From BERT to TinyBERT:

Energy Considerations for Deep Learning in NLP and Solutions

C.Feng@AMC Oct, 04

Energy and Policy Considerations for Deep Learning in NLP

Strubell, Emma et al. “Energy and Policy Considerations for Deep Learning in NLP.” *ACL* (2019).

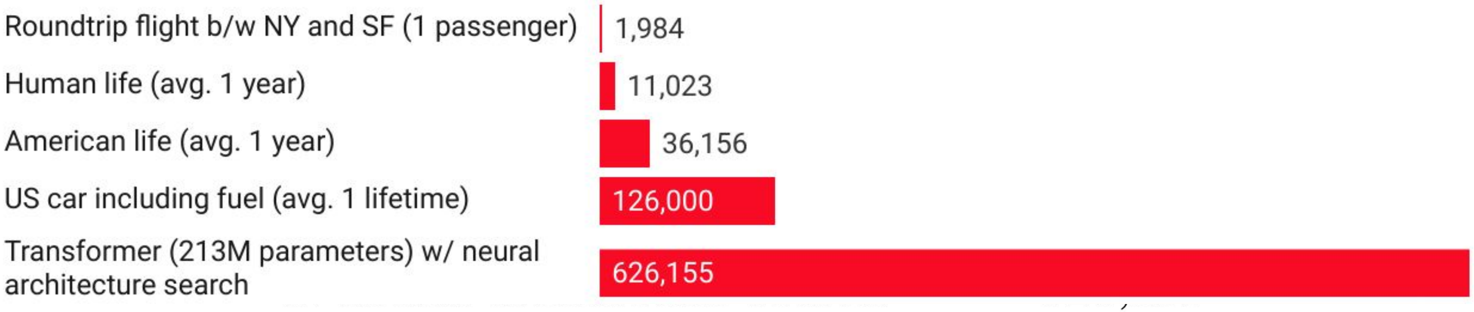
The larger the model training process, the more carbon dioxide is produced.

Model	Hardware	Power (W)	Hours	kWh·PUE	CO_2e	Cloud compute cost
$\text{Transformer}_{base}$	P100x8	1415.78	12	27	26	\$41–\$140
Transformer_{big}	P100x8	1515.43	84	201	192	\$289–\$981
ELMo	P100x3	517.66	336	275	262	\$433–\$1472
BERT_{base}	V100x64	12,041.51	79	1507	1438	\$3751–\$12,571
BERT_{base}	TPUv2x16	—	96	—	—	\$2074–\$6912
NAS	P100x8	1515.43	274,120	656,347	626,155	\$942,973–\$3,201,722
NAS	TPUv2x1	—	32,623	—	—	\$44,055–\$146,848
GPT-2	TPUv3x32	—	168	—	—	\$12,902–\$43,008

Table 3: Estimated cost of training a model in terms of CO_2 emissions (lbs) and cloud compute cost (USD).[7] Power and carbon footprint are omitted for TPUs due to lack of public information on power draw for this hardware.

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 a single AI model can emit as much carbon as five cars in their lifetimes.



Researchers should prioritize computationally efficient hardware and algorithms

TinyBERT: Distilling BERT for Natural Language Understanding

Jiao, Xiaoqi et al. “TinyBERT: Distilling BERT for Natural Language Understanding.” ArXiv abs/1909.10351 (2019): n. pag.

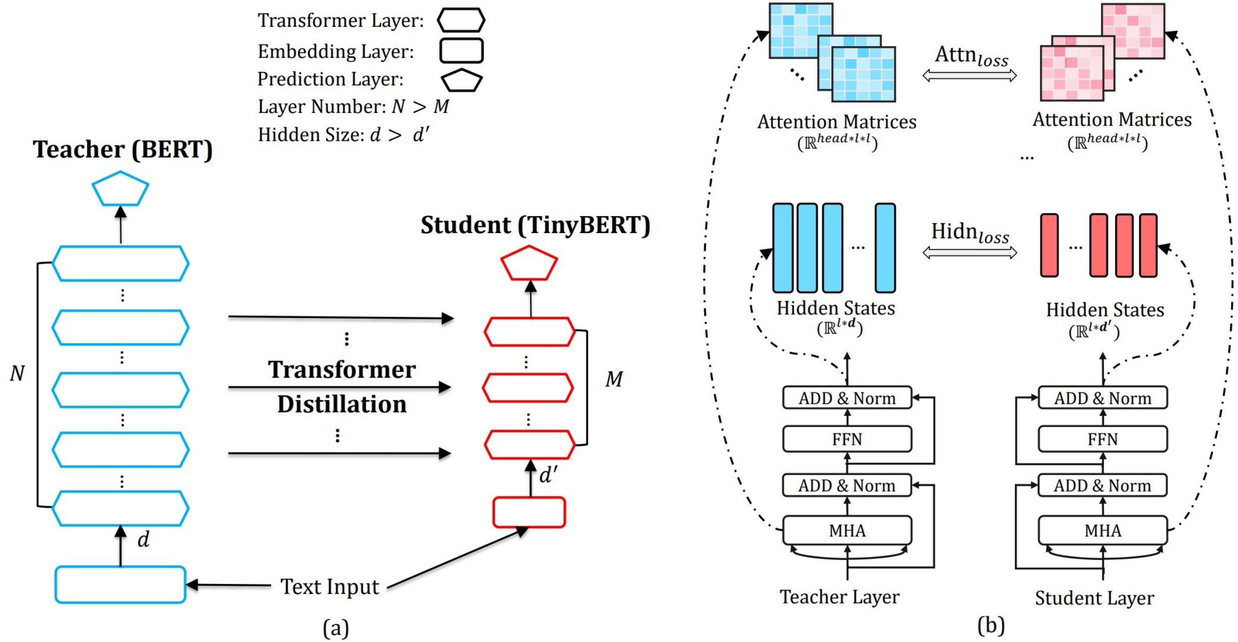


Figure 1: An overview of Transformer distillation: (a) the framework of Transformer distillation, (b) the details of Transformer-layer distillation consisting of Attn$_{loss}$(attention based distillation) and Hidn$_{loss}$(hidden states based distillation).

The first paper, from researchers at Huawei, produces a model called TinyBERT that is less than a seventh the size of the original and nearly 10 times faster. It also performs nearly as well in language understanding as the original.

Extreme Language Model Compression with Optimal Subwords and Shared Projections

Zhao, Sanqiang et al. “Extreme Language Model Compression with Optimal Subwords and Shared Projections.” ArXiv abs/1909.11687 (2019): n. pag.

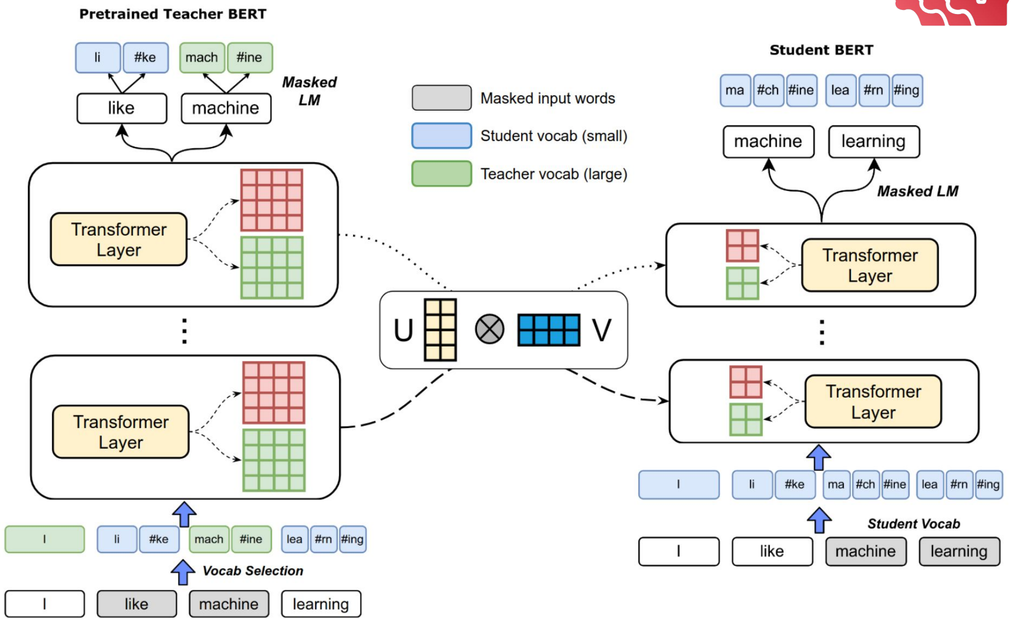


Figure 1: Knowledge Distillation on BERT with smaller student vocabulary. (Left) A pre-trained teacher BERT model with default BERT parameters (e.g., 30K vocab, 768 hidden state dimension). (Right) A student BERT model trained from scratch with smaller vocab (5K) and hidden state dimension (e.g., 48). During distillation, the teacher model randomly selects a vocabulary to segment each input word. The red and green square nexts to the transformer layers indicate trainable parameters for both the student and teacher models - note that our student models have smaller model dimensions. The projection matrices **U** and **V**, shown as having representative shapes, are shared across all layers for model parameters that have the same dimensions.

The second, from researchers at Google, produces another that is smaller by a factor of more than 60, but its language understanding is slightly worse compared with the original.