Human Language Engineering

Master in Artificial Intelligence
Universitat Politècnica de Catalunya

Salvador Medina Herrera (salvador.medina.herrera@upc.edu)

Outline

- Who are you?
- What is this all about?
- Why should I bother?
- What am I supposed to do?
- I still have some doubts…
- Let’s get started!

Who are you?

Salvador Medina Herrera

Telecommunications Engineer, Computer Scientist, MSc in Artificial Intelligence, PhD wannabe

Head of the Data Team @ Kompyte (SemRUSH)

Researcher & Associate Professor @ GPLN (UPC)

そして日本語が話せます、何となく

@ Kompyte: salva@kompyte.com |@ UPC: salvador.medina.herrera@upc.edu | @ Whatsapp: (+34) 659146008

What does Kompyte do?

The Leading Competitive Intelligence Tool:

- Automated competitive research
- Real-time insight analysis
- And more boring marketing stuff…

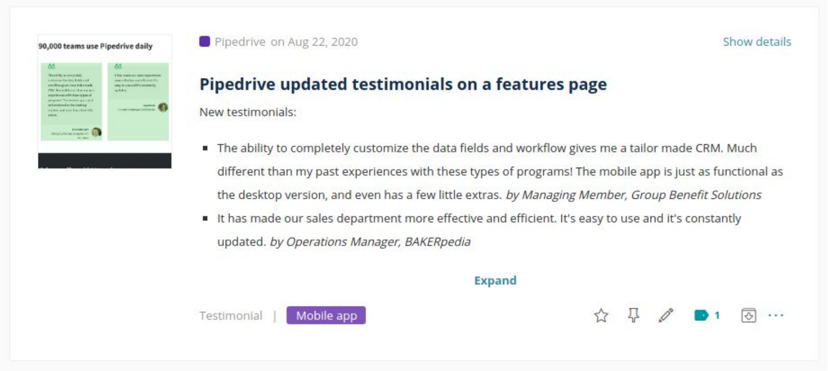
Kompyte.
by Semrush

Go check it out at https://www.kompyte.com/ if you are interested…

But the interesting thing is...

So what do YOU do at Kompyte?

Webzone Classification



Pipedrive on Aug 22, 2020

Show details

Pipedrive updated testimonials on a features page

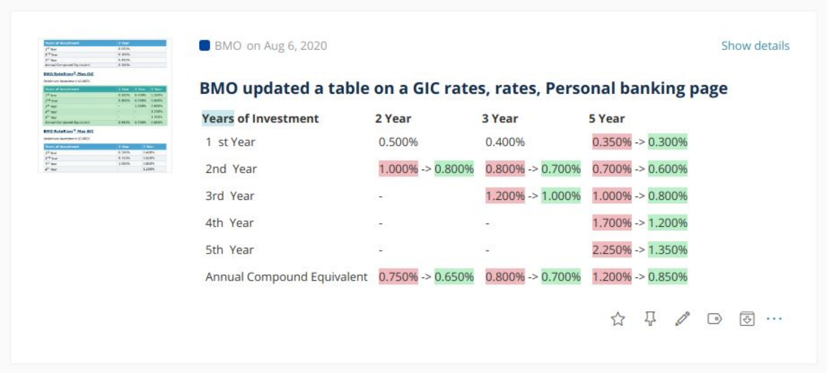
New testimonials:

- The ability to completely customize the data fields and workflow gives me a tailor made CRM. Much different than my past experiences with these types of programs! The mobile app is just as functional as the desktop version, and even has a few little extras. *by Managing Member, Group Benefit Solutions*
- It has made our sales department more effective and efficient. It's easy to use and it's constantly updated. *by Operations Manager, BAKERpedia*

Expand

Testimonial | Mobile app

Knowledge Extraction



BMO on Aug 6, 2020

Show details

BMO updated a table on a GIC rates, rates, Personal banking page

Years of Investment	2 Year	3 Year	5 Year
1 st Year	0.500%	0.400%	0.350% -> 0.300%
2nd Year	1.000% -> 0.800%	0.800% -> 0.700%	0.700% -> 0.600%
3rd Year	-	1.200% -> 1.000%	1.000% -> 0.800%
4th Year	-	-	1.700% -> 1.200%
5th Year	-	-	2.250% -> 1.350%
Annual Compound Equivalent	0.750% -> 0.650%	0.800% -> 0.700%	1.200% -> 0.850%

Interaction Detection & Automatization

NLP
NLP EVERYWHERE
LIGHTYEAR
imgflip.com

And what about Research?

GRAPH-MED: Semantic Graph Extraction from Electronic Health Documents*

- Information extraction from Electronic Health Documents
- Relation extraction
- ...

*Now part of TADIA-MED, including: negation, speculation and risk prediction

What is this all about?

- **Real-life** applications of Natural Language Processing
- **Recent trends** in Research and the Industry
- **Testimonials** from Companies and Researchers

In short, help you decide which way to go as future natural language processing engineers!

EXPECTATION
Coca-Co

REALITY

What is this all about?

01 Information Extraction

- Entity and Relation Extraction
- Event and Time Extraction
- Sentiment Analysis
- Summarisation

02 Machine Translation

- Classical Machine Translation
- Statistical Machine Translation
- Neural Machine Translation
- Resources, Models and Evaluation

03 Dialogue Systems

- Question Answering
- Conversational Agents
- Chatbots
- Virtual Assistants

Why should I bother?

Lots of Startups and Investment

verbit

$ 569M💰

GONG

$ 583M💰

PRIMER

$ 168M💰

TaskUs

$ 279M💰

grammarly

$ 400M💰

Moveworks

$ 305M💰

algolia

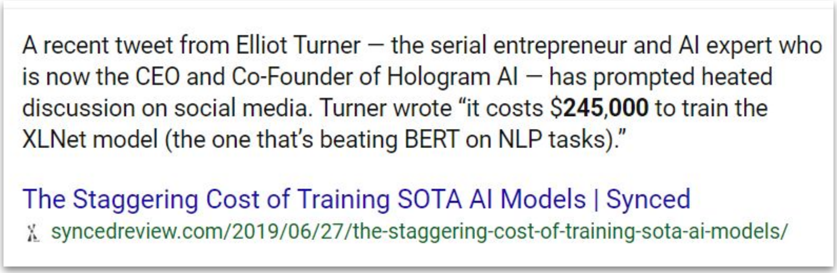
$ 334.2M💰

INVOCA

$ 201.5M💰

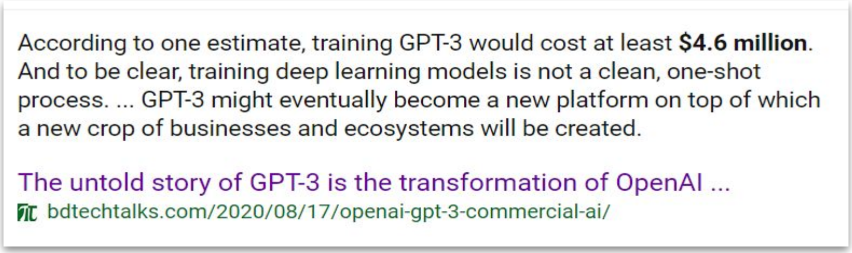
Why should I bother?

Huge investment by top-tier companies



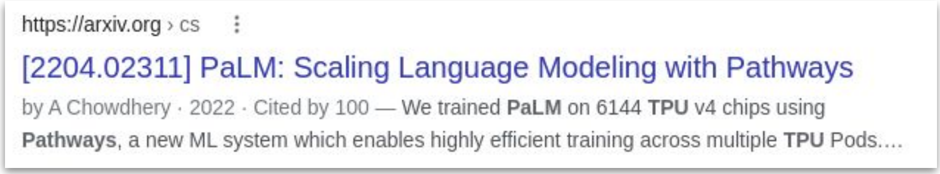
A recent tweet from Elliot Turner — the serial entrepreneur and AI expert who is now the CEO and Co-Founder of Hologram AI — has prompted heated discussion on social media. Turner wrote "it costs $**245,000** to train the XLNet model (the one that's beating BERT on NLP tasks)."

The Staggering Cost of Training SOTA AI Models | Synced
syncedreview.com/2019/06/27/the-staggering-cost-of-training-sota-ai-models/



According to one estimate, training GPT-3 would cost at least **$4.6 million**. And to be clear, training deep learning models is not a clean, one-shot process. ... GPT-3 might eventually become a new platform on top of which a new crop of businesses and ecosystems will be created.

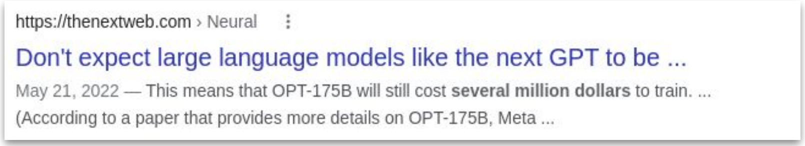
The untold story of GPT-3 is the transformation of OpenAI ...
bdtechtalks.com/2020/08/17/openai-gpt-3-commercial-ai/



https://arxiv.org › cs

[2204.02311] PaLM: Scaling Language Modeling with Pathways

by A Chowdhery · 2022 · Cited by 100 — We trained **PaLM** on 6144 **TPU** v4 chips using **Pathways**, a new ML system which enables highly efficient training across multiple **TPU** Pods....

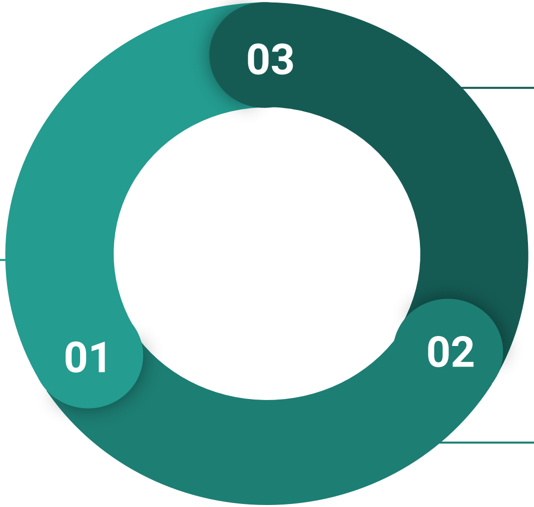


https://thenextweb.com › Neural

Don't expect large language models like the next GPT to be ...

May 21, 2022 — This means that OPT-175B will still cost **several million dollars** to train. ... (According to a paper that provides more details on OPT-175B, Meta ...

What am I supposed to do?



Write synthesis reports

2 Written Synthesis Reports
+ 1 Presentation (30%)

1 Paper Review or
Implementation (70%)

**Attend Lectures and
Speaker Sessions**

**Read referenced papers
(or from your choice)**

What am I supposed to do?

2 Written Synthesis Reports (10% of final grade each)

- Summarise the **key** ideas of a presentation or scientific paper from the bibliography
- up to 5 pages, handed in **individually** throughout the semester
- References to scientific papers are **mandatory**

1 Oral presentation about one of these two synthesis reports (10% of final grade)

What am I supposed to do?

1 Paper Review or Implementation (70% of final grade):

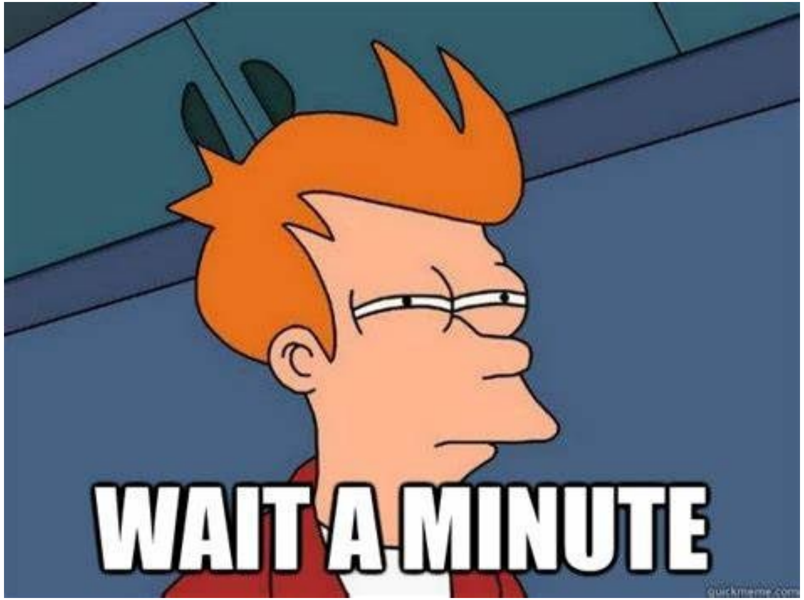
- Preferably in **pairs** (can be done individually)
- **Initial** and **final presentations** are 10% each, final **report** is 50%.
- You have 3 options, start making your own proposals ASAP, and ask for help:
    a. Deep study of a specific HLE application or a comparative study of HLE applications
    b. Development of a HLE application
    c. Development of a proposal to solve a specific real challenge

What am I supposed to do?

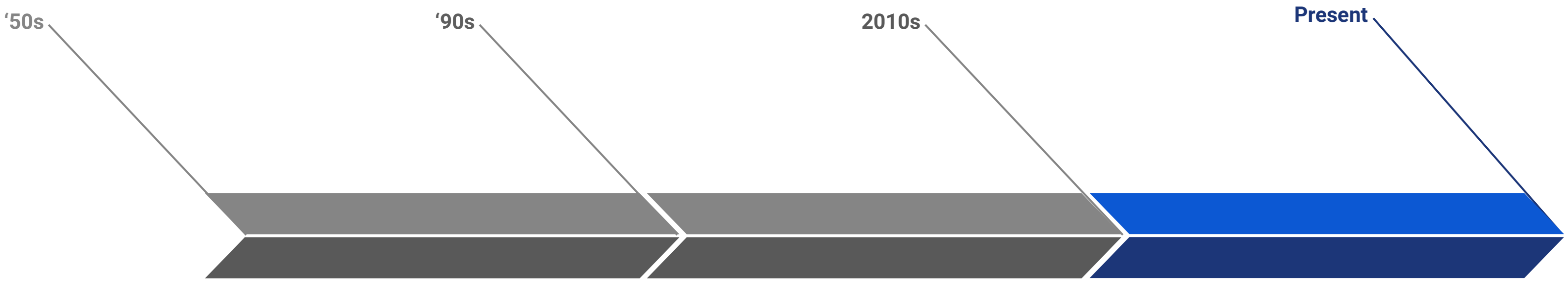
To sum up… Just be up to date, read scientific papers, read Reddit, read Medium, be involved!

https://towardsdatascience.com/, https://ai.googleblog.com/, https://paperswithcode.com/methods/area/natural-language-processing, https://www.reddit.com/r/LanguageTechnology/, and lots more…

I still have some doubts...



Let’s get started



‘50s

‘90s

2010s

Present

Symbolic NLP

“Given a collection of rules, the computer emulates natural language understanding (or other NLP tasks) by applying those rules to the data it is confronted with.”

Statistical NLP

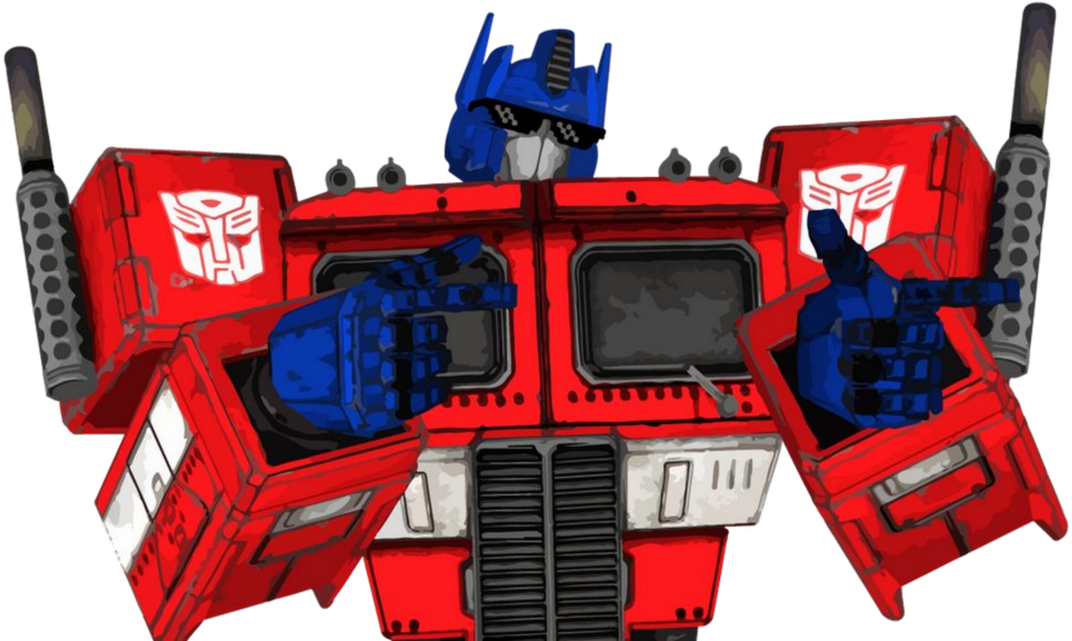
Use statistical methods to learn from existing textual corpora. Huge breakthrough in machine translation thanks to multilingual corpora. Use of unsupervised and semi-supervised algorithms.

Neural NLP

Use of deep artificial neural networks for most natural language processing tasks. Use of word and sentence embeddings. Large amounts of data, large models.

Where are we?

It is the year 2022, and Transformers rule the world… of NLP



Where are we?

State-of-the-art in most NLP tasks are achieved by Transformer and Attention-based models:

- ➔ 2017: **Transformer** was introduced by Google
- ➔ 2018: Bidirectional Encoder Representations from Transformers (**BERT**) by Google, outperformed state-of-the-art models in language understanding and question answering
- ➔ 2019: Several BERT-inspired models such as **RoBERTa** and BART by Facebook, **XLNet** by Google and CMU, and many more.
- ➔ 2020: Generative Pre-trained Transformer 3 (**GPT-3**) by OpenAI, achieves state-of-the-art in language generation. Still in active research.
- ➔ 2022: Highly optimized huge language models: **DeepSpeed** by Facebook, **PaLM** by Google

So why are we here?

State-of-the-art Transformer models are HUGE and extremely expensive to train. The full GPT-3 model is estimated to require:

- 175 Billion parameters
- 335 years to train with a 28 TFLOP GPU ($4.6 million using Tesla V100 GPUs in Lambda GPU Cloud, could potentially be reduced to ¼ by using optimized TensorCores)
- 350 GB of VRAM for 16-bit Float parameters (11 inter-connected Tesla V100 GPUs)

* Pointed out by Chuan Li at https://lambdalabs.com/blog/demystifying-gpt-3/

So why are we here? (II)

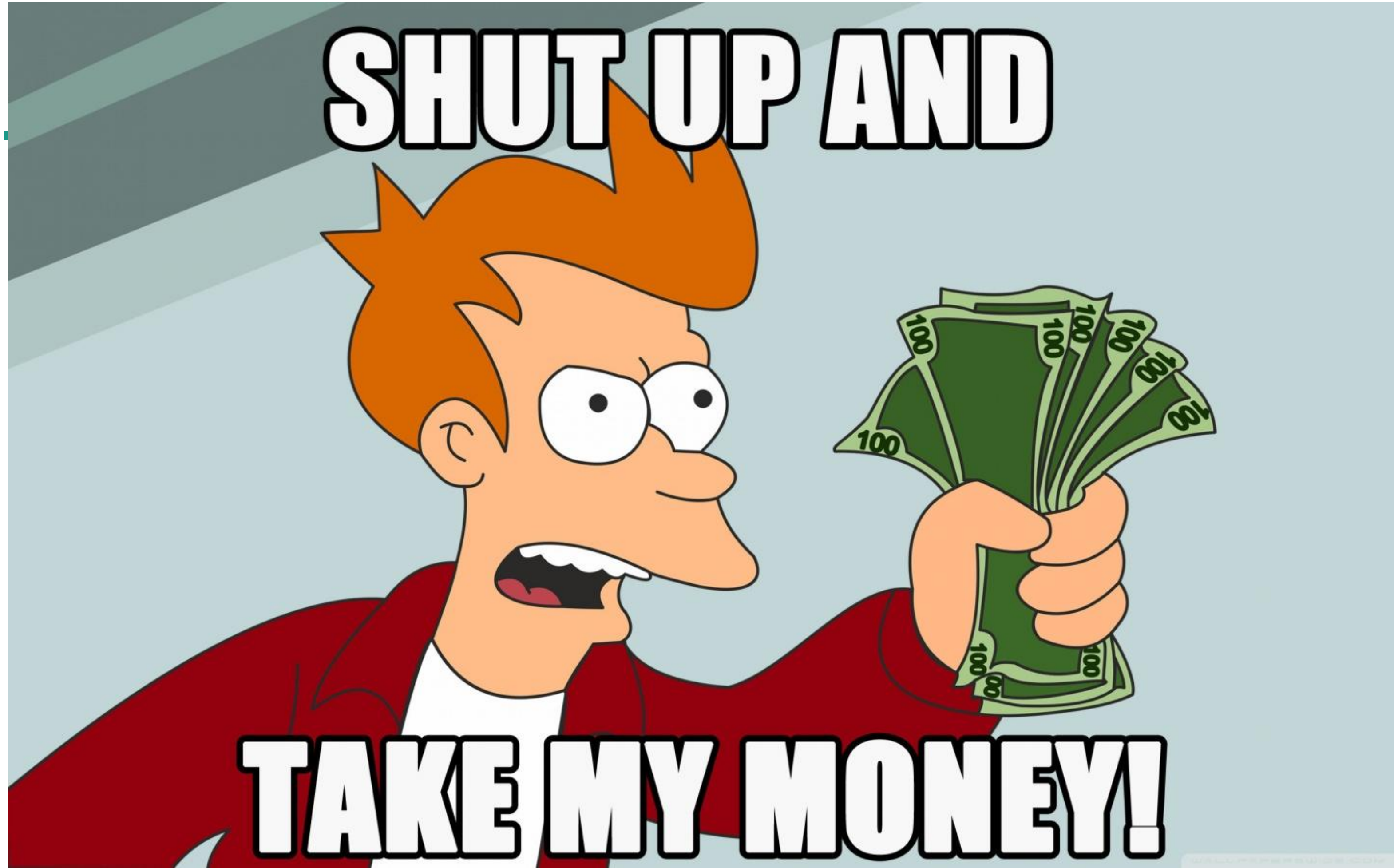
Since 2021, GPT-3 has been vastly surpassed:

- BAAI’s Wu Dao 2.0, a **multi-modal model** with **1.75 trillion** parameters trained with:
    - 1.2 terabytes of English text data in the Pile dataset + 1.2 terabytes of Chinese text in Wu Dao Corpora
    - 2.5 terabytes of Chinese graphic data
- Google’s Switch Transformer, **1.6 trillion** parameters
    - Distributed and sharded model (Switch Routing, **Mixture-of-Experts**)
    - Optimized for TPUs
    - Evolution from Google’s **GShard** (Sparsely-Gated Mixture-of-Experts)

So why are we here? (III)

In 2022, we want better yet well optimized Language Models:

- Facebook's DeepSpeed set of optimizations:
  - Optimized for GPU utilisation: training and inference
  - Can be applied to different LM: Megatron-Turing NLG (**530B**), BLOOM (176B), GPT-NeoX (20B), ,...
  - Innovations: ZeRO, 3D-Parallelism, DeepSpeed-MoE, ZeRO-Infinity…
- Google's Pathways Language Model (PaLM), **540 billion** parameters
  - Trained on 6144 TPUs
  - Very efficient FLOPs utilization (57.8%) for TPUs
  - Reformulated the Transformer + Feed Forward layers for parallelization


SHUT UP AND
TAKE MY MONEY!
100
100
100
100
100
100
100
100
100

So why are we here?

From-scratch training has become unfeasible. Moreover, these models require huge training corpora. Research and development by smaller companies and research institutes now revolves around:

- **Transfer learning** (Eg: fine-tune general purpose pretrained models)
- Use **pre-trained word or sentence representations** with simpler models (Eg: Word2Vec, FastText, eLMO or BERT embedding layer before LSTM / GRU and Attention layers)
- Model reduction strategies such as **distillation** (DistilBERT, DistilRoBERTa, ...) and **pruning**.
- Other simplified BERT-like models such as A Lite BERT (**ALBERT**), MobileBERT, ...

So why are we here? (II)

Transformers replacing costly components:

- Google’s FNet:
  - From “FNet: Mixing Tokens with Fourier Transforms”
  - Replaces Attention layers by Fourier Transforms: No trainable parameters, optimized for GPUs
- GroupBERT:
  - From “GroupBERT: Enhanced Transformer Architecture with Efficient Grouped Structures”
  - Mixes Convolutional Layers with Attention Layers
- FMMformer:
  - From “FMMformer: Efficient and Flexible Transformer via Decomposed Near-field and Far-field Attention”

Some References (I)

- Young, Tom, et al. "Recent trends in deep learning based natural language processing." *ieee Computational intelligenCe magazine* 13.3 (2018): 55-75.
- Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems.* 2017.
- Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).
- Liu, Yinhan, et al. "Roberta: A robustly optimized bert pretraining approach." *arXiv preprint arXiv:1907.11692* (2019).
- Yang, Zhilin, et al. "Xlnet: Generalized autoregressive pretraining for language understanding." *Advances in neural information processing systems*. 2019.
- Brown, Tom B., et al. "Language models are few-shot learners." *arXiv preprint arXiv:2005.14165* (2020).

Some References (II)

- Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." *arXiv preprint arXiv:1301.3781* (2013).
- Peters, Matthew E., et al. "Deep contextualized word representations." *arXiv preprint arXiv:1802.05365* (2018).
- Bojanowski, Piotr, et al. "Enriching word vectors with subword information." *Transactions of the Association for Computational Linguistics* 5 (2017): 135-146.
- Sanh, Victor, et al. "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter." *arXiv preprint arXiv:1910.01108* (2019).
- Lan, Zhenzhong, et al. "Albert: A lite bert for self-supervised learning of language representations." *arXiv preprint arXiv:1909.11942* (2019).
- Cheong, Robin, and Robel Daniel. *transformers. zip: Compressing Transformers with Pruning and Quantization*. tech. rep., Stanford University, Stanford, California, 2019.

Some References (III)

- Raffel, Colin, et al. "Exploring the limits of transfer learning with a unified text-to-text transformer." *arXiv preprint arXiv:1910.10683* (2019).
- Lepikhin, Dmitry, et al. "Gshard: Scaling giant models with conditional computation and automatic sharding." *arXiv preprint arXiv:2006.16668* (2020).
- Fedus, William, Barret Zoph, and Noam Shazeer. "Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity." *arXiv preprint arXiv:2101.03961* (2021).
- Lee-Thorp, James, et al. "FNet: Mixing Tokens with Fourier Transforms." *arXiv preprint arXiv:2105.03824* (2021).
- Chelombiev, Ivan, et al. "GroupBERT: Enhanced Transformer Architecture with Efficient Grouped Structures." *arXiv preprint arXiv:2106.05822* (2021).
- Nguyen, Tan M., et al. "FMMformer: Efficient and Flexible Transformer via Decomposed Near-field and Far-field Attention." *arXiv preprint arXiv:2108.02347* (2021).

Some References (IV)

- Smith, Shaden, et al. "Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model." *arXiv preprint arXiv:2201.11990* (2022).
- Chowdhery, Aakanksha, et al. "Palm: Scaling language modeling with pathways." *arXiv preprint arXiv:2204.02311* (2022).
- Zhang, Susan, et al. "Opt: Open pre-trained transformer language models." *arXiv preprint arXiv:2205.01068* (2022).
- Rae, Jack W., et al. "Scaling language models: Methods, analysis & insights from training gopher." *arXiv preprint arXiv:2112.11446* (2021).

Other Interesting Links:

- https://nlpprogress.com/
- https://paperswithcode.com/
- https://sites.google.com/view/iberlef2020/home
- http://alt.qcri.org/semeval2020/
- https://huggingface.co/transformers/
- https://radimrehurek.com/gensim/
- https://fasttext.cc/
- https://jalammar.github.io/
- https://www.deepspeed.ai/
- https://bdtechtalks.com/