Tiny CamemBERT and CamemBERT -

A Comparative Study of Two French BERT-Based Models

Bouchet Elliot, Brun Samuel, Gauci Corentin, Lithaud Alexandre

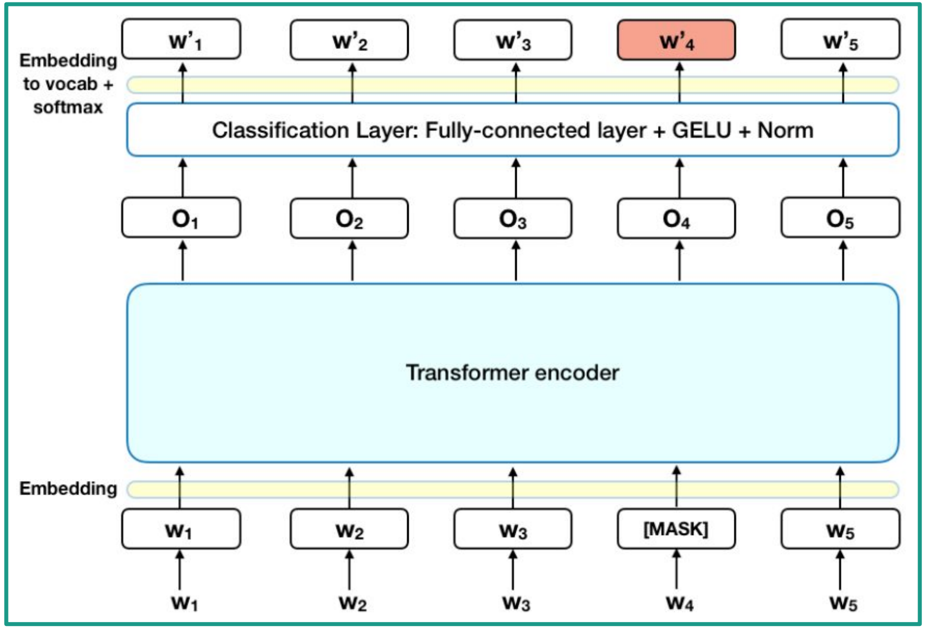
POLYTECH° GRENOBLE

LINKÖPING UNIVERSITY

Tiny BERT and CamemBERT

LINKÖPING
UNIVERSITY

The BERT Model



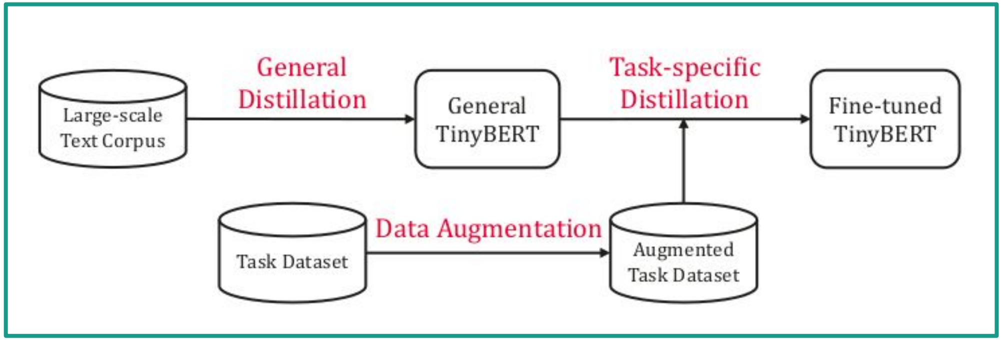
- BERT is pre-trained on large corpora of text data using unsupervised learning tasks like masked language modeling and next sentence prediction

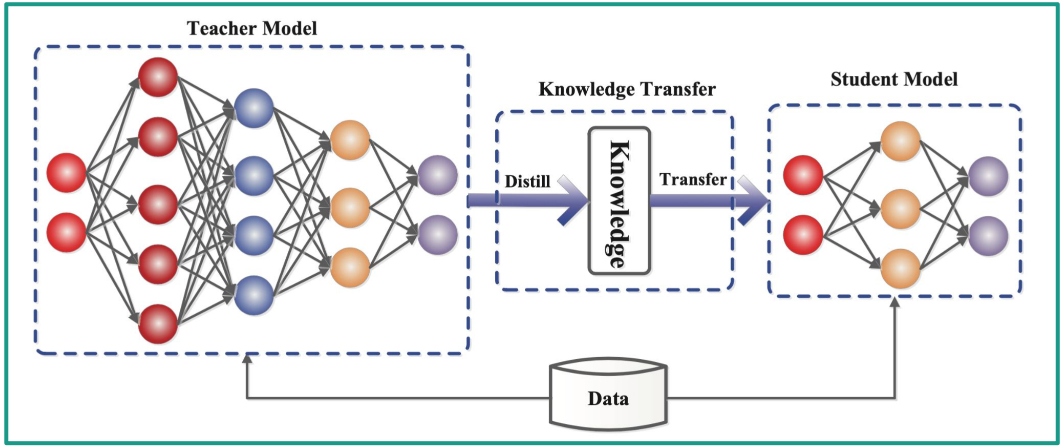
- BERT is built upon the Transformer architecture, which enables it to capture long-range dependencies in text efficiently through self-attention mechanisms.

The CamemBERT Model

- CamemBERT is a state-of-the-art language model for French based on the RoBERTa architecture
- A model that aims to investigate the feasibility of training monolingual Transformer-based language models for other languages
- Shows that a relatively small web crawled dataset (4GB) leads to results that are as good as those obtained using larger datasets (130+GB).

The arrangements of TinyBERT

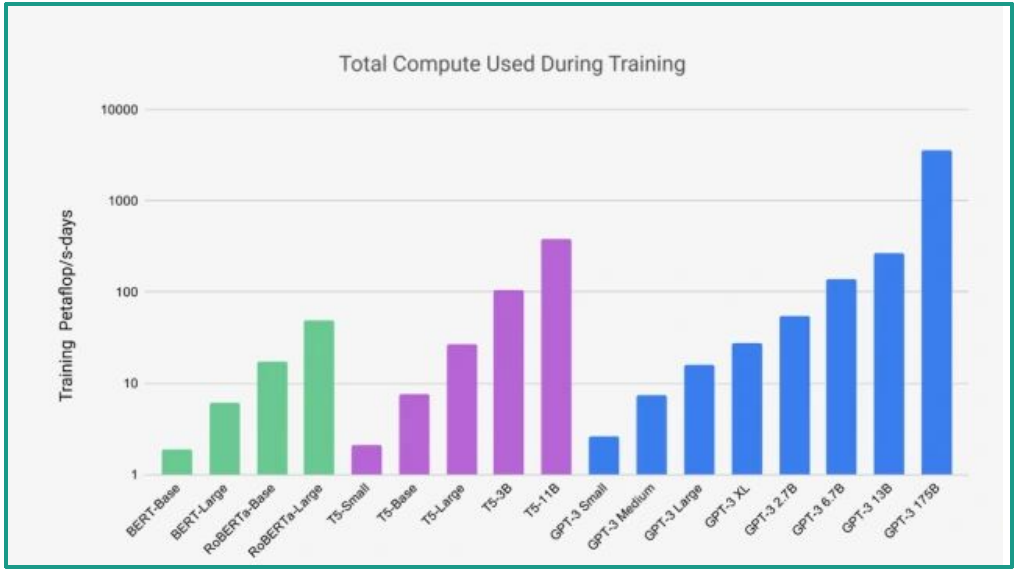


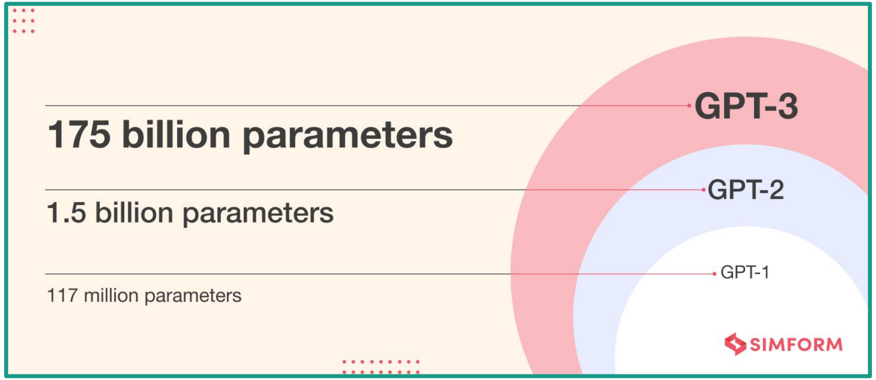


- Use a technique of model compression known as Knowledge Distillation.
- Use a teacher in order to transfer the linguistic knowledge with fewer parameters.
- Faster inference speed and overall steady accuracy.

Context of our project

The Problem of Newer and Bigger models



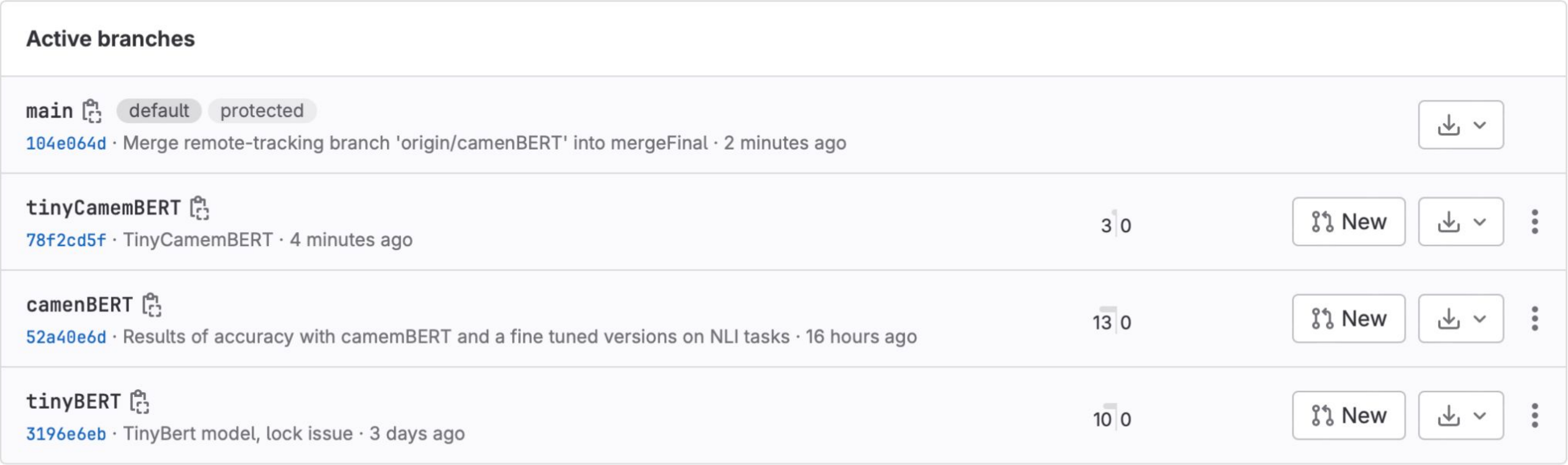


Sam Altman Seeks Trillions of Dollars to Reshape Business of Chips and AI

Organisation

LINKÖPING
UNIVERSITY

GIT


Active branches
main
default
protected
104e064d · Merge remote-tracking branch 'origin/camenBERT' into mergeFinal · 2 minutes ago
tinyCamemBERT
78f2cd5f · TinyCamemBERT · 4 minutes ago
3 0
New
camenBERT
52a40e6d · Results of accuracy with camemBERT and a fine tuned versions on NLI tasks · 16 hours ago
13 0
New
tinyBERT
3196e6eb · TinyBert model, lock issue · 3 days ago
10 0
New

Experiments with CamemBERT

LINKÖPING UNIVERSITY

Exploration of CamemBERT

For CamemBERT we aimed to reproduce the Natural Language Inference performances described in the paper. We used XNLI which is a subset of a few thousand examples from MNLI.

P[a]	A senior is waiting at the window of a restaurant that serves sandwiches.	Relationship
H[b]	A person waits to be served his food.	Entailment
	A man is looking to order a grilled cheese sandwich.	Neutral
	A man is waiting in line for the bus.	Contradiction

[a] P, Premise.
[b] H, Hypothesis.

Example of Natural language Inference

Model	Acc.	#Params
mBERT (Devlin et al., 2019)	76.9	175M
$XLM_{MLM\text{-}TLM}$ (Lample and Conneau, 2019)	80.2	250M
$XLM\text{-}R_{BASE}$ (Conneau et al., 2019)	80.1	270M
CamemBERT (fine-tuned)	**82.5**	110M
Supplement: LARGE models		
$XLM\text{-}R_{LARGE}$ (Conneau et al., 2019)	85.2	550M
$CamemBERT_{LARGE}$ (fine-tuned)	**85.7**	335M

Table showing NLI accuracy on the French XNLI test set

CamemBERT NLI	Accuracy
Before fine-tuning	33.25%
After fine-tuning	81.48%

CamemBERT performances with sentiment analysis

French Twitter Sentiment Analysis

1.5 million tweets in French and their sentiment.
label: Polarity of the tweet (0 = negative, 1 = positive)

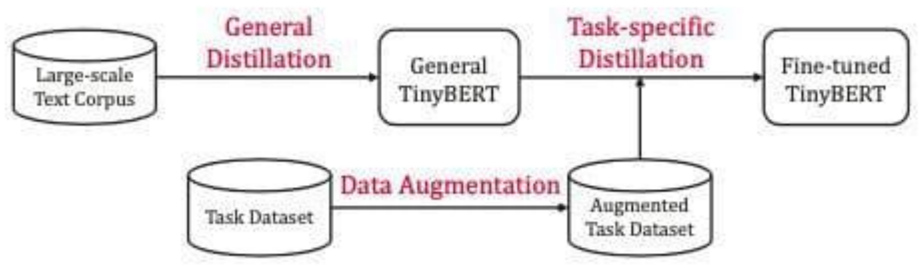
CamemBERT (sentiment analysis)	Loss	Accuracy
Before fine-tuning	0.6937	49.43%
After fine-tuning	0.3702	83.62%
CamemBERT accuracy in the paper concerning NLI	**Loss**	**Accuracy**
After fine-tuning	...	82.50%

Experiments with TinyBERT

LINKÖPING
UNIVERSITY

Exploration of TinyBERT : Source Code

How to use :



Steps :

1) pregenerate_training_data.py
2) general_distill.py -> pre-train
3) data_augmentation.py
4) task_distill.py -> train & evaluate

Source :
https://github.com/huawei-noah/Pretrained-Language-Model/tree/master/TinyBERT

Advantages :

- Precise control over the tools
- Choice of datasets/parts of datasets
- Architecture strictly following those described in the paper

Disadvantages :

- Heavy and complex code
- Obsolete code (last update 3 years ago)
- Very little documentation
- Software and hardware requirements

Exploration of TinyBERT : Hugging Face Version

Advantages :

- Classic hugging face interface
- Easily compatible with notebooks
- Already pre-train

Disadvantages :

- Reduced architecture control

Model Name :
“huawei-noah/TinyBERT_General_4L_312D”

Hugging Face

Hugging Face Provide :

- Tokenizer (english version)
- Model (pre-train on some data)
- Easy access to databases

Experiments with TinyCamemBERT

TinyCamemBERT : Our buildings

Already Pre-trained Version :

- pre-train on english data

Results :

Parameters : 14 M
Time : 20 minutes

TinyBERT (sentiment analysis)	Accuracy
Before fine-tuning	49.23%
After fine-tuning	79.62%

Our Pre-trained Version :

- pre-train on few french data
  > same data as CamemBERT

Results :

<u>pre-train:</u>

1k data, 15 min

Parameters : 14 M
Time : 25 minutes

TinyCamemBERT (sentiment analysis)	Accuracy
Before fine-tuning	50.20%
After fine-tuning	75.62%

CamemBERT Vs TinyCamemBERT

LINKÖPING
UNIVERSITY

Comparison CamemBERT & TinyBERT

TinyCamemBERT (sentiment analysis)	Accuracy	CamemBERT (sentiment analysis)	Accuracy
Before fine-tuning	50.20%	Before fine-tuning	49.43%
After fine-tuning	75.62%	After fine-tuning	83.62%

TinyBERT (sentiment analysis)	Accuracy
Before fine-tuning	49.23%
After fine-tuning	79.62%

Differences :

- Training time
- Datasets
- Tokenizer

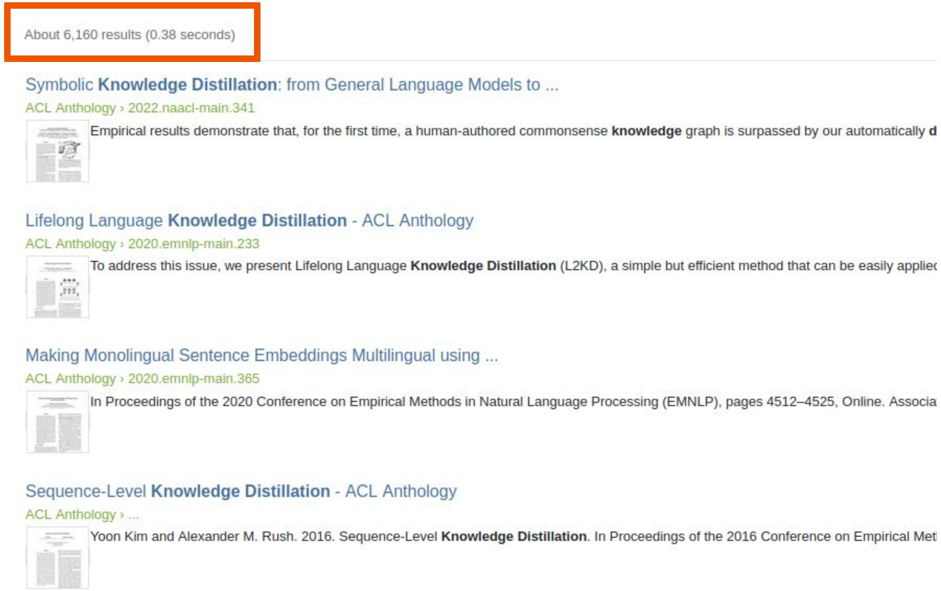
Conclusion

LINKÖPING UNIVERSITY

Future of TinyBERT

Knowledge distillation remains a vibrant area of research in machine learning, with ongoing efforts to develop more effective distillation techniques and apply them to a wide range of tasks and domains

I today's experience of NLP, it is likely that knowledge distillation and model compression techniques will be very important



The Other Model Compression Techniques

- Quantization
- Weights Pruning
- Low-rank approximation
- Sparse matrices

Y. Gong, L. Liu, M. Yang, and L. Bourdev. 2014. Compressing deep convolutional networks using vector quantization. arXiv preprint arXiv:1412.6115.

S Han, J. Pool, J. Tran, and W. Dally. 2015. Learning both weights and connections for efficient neural network. In NIPS.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, & Weizhu Chen. (2021). LoRA: Low-Rank Adaptation of Large Language Models.

….

Bibliography

- Martin, L., Muller, B., Ortiz Suárez, P., Dupont, Y., Romary, L., Clergerie, É., Seddah, D., & Sagot, B. (2020). CamemBERT: a Tasty French Language Model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

- Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, & Qun Liu. (2020). TinyBERT: Distilling BERT for Natural Language Understanding.

- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, & Dario Amodei. (2020). Language Models are Few-Shot Learners.

	Claude 3 Opus	Claude 3 Sonnet	Claude 3 Haiku	GPT-4	GPT-3.5	Gemini 1.0 Ultra	Gemini 1.0 Pro
Undergraduate level knowledge *MMLU*	86.8% 5 shot	79.0% 5-shot	75.2% 5-shot	86.4% 5-shot	70.0% 5-shot	83.7% 5-shot	71.8% 5-shot
Graduate level reasoning *GPQA, Diamond*	50.4% 0-shot CoT	40.4% 0-shot CoT	33.3% 0-shot CoT	35.7% 0-shot CoT	28.1% 0-shot CoT	—	—
Grade school math *GSM8K*	95.0% 0-shot CoT	92.3% 0-shot CoT	88.9% 0-shot CoT	92.0% 5-shot CoT	57.1% 5-shot	94.4% Maj1@32	86.5% Maj1@32
Math problem-solving *MATH*	60.1% 0-shot CoT	43.1% 0-shot CoT	38.9% 0-shot CoT	52.9% 4-shot	34.1% 4-shot	53.2% 4-shot	32.6% 4-shot
Multilingual math *MGSM*	90.7% 0-shot	83.5% 0-shot	75.1% 0-shot	74.5% 8-shot	—	79.0% 8-shot	63.5% 8-shot
Code *HumanEval*	84.9% 0-shot	73.0% 0-shot	75.9% 0-shot	67.0% 0-shot	48.1% 0-shot	74.4% 0-shot	67.7% 0-shot
Reasoning over text *DROP, F1 score*	83.1 3-shot	78.9 3-shot	78.4 3-shot	80.9 3-shot	64.1 3-shot	82.4 Variable shots	74.1 Variable shots
Mixed evaluations *BIG-Bench-Hard*	86.8% 3-shot CoT	82.9% 3-shot CoT	73.7% 3-shot CoT	83.1% 3-shot CoT	66.6% 3-shot CoT	83.6% 3-shot CoT	75.0% 3-shot CoT
Knowledge Q&A *ARC-Challenge*	96.4% 25-shot	93.2% 25-shot	89.2% 25-shot	96.3% 25-shot	85.2% 25-shot	—	—
Common Knowledge *HellaSwag*	95.4% 10-shot	89.0% 10-shot	85.9% 10-shot	95.3% 10-shot	85.5% 10-shot	87.8% 10-shot	84.7% 10-shot

Credit: Anthropic