How NLP is reshaping Machine Learning

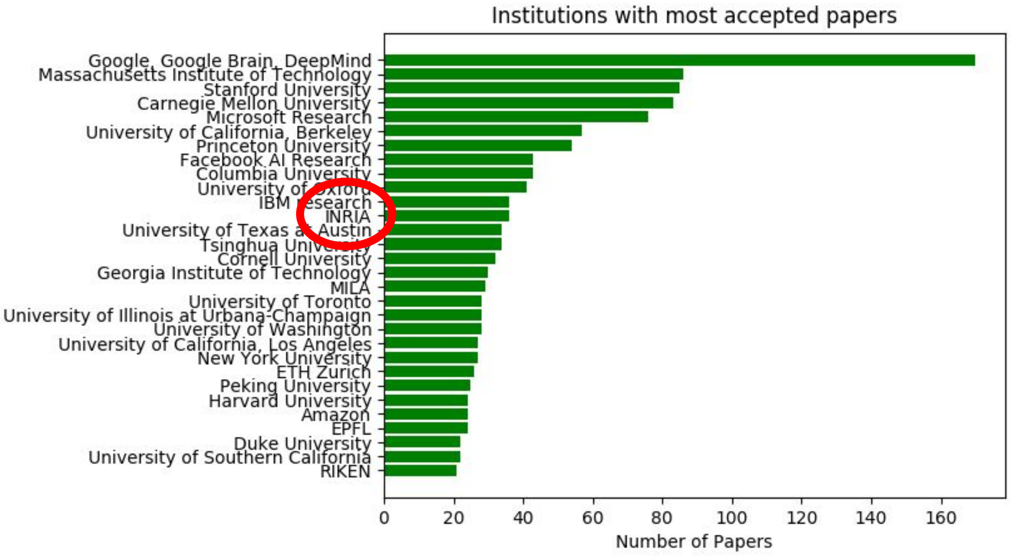
Grégoire Mialon - Paris NLP Meetup S.5#3

04/02/21

INVENTEURS DU MONDE NUMÉRIQUE

Introduction:
A few words on Inria

- A national research institute.
- We work on machine learning, cryptography, quantum computing, cognitive science...
- Well-funded, with some of the best researchers in machine learning.



INVENTEURS DU MONDE NUMÉRIQUE

Introduction:
A few words on Inria

- A national research institute.
- We work on machine learning, cryptography, quantum computing, cognitive science…
- Well-funded, with some of the best researchers in machine learning.
- Collaborations with start-ups and public organizations.

Join us!

Introduction:
Who am I and what I would like to talk about

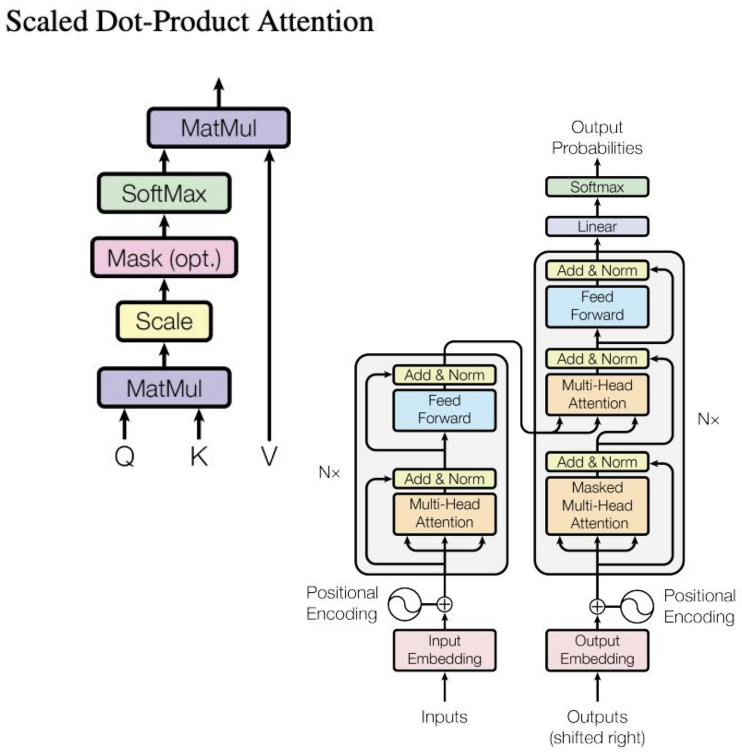
- Me: not a NLP researcher, although I did some prior to my PhD.
- My PhD: learning competitive models when labeled data is scarce.
- How? Integrating priors adapted to the data.
- Pre-trained (Transformer) models: outperform this approach in various domains. “The bitter lesson of machine learning” (Richard Sutton, 2019).
- In this talk, NLP “=” Transformer pre-trained language models.

A high level review on how Pre-trained Transformers are changing the practice of machine learning.

I - What happened to NLP?

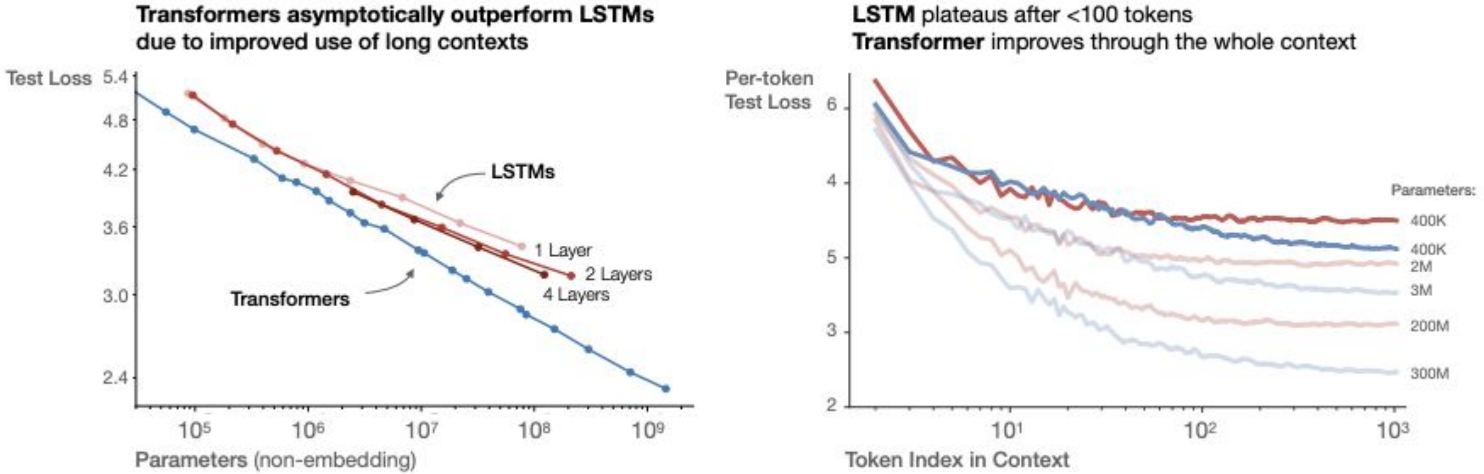
- **2018: the ImageNet moment of NLP.**
  - Transformer (Vaswani et al., 2017).
  - BERT (Devlin et al., 2019).

- How?
  - Transformer, an attention-based neural network architecture with the right inductive bias?
  - When learning language models, any piece of text is training data (Books, Wikipedia, the Internet…).



I - What happened to NLP?

- **2018: the ImageNet moment of NLP.**



  - Transformers scale better than LSTMs when its comes to bigger models (Kaplan et al., 2020).

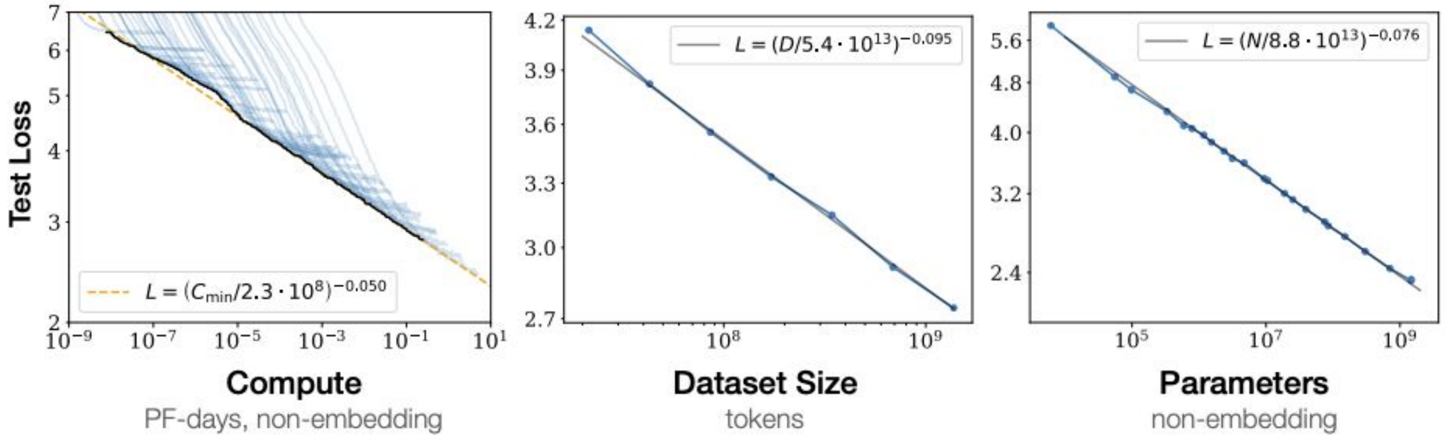
I - What happened to NLP?

- **Consequences**

    - NLP is much more accessible than ever.
        - Models: CamemBERT (Martin et al., 2020), FlauBERT, (Le et al., 2020)
        - Open source building blocks: transformers, (Wolf et al., 2019), spaCy ~2017.
        - BERTology, GPT-2 & 3.

    - The gap between academia and organizations such as FAANGs grows bigger.
        - GPT-3 ~$4.6 Million? Still behind an API.
        - Few papers studying GPT-3 for now, through the API only. None at ICLR: transparency issue.

II - How Transformers and self-supervised learning are influencing machine learning



- Transformers scaling follows a power law without plateauing, yet (Kaplan et al., 2020).
- What if we trained a Transformer "language" model on other types of data?

II - How Transformers and self-supervised learning are influencing machine learning

- **Language models out of NLP: the bioinformatics case.**
    - Proteins are sequences of amino acids.

Nucleotide triplet	CUU	GAC	AAA	GUU	GAG	GCU	GAA	GUG	CAA	AUU	GAU	AGG	UUG	AUC	ACA	GGC
Amino acid	L	D	K	V	E	A	E	V	Q	I	D	R	L	I	T	G

II - How Transformers and self-supervised learning are influencing machine learning

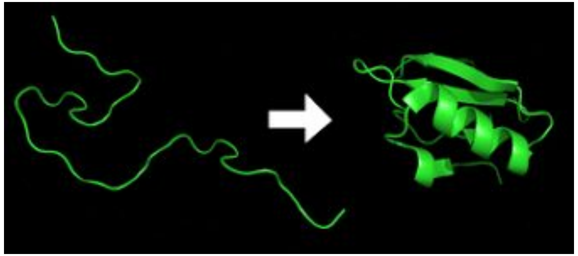
- **Language models out of NLP: the bioinformatics case.**
  - Proteins are sequences of amino acids.

Nucleotide triplet	CUU	GAC	AAA	GUU	GAG	GCU	GAA	GUG	CAA	AUU	GAU	AGG	UUG	AUC	ACA	GGC
Amino acid	L		K	V	E		E	V	Q		D	R		I	T	G

  - Transformer based language models can therefore be trained by masking some amino acids!

II - How Transformers and self-supervised learning are influencing machine learning

- **Language models out of NLP: the bioinformatics case.**
    - (Rives et al., 2019), 250 M sequences, BERT-like architecture.
    - One important task: fold prediction.



II - How Transformers and self-supervised learning are influencing machine learning

- **Language models out of NLP: the bioinformatics case.**
  - (Rives et al., 2019), 250 M sequences, BERT-like architecture.
  - One important task: fold prediction.
  - Outperforming current models (Mialon et al., 2021), AlphaFold2?

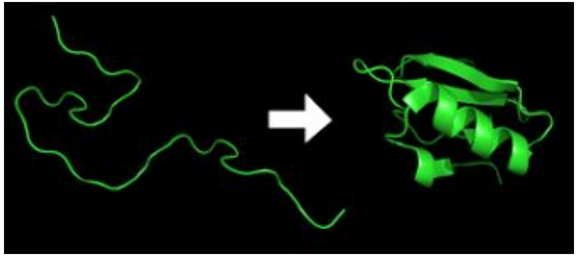


Table 2: Classification accuracy (top 1/5/10) on test set for SCOP 1.75 for different unsupervised and supervised baselines, averaged from 10 different runs (q re

Method	Unsupervised	
DeepSF (Hou et al., 2019)	Not available.	
CKN (Chen et al., 2019a)	81.8±0.8/92.8±0.2/9	
RKN (Chen et al., 2019b)	Not available.	
Set Transformer (Lee et al., 2019)	Not available.	
Approximate Rep the Set (Skianis et al., 2020)	Not available.	84.5±0.6/94.0±0.4/95.7±0.4
Ours (dot-product instead of OT)	78.2±1.9/93.1±0.7/96.0±0.4	87.5±0.3/95.5±0.2/96.9±0.1
Ours (Unsup.: 1 × 100 / Sup.: 5 × 10)	**85.8±0.2/95.3±0.1/96.8±0.1**	**88.7±0.3/95.9±0.2/97.3±0.1**

Table 5: Classification accuracy (top 1/5/10) results of our unsupervised embedding for SCOP 1.75 with pre-trained ESM models (Rives et al., 2019).

Model	Nb parameters	Mean Pooling	Unsupervised OTKE
ESM1-t6-43M-UR50S	43M	84.01/93.17/95.07	85.91/93.72/95.30
ESM1-t34-670M-UR50S	670M	94.95/97.32/97.91	95.22/97.32/98.03

II - How Transformers and self-supervised learning are influencing machine learning

- **Transformers also work with labels: the computer vision case.**
    - Training data: JFT: 300M labeled images whereas ImageNet is 15M.
    - ViT (Dosovitskyi et al., 2021).
    - Outperforms competitive CNNs (ResNet) trained on the same huge amount of data.

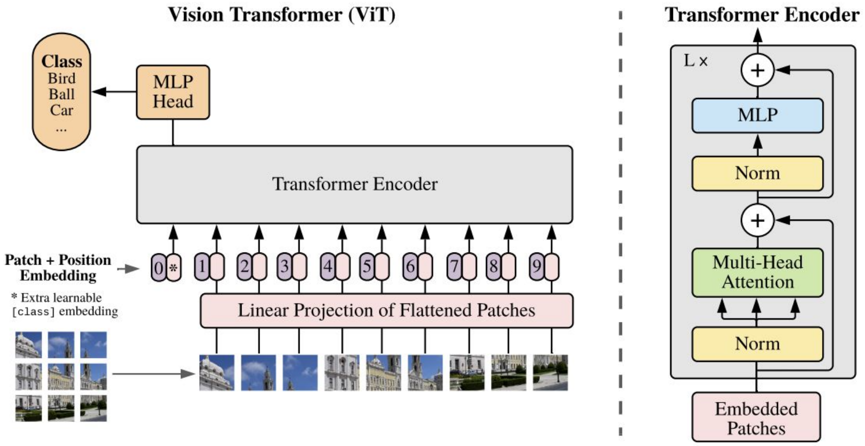
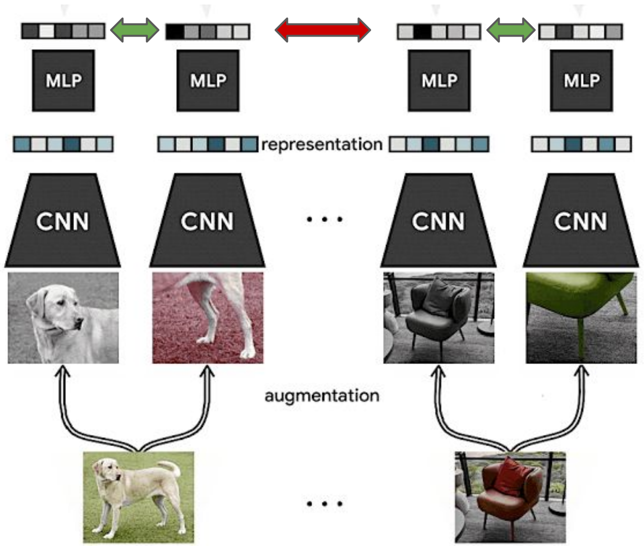
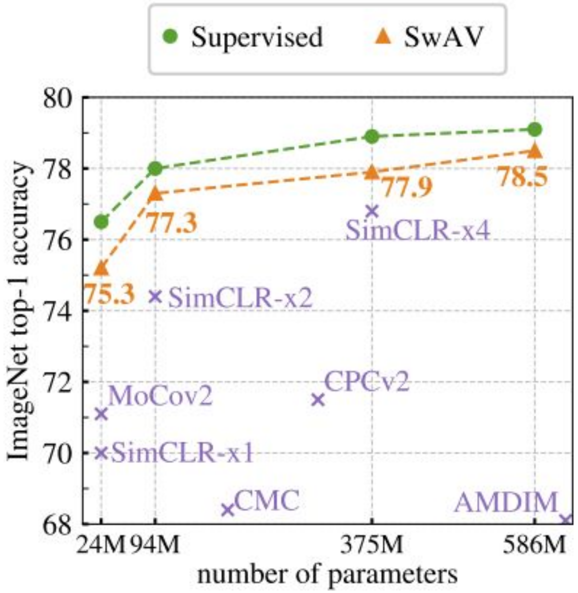


Figure 1: Model overview. We split an image into fixed-size patches, linearly embed each of them, add position embeddings, and feed the resulting sequence of vectors to a standard Transformer encoder. In order to perform classification, we use the standard approach of adding an extra learnable "classification token" to the sequence. The illustration of the Transformer encoder was inspired by Vaswani et al. (2017).

II - How Transformers and self-supervised learning are influencing machine learning

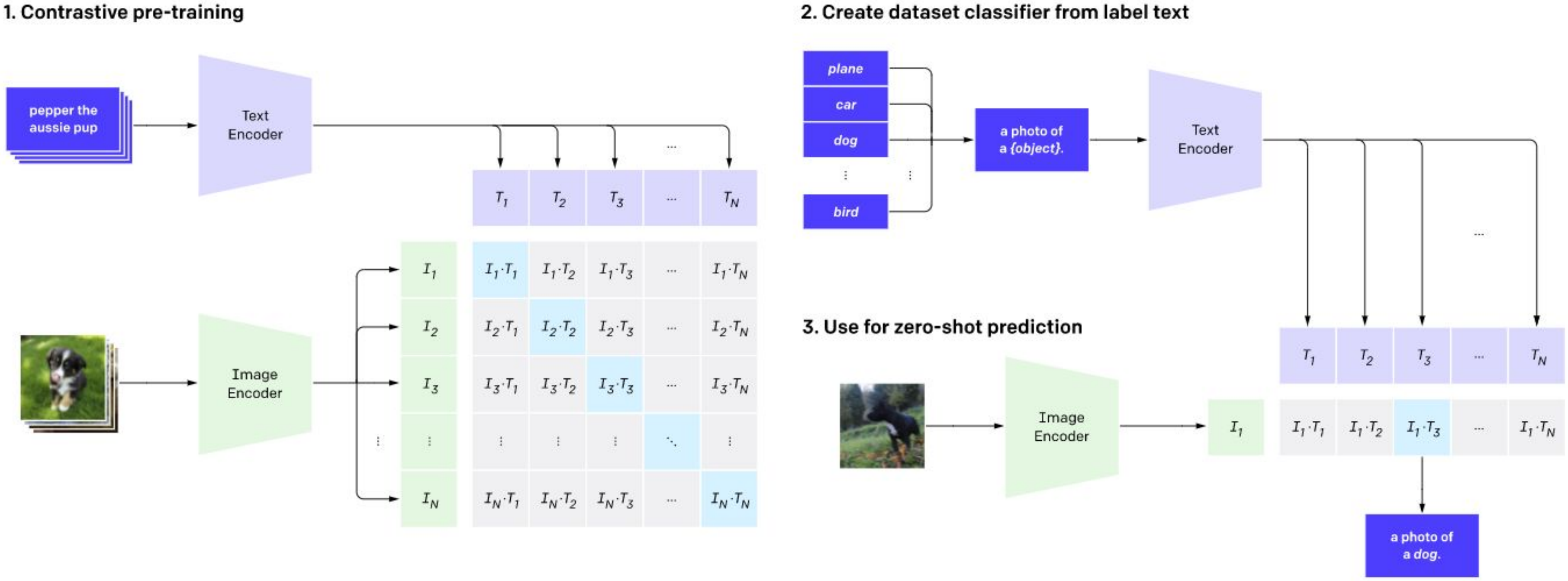
- **But, will labels always be needed?**
  - SwAV (Caron et al., 2020), BYOL (Grill et al., 2020) for learning visual features without labels.
  - Building on the idea of contrastive learning.





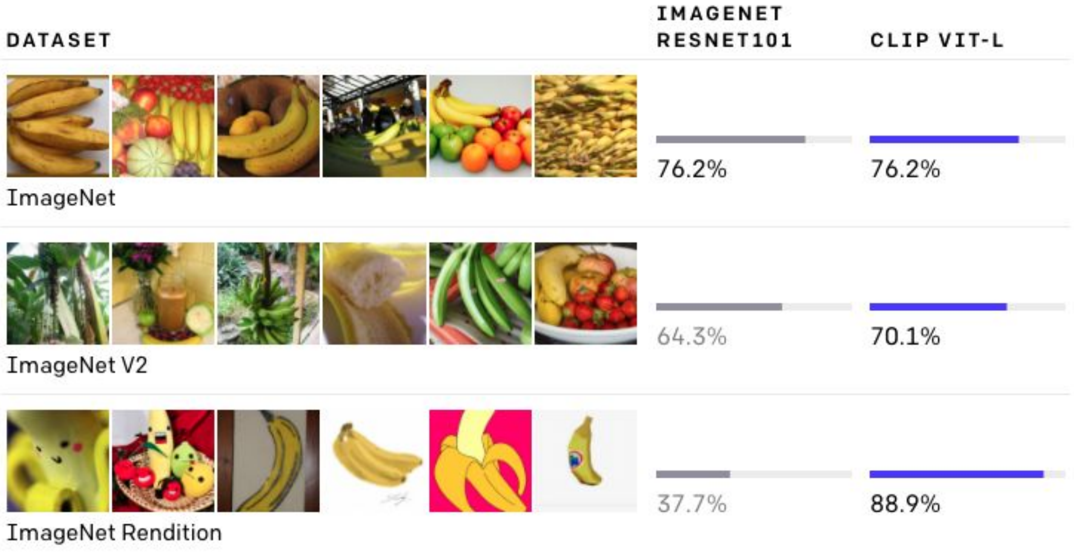
II - How Transformers and self-supervised learning are influencing machine learning

- **But, will labels always be needed? CLIP (Radford et al., 2021).**



II - How Transformers and self-supervised learning are influencing machine learning

- But, will labels always be needed? CLIP (Radford et al., 2021).



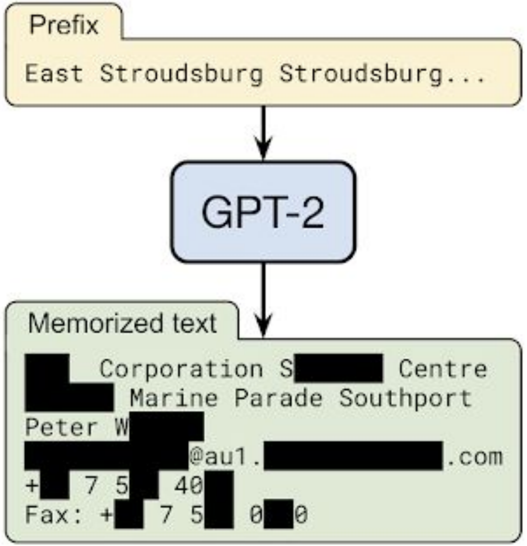
III - Pre-trained Transformer models underline challenges machine learning has to solve

- **Energy efficiency of language models.**
  - Strubell et al., 2019.
  - Be wary about the numbers (energy mix, hardware, implementation, etc.).
- However:
  - Many will want to train their own model.
  - What if we train Transformers on images or videos?
    GPT-3 data: 570GB. JFT300M: 45TB? Existing open video datasets ~1-10 TB.
  - What about the cost of deploying these models in real products?

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 one model (GPU)	
BERT_{base}	1438

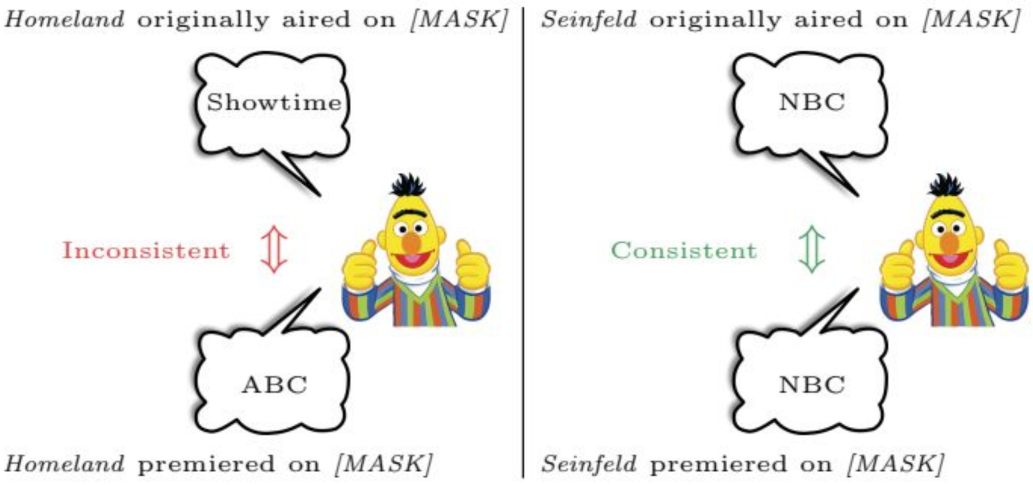
III - Pre-trained Transformer models underline challenges machine learning has to solve

- **Robustness of language models: “all the bad things that can happen when the model is deployed”.**
  - Learned data sets (Carlini et al., 2020).



III - Pre-trained Transformer models underline challenges machine learning has to solve

- **Robustness of language models.**
  - Learned data sets (Carlini et al., 2020).
  - Bias, offensive content (Kurita et al., 2019, Bender et al., 2021).
  - Inconsistency (Elazar et al., 2021).



- And also many other challenges more specific to Transformers.

Conclusion

- In the short/middle term, we can expect success of Transformers in new domains of machine learning.
- But machine learning is still far from being solved.
- Some organizations exhibit secretive behaviors when it comes to releasing their models. But one of the reason for the recent success of machine learning is its open source culture.

“In computing, the phenomenon when certain algorithms win not because they are ideally suited to solve certain problems, but because they run well on the existing hardware is called Hardware Lottery (Hooker, 2020) - and this is the case with Transformers running on GPUs”. - Michael Bronstein. Three years ago, Transformers were barely used: let’s keep an open mind!

Thank you!

And thanks to Thomas Eboli, Yana Hasson, Geoffrey Negiar, and Robin Strudel for their comments.

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models, 2020.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2019.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. Measuring and improving consistency in pretrained language models, 2021.

References

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Sara Hooker. The hardware lottery. In *arXiv 2009.06489*, 2020.

Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. Exploring the limits of language modeling, 2016.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. Measuring bias in contextualized word representations. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 166–172, 2019.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoît Crabbé, Laurent Besacier, and Didier Schwab. Flaubert: Unsupervised language model pre-training for french. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2479–2490, Marseille, France, May 2020. European Language Resources Association. URL `https://www.aclweb.org/anthology/2020.lrec-1.302`.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de la Clergerie, Djamé Seddah, and Benoît Sagot. Camembert: a tasty french language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

Grégoire Mialon, Dexiong Chen, Alexandre d'Aspremont, and Julien Mairal. A trainable optimal transport embedding for feature aggregation and its relationship to attention. In *International Conference on Learning Representations (ICLR)*, 2021.

References

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C. Lawrence Zitnick, Jerry Ma, and Rob Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *bioRxiv*, 2019. doi: 10.1101/622803. URL `https://www.biorxiv.org/content/10.1101/622803v4`.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in nlp, 2019.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. Glue: a multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations (ICLR)*, 2019.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. Huggingface's transformers: State-of-the-art natural language processing. In *arXiv preprint arXiv: 1910.03771*, 2019.