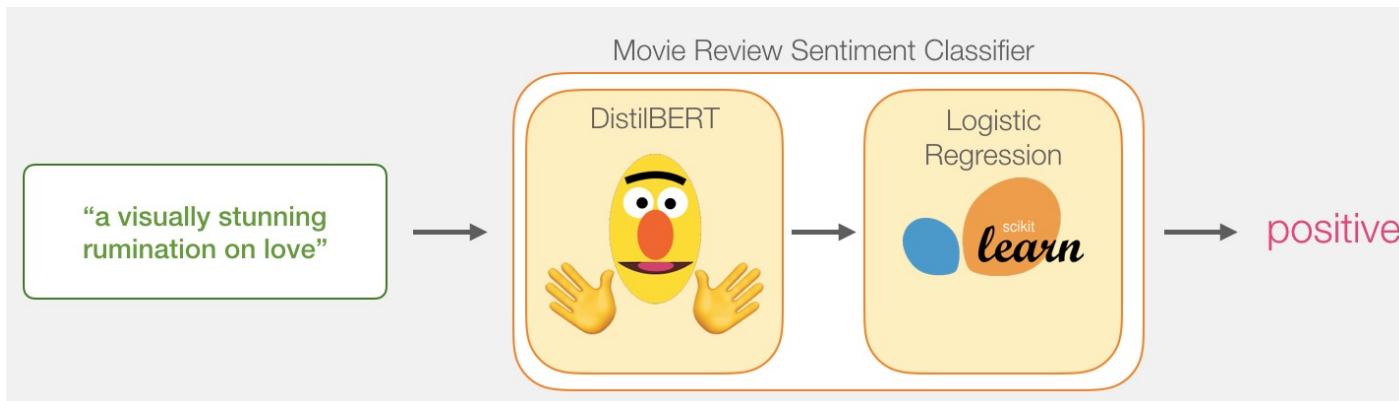
## CLASSIC APPROACH



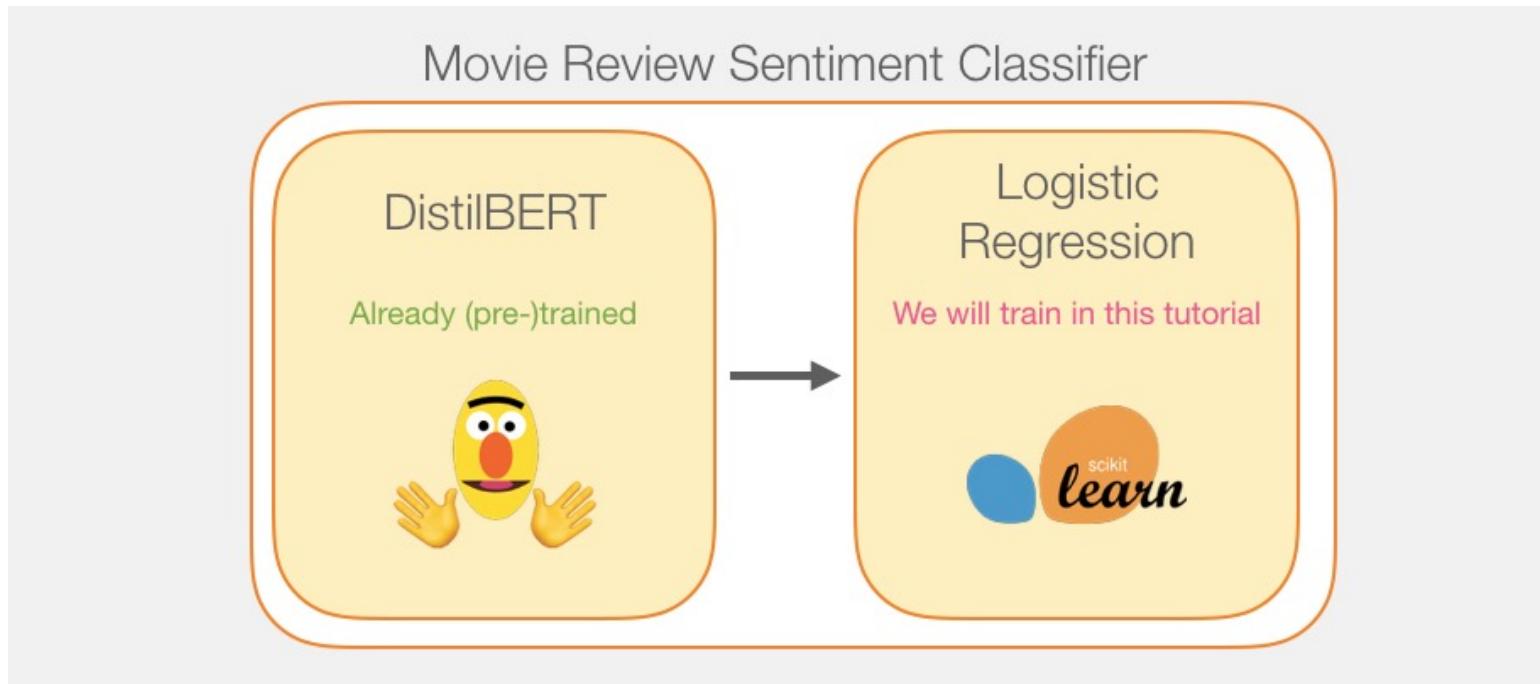
Example:  
LOGISTIC REGRESSION  
(or other classifier)

...

## NOW

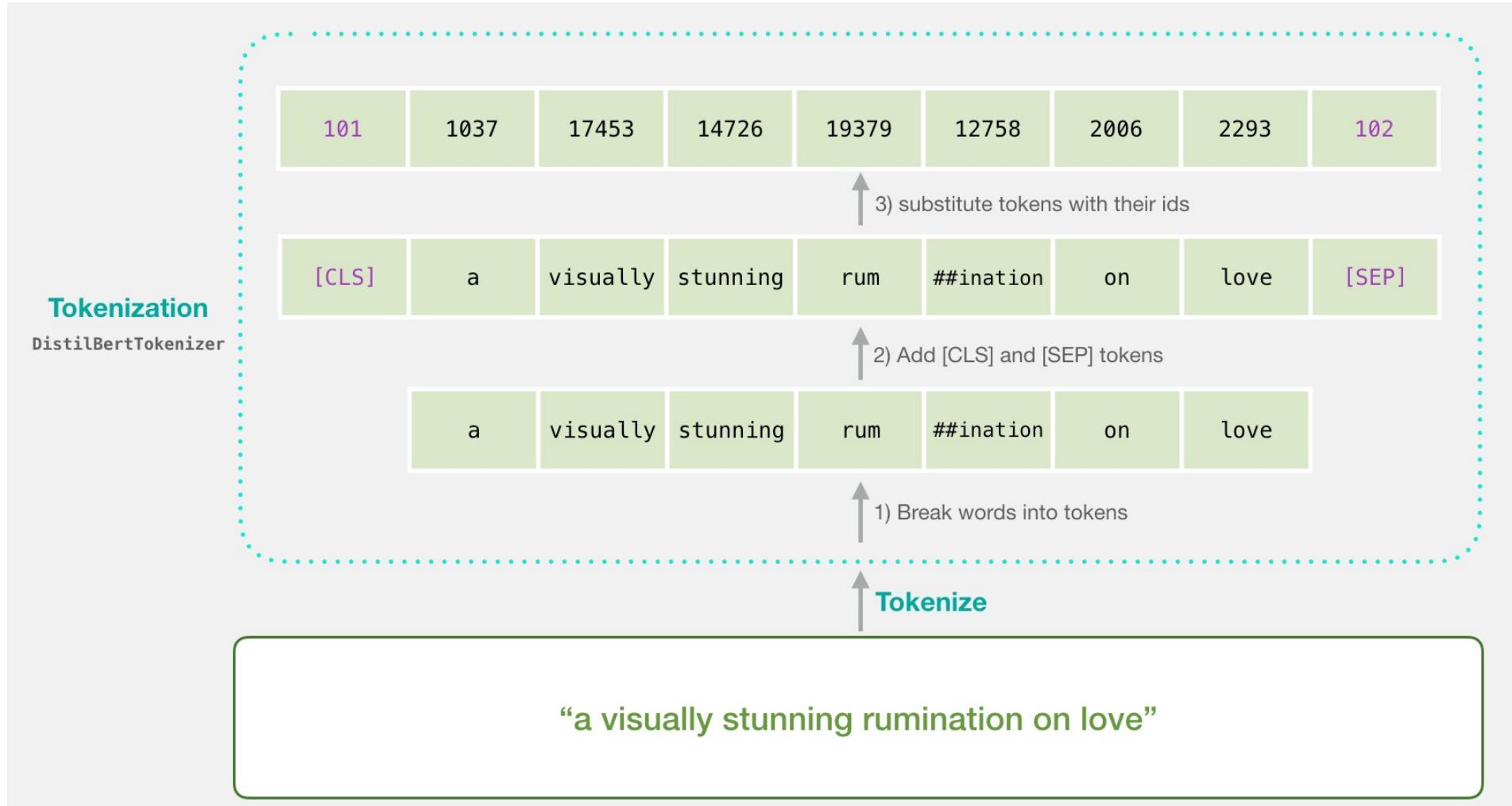


# FEATURE-BASED TRANSFER LEARNING

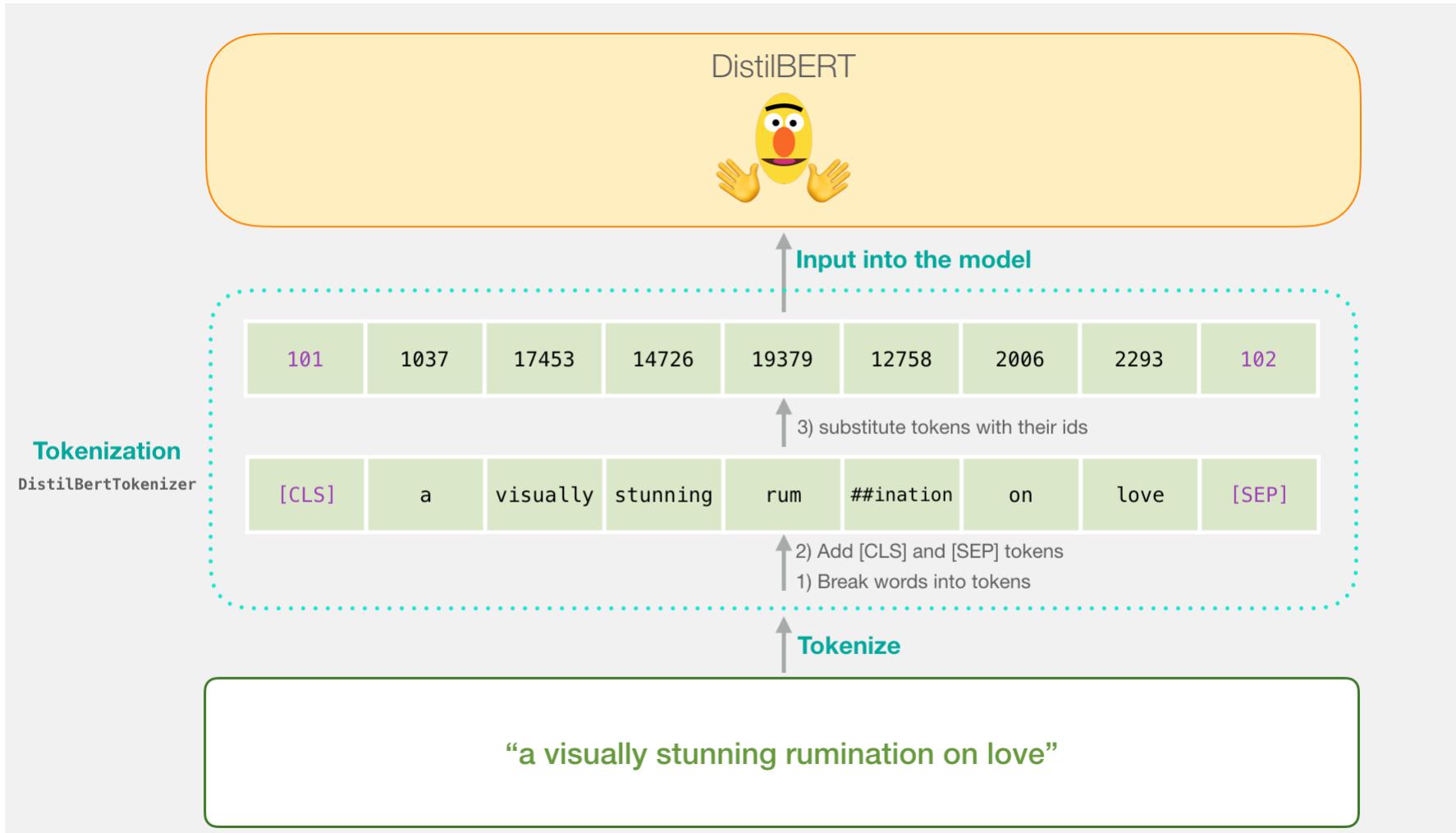


Note that DistilBERT is trained on the English language, but it was not trained to perform sentence classification.

# FEATURE-BASED TRANSFER LEARNING

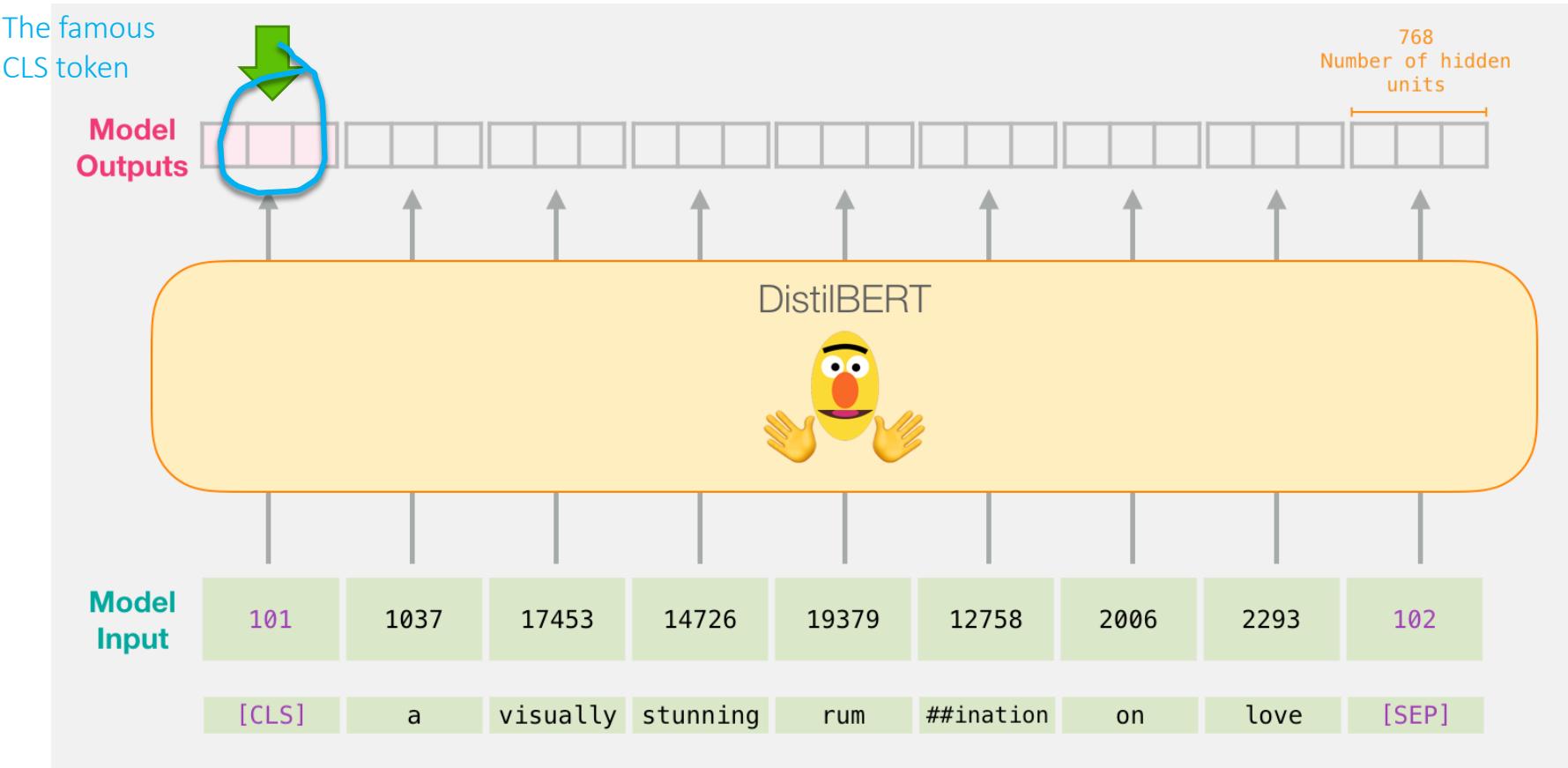


# FEATURE-BASED TRANSFER LEARNING

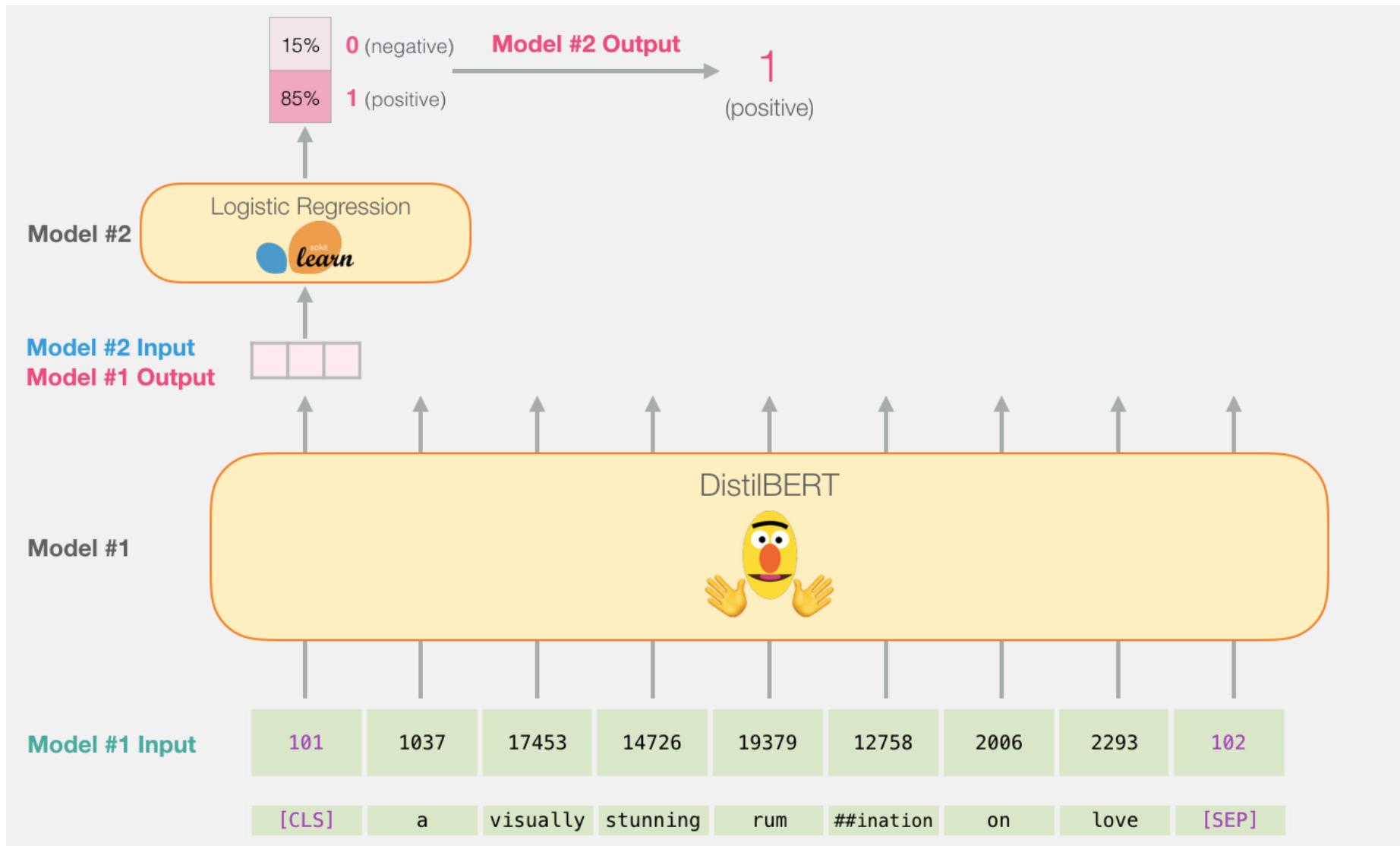


Note: Each token has an ID. Each ID is associated with a raw in an embedding matrix in BERT

# FEATURE-BASED TRANSFER LEARNING



# FEATURE-BASED TRANSFER LEARNING



# OVERVIEW

- Learning objectives
- Topics
  - Pre-trained models
  - How to use pre-trained models
    - Direct Use
    - Transfer Learning
      - Feature-based Transfer Learning
      - **Fine-tuning**
  - Parameter-efficient fine-tuning methods
  - Multi-task learning
  - Compressing Techniques
  - Applications
- key takeaways
- Suggested readings

## FINE-TUNING

- Fine-tuning is specific form of [transfer learning](#) in which the parameters of a model are adjusted to fit with specific observations
  - Involves taking a pre-trained model and continuing the training process on a new, typically smaller, dataset (usually, new layer(s) is(are) added to fit the targets specified in the new domain/task).

# FINE-TUNING TECHNIQUES

- We can:
  - Train/freeze the entire pre-trained model and/or just train some layers while freezing others
  - Freeze in the beginning some layers and then train the whole architecture
  - ...

# FINE-TUNING TECHNIQUES

- In theory, you will not need much data, and it will not be (super) expensive...
- But:
  - fine-tuning these models for a particular task can be (very) costly, both in terms of time and computational resources



- Parameter-efficient fine-tuning methods are needed!

# OVERVIEW

- Learning objectives
- Topics
  - Pre-trained models
  - How to use pre-trained models
    - Direct Use
    - Transfer Learning
      - Feature-based Transfer Learning
      - Fine-tuning
  - Parameter-efficient fine-tuning methods
  - Multi-task learning
  - Compressing Techniques
  - Applications
- key takeaways
- Suggested readings

# PARAMETER-EFFICIENT FINE-TUNING METHODS

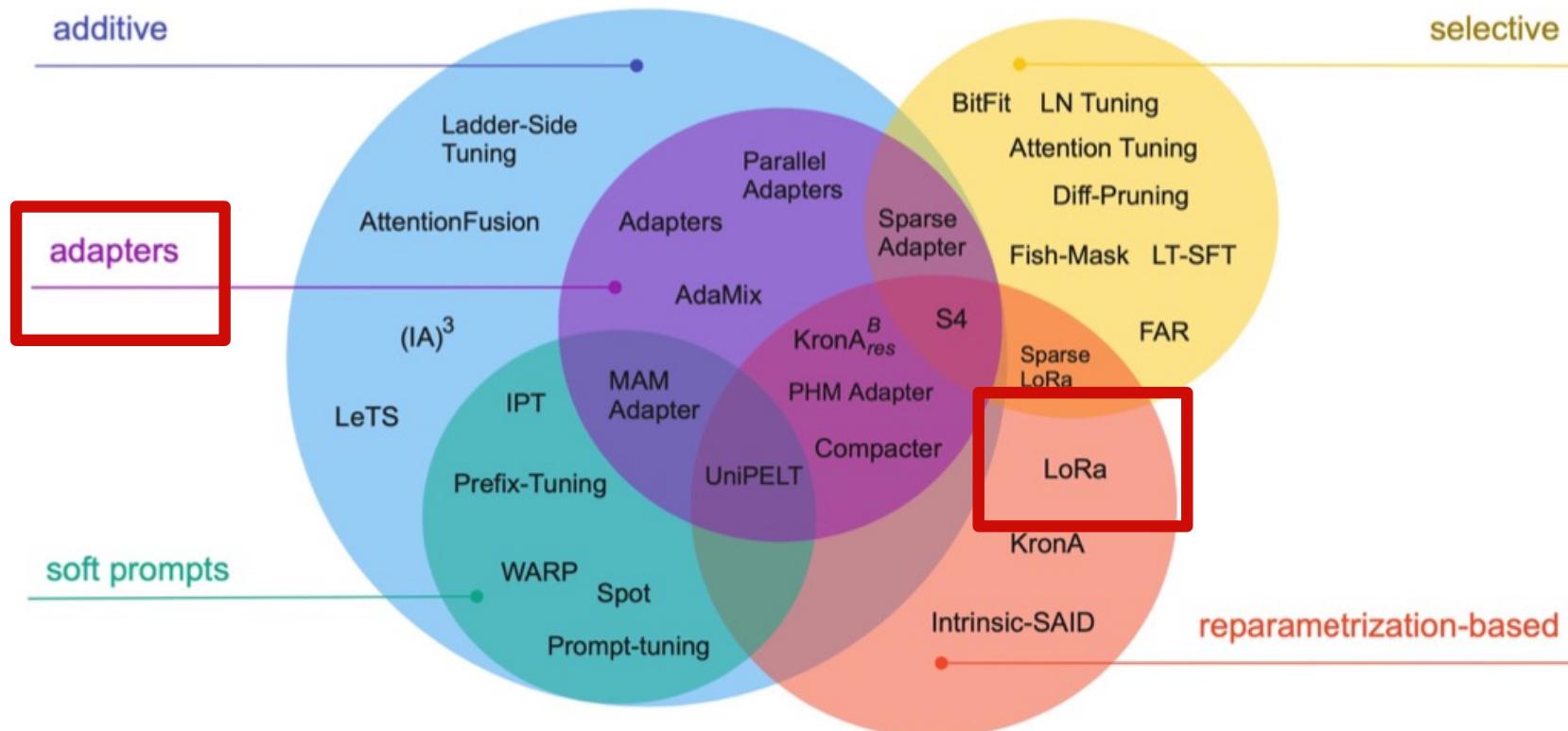
- Parameter-efficient fine-tuning (**PEFT**) focuses on fine-tuning pre-trained models efficiently by **training only a small number of extra parameters**

Small number  
of **\*\*\*EXTRA\*\*\***  
parameters?  
Hum...



# PARAMETER-EFFICIENT FINE-TUNING METHODS

- There is a world of parameter-efficient fine-tuning techniques!



# ADAPTERS

- Adapters are small, trainable layers inserted into a pre-trained model's architecture
  - These layers are designed to capture task-specific information while keeping the pre-trained model's weights frozen
    - thus, no need for extensive retraining of the entire model

# ADAPTERS

- How Adapters work:
  - Start with a pre-trained model that has learned general features from a large dataset
  - Insert adapter modules (weights randomly initialized) into specific layers of the pre-trained model
    - These adapters have a small number of parameters compared to the full model
  - Only the adapter layers are trained on the new task's dataset, while the rest of the model remains unchanged
    - This reduces the computational cost
  - During inference, the pre-trained model, along with the adapter layers, is used to make predictions on new data
    - They are “loaded” to augment the frozen base model

# ADAPTERS

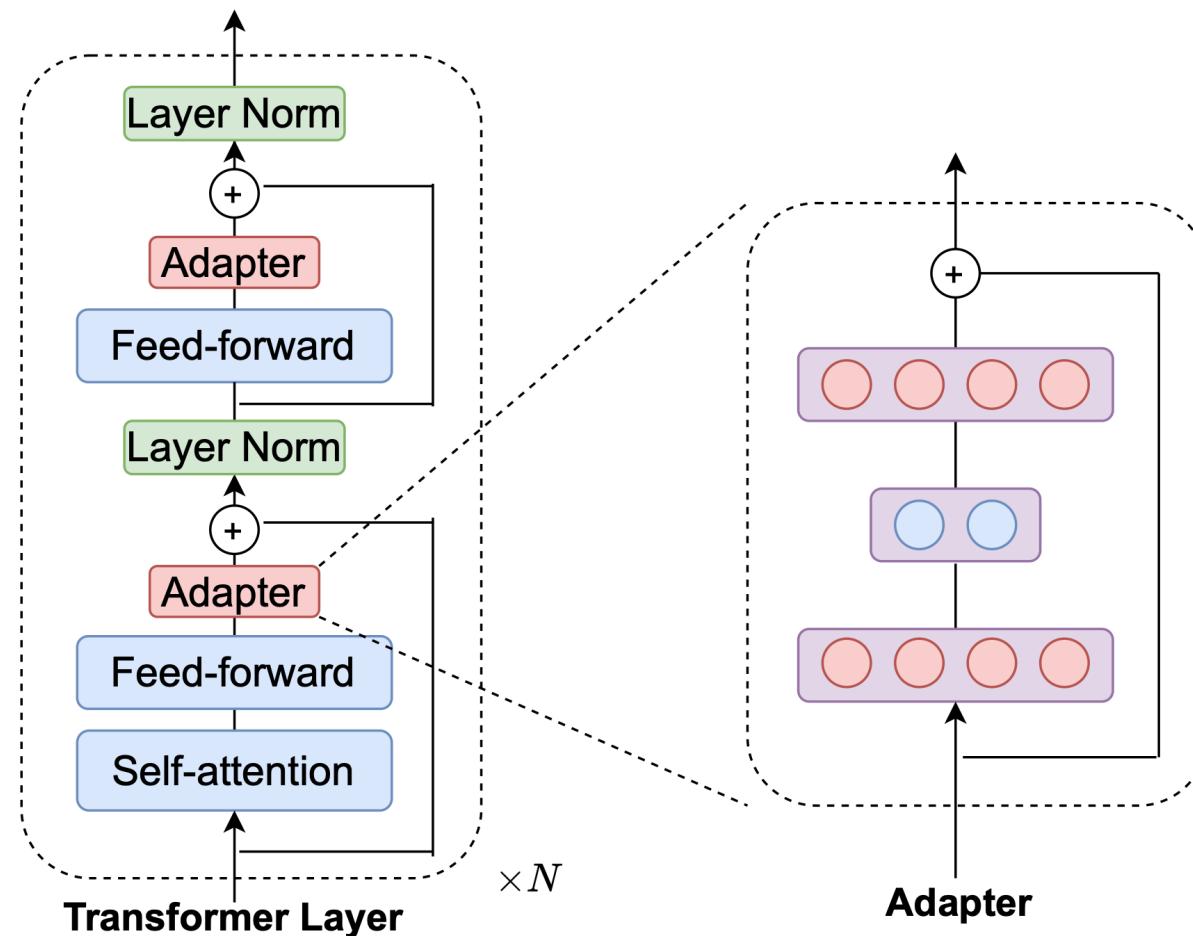


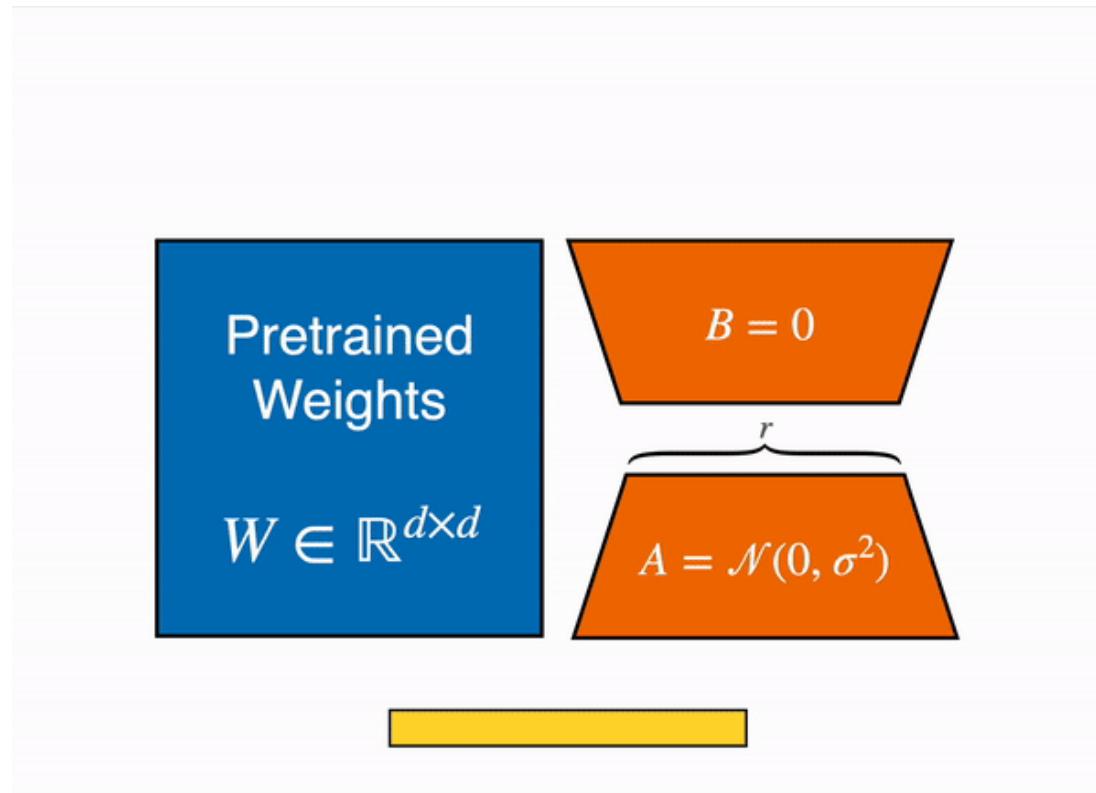
Image from <https://aclanthology.org/2021.acl-long.172/>

# ADAPTERS

- In some scenarios, only 3% of task-specific parameters are needed to almost match the results of the 100% task-specific parameters used by the fully fine-tuned model
  - Sometimes adapter-based tuning outperforms fine-tuning on low-resource and cross-lingual tasks (he et al. 2021)

# LoRA

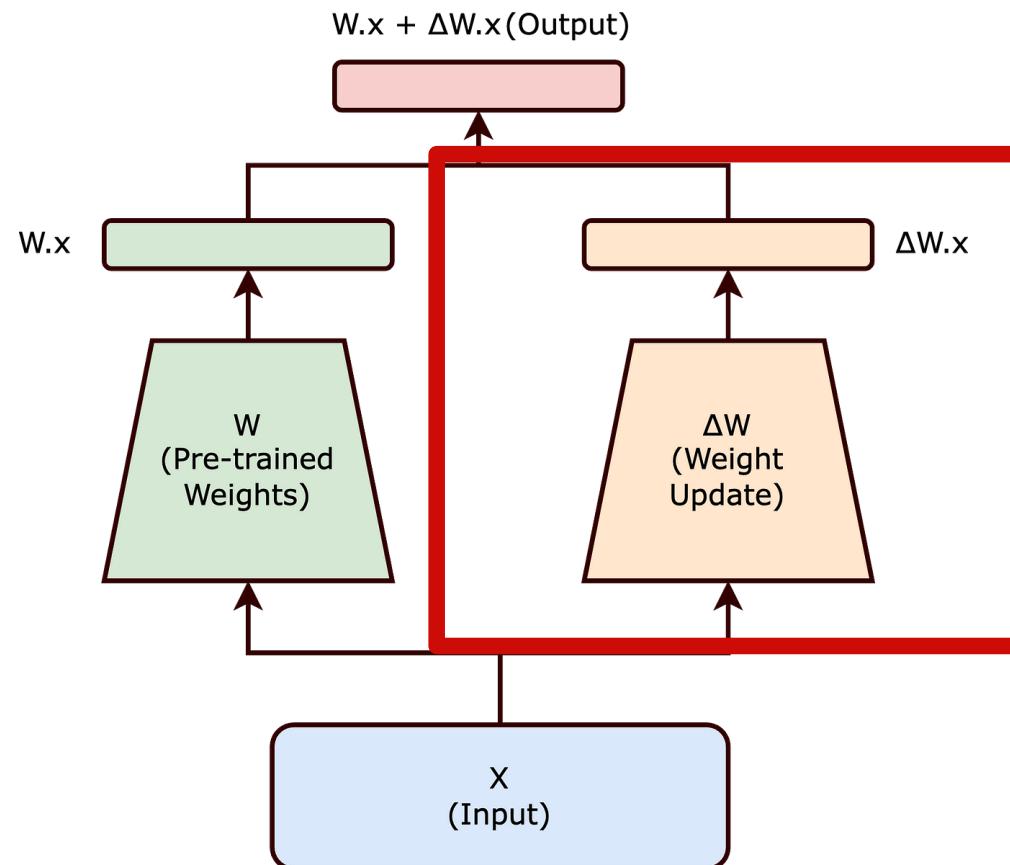
- Low-Rank Adaptation (LoRA) decomposes the model weight matrices using low-rank decomposition



From <https://medium.com/@manindersingh120996/understanding-low-rank-adaptation-lora-for-efficient-fine-tuning-of-large-language-models-082d223bb6db>

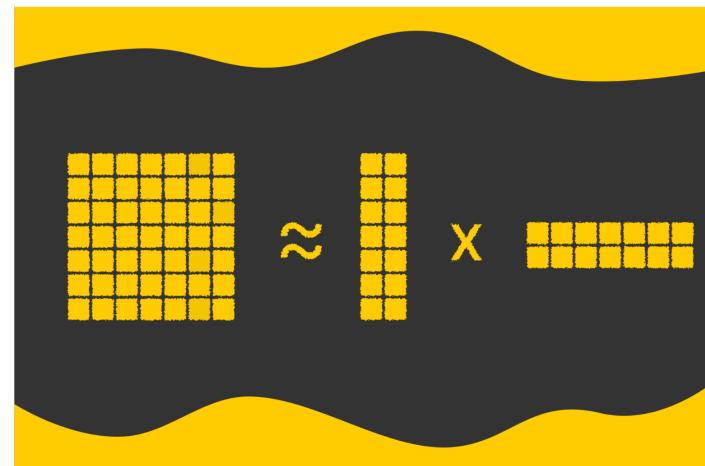
# LoRA

- $h = Wx + \Delta Wx$ 
  - $W$  are the model weights
  - $\Delta W$  its **accumulated gradient update during adaptation**



# LoRA

- $h = Wx + \Delta Wx = Wx + BAx$ 
  - $W$  are the model weights
  - $\Delta W$  its accumulated gradient update during adaptation
  - $BAx$  are the LoRA weight changes



- Question: which should be A and B dimensions?

## LoRA (SOME MATH)

- The concept of low rank
  - The rank of a matrix is given by the number of its linearly independent columns (or rows).
    - Note: It can be proven that the number of independent columns (known as column rank) is always equal to the number of independent rows (called row rank).
  - Example:

$$\text{rank} \left( \begin{bmatrix} 1 & 2 & 3 \\ 1 & 1 & 0 \\ 3 & 0 & 4 \end{bmatrix} \right) = 3$$

## LoRA (SOME MATH)

- But:

$$\text{rank} \begin{pmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \end{pmatrix} = 1$$

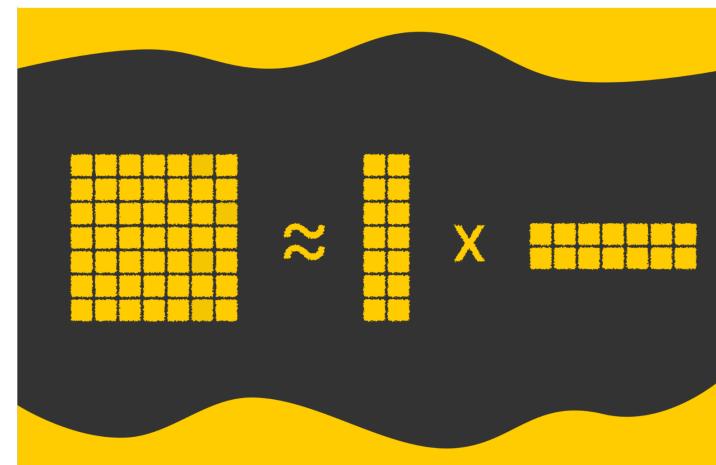
- the second row is just the first row multiplied by 2.
  - This means **the rows are not linearly independent**
  - The rank of this matrix is 1
  - So, this is a **low-rank matrix**

## LoRA (SOME MATH)

- Rank decomposition of the matrix  $\Delta W$  is the factorization of the form

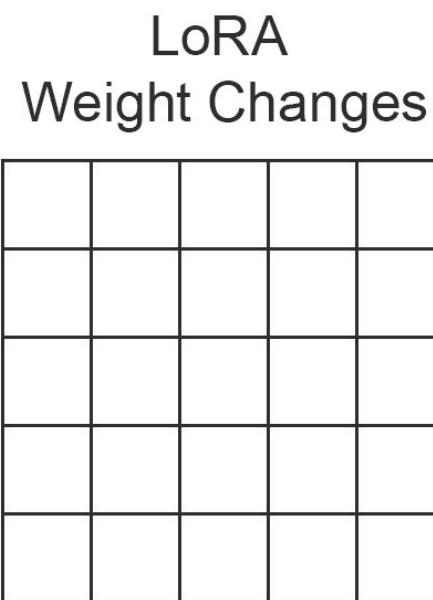
$$\Delta W = A B$$

where  $\text{rank}(\Delta W) = r$

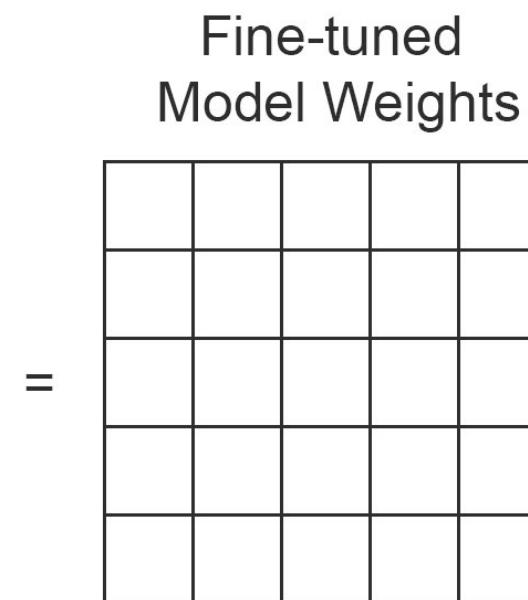


Note: It can be proven that every (finite) matrix has a rank decomposition. Techniques like SVD (Singular Value Decomposition – remember?) can be used to construct such a decomposition

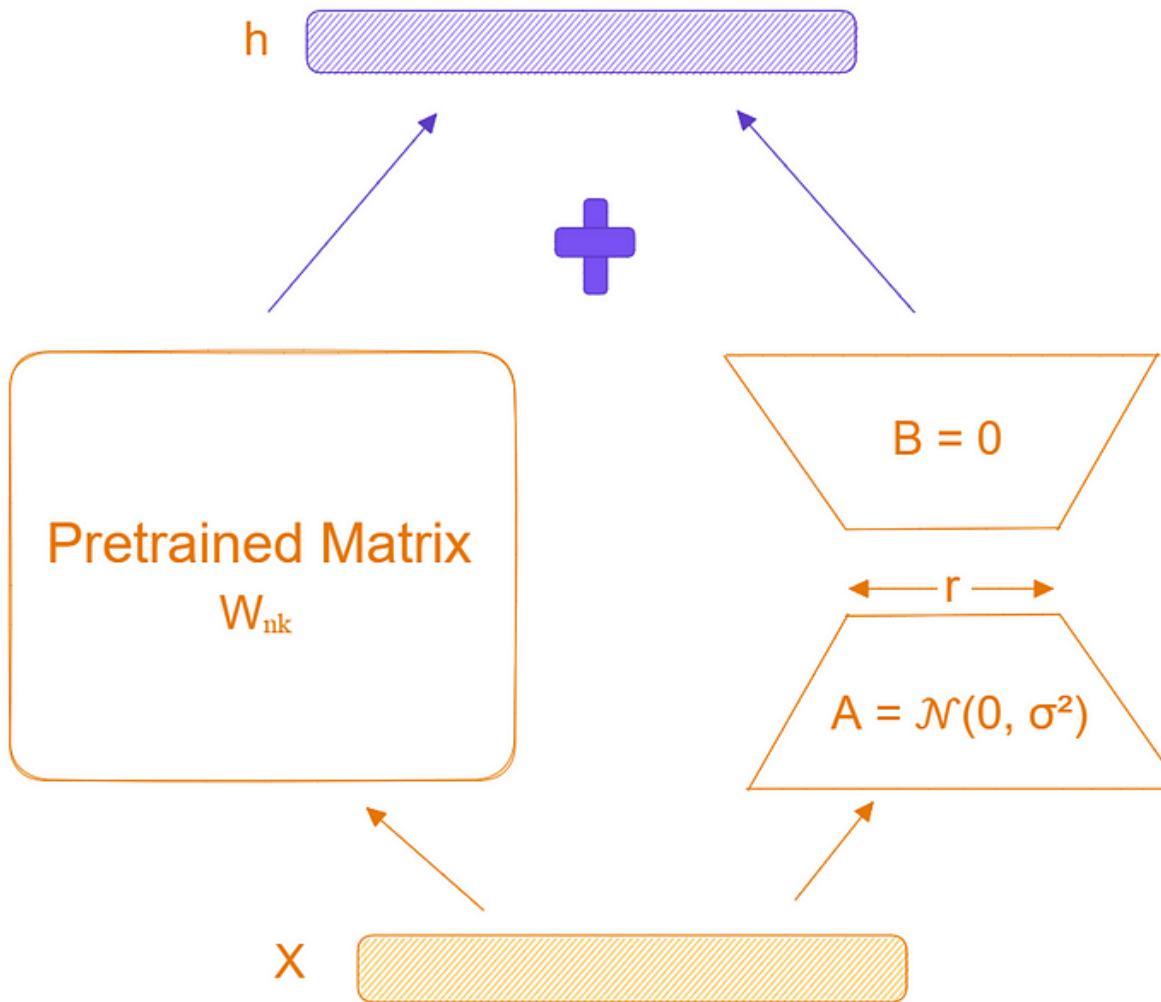
# LoRA (MAIN IDEA)



+ Model Weights



# LoRA (MAIN IDEA)



By the way:

- $A = \mathcal{N}(0, \sigma^2)$  refers to a random matrix  $A$  that is initialized from a normal distribution (or Gaussian distribution) with a mean of 0 and a variance of  $\sigma^2$ .
- $B = 0$  means that the matrix  $B$  is initialized to all zeros.

# LoRA

- Why does this make sense?
  - Large models are trained to capture the general representation of their domain. These models capture a variety of features which allow them to be used for diverse tasks with reasonable zero-shot accuracy
    - However, when adapting such a model to a specific task or dataset, only a few features need to be emphasized or re-learnt. This means that the update matrix ( $\Delta W$ ) can be a low-rank matrix.

From <https://www.ml6.eu/blogpost/low-rank-adaptation-a-technical-deep-dive>

# LoRA

- For a model like GPT-3, trainable parameters are reduced by 10000 times.

From your colleagues (LoRA and Fine-Tuning):

[https://www.youtube.com/watch?v=rcMYq\\_c5bMg](https://www.youtube.com/watch?v=rcMYq_c5bMg)

<https://youtu.be/fRUPzzlEHU4>

<https://www.youtube.com/watch?v=liNX8JsBzxA>

Generated by DALL-E



# OVERVIEW

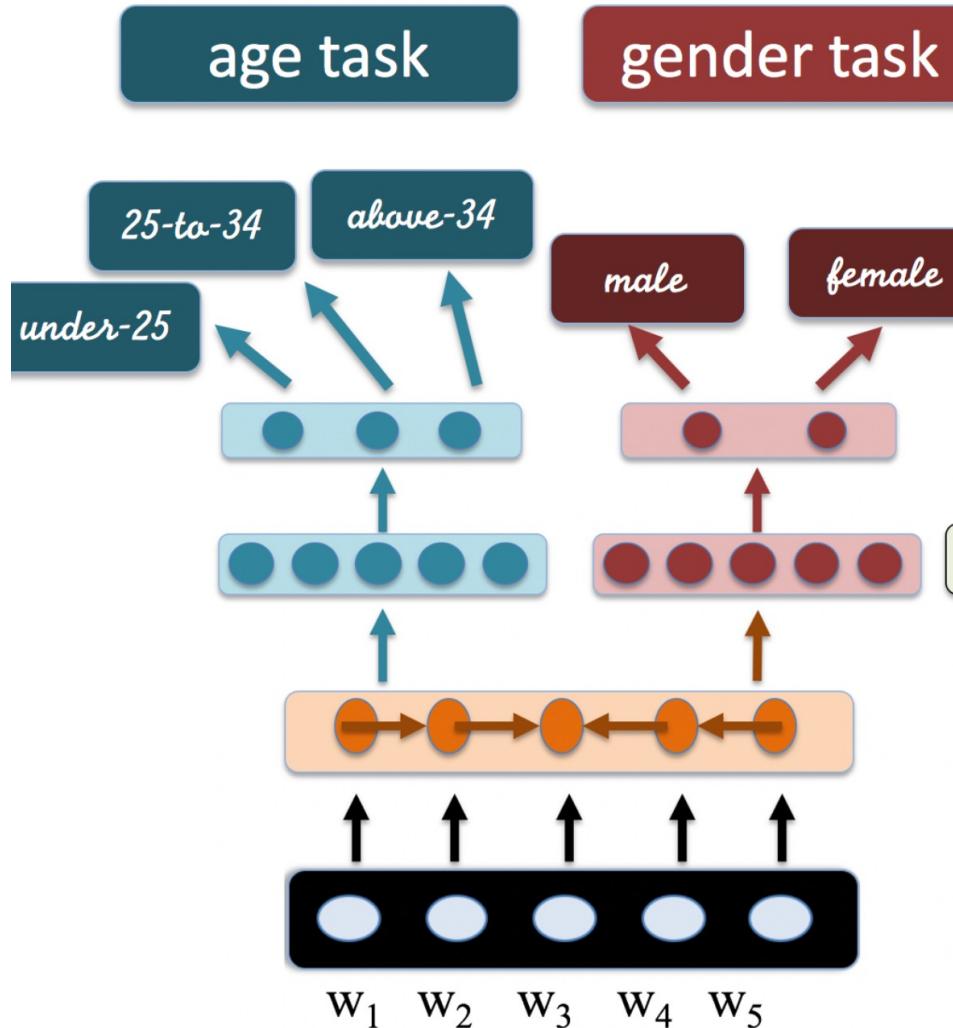
- Learning objectives
- Topics
  - Pre-trained models
  - How to use pre-trained models
    - Direct Use
    - Transfer Learning
      - Feature-based Transfer Learning
      - Fine-tuning
  - Parameter-efficient fine-tuning methods
  - Multi-task learning
  - Compressing Techniques
  - Applications
- key takeaways
- Suggested readings

## MULTI-TASK LEARNING

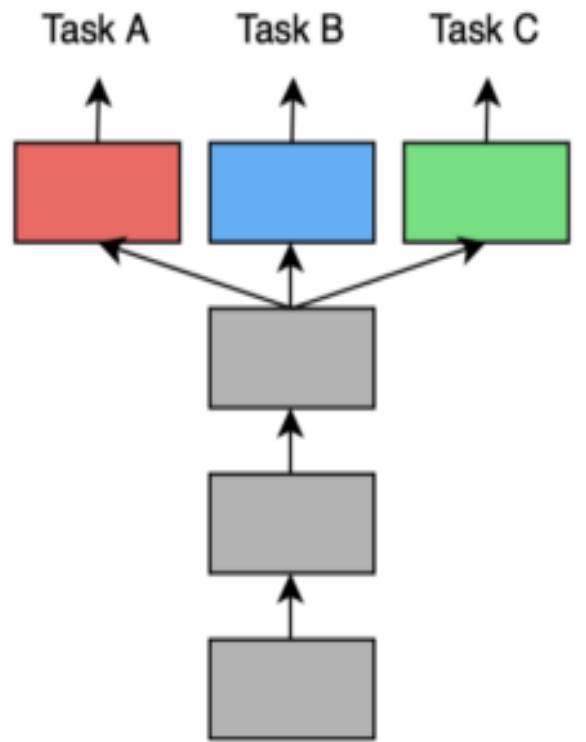
- Multi-task learning is a subfield of machine learning in which multiple learning tasks are solved at the same time, leveraging the shared knowledge across these tasks to improve overall performance

# MULTI-TASK LEARNING

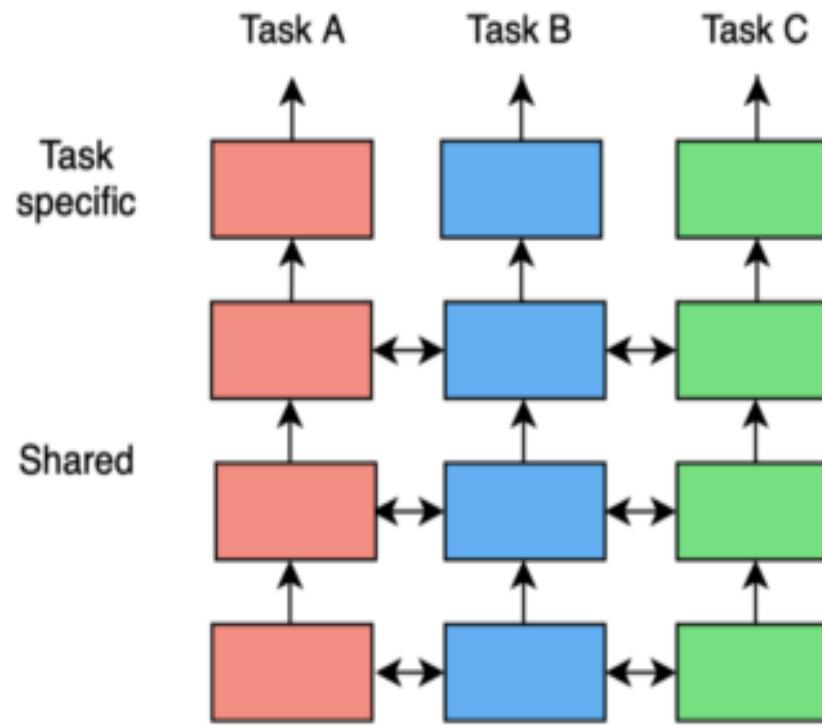
- Multi-task learning models:
  - learn a shared representation of the input data that captures the common features and patterns across the tasks
  - learn task-specific information. Each task has its own set of parameters that capture the unique characteristics of that task



# MULTI-TASK LEARNING



(a) Hard parameter sharing



(b) Soft parameter sharing

## MULTI-TASK LEARNING

- As usual, besides soft/hard parameter sharing, there are many other approaches

## EXAMPLE: DECANLP TASKS

- The Natural Language Decathlon (decaNLP) is a 10-task challenge:
  - Question Answering,
  - Machine Translation,
  - Summarization,
  - Natural Language Inference,
  - Sentiment Analysis,
  - Semantic Role Labeling,
  - Relation Extraction,
  - Goal-Oriented Dialogue,
  - Semantic Parsing,
  - Common sense Reasoning

# ACTIVE LEARNING MOMENT: THINK-PAIR-SHARE



# THINK-PAIR-SHARE

- Any idea of how to train a model in all these tasks (with so many different inputs/outputs)?

## Examples

<u>Question</u>	<u>Context</u>	<u>Answer</u>
What is a major importance of Southern California in relation to California and the US?	...Southern California is a <b>major economic center</b> for the state of California and the US....	major economic center
What is the translation from English to German?	Most of the planet is ocean water.	Der Großteil der Erde ist Meerwasser
What is the summary?	Harry Potter star Daniel Radcliffe gains access to a reported £320 million fortune...	Harry Potter star Daniel Radcliffe gets £320M fortune...
Hypothesis: Product and geography are what make cream skimming work. <b>Entailment</b> , neutral, or contradiction?	Premise: Conceptually cream skimming has two basic dimensions – product and geography.	Entailment
Is this sentence <b>positive</b> or negative?	A stirring, funny and finally transporting re-imagining of Beauty and the Beast and 1930s horror film.	positive



## EXAMPLE: DECANLP TASKS

- Key Idea: one model, one format, ten tasks
  - DecaNLP reformulates all tasks as question answering:
    - Input = question + question
    - Output = answer text
  - The model just learns to “answer questions”, whether that question is about translation, sentiment, or inference.

# OVERVIEW

- Learning objectives
- Topics
  - Pre-trained models
  - How to use pre-trained models
    - Direct Use
    - Transfer Learning
      - Feature-based Transfer Learning
      - Fine-tuning
  - Parameter-efficient fine-tuning methods
  - Multi-task learning
  - Compressing Techniques
  - Applications
- key takeaways
- Suggested readings

# COMPRESSION TECHNIQUES

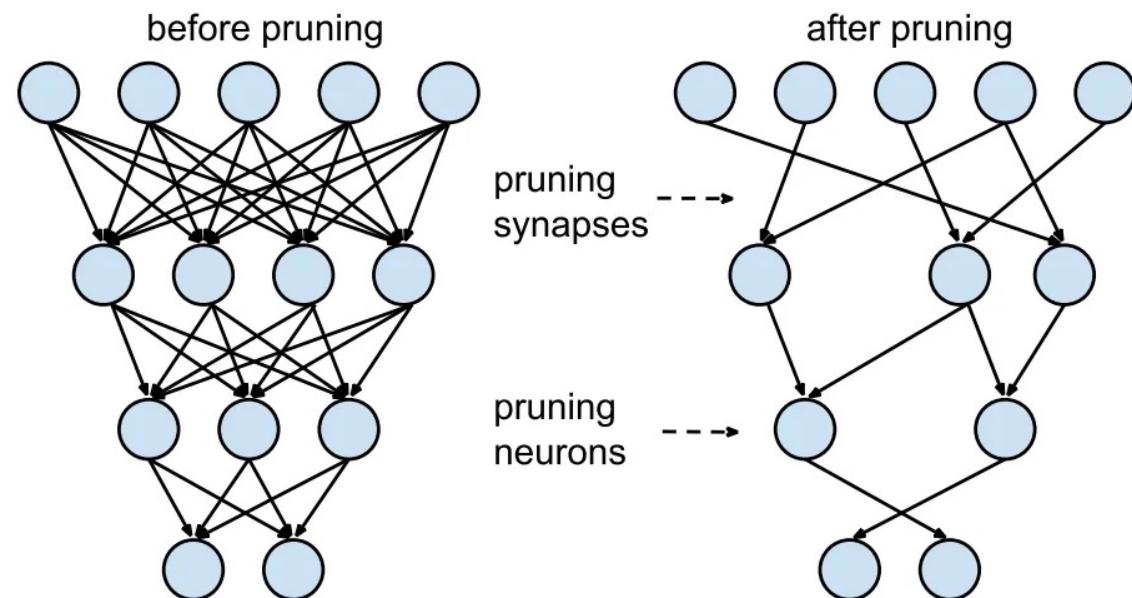
- Compression techniques are a set of methods used to reduce the size or complexity of neural network models without significantly sacrificing performance
  - Examples:
    - Pruning
    - Quantization
    - Teacher-student model (knowledge distillation)

# PRUNING

- Pruning involves removing unnecessary weights or neurons from the neural network.

- Examples:

- Prune weights: prune connections that are below some predefined thresholds
- Prune neurons
- Prune layers
- ...



# QUANTIZATION

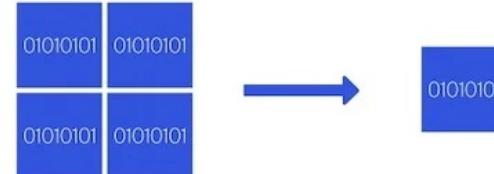
- Quantization reduces the precision of the weights and activations in the neural network, typically from 32-bit floating-point numbers to lower bit-width integers. This reduces the memory and computational requirements of the model

## Quantization

Floating point                      Integer

3452.3194      →      3452

32 bit                      8 bit



# TEACHER-STUDENT MODEL (KNOWLEDGE DISTILLATION)

- Knowledge distillation involves training a smaller “student” model to mimic the predictions of a larger “teacher” model
  - In teacher-student training, the dataset provides hard targets (a single target label) and the teacher provides soft targets (a distribution over all labels – logits are used)

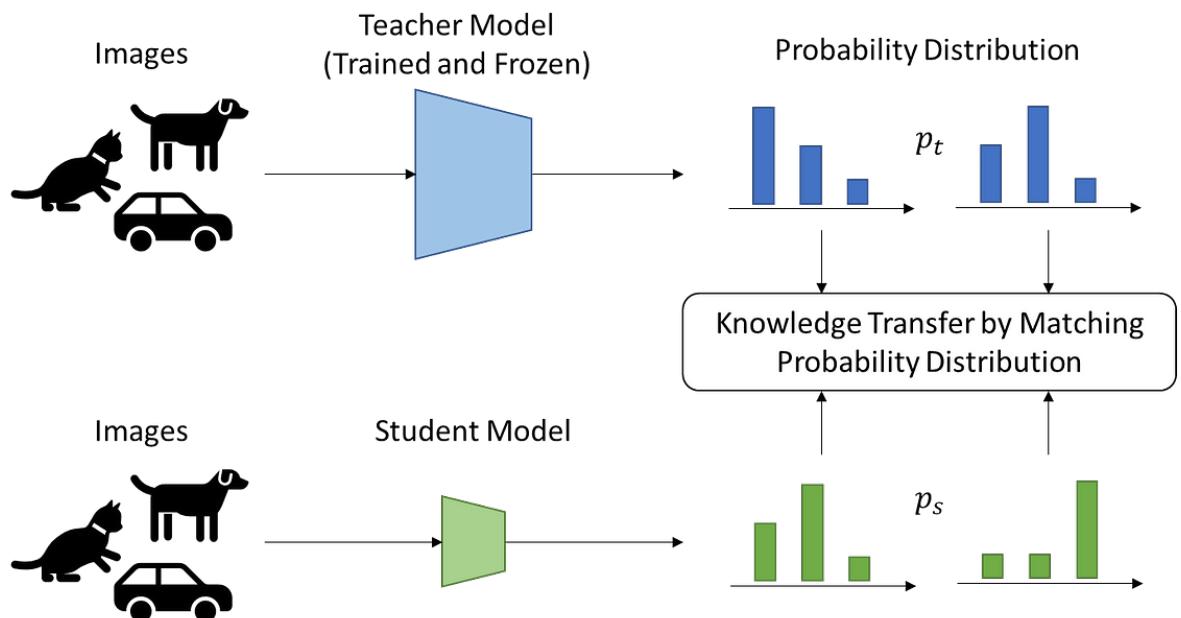
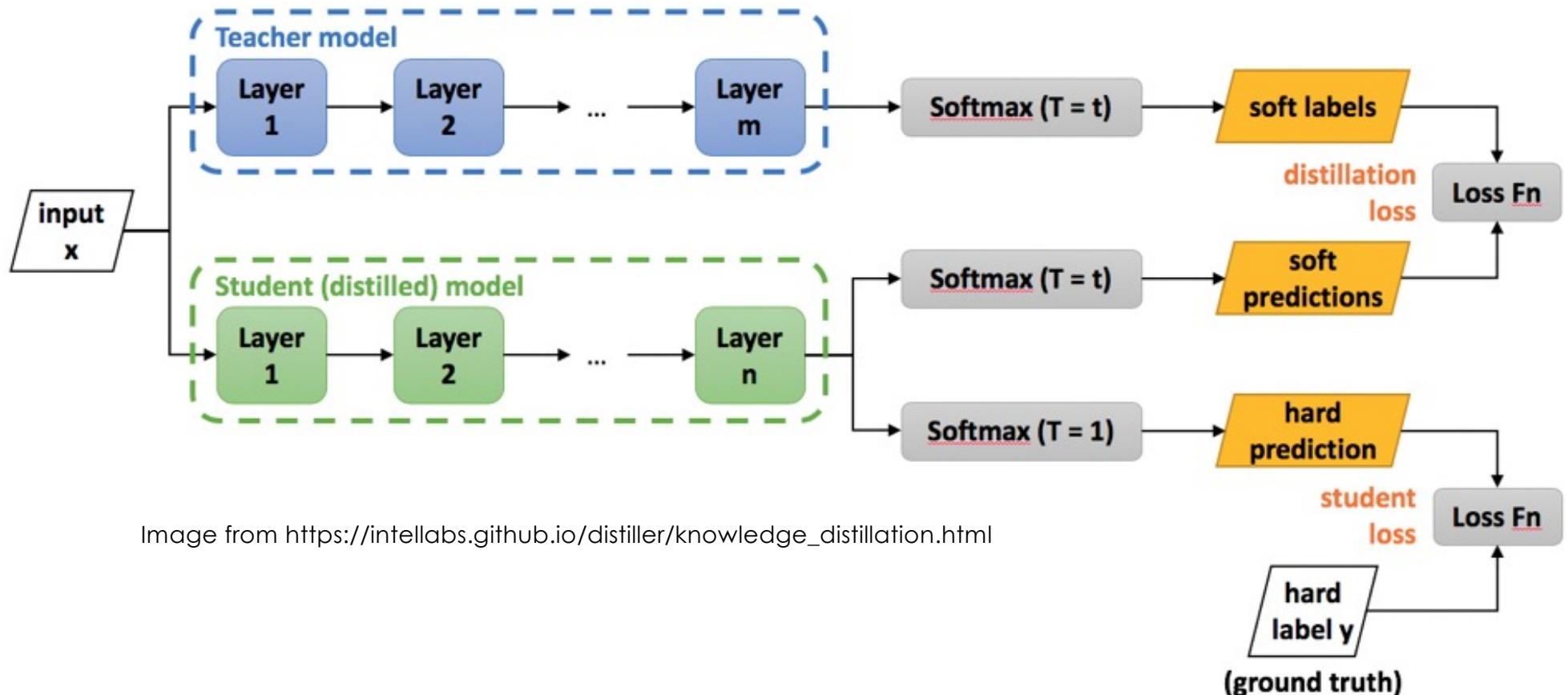


Image from <https://medium.com>

# TEACHER-STUDENT MODEL (KNOWLEDGE DISTILLATION)



T for temperature: we will talk about this in the next class

# OVERVIEW

- Learning objectives
- Topics
  - Pre-trained models
  - How to use pre-trained models
    - Direct Use
    - Transfer Learning
      - Feature-based Transfer Learning
      - Fine-tuning
  - Parameter-efficient fine-tuning methods
  - Multi-task learning
  - Compressing Techniques
  - Applications
- key takeaways
- Suggested readings

# APPLICATIONS: QUESTION/ANSWERING

- Question Answering (QA): receive a question and a context that contains information necessary to output the desired answer ← this is a new definition (in the early days, no context was provided)
  - Widely used dataset: SQuAD

## Passage Segment

...The European Parliament and the Council of the European Union have powers of amendment and veto during the legislative process...

## Question

Which governing bodies have veto power?

# APPLICATIONS: QUESTION/ANSWERING

- Interesting: this is an example of Span-based QA
  - we "only" need to find the beginning and the end of the answer

## Passage Segment

...The European Parliament and the Council of the European Union have powers of amendment and veto during the legislative process...

## Question

Which governing bodies have veto power?

## APPLICATIONS: QUESTION/ANSWERING

- But:
  - QA could be very complicated!
  - From the shared task PÁGICO (LINGUAMÁTICA):
    - (PT) Quais os jogadores de futebol de língua portuguesa que passaram por mais de três países estrangeiros na sua vida profissional?
    - (EN) Which Portuguese-speaking football players have played in more than three foreign countries during their professional careers?

# ACTIVE LEARNING MOMENT: THINK-PAIR-SHARE



# EXERCISE

- How do you think QA was implemented 20 years ago?

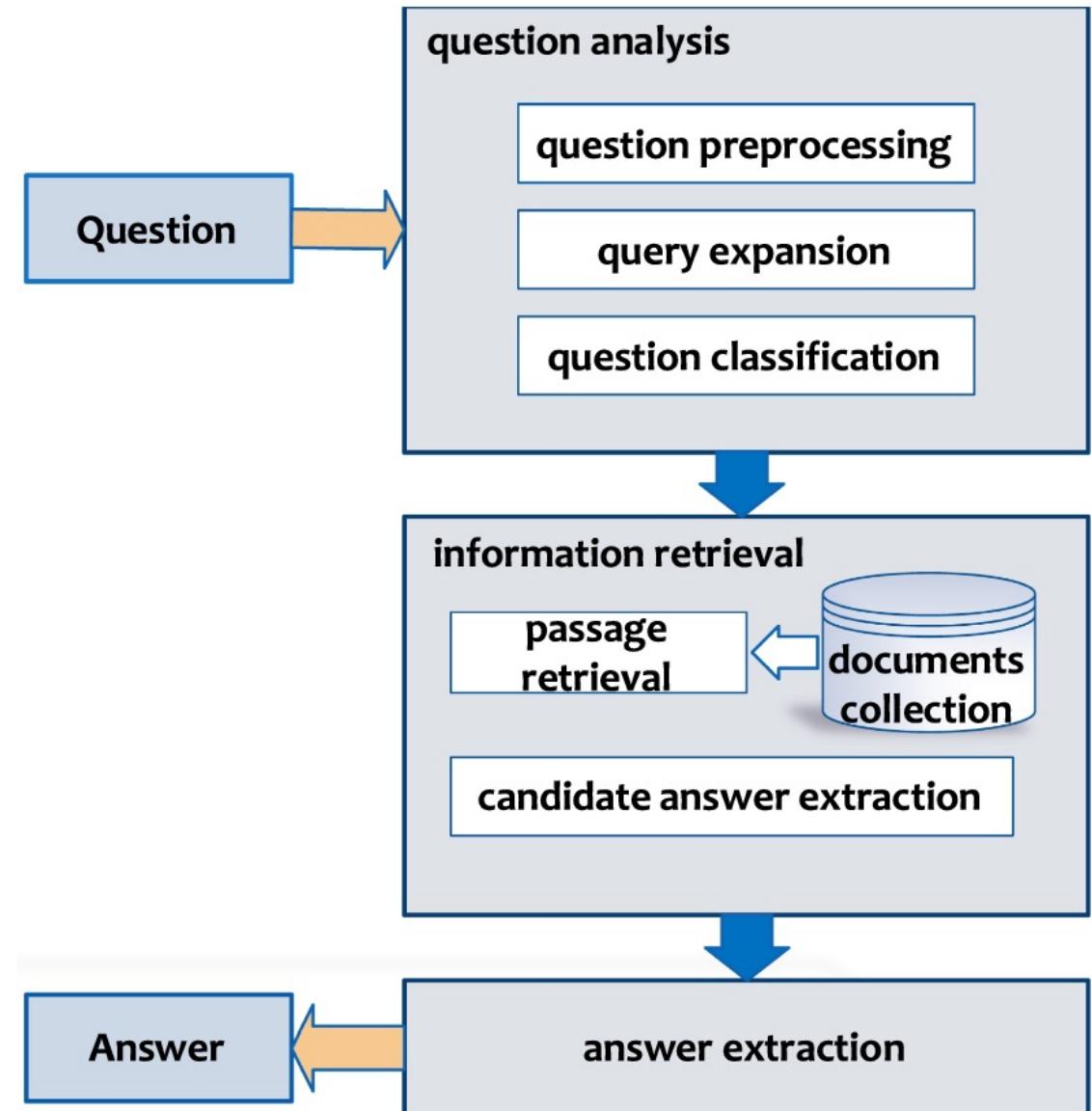


Image from the paper: Arabic question answering system: a survey

# APPLICATIONS: QUESTION/ANSWERING

- Now (answer by ChatGPT (did not check it)):
  - Which Portuguese-speaking football players have passed through more than three foreign countries in their professional lives?

[Cristiano Ronaldo](#): [...] Ronaldo has played in Portugal (Sporting CP), England (Manchester United), Spain (Real Madrid), Italy (Juventus), and Saudi Arabia (Al Nassr).

[Deco](#): Born in Brazil and later naturalized as a Portuguese citizen, Deco played in Brazil (Corinthians, Corinthians Alagoano), Portugal (Benfica, Alverca, Salgueiros, Porto), Spain (Barcelona), and England (Chelsea).

[Pepe](#): Another Brazilian-born player who represents Portugal, Pepe has played in Portugal (Marítimo, Porto), Spain (Real Madrid), Turkey (Beşiktaş), and returned to Portugal with Porto.

[Nani](#): [...] played in Portugal (Sporting CP), England (Manchester United), Turkey (Fenerbahçe), Spain (Valencia), Italy (Lazio), and the United States (Orlando City).

[Eusébio](#): played in Portugal (Benfica), the United States (Boston Minutemen, Las Vegas Quicksilvers), Mexico (Monterrey), and Canada (Toronto Metros-Croatia).

# APPLICATIONS: SUMMARIZATION

- Summarization: take in a document and output a summary of that document
- Then:

## (a) Extractive Summarization

Source Text: Peter and Elizabeth took a taxi to attend the night party in the city.

While in the party, Elizabeth collapsed and was rushed to the hospital.

Summary: Peter and Elizabeth attend party city. Elizabeth rushed hospital.

## (b) Abstractive Summarization

Source Text: Peter and Elizabeth took a taxi to attend the night party in the city.

While in the party, Elizabeth collapsed and was rushed to the hospital.

Summary: Elizabeth was hospitalized after attending a party with Peter.

## APPLICATIONS: SUMMARIZATION

- Now (answer by ChatGPT):
  - Summarize in 5 sentences Stranger Things season 3
    - In "Stranger Things" Season 3, the story is set in the summer of 1985 in Hawkins, Indiana. The newly opened Starcourt Mall becomes the town's main attraction, and it's secretly a front for Russians trying to open a gate to the Upside Down. The Mind Flayer, an entity from the Upside Down, returns to possess Billy and other residents to create a monstrous creature. Eleven and her friends uncover the Russians' plot and fight against the Mind Flayer, leading to a climactic battle at the mall. The season ends with Hopper's apparent death and the Byers family, along with Eleven, moving out of Hawkins, setting the stage for future conflicts.

# APPLICATIONS: NATURAL LANGUAGE INFERENCE

- Natural Language Inference (NLI): receive two input sentences: a premise and a hypothesis, and outputs whether the premise entails, contradicts or is neutral with respect to the hypothesis.

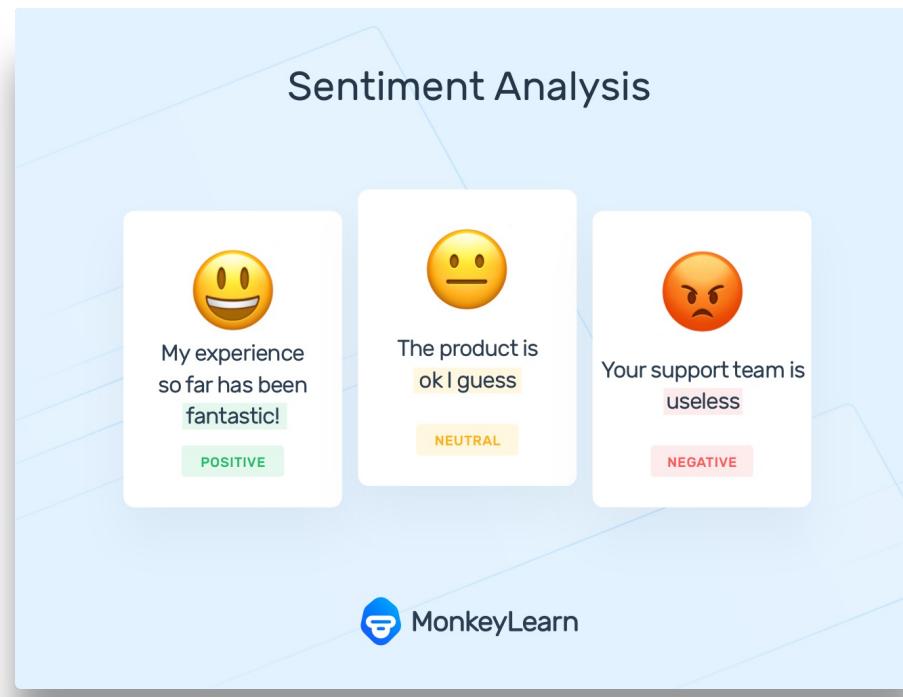
---

Premise	A woman selling bamboo sticks talking to two men on a loading dock.
Entailment	There are <b>at least three people</b> on a loading dock.
Neutral	A woman is selling bamboo sticks <b>to help provide for her family</b> .
Contradiction	A woman is <b>not</b> taking money for any of her sticks.

---

# APPLICATIONS: SENTIMENT ANALYSIS

- Sentiment Analysis: models trained to classify the sentiment expressed by input text.



<https://monkeylearn.com/sentiment-analysis/>

## BY THE WAY: NOT THAT EASY

- “The characters are so real and handled so carefully, that [...]. There were several times **where I actually dropped the book** and was too scared to pick it back up. Intellectually, you know it's not real. [...]. Still, whenever I go into the bathroom late at night, I have to pull back the shower curtain just to make sure.”



## BY THE WAY: NOT THAT EASY

- “It took a couple of goes to get into it, but once the story hooked me, I found it difficult to put the book down – except for those moments when I had to stop and shriek at my friends, "SPARKLY VAMPIRES!" or "VAMPIRE BASEBALL!" or "WHY IS BELLA SO STUPID?" These moments came increasingly often as I reached the climactic chapters, until I simply reached the point where I had to stop and flail around laughing.”



# APPLICATIONS: COMMON SENSE REASONING

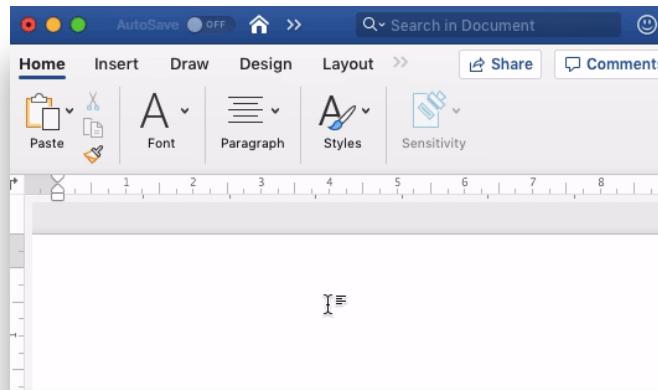
- Common sense Reasoning: models are required to answer questions that request some reasoning.
- Example of a dataset: [Modified Winograd Schema Challenge, MWSC](#)

## Winograd Schema Data

*The trophy doesn't fit in the suitcase because it is too **big**. What is too big?*

*Answer 0: **the trophy**. Answer 1: **the suitcase***

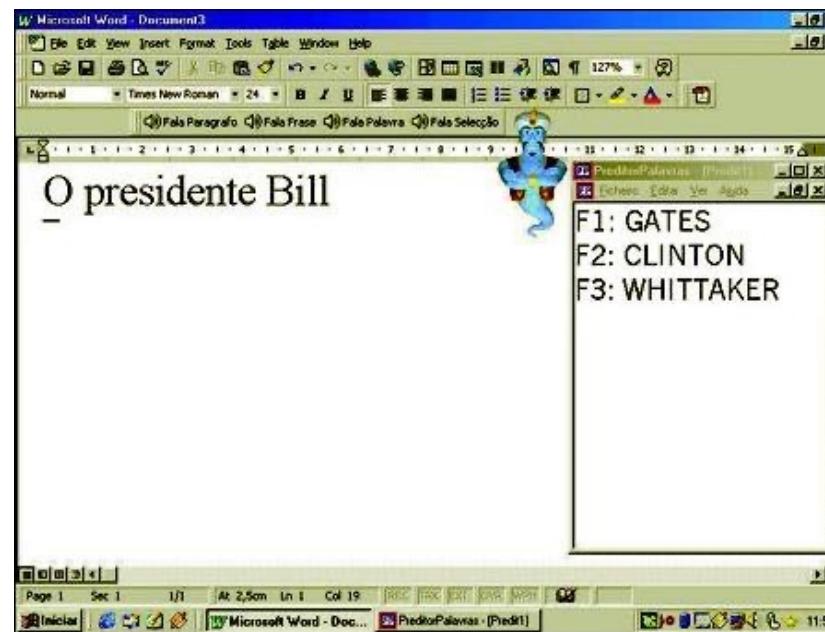
# MORE NLP APPLICATIONS



20% of kids cyberbullied think about suicide,  
and 1 in 10 attempt it.  
4500 kids commit suicide each year

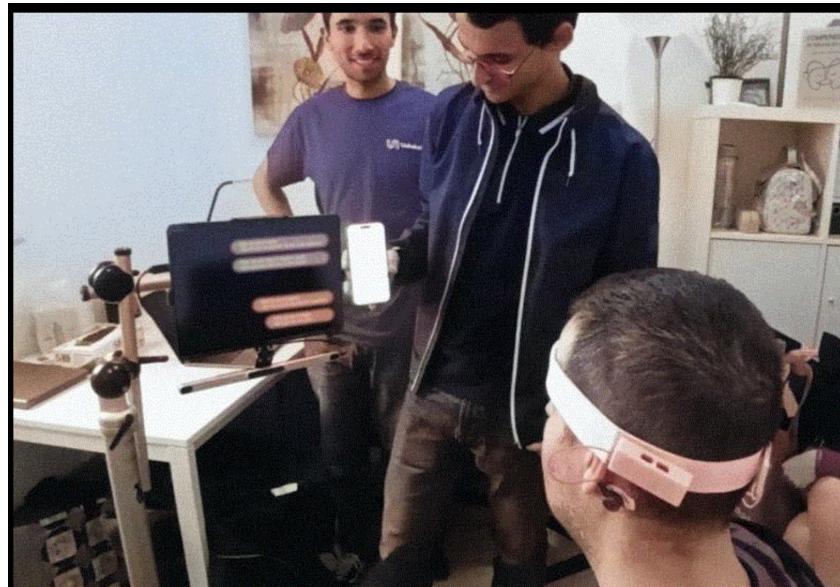
# MORE NLP APPLICATIONS

- Assistive Technology
  - Previously (last century at INESC-ID):
    - Target: Cerebral palsy (for instance)
      - TTS with virtual keyboards and word prediction



# MORE NLP APPLICATIONS

- Assistive Technology
  - Target: Amyotrophic lateral sclerosis (ALS).
    - Halo: uses EMG (electromyography) sensors and large language models (LLM)
    - Could replace the current communication models for patients with speech difficulties, which are based on eye tracking.



# KEY TAKEAWAYS

## KEY TAKEAWAYS

- There are several ways to take advantage of pre-trained models
- There are several ways to perform transfer learning
- Multi-task learning is a recent trend in NLP
- There are several compression Techniques that can be applied to compress neural networks
- There are many NLP applications. We have seen several along the course; now we focus on some of the ones from DecaNLP
- Concepts: pre-trained models, inference, prompting, feature-based Transfer Learning, fine-tuning, multi-task learning, compression techniques, ...

## SUGGESTED READINGS

## SUGGESTED READINGS

- PEFT:
  - <https://arxiv.org/pdf/2303.15647>
- LoRA:
  - Paper: LoRA: low-rank adaptation of large language models (Hu et al. 2021)
  - Low-rank Adaption of Large Language Models: Explaining the Key Concepts Behind LoRA (YouTube)
  - <https://www.ml6.eu/blogpost/low-rank-adaptation-a-technical-deep-dive>