Transfer Learning and Transformers in NLP

Ehsan Taher

March 2021
Farvardin 1400

Contents

- What is Transfer Learning?
- Downstream tasks and Model Adaptation: Quick Examples
- Trends and limits of Transfer Learning in NLP
- Transformers: BERT, GPT (in english and persian)

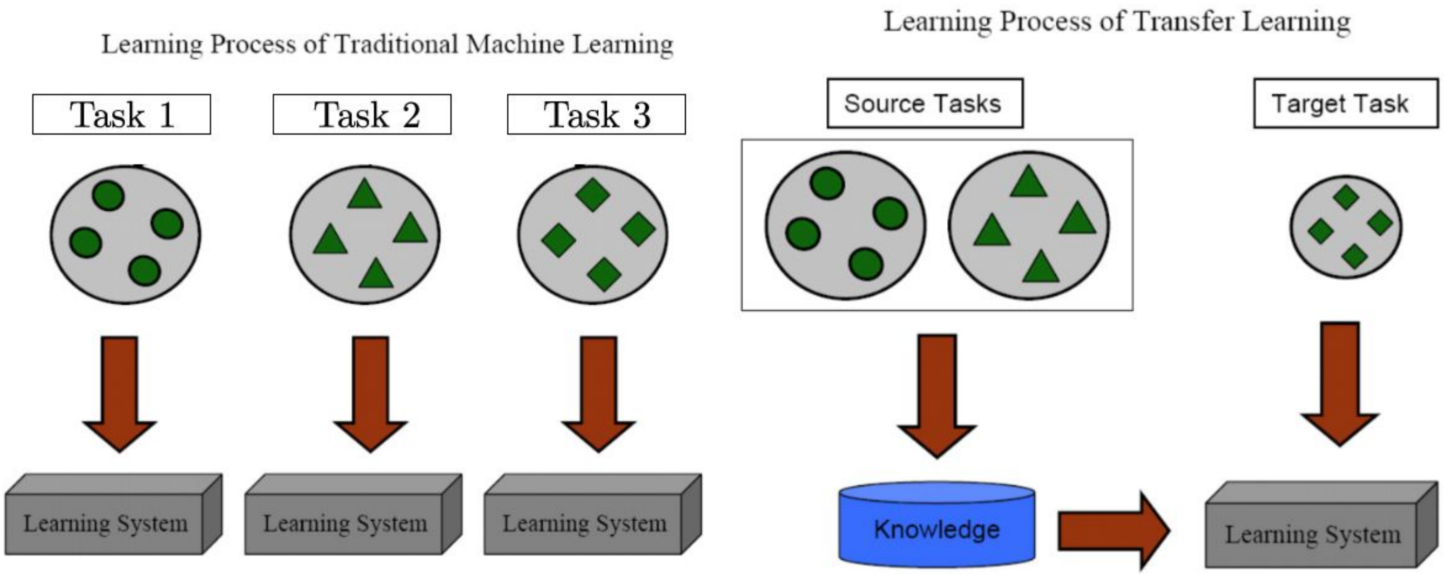
- Introduction to Hugging Face Transformers Library
- Introduction to pytorch , models and training procedure
- Training and Fine tuning a transformer model

Prerequisites

- Basic knowledge in:
  - Machine Learning
  - NLP
  - Python

What is Transfer Learning?

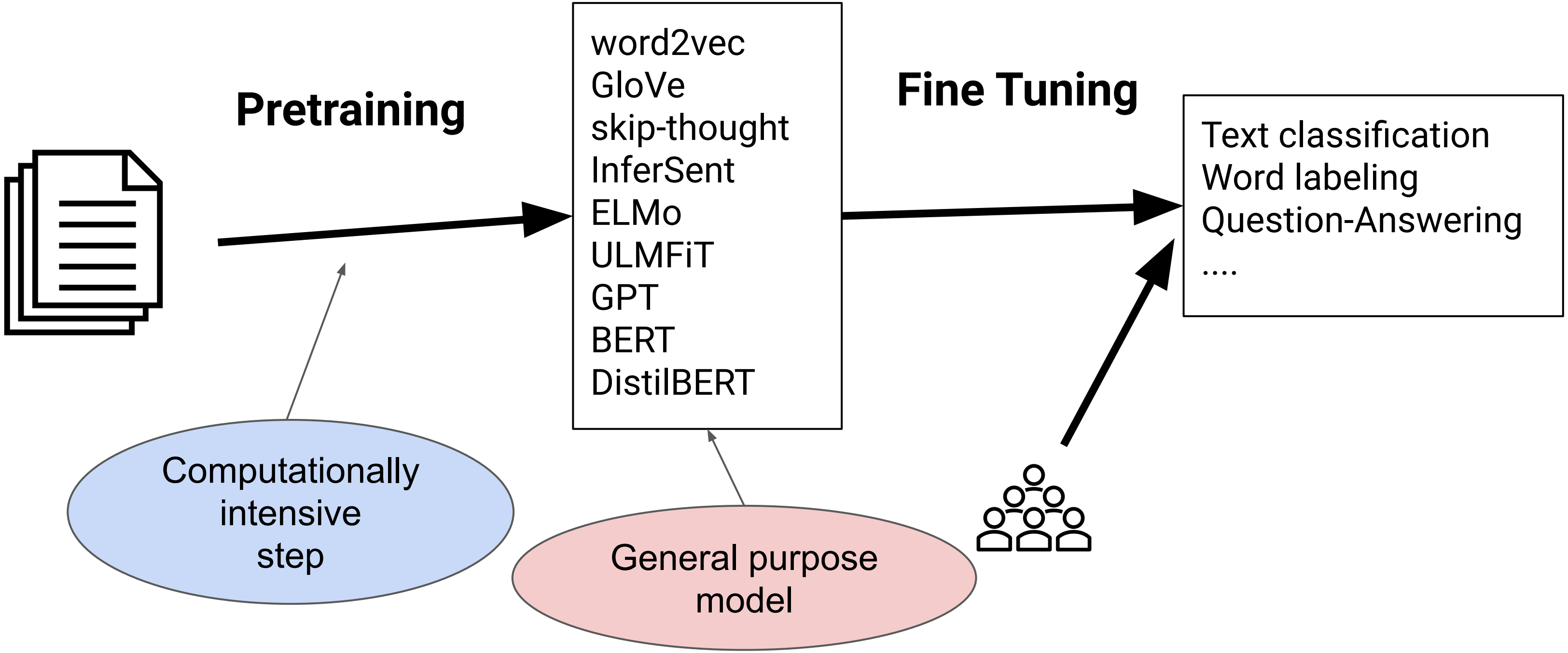
What is Transfer Learning?



(a) Traditional Machine Learning (b) Transfer Learning

Sequential Transfer Learning

Learn on one task/dataset, transfer to another task/dataset



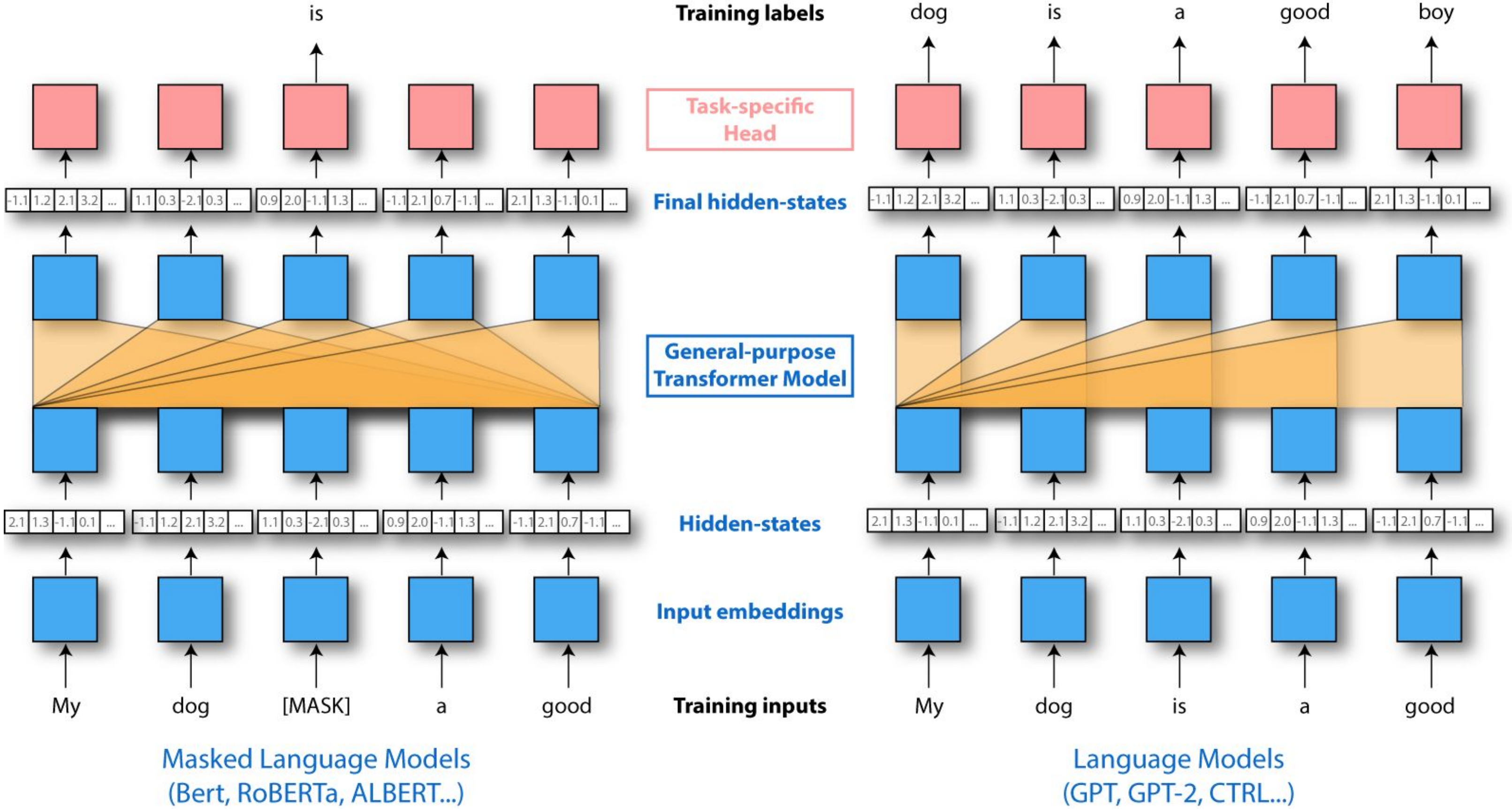
Pretraining: Language modeling

Many currently successful **pretraining** approaches are based on **language modeling**: learning to predict $P_\theta(\mathit{text})$ or $P_\theta(\mathit{text} \mid \mathit{other\ text})$

Advantages:

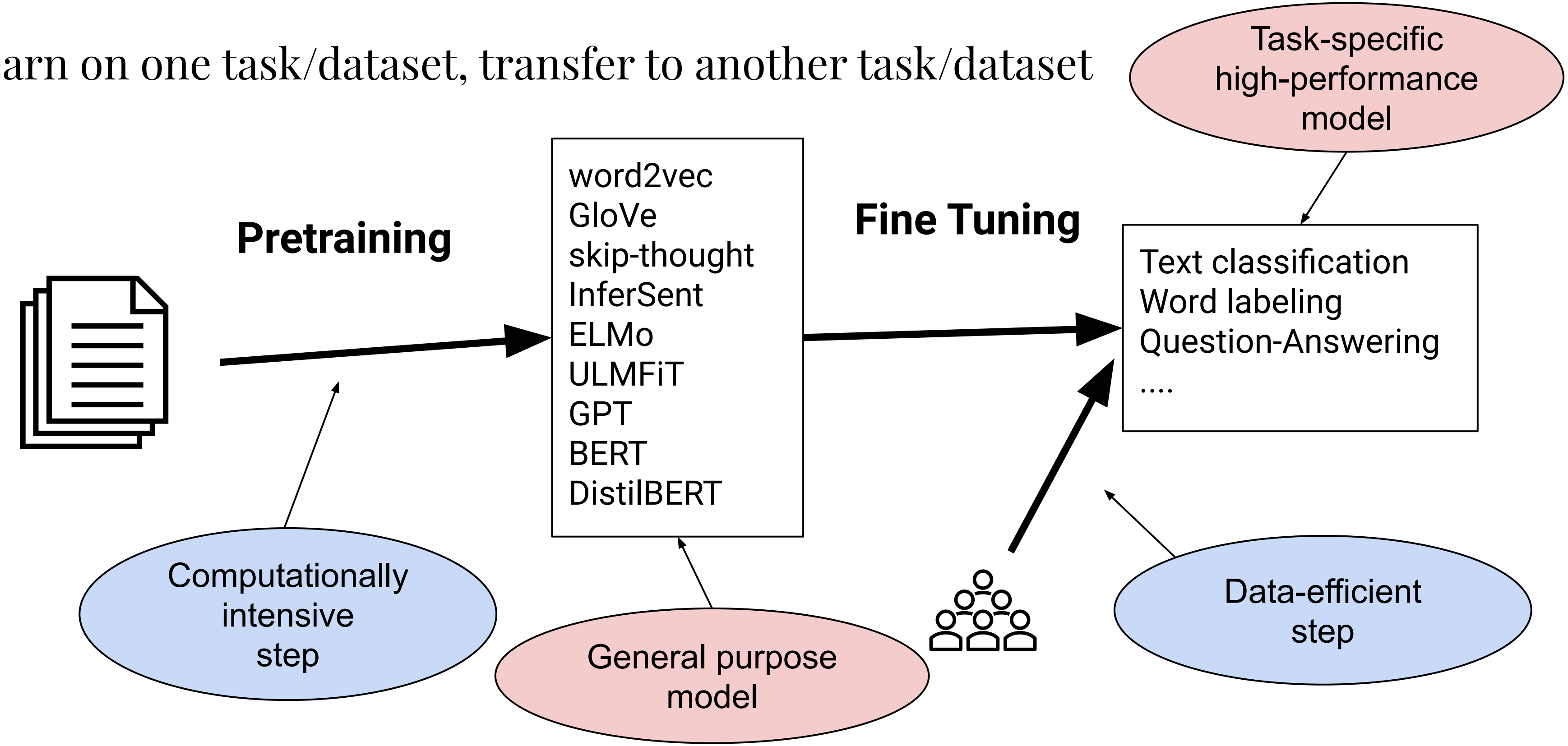
- Doesn’t require human annotation – **self-supervised**
- Many languages have **enough text** to learn high capacity model
- **Versatile** – can be used to learn both sentence and word representations with a variety of objective functions

Pretraining Transformers models (BERT, GPT...)



Sequential Transfer Learning

Learn on one task/dataset, transfer to another task/dataset



Model: Adapting for target task

General workflow:

1. **Remove pretraining task head** (if not used for target task)

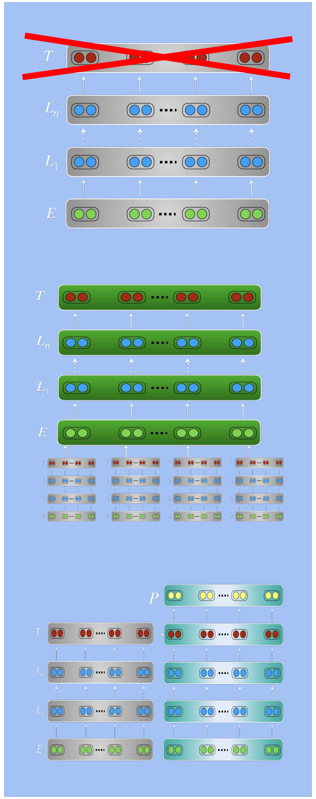
2. **Add target task-specific elements** on top/bottom:
   - **simple**: linear layer(s)
   - **complex**: full LSTM on top

Sometimes very complex: Adapting to a structurally different task

Ex: Pretraining with a **single input sequence** and adapting to a task with **several input sequences** (ex: translation, conditional generation...)

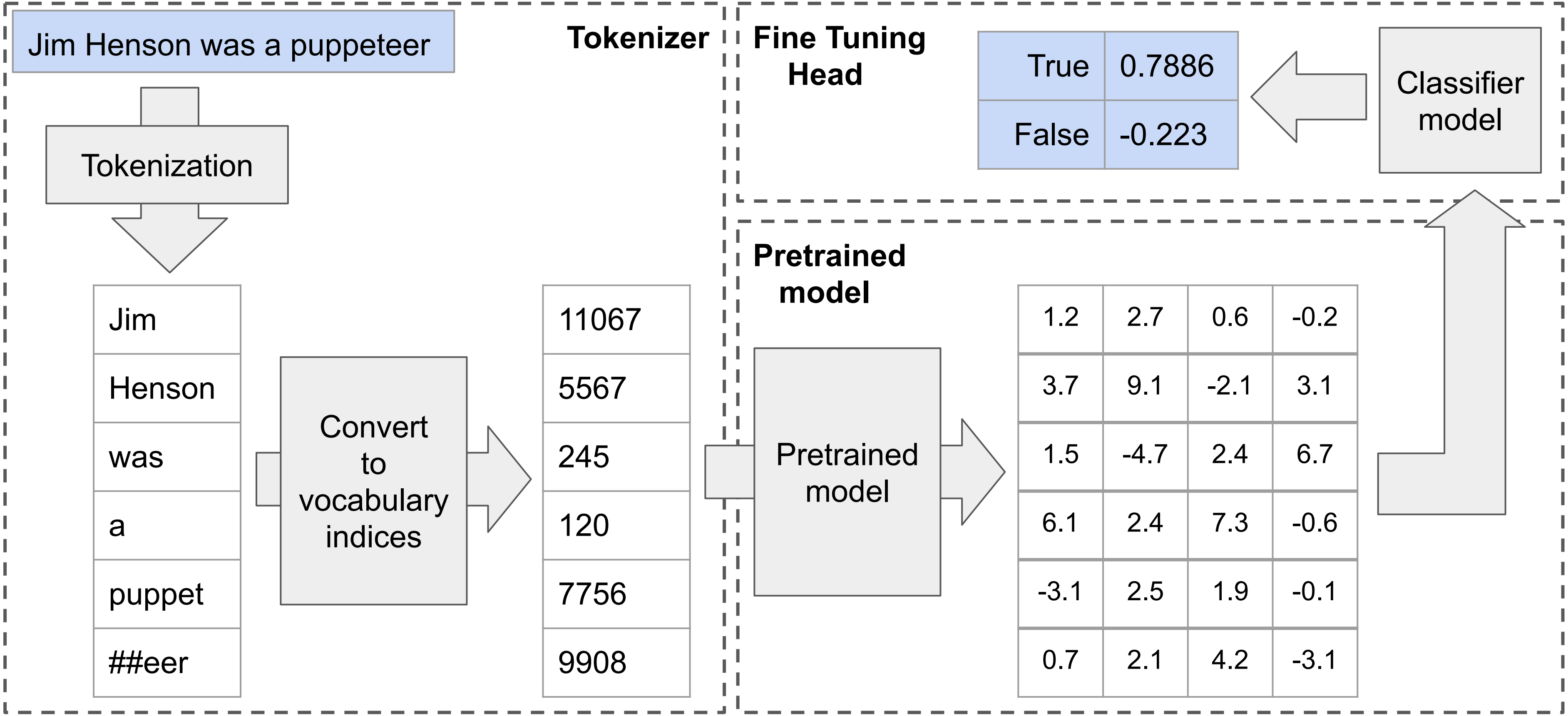
⇨ Use pretrained model to initialize as much as possible of target model

⇨ Ramachandran et al., EMNLP 2017; Lample & Conneau, 2019



Downstream tasks and Model Adaptation: Quick Example

Transfer Learning for text classification



Transfer Learning for text classification

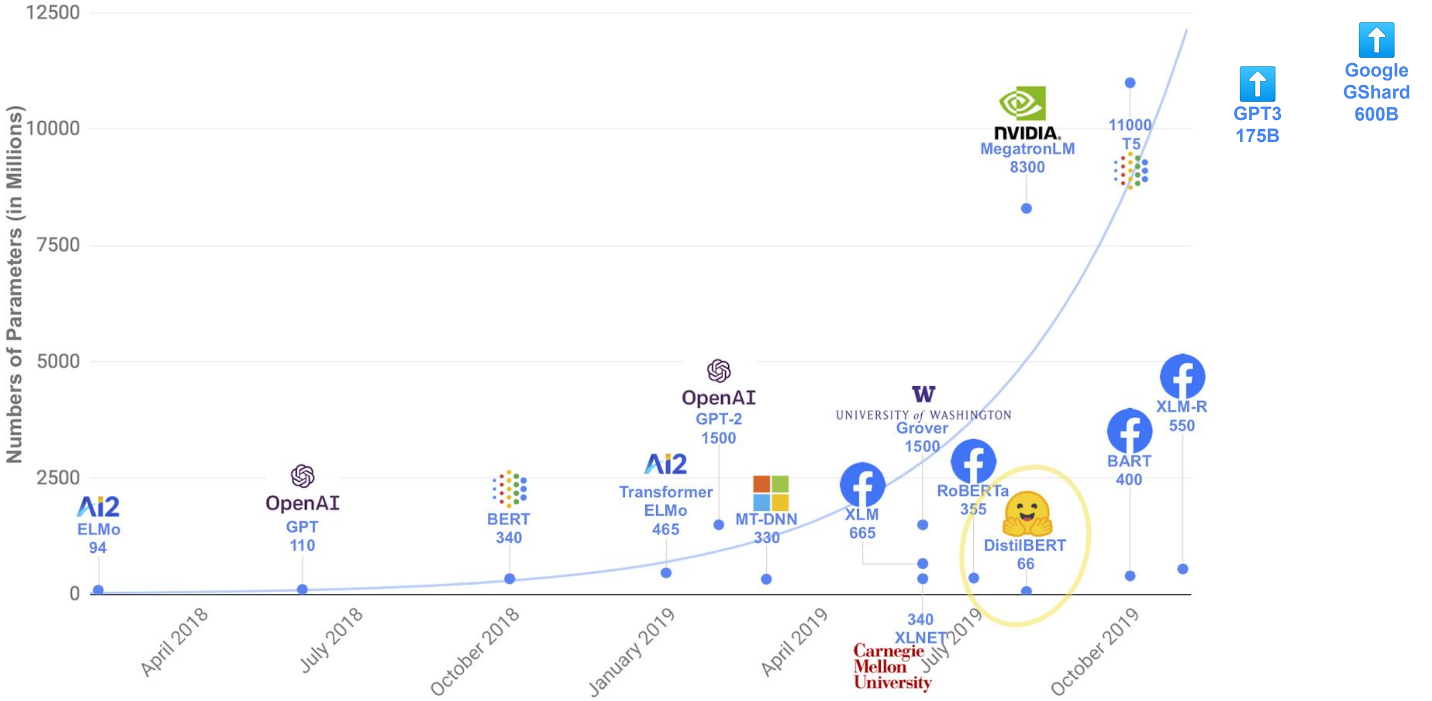
Remarks:

- The error rate goes down quickly! After one epoch we already have good accuracy.
  - Fine-tuning is highly data efficient in Transfer Learning
- We took our pre-training & fine-tuning hyper-parameters straight from the literature on related models.
  - Fine-tuning is often robust to the exact choice of hyper-parameters

Trends and limits of Transfer Learning in NLP

Model size and Computational efficiency

Going big on model sizes - over 1 billion parameters as become the norm for SOTA



In Persian!

- Pars BERT :
  - https://huggingface.co/HooshvareLab/bert-fa-zwnj-base
- GPT:
  - https://huggingface.co/HooshvareLab/gpt2-fa
  - https://huggingface.co/HooshvareLab/gpt2-fa-poetry
  - https://huggingface.co/HooshvareLab/gpt2-fa-comment

Model size and Computational efficiency

Why is this a problem?

- Narrowing the research competition field
  - what is the place of academia in today's NLP?
  - fine-tuning? analysis and BERTology? critics?
- Environmental costs

Consumption	**CO_2e (lbs)**
Air travel, 1 passenger, NY↔SF	1984
Human life, avg, 1 year	11,023
American life, avg, 1 year	36,156
Car, avg incl. fuel, 1 lifetime	126,000
Training one model	
SOTA NLP model (tagging)	13
w/ tuning & experimentation	33,486
Transformer (large)	121
w/ neural architecture search	394,863

- Is bigger-is-better a scientific research program?

Model size and Computational efficiency

Reducing the size of a pretrained model

Three main techniques currently investigated:

- Distillation
  - DistilBert: 95% of Bert performances in a model 40% smaller and 60% faster
- Pruning
- Quantization
  - From FP32 to INT8

The inductive bias question

The generalization problem:

- Models are brittle: fail when text is modified, even with meaning preserved
- Models are spurious: memorize artifacts and biases instead of truly learning

Article: Super Bowl 50
Paragraph: "*Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.*"
Question: "*What is the name of the quarterback who was 38 in Super Bowl XXXIII?*"
Original Prediction: John Elway
Prediction under adversary: Jeff Dean

Hands on Code

Google Colab

- Free
- Good GPU
- Limits:
  - 12 hours

Python libraries

- pandas
- Pytorch
- Transformers
- sklearn

Model Report (with pretrained PARS BERT)

	precision	recall	f1-score	support
0	0.98	0.98	0.98	765
1	1.00	1.00	1.00	4151
accuracy			0.99	4916
macro avg	0.99	0.99	0.99	4916
weighted avg	0.99	0.99	0.99	4916

	precision	recall	f1-score	support
0	0.76	0.67	0.72	335
1	0.94	0.96	0.95	1773
accuracy			0.91	2108
macro avg	0.85	0.82	0.83	2108
weighted avg	0.91	0.91	0.91	2108

Model Report (with pretrained Multilingual BERT)

0.23090923121307373

	precision	recall	f1-score	support
0	0.89	0.89	0.89	765
1	0.98	0.98	0.98	4151
accuracy			0.96	4916
macro avg	0.93	0.93	0.93	4916
weighted avg	0.96	0.96	0.96	4916

	precision	recall	f1-score	support
0	0.58	0.70	0.63	335
1	0.94	0.90	0.92	1773
accuracy			0.87	2108
macro avg	0.76	0.80	0.78	2108
weighted avg	0.88	0.87	0.88	2108