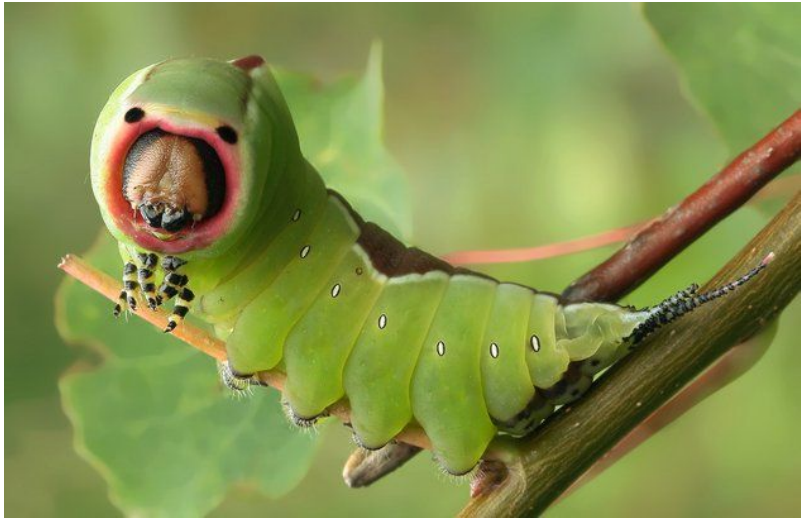
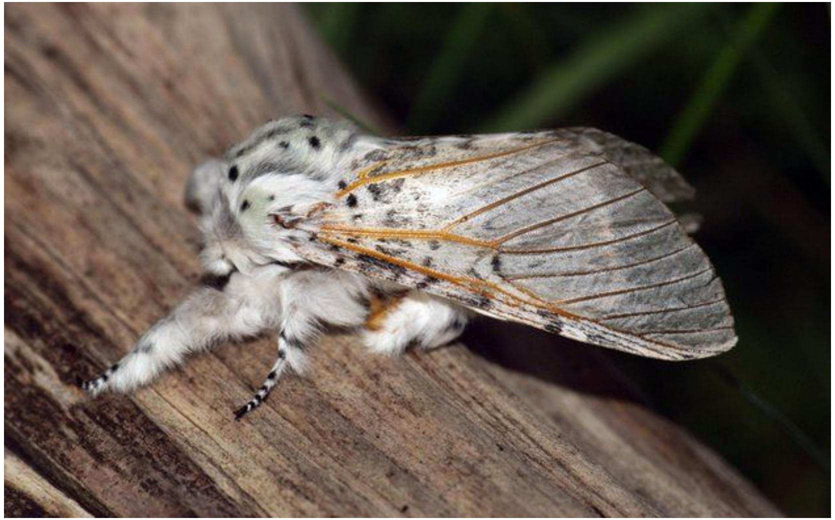
Lecture 15:
Knowledge Distillation

Radoslav Neychev

Extra: Knowledge Distillation

Cerura Vinula in caterpillar and butterfly forms



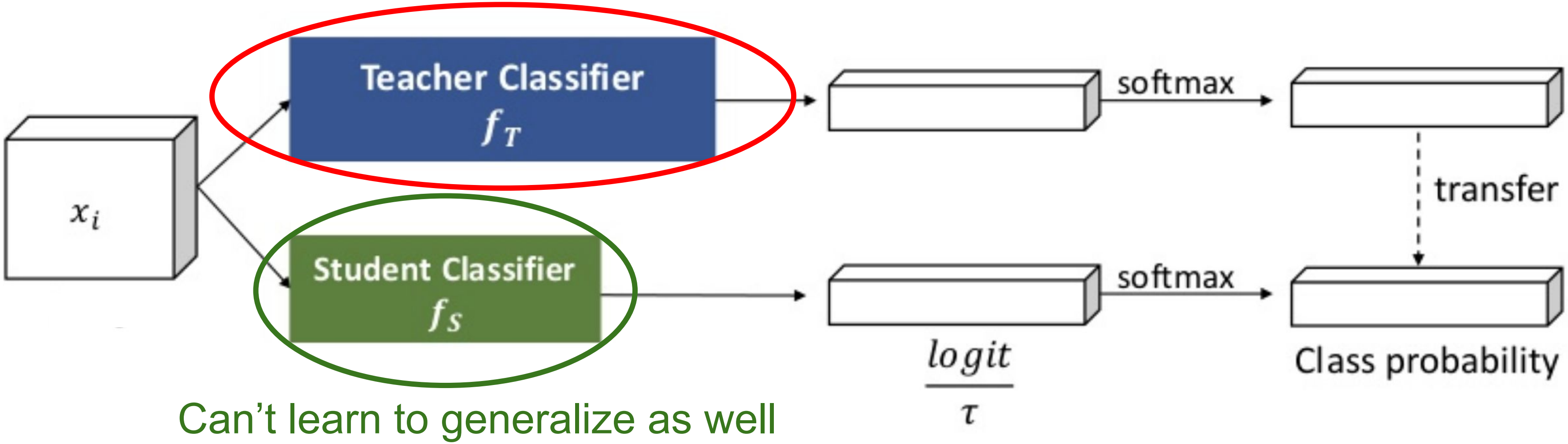


Do they have the same “life purpose”
and solve the same problems?

Image credits: commons.wikimedia.org (left), Ben Sale/Flickr.com (right), copied from
https://nat-geo.ru/nature/insects/bylo-stalo-kak-vyglyadyat-babochki-i-ikh-gusenitsy/

Knowledge distillation

Learns to generalize
and perform well



Can’t learn to generalize as well
as the teacher,
but still can perform well

Source: https://www.slideshare.net/NaverEngineering/relational-knowledge-distillation

Knowledge distillation

Denote **teacher** and **student** models.

Student model has logits z_i and corresponding probabilities q_i , derived with the softmax operation:

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

where *T* stays for the temperature.

Teacher model has logits v_i and corresponding probabilities p_i .

Source: Distilling the Knowledge in a Neural Network

Knowledge distillation

Let’s derive the cross-entropy gradient on **student** logits using the **teacher** predictions as targets:

$$\frac{\partial C}{\partial z_i} = \frac{1}{T}\left(q_i - p_i\right) = \frac{1}{T}\left(\frac{e^{z_i/T}}{\sum_j e^{z_j/T}} - \frac{e^{v_i/T}}{\sum_j e^{v_j/T}}\right)$$

If the temperature is high, the following equation takes place:

$$\frac{\partial C}{\partial z_i} \approx \frac{1}{T}\left(\frac{1 + z_i/T}{N + \sum_j z_j/T} - \frac{1 + v_i/T}{N + \sum_j v_j/T}\right)$$

Source: Distilling the Knowledge in a Neural Network

Knowledge distillation

Logits can be centered, so

$$\sum_j z_j = \sum_j v_j = 0$$

Then the gradient takes form:

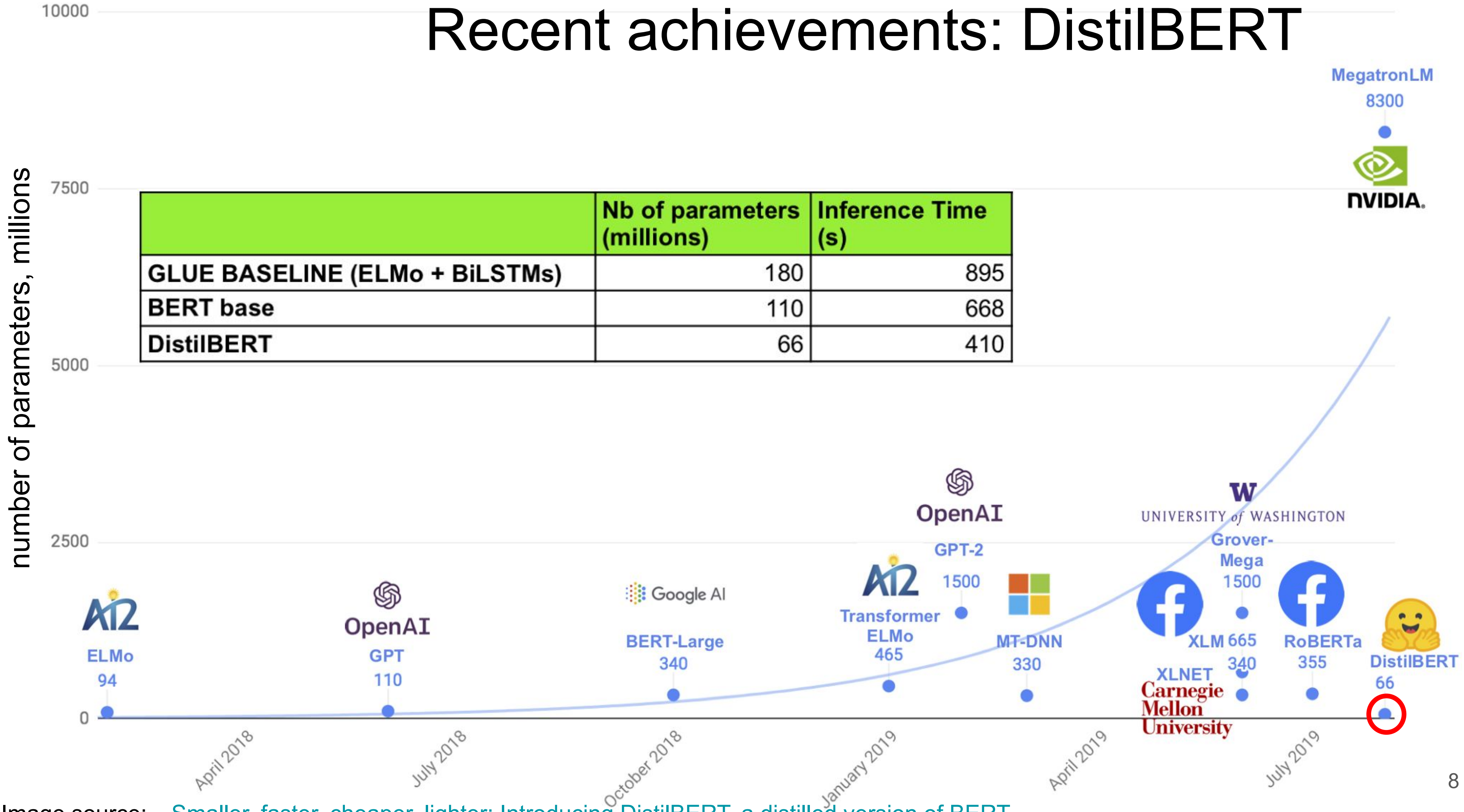
$$\frac{\partial C}{\partial z_i} \approx \frac{1}{T}\left(\frac{1 + z_i/T}{N + \sum_j z_j/T} - \frac{1 + v_i/T}{N + \sum_j v_j/T}\right) \approx \frac{1}{NT^2}\left(z_i - v_i\right)$$

Constant

$$\frac{dC}{dz_i} = \frac{1}{NT^2}(z_i - v_i) \sim (z_i - v_i) = M\frac{d(z_i - v_i)^2}{dz_i}$$

Source: Distilling the Knowledge in a Neural Network

Recent achievements: DistilBERT



	Nb of parameters (millions)	Inference Time (s)
GLUE BASELINE (ELMo + BiLSTMs)	180	895
BERT base	110	668
DistilBERT	66	410

Image source: Smaller, faster, cheaper, lighter: Introducing DistilBERT, a distilled version of BERT

Main ideas

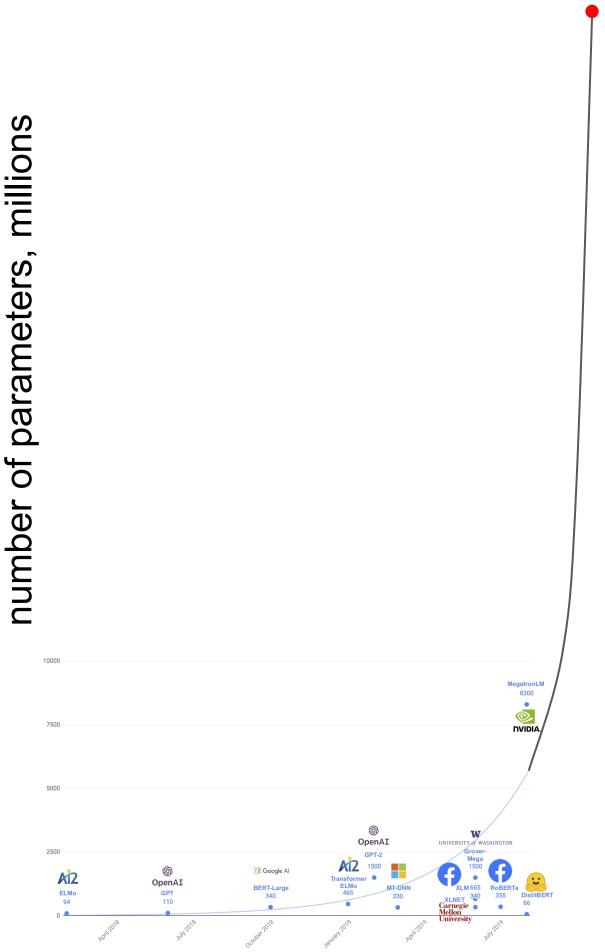
- DistilBERT is initialized from its teacher, BERT, by taking one layer out of two, leveraging the common hidden size.
    - *Comment: Training a sub-network is not only about the architecture. It is also about finding the right initialization for the sub-network to converge.*

- DistilBERT is trained on very large batches leveraging gradient accumulation (up to 4000 examples per batch), with dynamic masking and removed the next sentence prediction objective.
    - *Comment: the way BERT is trained is crucial for its final performance.*

- DistilBERT was trained on eight 16GB V100 GPUs for approximately three and a half days using the concatenation of Toronto Book Corpus and English Wikipedia (same data as original BERT).

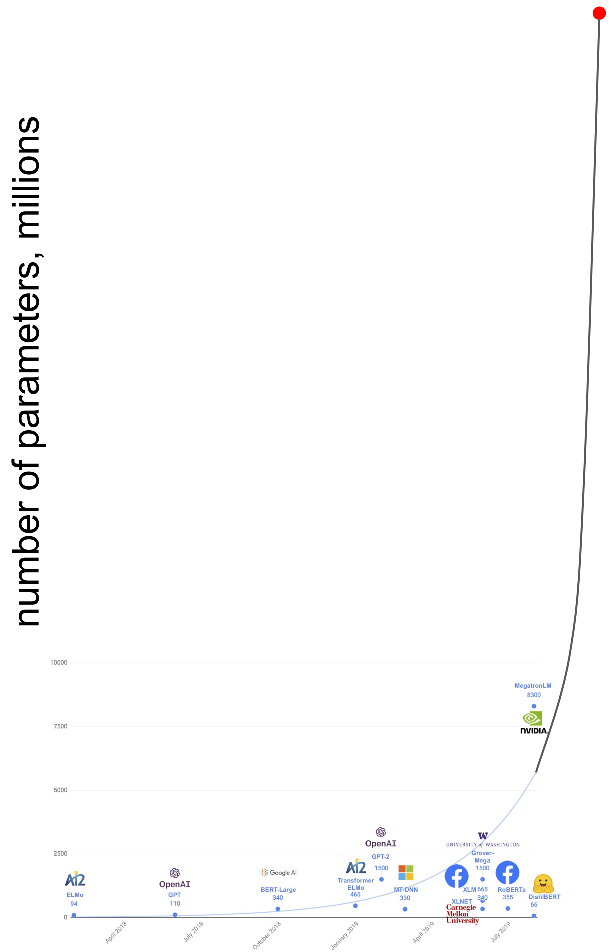
Image source: [Smaller, faster, cheaper, lighter: Introducing DistilBERT, a distilled version of BERT]

Recent achievements: GPT-3

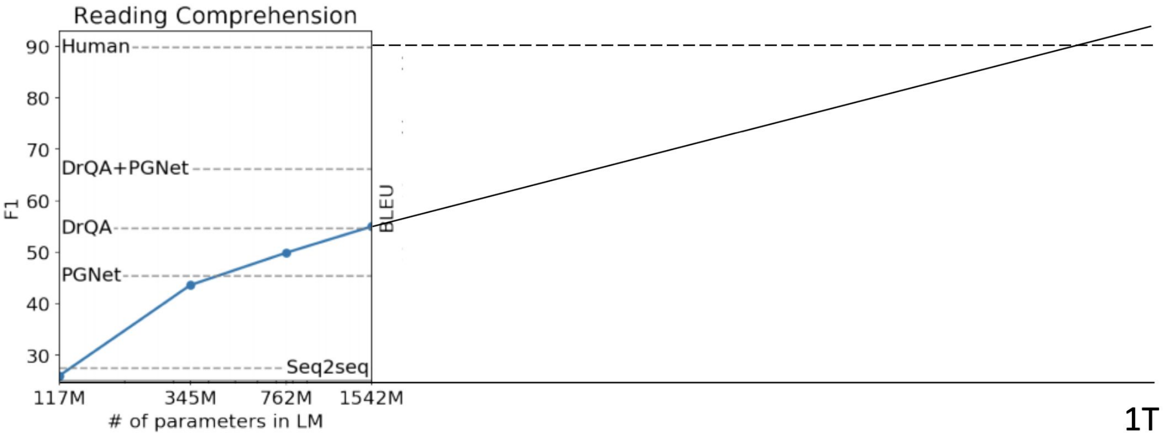


GPT-3, May 2020
175B parameters
(proportions are incorrect for visual sake)

Recent achievements: GPT-3



GPT-3, May 2020
175B parameters
(proportions are incorrect for visual sake)



Hypothesis from Stanford CS224n (2019) lecture 20