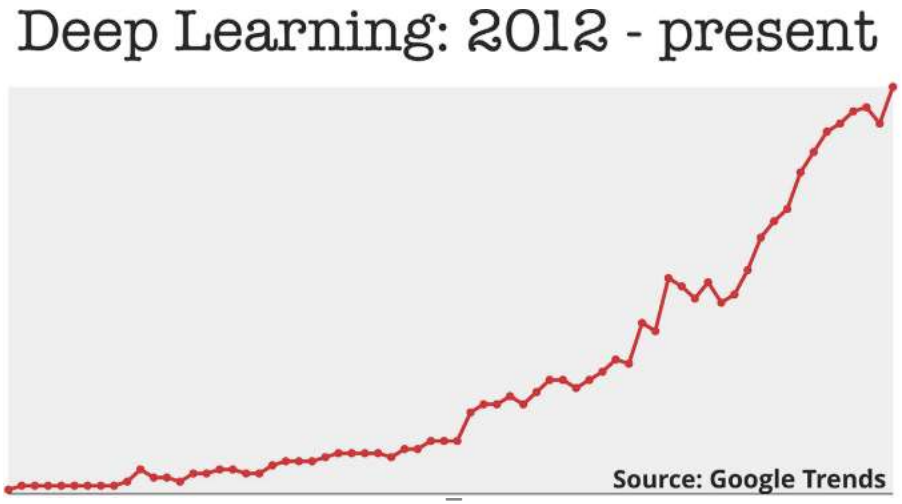
Landscape

Kyong-Ha Lee

A Few Quates

- “A breakthrough in machine learning would be worth ten Microsofts” Bill Gates, Chairman, Microsoft)
- “Machine learning is the next Internet” (Tony Tether, Director, DARPA)
- Machine learning is the hot new thing” (John Hennessy, President, Stanford University)
- “Web rankings today are mostly a matter of machine learning” (Prabhakar Raghavan, Dir. Research, Yahoo)



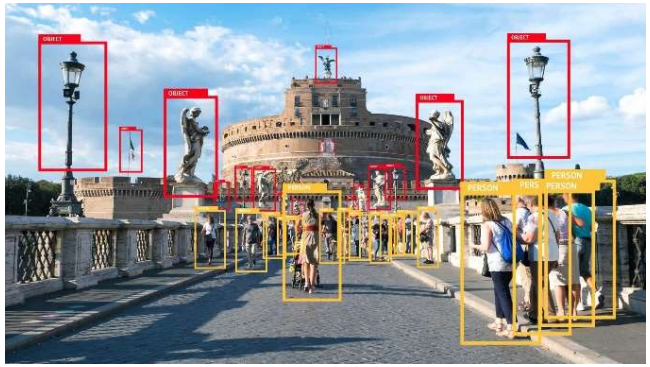
총논문 : 660 편

* 고성능컴퓨팅	25편 (3.79%)
* 국방소프트웨어	8편 (1.21%)
* 데이터베이스	43편 (6.52%)
* 모바일응용및시스템	39편 (5.91%)
* 사물인터넷	22편 (3.33%)
* 소프트웨어공학	27편 (4.09%)
* 스마트시티	18편 (2.73%)
* 언어공학	51편 (7.73%)
* 오픈소스소프트웨어	9편 (1.36%)
* 인공지능	233편 (35.30%)
* 전산교육시스템	10편 (1.52%)
* 정보보안및고신뢰컴퓨팅	43편 (6.52%)
* 정보통신	40편 (6.06%)
* 컴퓨터그래픽스및상호작용	31편 (4.70%)
* 컴퓨터시스템	54편 (8.18%)
* 컴퓨터이론	3편 (0.45%)
* 프로그래밍언어	4편 (0.61%)

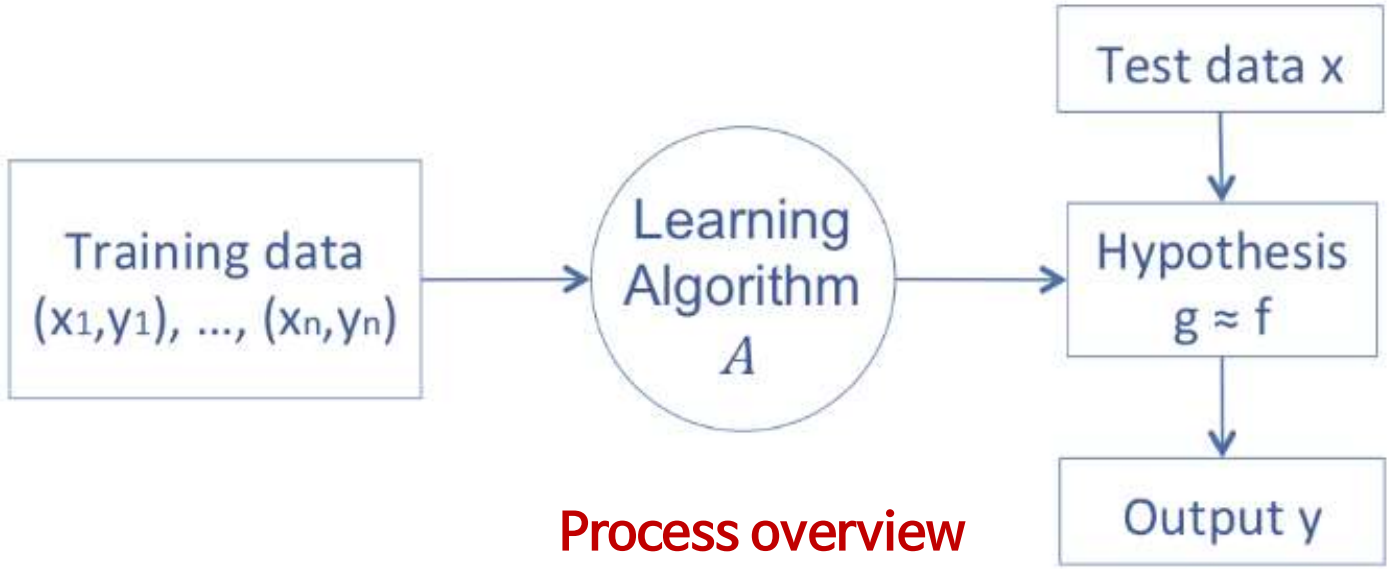
Submission stats. in KCC 2020

Sample applications

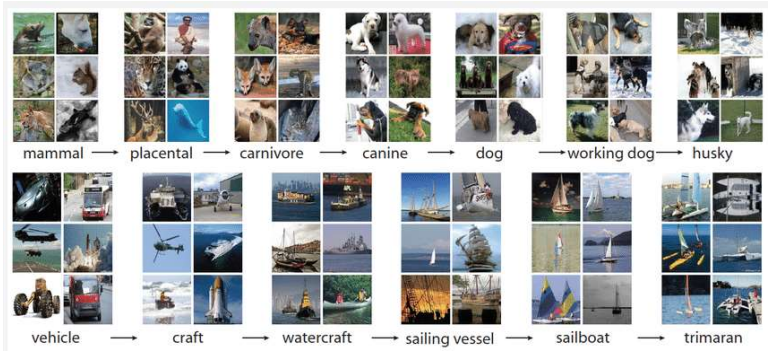
- Voice recognition
  - e.g., Apple SIRI, Amazon echo, …
- Machine translation
  - Google translate, Naver papago, …
- Visual recognition
  - Face recognition, license plate recognition, automated surveillance cameras, …
- Robotics
- Recommender systems
- Automated driving systems



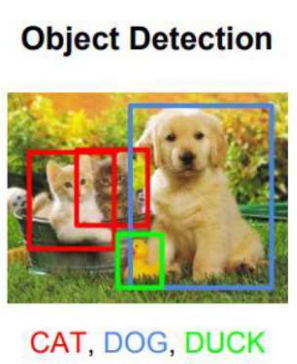
Overview: Deep Learning



Process overview

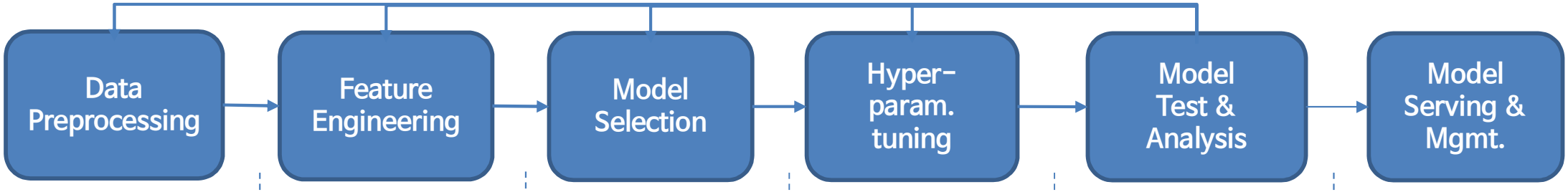


Massive training (labeled) data



Applications

Challenge 1: Human-In-The-Loop Process



- Data transforming
- Data cleansing
- Data labeling
- Problem
  - Inconsistency
  - Noise
  - Missing values
  - Imbalanced data
  - Outliers
  - Highly correlated data
  - Target leakage
  - Data size vs. memory size

- Estimating feature importance
- Feature selection
- Data vectorization
- Removing Highly correlated features
- Domain-specific features like POS in NLP

- Selecting a proper model from model collection
- Using pre-trained model
- Model design/ modification

- Selecting right values for various tuning knobs :
  - learning rate
  - regularization (weight decay)
  - batch size
  - # of epoch
  - dropout rate
  - optimizer
  - no. of layers
  - no. of units
  - act. functions

  . . .

- Model accuracy
- Train loss
- Model fairness

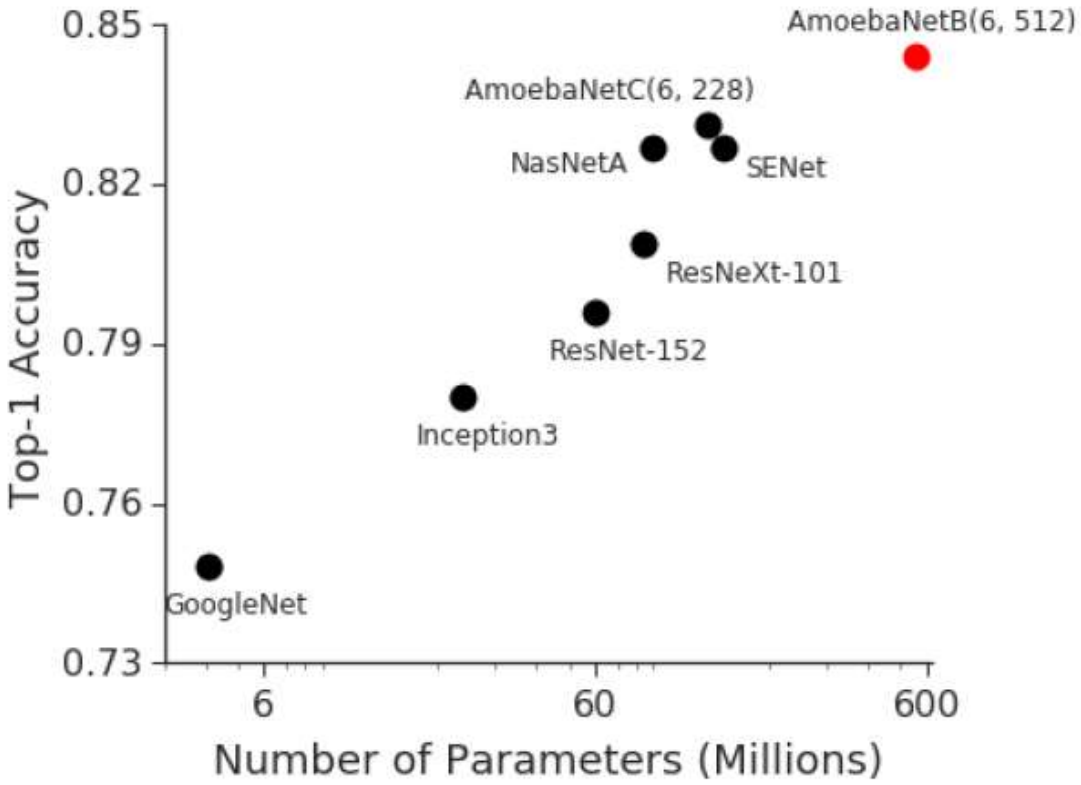
. . . .

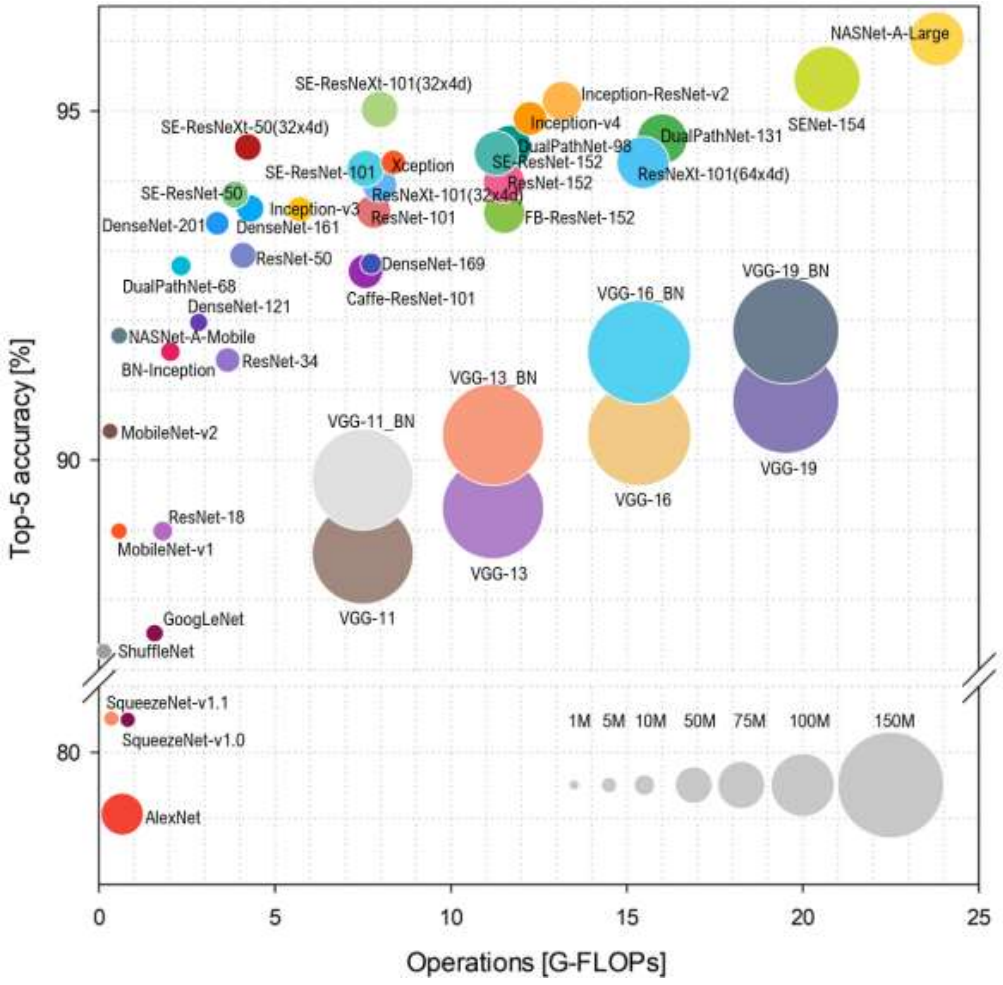
- Deployment
- Metadata mgmt.
- Model version control & mgmt.
- Lightening model for Mobile env.

DL developments require many *Iterative* tasks and human interventions

Challenge 2: Size Matters!

- CNN-based image classifiers



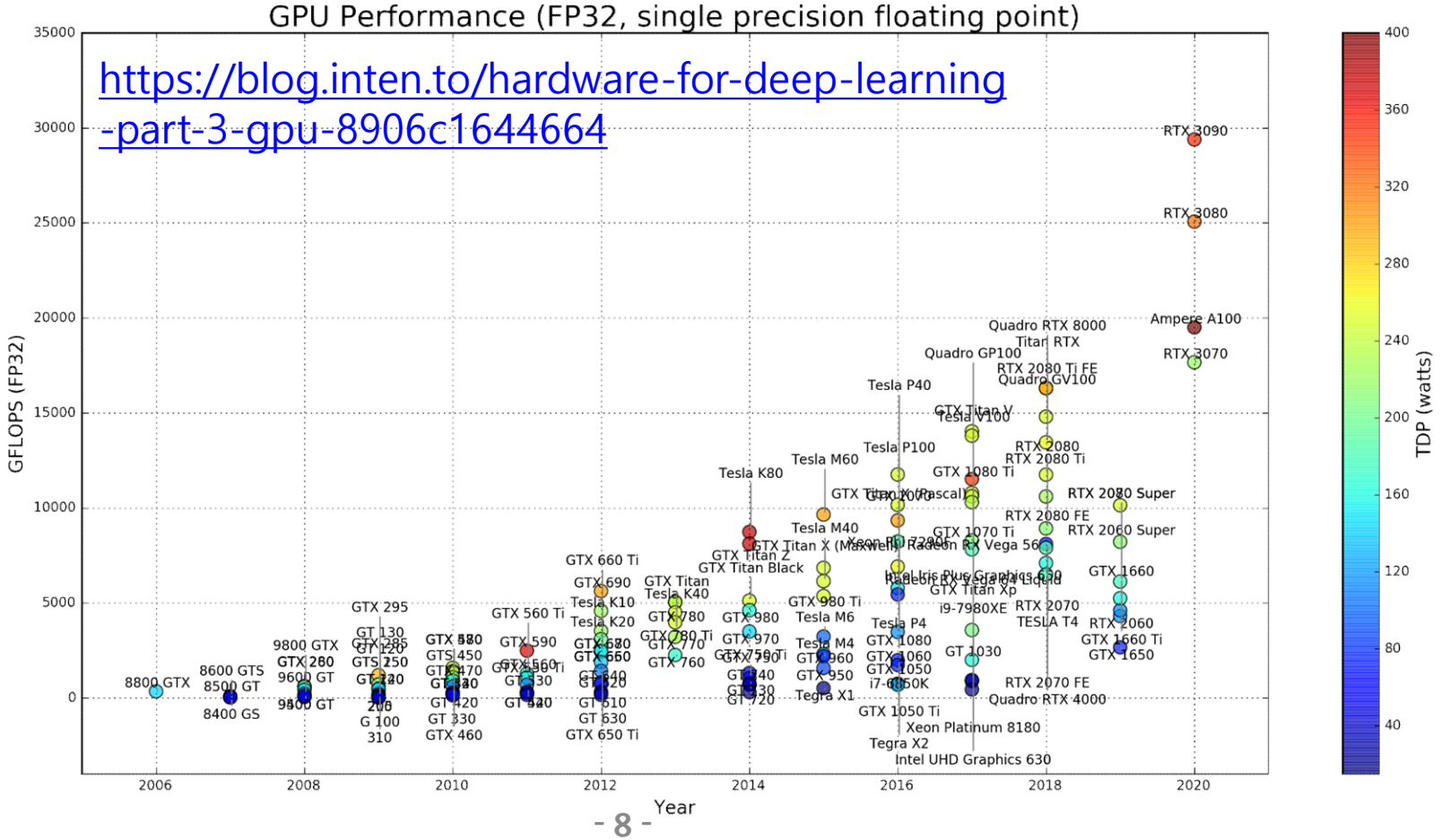


Recent advances have shown that ever-larger DNN models lead to better task performance and past progress in visual recognition tasks has also shown **a strong correlation between the model size and accuracy**

* An analysis of Deep Neural Network Models for Practical Applications, A. Canziani et al., April 2017
** Benchmark analysis of representative deep neural network architecture, Blanco et al., Oct. 2018

Challenge 2: Gap in Technical Progress

- Performance of GPUs increases by up to 3 times for 5 years (’15~’20)
- GPU Memory increases by up to 1.5 times for 5 years (16→24GB, ’15~’20)
- # operations for inferencing w/ ML models increases by about 25 times or more for 6 years(’12~’18)



Challenge 3: Time & Costs

KiSTi 한국과학기술정보연구원 Korea Institute of Science and Technology Information www.kisti.re.kr

Long training time limits ML researcher's productivity

Correlation btw. #layers and time

Model	Error rate	Training time
ResNet18	10.76%	2.5 days
ResNet50	7.02%	5 days
ResNet101	6.21%	1 week
ResNet150	6.16%	1.5 weeks

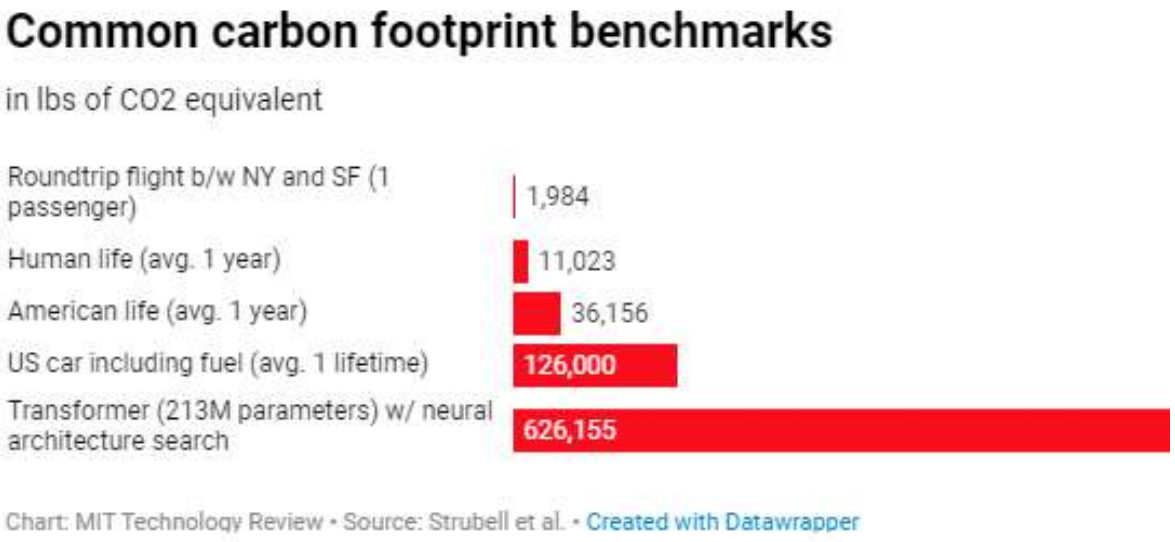
* M40 GPU, fb.resnet.torch

- KoBERT (SKT, Oct. 2019)
  - 24 layers, 340M parameters
  - 1 month with 32 V100 GPUs interconnected with Horovod(w/ infiniBand)

- XLNet (Yang, arXiv 19 Jun 2019)
  - 340 million parameters
  - Training : 2.5 days with 512 TPU v3 chips for 500k steps
  - 512 TPU x 2.5 days x $8 = $ 245,000
- Gpipe (Huang, NIPS Dec. 2019)
  - 556 million parameters
- NASNet (Barret, CVPR June 2018)
  - 800 GPU, 28 days training
- GPT-3 (OpenAI, 2020)
  - 175B parameters, required 3.14E23 FLOPS for training
  - At theoretical 28 TFLOPS for V100, 355 GPU-years and cost $4.6M for a single training run
  - 700GB memory to store it in FP32

Challenge 4: Energy Efficiency

- AlphaGo : 1,920 CPUs and 280 GPUs, $3,000 electric bill per game
- “Training a single DL model can emit as much carbon as 5 cars in their lifetimes” - MIT Tech. Review, 2019



The estimated costs of training a model

	Date of original paper	Energy consumption (kWh)	Carbon footprint (lbs of CO2e)	Cloud compute cost (USD)
Transformer (65M parameters)	Jun, 2017	27	26	$41-$140
Transformer (213M parameters)	Jun, 2017	201	192	$289-$981
ELMo	Feb, 2018	275	262	$433-$1,472
BERT (110M parameters)	Oct, 2018	1,507	1,438	$3,751-$12,571
Transformer (213M parameters) w/ neural architecture search	Jan, 2019	656,347	626,155	$942,973-$3,201,722
GPT-2	Feb, 2019	-	-	$12,902-$43,008

Note: Because of a lack of power draw data on GPT-2's training hardware, the researchers weren't able to calculate its carbon footprint.

Table: MIT Technology Review • Source: Strubell et al. • Created with Datawrapper

Overview of Challenges

- **DL development is naturally an Iterative HITL process**
  - Learning and test processes are performed in a batch job
  - Usually developed in a trial-and-error fashion
  - **Massive labeled data are required** for learning accurate DL models

- **Recent models become much larger**
  - Even a single model is composed of 175 billion parameters (i.e., GPT-3, ~652GBs)
  - All the SIZE, TIME, and COST grow very fast beyond our resources
    - Learning models tend to become harder with a limited budget and time

- **Expensive computational resources are required**
  - Energy efficiency is also an issue