Knowledge Distillation with a focus on DistilBert

Mayank Musaddi
Sprinklr Intern ML Team

Flow

- Motivation
- Knowledge Distillation
  - vs Transfer Learning
  - Student Teacher Paradigm
  - Dark Knowledge
  - Temperature
  - Learning Architecture
- DistilBert
  - Loss
  - Performance
  - Experiments

Motivation

Lonnngggg Inference Time!

Deeper model were giving more accuracy
but..

Deeper the model, higher the inference time

Resnet 11M

GPT 117M

BERT 345M

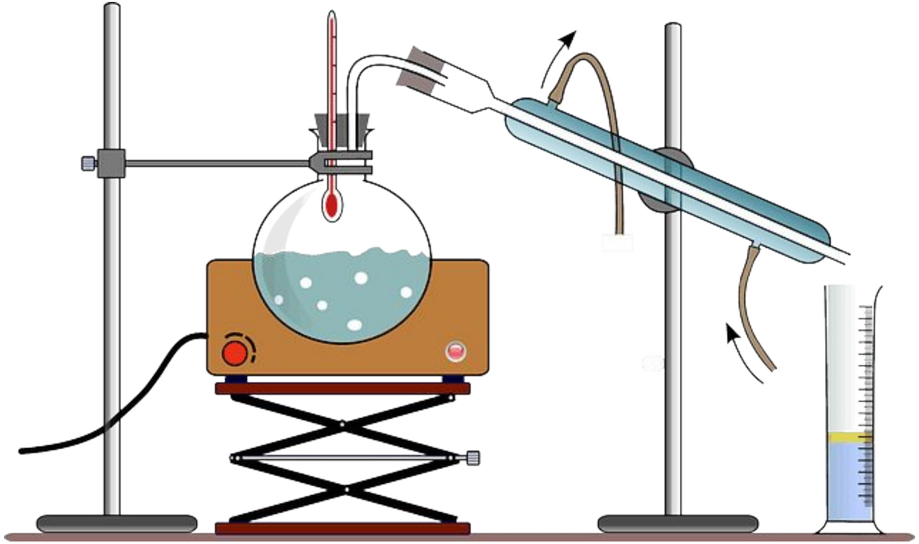
XLNET 340M

Ensemble of Models 😵

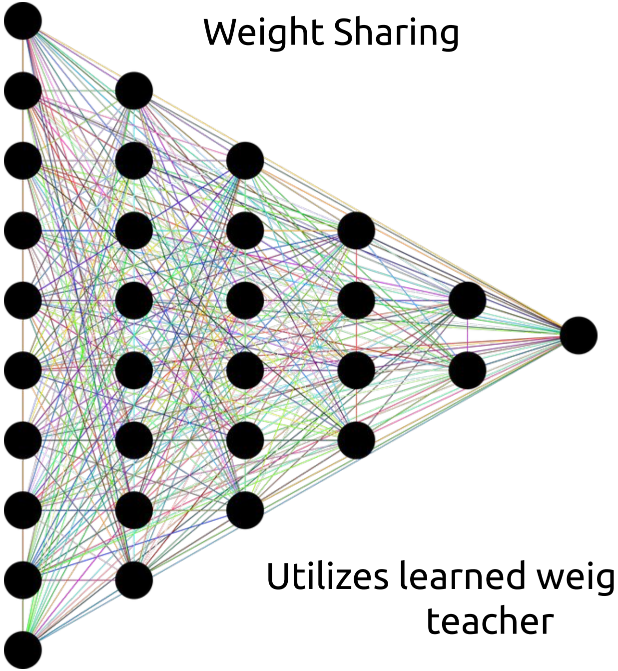
Tradeoff Between Accuracy and Inference Time

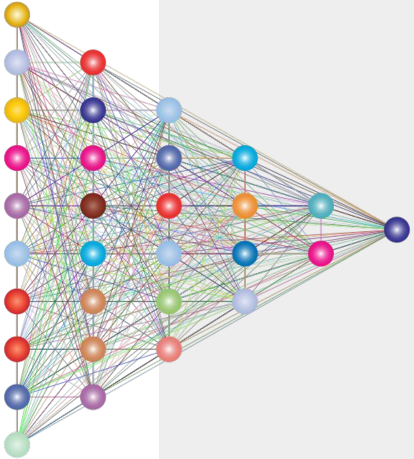
Smaller Model with Higher Accuracy

Knowledge Distillation : Smaller Model learns from Larger Model



Transfer Learning vs Knowledge Distillation

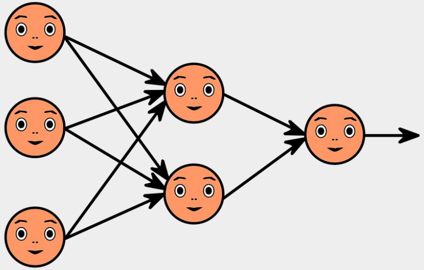




Weight Sharing

Utilizes learned weights of teacher

Student (a smaller model) learns from Teacher



Utilizes the output of Teacher when training

Significance of Output: Dark Knowledge

Softmax Gives Probability of Output Classes

$$\sigma(\vec{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$$

```
Input: ['[CLS]', 'i', 'think', 'this', 'is', 'the', 'beginning', 'of', 'a', 'beautiful', '[MASK]', '.', '[SEP]']
```

Output	Label
Rank 0 – Token: day – Prob: 0.21348	1
Rank 1 – Token: life – Prob: 0.18380	0
Rank 2 – Token: future – Prob: 0.06267	0
Rank 3 – Token: story – Prob: 0.05854	0
Rank 4 – Token: world – Prob: 0.04935	0
Rank 5 – Token: era – Prob: 0.04555	0
Rank 6 – Token: time – Prob: 0.03210	0
Rank 7 – Token: year – Prob: 0.01722	0
Rank 8 – Token: history – Prob: 0.01663	0
Rank 9 – Token: summer – Prob: 0.01335	0
Rank 10 – Token: adventure – Prob: 0.01233	0
Rank 11 – Token: dream – Prob: 0.01209	0
Rank 12 – Token: moment – Prob: 0.01129	0
Rank 13 – Token: night – Prob: 0.01084	0
Rank 14 – Token: beginning – Prob: 0.00937	0
Rank 15 – Token: season – Prob: 0.00664	0
Rank 16 – Token: journey – Prob: 0.00621	0
Rank 17 – Token: period – Prob: 0.00553	0
Rank 18 – Token: relationship – Prob: 0.00517	0
Rank 19 – Token: thing – Prob: 0.00508	0

Compared with one-hot encoded labels

Cross Entropy Loss
:
-1*log(0.21348)

Distillation Loss

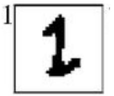
https://arxiv.org/pdf/1503.02531.pdf

Student instead of comparing its result from one-hot encoded labels, can now compare it with the output of the teacher

Distillation Loss

$$L = -\sum_i t_i * log(s_i)$$

With **t** the logits from the teacher and **s** the logits of the student



Teacher Output	Student Output	One-Hot Labels
0.00	0.08	0
0.52	0.27	1
0.21	0.21	0
0.02	0.03	0
0.04	0.08	0
0.03	0.10	0
0.01	0.01	0
0.12	0.17	0
0.00	0.03	0
0.05	0.02	0

Teacher Output ⟺ Student Output: Distillation Loss

Student Output ⟺ One-Hot Labels: Student Loss

Temperature

https://arxiv.org/pdf/1503.02531.pdf

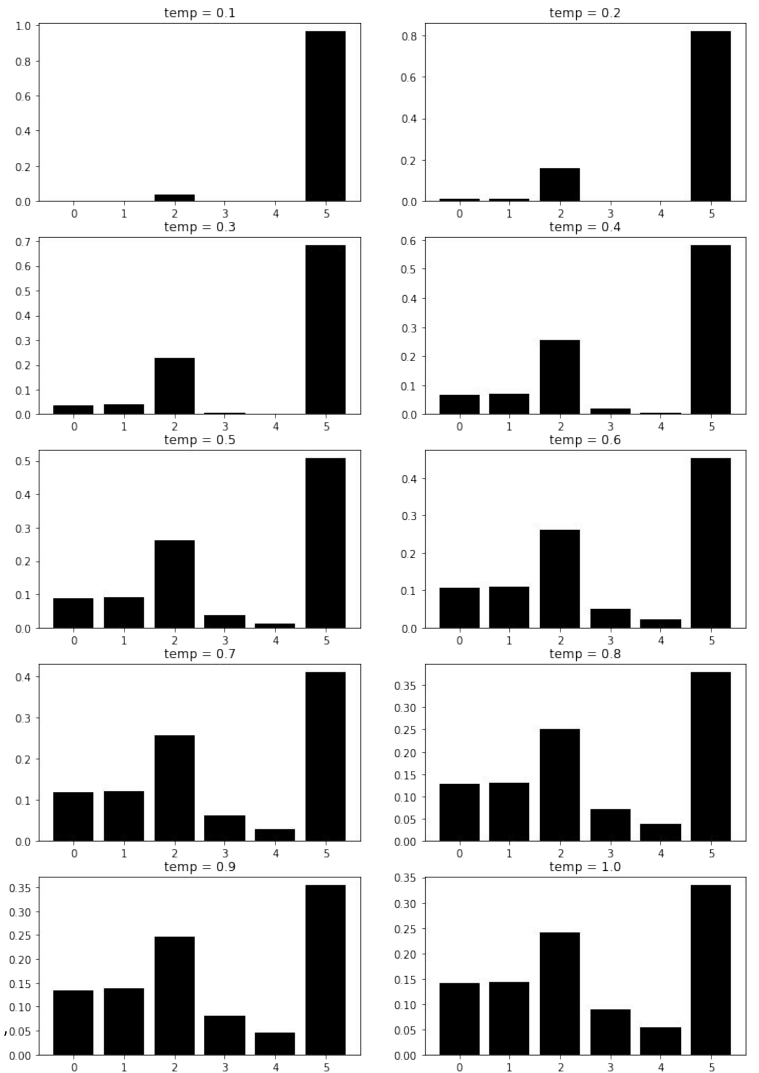
Additional Parameter Temperature (T) is applied in the softmax function to make the output more informative by making the outputs softer

$$p_i = \frac{exp(z_i/T)}{\sum_j exp(z_j/T)}$$

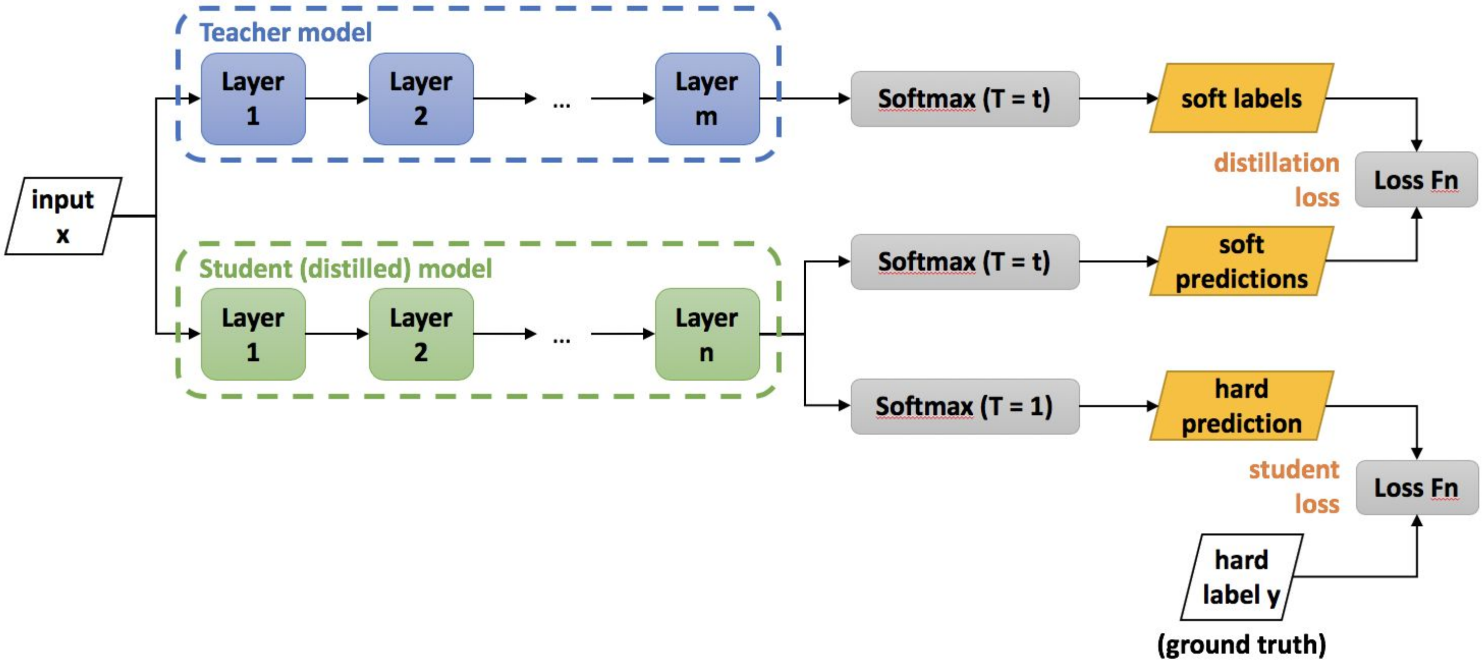
T is the temperature parameter.

$$p_i = 1 \, (z_i = max(z_j) \,)$$
$$T \to 0$$

$$p_i = 1/J$$
$$T \to +\infty$$



Learning Architecture

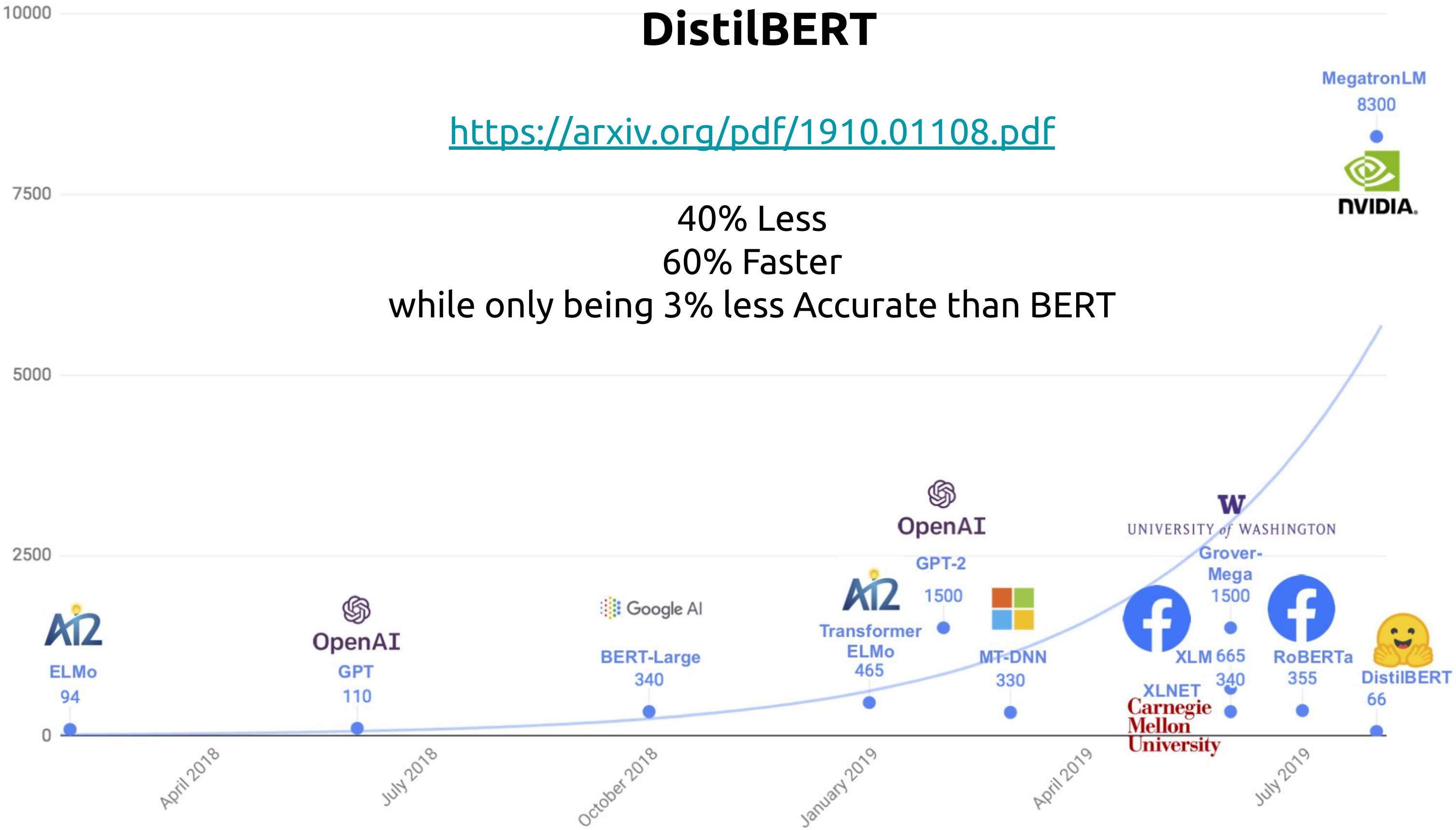


https://intellabs.github.io/distiller/knowledge_distillation.html

DistilBERT

https://arxiv.org/pdf/1910.01108.pdf

40% Less
60% Faster
while only being 3% less Accurate than BERT



Details

Architecture: Same as BERT except,
The token-type embeddings and the pooler are removed while the number of layers is reduced by a factor of 2

Dataset: Same as BERT (English Wikipedia and Toronto Book Corpus)

Initialisation: Student take one layer out of two from teacher

Additional
Large batches
Dynamic masking
no NSP

Experiment: Adding another step of distillation for SQuAD task during the adaptation phase
Fine-tuning DistilBERT on SQuAD
Teacher: BERT model previously fine-tuned on SQuAD

Loss

Learning Inductive Biases of Teacher:

Masked Language Modelling Loss L_{mlm}
Distillation Loss L_{ce}
Cosine Distance Loss L_{cos}

tend to align the directions of the student and teacher hidden states vectors

Ablation Study showed that MLM Loss has least impact

Table 4: **Ablation study.** Variations are relative to the model trained with triple loss and teacher weights initialization.

Ablation	**Variation on GLUE macro-score**
$\emptyset$ - L_{cos} - L_{mlm}	-2.96
L_{ce} - $\emptyset$ - L_{mlm}	-1.46
L_{ce} - L_{cos} - $\emptyset$	-0.31
Triple loss + random weights initialization	-3.69

🎢 Performance

Table 1: **DistilBERT retains 97% of BERT performance.** Comparison on the dev sets of the GLUE benchmark. ELMo results as reported by the authors. BERT and DistilBERT results are the medians of 5 runs with different seeds.

Model	**Score**	CoLA	MNLI	MRPC	QNLI	QQP	RTE	SST-2	STS-B	WNLI
ELMo	68.7	44.1	68.6	76.6	71.1	86.2	53.4	91.5	70.4	56.3
BERT-base	79.5	56.3	86.7	88.6	91.8	89.6	69.3	92.7	89.0	53.5
DistilBERT	77.0	51.3	82.2	87.5	89.2	88.5	59.9	91.3	86.9	56.3

Table 2: **DistilBERT yields to comparable performance on downstream tasks.** Comparison on downstream tasks: IMDb (test accuracy) and SQuAD 1.1 (EM/F1 on dev set). D: with a second step of distillation during fine-tuning.

Model	IMDb (acc.)	SQuAD (EM/F1)
BERT-base	93.46	81.2/88.5
DistilBERT	92.82	77.7/85.8
DistilBERT (D)	-	79.1/86.9

Table 3: **DistilBERT is significantly smaller while being constantly faster.** Inference time of a full pass of GLUE task STS-B (sentiment analysis) on CPU with a batch size of 1.

Model	# param. (Millions)	Inf. time (seconds)
ELMo	180	895
BERT-base	110	668
DistilBERT	66	410

Bonus

http://ralphtang.com/papers/deeplo2019.pdf
Ralph Tang et al.

- Student can be tiny
- Student Architecture doesn't matter
- Lots of data required for learning
- Semantic Knowledge is hard to distill (CoLA)
- Layer Width > Number of Layers

Task Specific vs Multi-task Knowledge Distillation
https://www.aclweb.org/anthology/2020.aacl-main.9.pdf
https://arxiv.org/pdf/1911.03588.pdf

Instead of Distilling on Individual Tasks, distilling on general language modelling problem seems to be a better approach as it captures semantic information

Thank you
Any QA?