NLP SUMMIT

intel labs

Few-shot learning in NLP
SetFit Model

Moshe Wasserblat
NLP/EAI/Intel Labs

April 2023

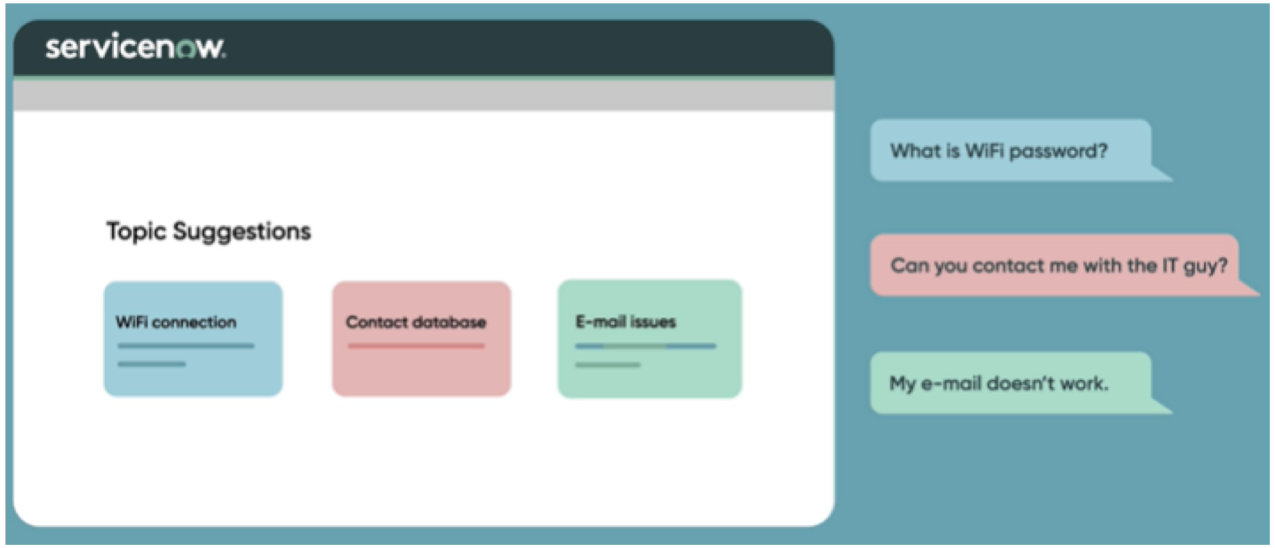
BIO

- **NICE Systems** 
  - Led NLP research group
  - First company to productize Speech2Text, ED, Text Analytics in Call-Center
- **INTEL** 
  - Applied NLP/DL Research – Increase adoption of DL in the industry
  - SoTA NLP research
  - Explore compute features that disrupt our HW

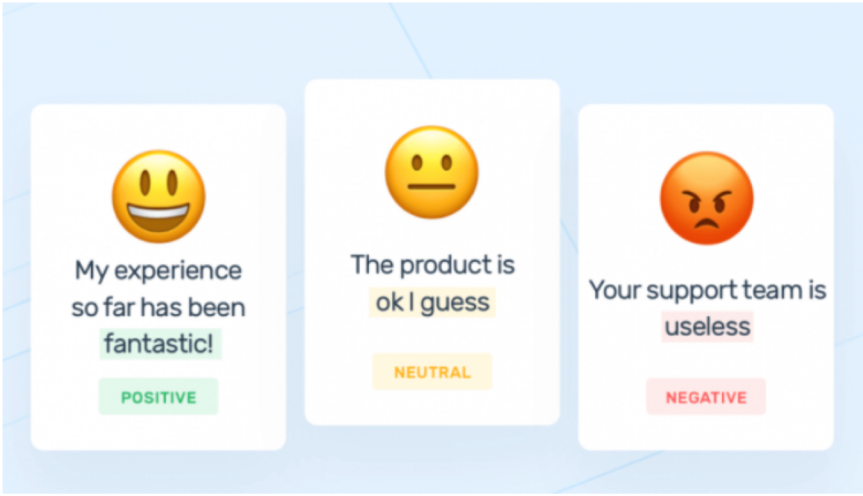
Text Classification - Usages

48% of Large companies use NLP for document classification

1. Topic classification for routing



2. Sentiment Analysis



3. Spam detection
4. Hate speech detection
5. Etc.

Text Classification in Healthcare

Label the following sentence based on whether it is related to an Adverse Drug Effect (***ADE)**

{**Sentence**: 'No regional side effects were noted'
Label: 'not ADE-related' }

Label the following publication **abstracts** based on whether it is related to one or many of **HoC** 10 labels.

{**Abstract**: 'ZD1839 increased the expression of p27(KIP1) ... and maturation markers (P <.001) '
Label: 'evading growth suppressors, resisting cell death' }

HOC (Hallmarks of Cancer)

Agenda

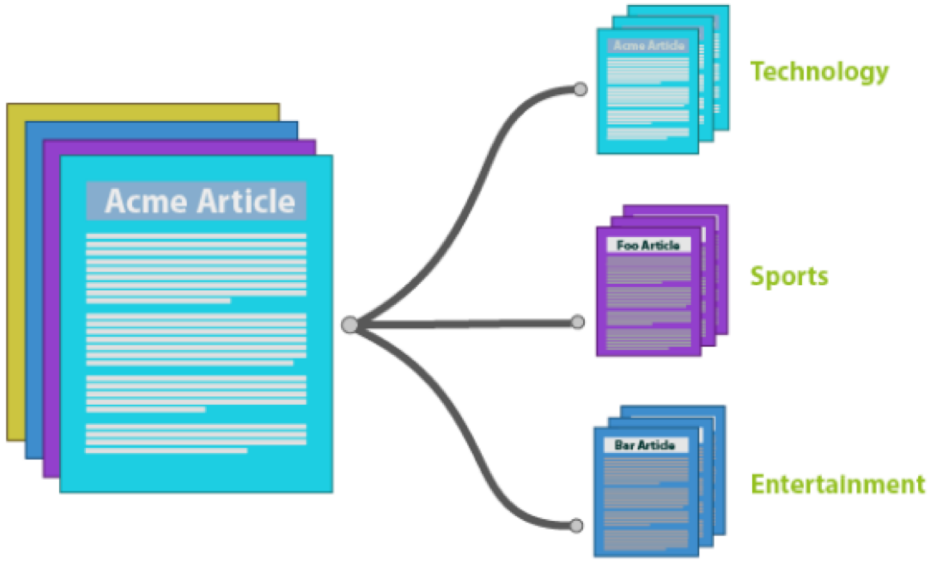
Few-shot learning overview

SetFit

How does SetFit work?

Few-shot Learning with Transformers

- Traditional fine-tuning -> **1000s** labeled training sample
- Fewshot learning -> **8-64** labeled training sample



Text Classification

Method 1 - In-context Learning

Language Models are Few-Shot Learners

Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

```
1  Translate English to French:          ← task description
2  sea otter => loutre de mer            ← examples
3  peppermint => menthe poivrée          ← examples
4  plush girafe => girafe peluche        ← examples
5  cheese =>                             ← prompt
```

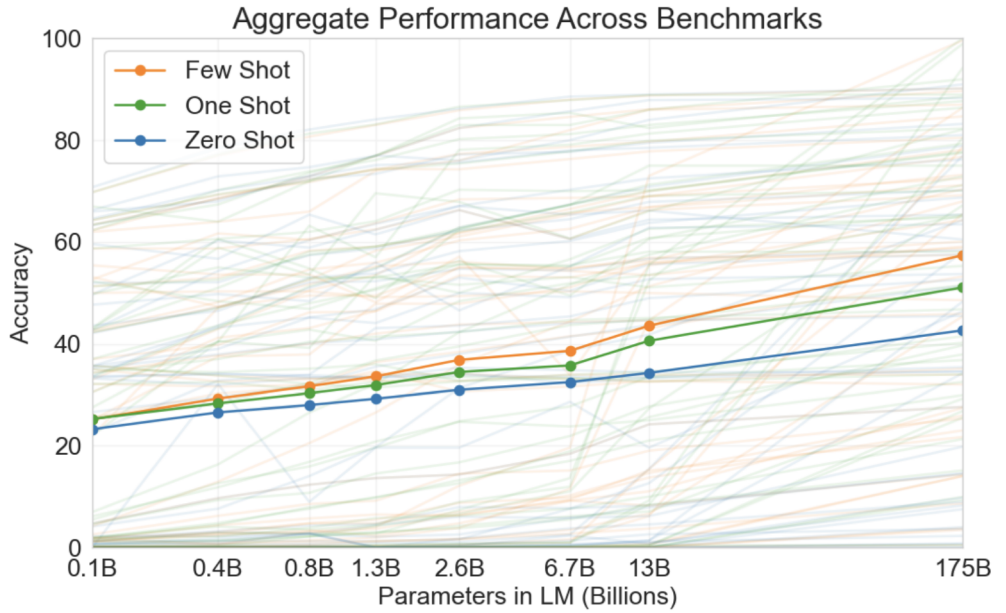


Image source: [Language Models are Few-Shot Learners]

- Popularised by GPT-3 with **"in context learning"**
- Performance increases with scale
- In practice:
  - performance sensitive to quality of **"prompt engineering"**
  - hard/expensive to deploy large models

Method 2 - Parameter Efficient Fine-tuning

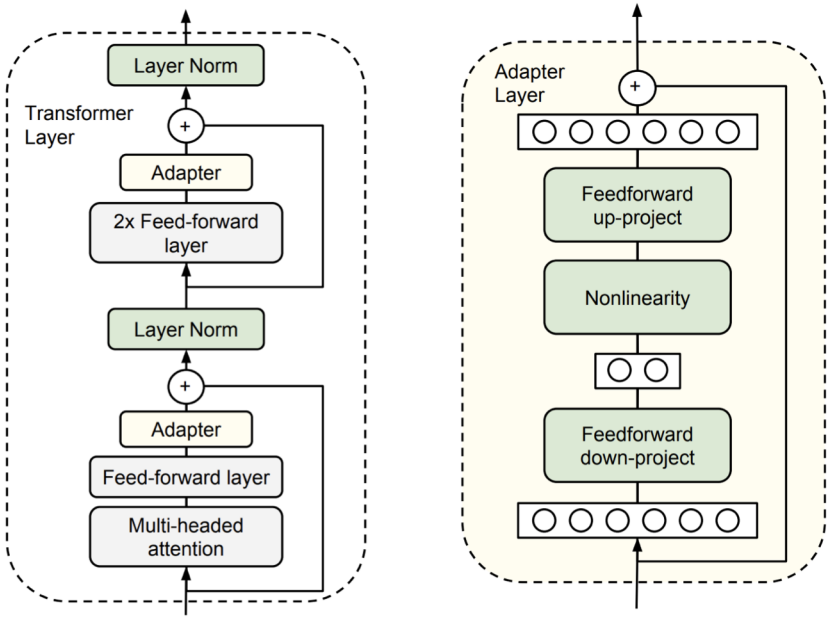


Image credit: Parameter-Efficient Transfer Learning for NLP

T-Few much more efficient / performant than GPT-3 💪

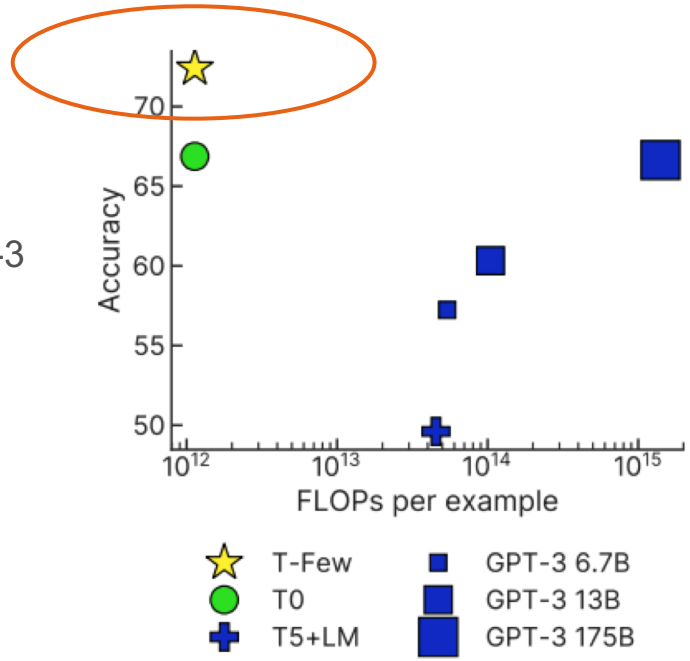


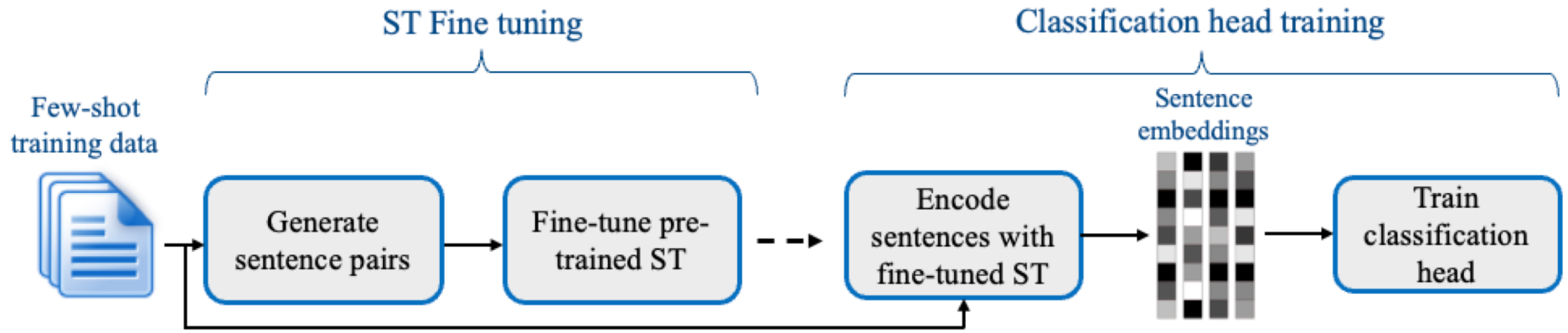
Image credit: T-Few paper

- Alternatives include **"parameter efficient fine-tuning" (PEFT)** which add/update a small number of parameters in a pretrained LM. **T-Few is current SOTA**
- In practice:
  - performance sensitive to quality of **"prompt engineering"** (like ICL)
  - SOTA performance relies on largish 11B (11.4GB) T0 model

Can we do better? 🤔

- No depending on large generative models
- Not depending on manually crafted prompts

SetFit – **Se**ntence **T**ransformer **Fi**ne-**T**uning

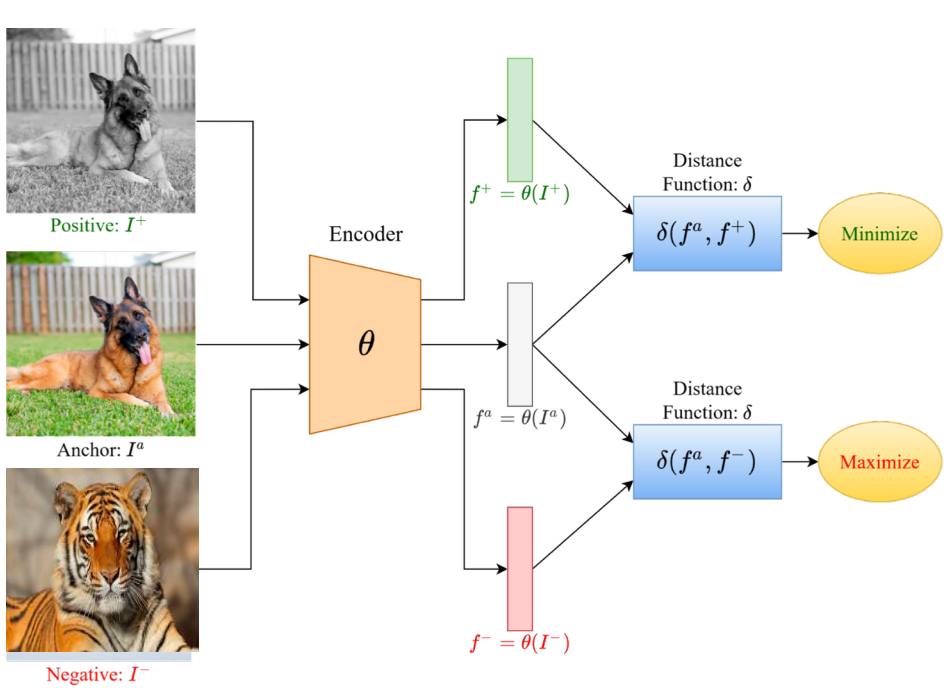


SetFit uses a simple, yet effective **2-stage training recipe**:

- Fine-tune a Sentence Transformer "body" with **contrastive learning** on small number of labelled examples (e.g. 8 examples per class)
- Train classification "head" using embeddings from fine-tuned Sentence Transformer

Works well for **text classification** (token classification WIP)

Contrastive learning primer



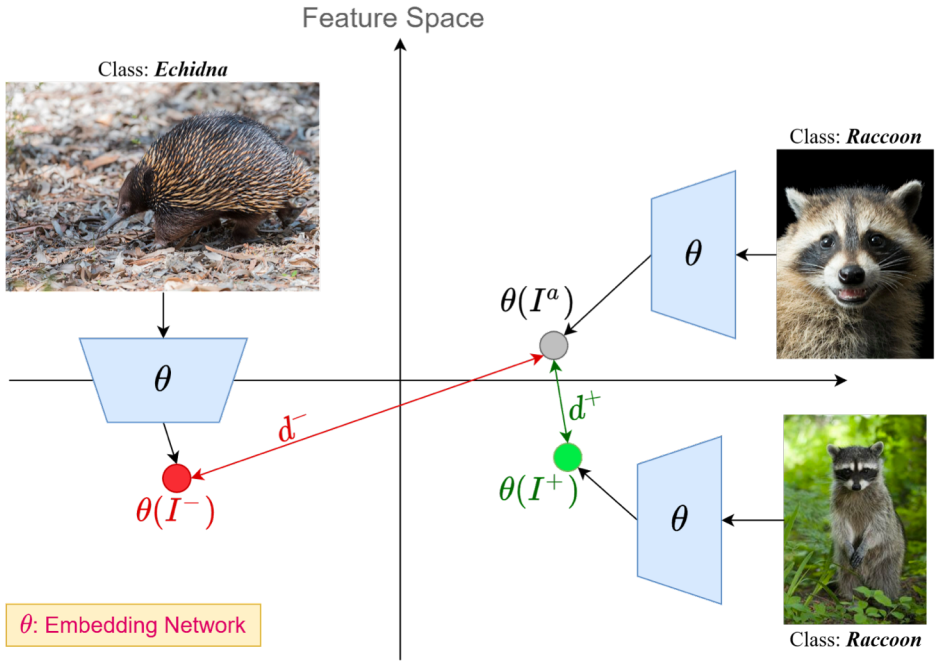
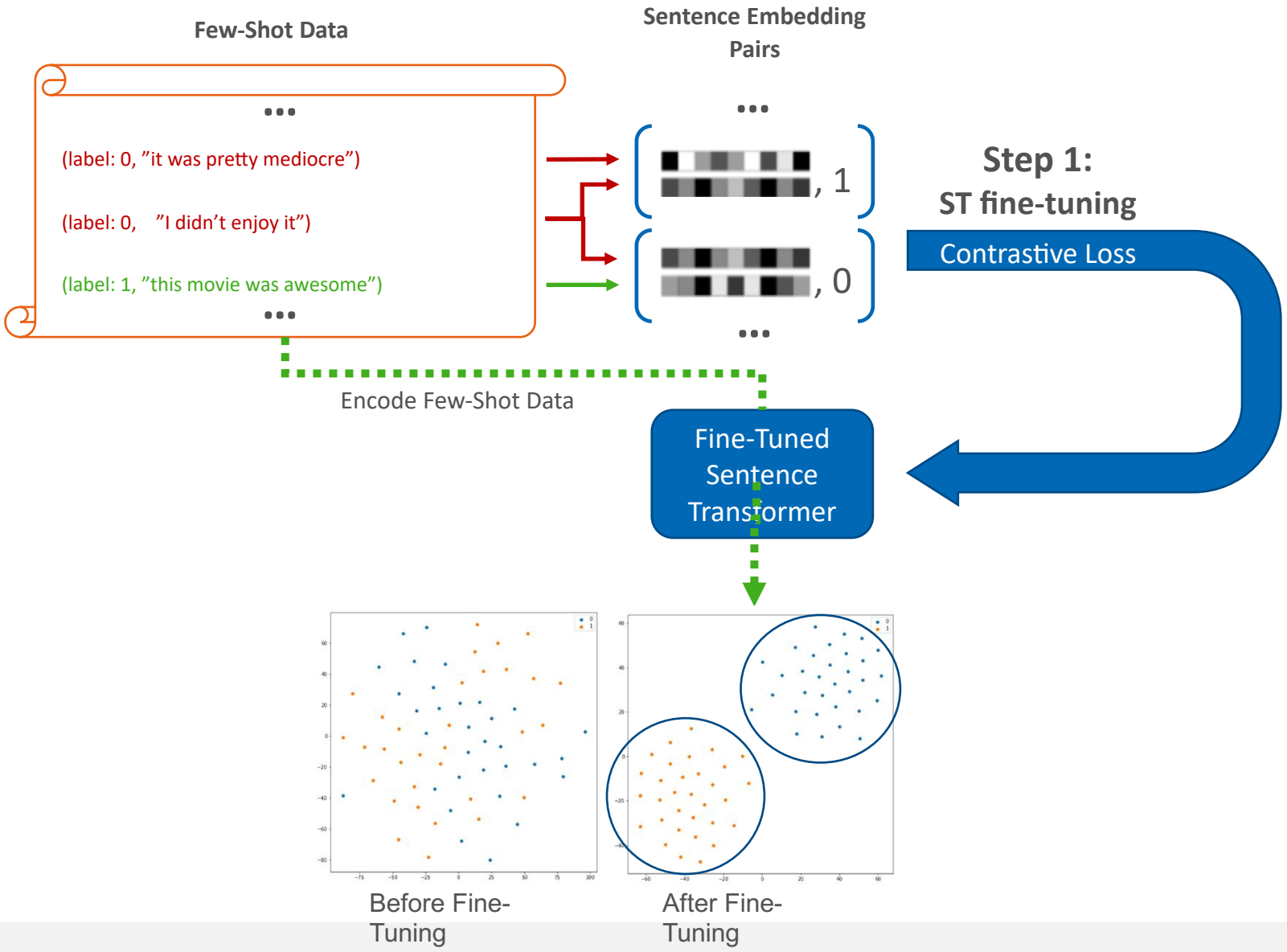


Image source: V7 Blog

Basic idea:

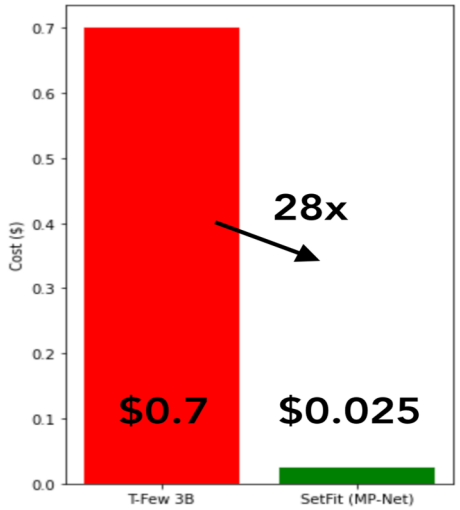
- Minimize (maximize) distance between “positive” (“negative”) examples
- Distance metric often Euclidean distance or cosine similarity
- Produces high quality representations for images / text

SetFit - Training



SetFit Results

Method	SST-5	AmazonCF	CR	Emotion	EnronSpam	AGNews	Average†
			$\lvert N\rvert = 8$*				
FINETUNE	$33.5_{2.1}$	$9.2_{4.9}$	$58.8_{6.3}$	$28.7_{6.8}$	$85.0_{6.0}$	$81.7_{3.8}$	$43.0_{5.2}$
PERFECT	$34.9_{3.1}$	$18.1_{5.3}$	$81.5_{8.6}$	$29.8_{5.7}$	$79.3_{7.4}$	$80.8_{5.0}$	$48.7_{6.0}$
ADAPET	$50.0_{1.9}$	$19.4_{7.3}$	$91.0_{1.3}$	$46.2_{3.7}$	$85.1_{3.7}$	$\mathbf{85.1}_{2.7}$	$58.3_{3.6}$
T-FEW 3B	$\mathbf{55.0}^{\star}_{1.4}$	$19.0_{3.9}$	$\mathbf{92.1}_{1.0}$	$\mathbf{57.4}_{1.8}$	$\mathbf{93.1}_{1.6}$	–	$\mathbf{63.4}_{1.9}$
SETFIT$_{\text{MPNET}}$	$43.6_{3.0}$	$\mathbf{40.3}_{11.8}$	$88.5_{1.9}$	$48.8_{4.5}$	$90.1_{3.4}$	$82.9_{2.8}$	$62.3_{4.9}$
			$\lvert N\rvert = 64$*				
FINETUNE	$45.9_{6.9}$	$52.8_{12.1}$	$88.9_{1.9}$	$65.0_{17.2}$	$95.9_{0.8}$	$88.4_{0.9}$	$69.7_{7.8}$
PERFECT	$49.1_{0.7}$	$\mathbf{65.1}_{5.2}$	$92.2_{0.5}$	$61.7_{2.7}$	$95.4_{1.1}$	$\mathbf{89.0}_{0.3}$	$72.7_{1.9}$
ADAPET	$54.1_{0.8}$	$54.1_{6.4}$	$92.6_{0.7}$	$72.0_{2.2}$	$96.0_{0.9}$	$88.0_{0.6}$	$73.8_{2.2}$
T-FEW 3B	$\mathbf{56.0}_{0.6}$	$34.7_{4.5}$	$\mathbf{93.1}_{1.0}$	$70.9_{1.1}$	$\mathbf{97.0}_{0.3}$	–	$70.3_{1.5}$
SETFIT$_{\text{MPNET}}$	$51.9_{0.6}$	$61.9_{2.9}$	$90.4_{0.6}$	$\mathbf{76.2}_{1.3}$	$96.1_{0.8}$	$88.0_{0.7}$	$\mathbf{75.3}_{1.3}$
			$\lvert N\rvert = \textit{Full}$**				
FINETUNE	59.8	80.1	92.4	92.6	99.0	93.8	84.8



CPU friendly

- SetFit performance comparable to other SOTA techniques like T-Few (3B) and ADAPET
- 1-2 orders of magnitude faster to train / run inference 🏎️!

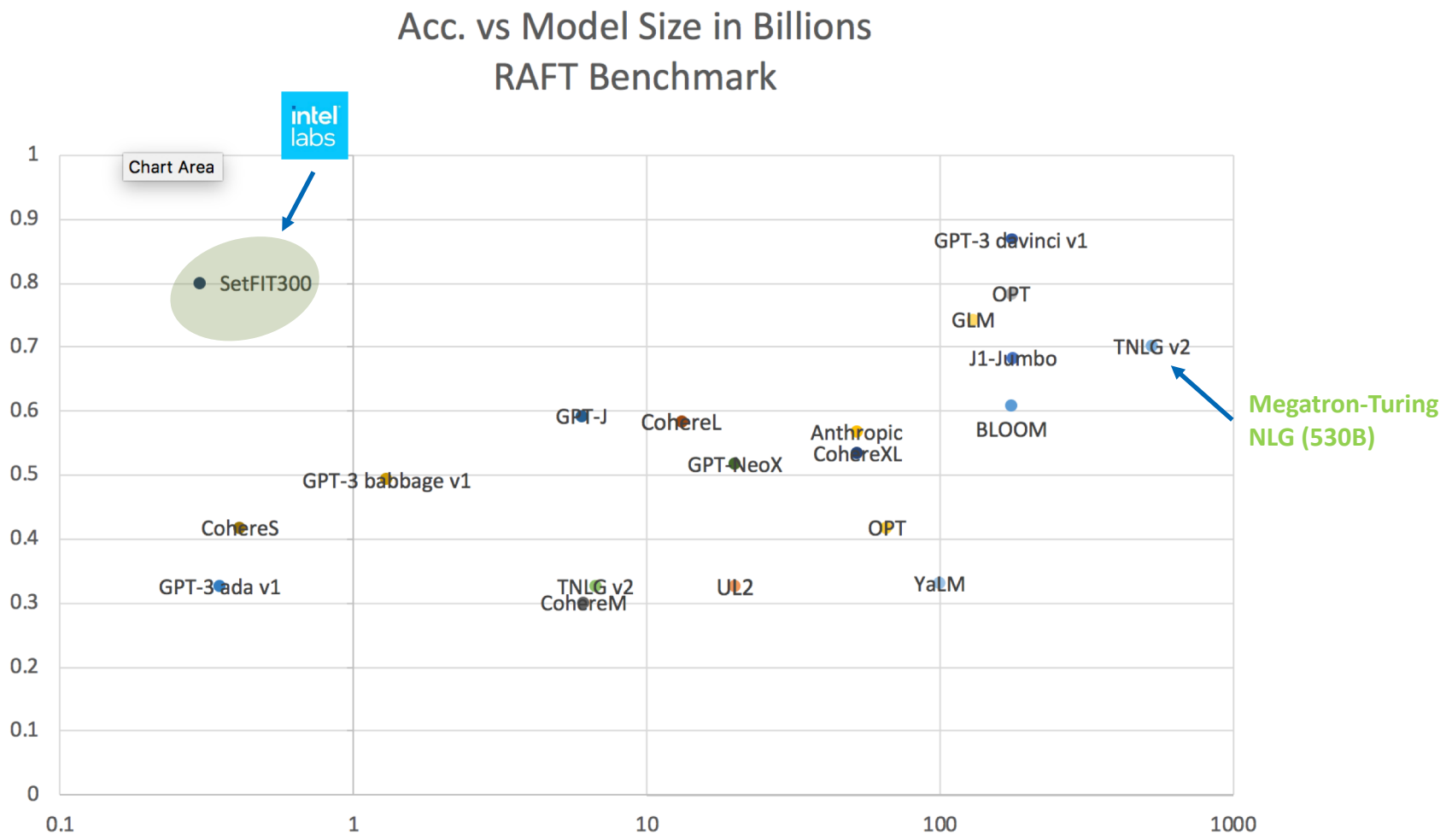
RAFT Benchmark

	Rank	Method	Score	Size*
	1	YIWISE	76.8	-
	2	T-FEW 11B	75.8	11B
	4	Human baseline	73.5	-
350M	6	SETFIT$_{\text{ROBERTA}}$	71.3	355M
	9	PET	69.6	235M
110M	11	SETFIT$_{\text{MPNET}}$	66.9	110M
	12	GPT-3	62.7	175B

Table 3: SETFIT compared to prominent methods on the RAFT leaderboard (as of Sept. 5, 2022). *Number of parameters.

- RAFT is a real-world challenging few-shot benchmark
- SetFit outperforms GPT-3, while being 1600x smaller and not using prompts
- SetFit surpasses the Human baseline in 7 out of 11 tasks

RAFT Benchmark – Healthcare use-case (EDA)



Stanford – Center for Research on Foundation Models

[Large Models Benchmark]

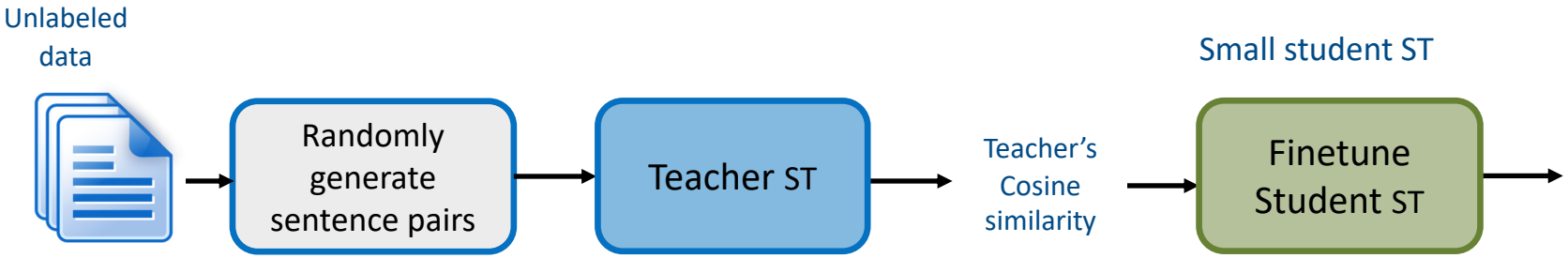
🎉 SetFit SOTA on Bio Text Classification (full-shot) 🎉

- 🏋️ Outperform models that were trained from scratch on Bio data
- 💾 3x times smaller
- 🏎️ Competitive with BioGPT

Model	#params[M]	F1	Pre-train Data
BioBERT[1]	110	81.5	Bio
PubMedBERT[2]	340	82.7	Bio
BioLinkBERT[3]	340	84.9	Bio
GPT-2	355	81.8	General
BioGPT[4]	347	85.1	Bio
SetFit	105	**85.1**	General

HOC (Hallmarks of Cancer)

Experiments - Knowledge Distillation



Sentece1	Sentence2	Teacher's Cosine Sim
POS	POS	1
NEG	NEG	0.81
POS	POS	0.95
POS	NEG	0.05
POS	POS	0.91
NEG	POS	0.1
NEG	NEG	1
POS	NEG	0.02

Basic idea:

- Train teacher and student Sentence Transformer models with 16 examples / class
- Student learn to mimic teacher cosine-similarity
- *Use student embeddings + teacher logits to train SetFit model

Experiments - Knowledge Distillation

	Method	Inf. FLOPs	Train FLOPs	Speed-up	Score
	T-FEW 3B	1.6e11	3.9e15	1x	$63.4_{1.9}$
110M	SETFIT$_{\text{MPNET}}$	8.3e9	2.0e14	19x	$62.3_{4.9}$
15M	SETFIT$_{\text{MINILM}}$ †	1.3e9	3.2e13	123x	$60.3_{1.6}$

Experiment details:

- 16 samples per class for teacher training + unlabeled data for distillation
- The score shown is an average across 5 datasets
- Setfit Student falls short in only 2 points from Setfit teacher

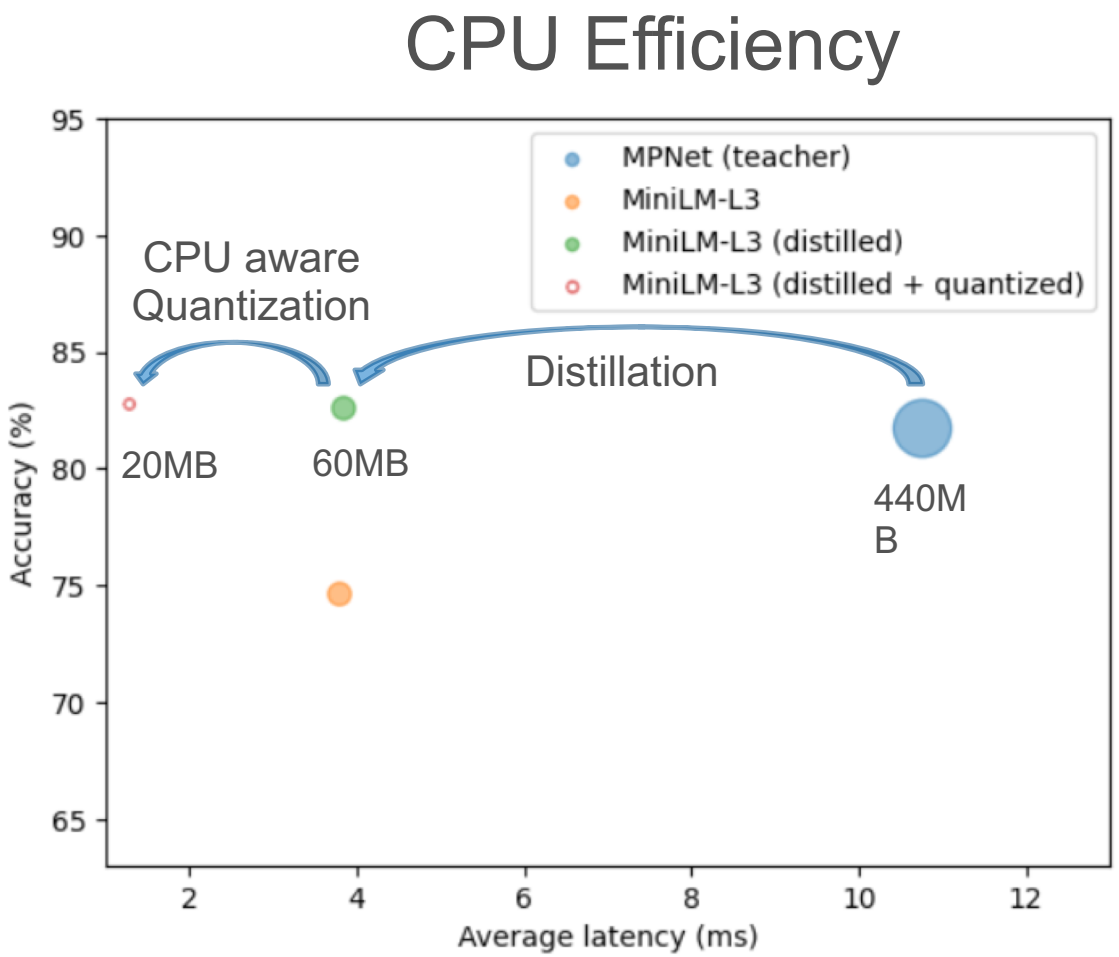
SetFit: Few-shot Text Classification

Efficiency steps:

- Distillation
- CPU aware quantization

Results:

- 22x model size reduction
- 10x latency reduction



Efficient Few-Shot Learning Without Prompts

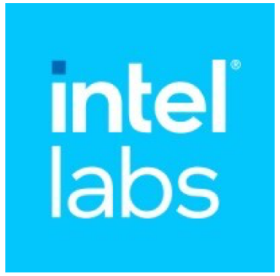
Lewis Tunstall[1], Nils Reimers[2], Unso Eun Seo Jo[1], Luke Bates[3], Daniel Korat[4], Moshe Wasserblat[4], Oren Pereg[4]

[1]Hugging Face [2]cohere.ai

[3]Ubiquitous Knowledge Processing Lab, Technical University of Darmstadt

[4]Emergent AI Lab, Intel Labs

\+



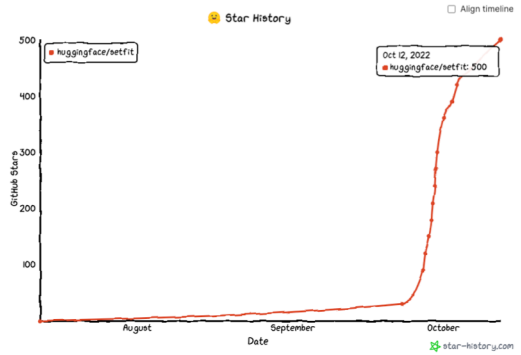
NLP/EAI Lab

\+

UBIQUITOUS
KNOWLEDGE
PROCESSING

🔮 Paper: https://arxiv.org/abs/2209.11055
✍️ Blog: https://huggingface.co/blog/setfit
🔥 Code: https://github.com/huggingface/setfit

Shout-out



150000+ downloads

rubrix.readthedocs.io

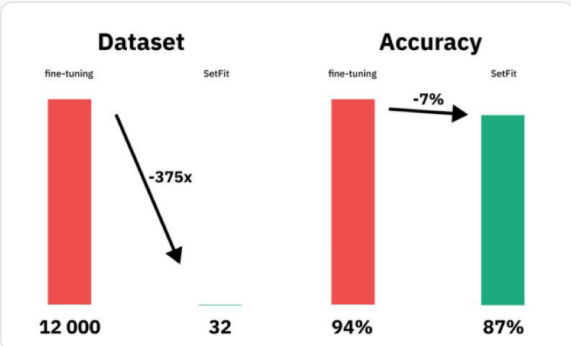
Few-shot classification with SetFit and a custom dataset

SetFit is an exciting open-source package for few-shot classification developed by teams at Hugging Face and Intel Labs. You can read all about it on the project repository. To showcase how powerfu... (485 kB)

linkedin.com

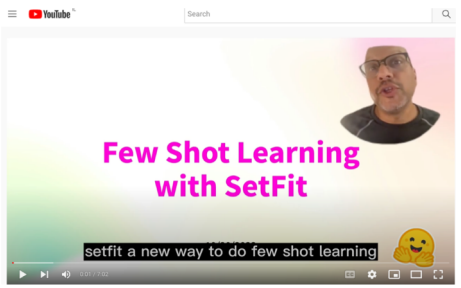
Philipp Schmid on LinkedIn: Outperform OpenAI GPT-3 with SetFit for text-classification

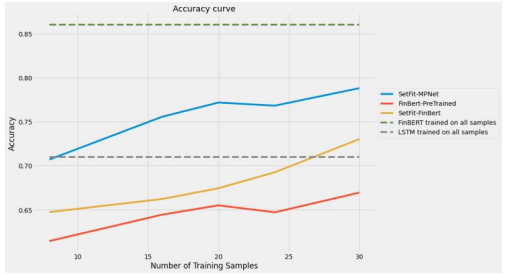
OpenAI GPT-3 is one of the most known Large Language Models (LLM) today. Those LLMs are achieving good performance in classification, where you only have a few... (18 kB)



"This is huge! SetFit will help so many companies to get started with transformers, without the need to label a lot of data and compute power."

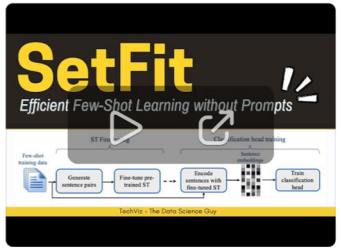
"Now anyone can build a few shot classifier like a pro on only a handful of labelled data"





YouTube | TechViz - The Data Science Guy

SetFit - Efficient Few-Shot Learning Without Prompts (Research Paper Walkthrough)



YouTube | Rohan-Paul-AI

SetFit (Sentence Transformer Fine-tuning) - Paper Discussion with Code | NLP | Machine Learning

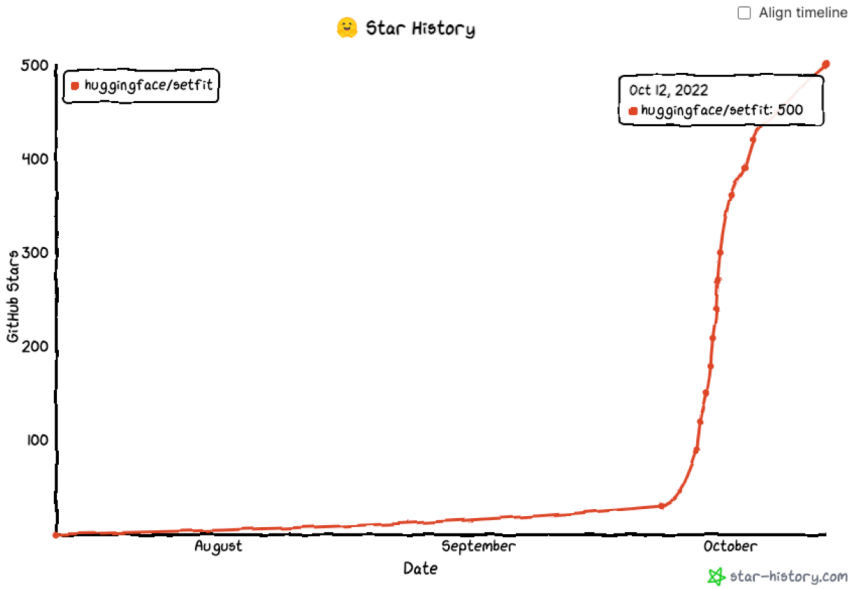
SetFit

150000+ installs

Outperform GPT-3, 1600x smaller

~20k faster inference

2min fine-tuning on SPR



Service providers that offer SetFit

Argilla mantis witty.works

NLP SUMMIT

intel labs

April 2023