“IA et langage naturel” 3

Final Thoughts

Prof. Stéphane DUPONT

2022-2023

OUTLINE

- Introduction
  - Historique
  - Catégories d'applications et exemples
  - Elements de linguistique: domaine, structure de phrase, ambiguïtés (lexicales et syntaxiques), contexte
- Algorithmes et outils
  - Descripteurs (sparse, denses, contextuels,...)
  - Réduction de dimension
  - Modèles de langage
  - Architectures des modèles (CNN, RNNs, Transformers, …)
  - Classification de texte
- Méthodologie
  - Databases, data augmentation, métriques (y compris NMT)
- Quelques complements
- Présentation du défi
- Conclusions et points de vues critiques

Compléments

Rappel- Taille des modèles

- Petits modèles (sm)

- … et gros modèles (lg)

  - Utilisez les petits pour commencer, afin de réduire les temps de download et de calacul,
  - mais ne vous étonnez donc pas si cela ne casse rien en termes de précision des résultats d’analyse du langage.

Evaluating Language Models

- The standard evaluation metric for Language Models is perplexity.

$$\text{perplexity} = \underbrace{\prod_{t=1}^{T} \left(\frac{1}{P_{\text{LM}}(\boldsymbol{x}^{(t+1)} | \, \boldsymbol{x}^{(t)}, \ldots, \boldsymbol{x}^{(1)})} \right)}_{\text{Inverse probability of corpus, according to Language Model}}^{1/T}$$

Normalized by number of words

- This is equal to the exponential of the cross-entropy loss $J(\theta)$:

$$= \prod_{t=1}^{T} \left(\frac{1}{\hat{\boldsymbol{y}}^{(t)}_{\boldsymbol{x}_{t+1}}} \right)^{1/T} = \exp\left(\frac{1}{T} \sum_{t=1}^{T} - \log \hat{\boldsymbol{y}}^{(t)}_{\boldsymbol{x}_{t+1}} \right) = \exp(J(\theta))$$

$\boxed{\textbf{Lower} \text{ perplexity is better!}}$

Perplexité

$$= \prod_{t=1}^{T} \left(\frac{1}{\hat{\boldsymbol{y}}^{(t)}_{\boldsymbol{x}_{t+1}}} \right)^{1/T} = \exp\left(\frac{1}{T} \sum_{t=1}^{T} -\log \hat{\boldsymbol{y}}^{(t)}_{\boldsymbol{x}_{t+1}} \right) = \exp(J(\theta))$$

```
nlls = []
for i in tqdm(range(0, encodings.input_ids.size(1), stride)):
    begin_loc = max(i + stride - context_length, 0)
    end_loc = min(i + stride, encodings.input_ids.size(1))
    if end_loc <= begin_loc:
      break

    trg_len = end_loc - i    # may be different from stride on last loop
    input_ids = encodings.input_ids[:,begin_loc:end_loc]
    target_ids = input_ids.clone()
    target_ids[:,:-trg_len] = -100

    with torch.no_grad():
        outputs = model(input_ids, labels=target_ids)
        neg_log_likelihood = outputs[0] * trg_len

    nlls.append(neg_log_likelihood)

ppl = torch.exp(torch.stack(nlls).sum() / end_loc)
print(ppl)
```

https://huggingface.co/docs/transformers/main_classes/output

The outputs object is a SequenceClassifierOutput, as we can see in the documentation of that class below, it means it has an optional loss, a logits an optional hidden_states and an optional attentions attribute. Here we have the loss since we passed along labels, but we don't have hidden_states and attentions because we didn't pass output_hidden_states=True or output_attentions=True.

Problème à résoudre pour le défi
Détection de désinformation

Détection de désinformation

Classification of complete sentences

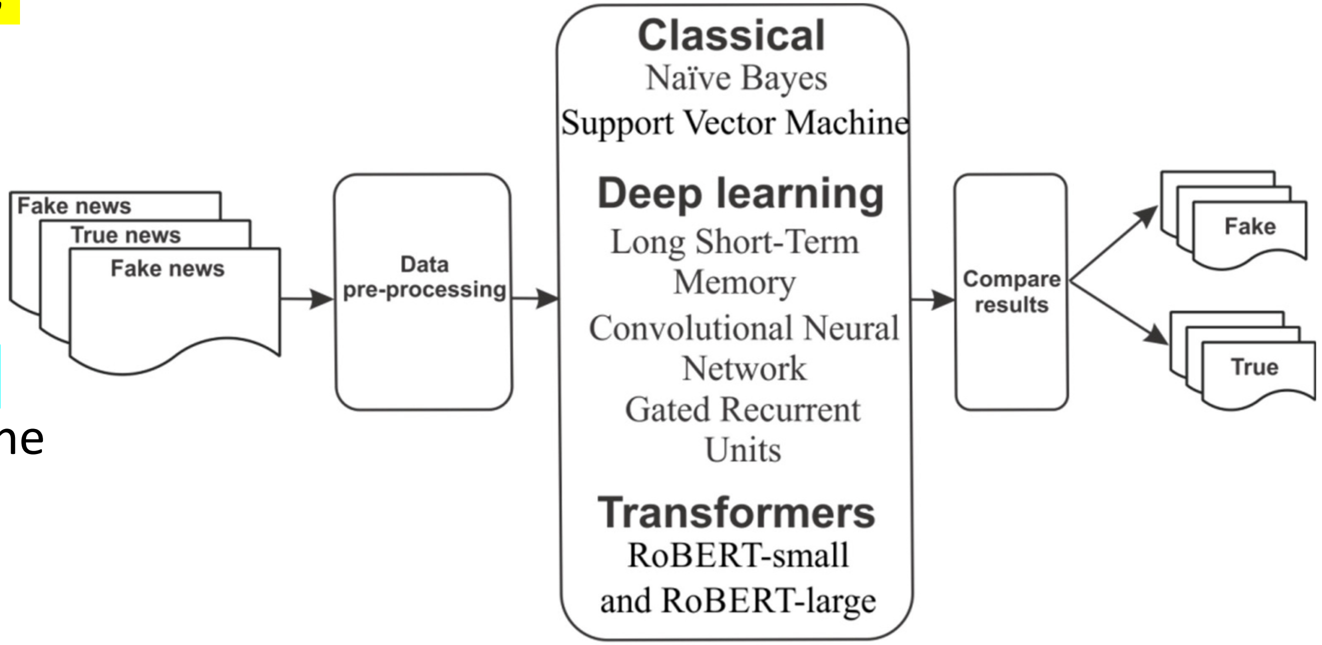
(Undertanding - NLU)

- Classification de texte à 2 classes: Fake / True

- The latest Pixar movie deserves some Oscars

⇒ classify sentence as [“positive”, “negative”, “neutral”]

- Tâche de classification de texte (phrase ou document)
  - Identifiation de la langue
  - Classification thématique (**20-newsgroups => 6**)
  - Détection de harcèlement et incitations à la haine
  - Analyse de sentiment et de polarité d’opinion
  - Détection de “fake news” (**Le défi**)
  - Détection de fraude
  - Triage d’appels ou de déclaration de sinistres
  - …



Data sets

- Dataset et notebook de chargement de celui-ci sur Moodle:
  - Train set de base: 1457 textes
  - Validation set: 487 textes
  - Test set gardé secret ????

- Votre travail:
  - développez au moins 2 pipelines différents sur base des composantes vues lors des séances
  - complétez éventuellement le training set par des textes que vous téléchargez sur le web.
  - Tester et optimiser vos modèles sur le ”validation set”.

- Notre travail:
  - Répondre à vos questions,
  - Tester vos modèles sur un test set gardé secret, afin de pouvoir évaluer leur capacité de généralisation.

Evaluation

Evaluation

Par groupe (même groupes que le défi 1)

à remettre via Moodle pour le

<u>30/12/22 17h00</u>

- Contenu du zip:
  - Rapport : description de la solution proposée, justification des choix, interprétation des résultats, au format libre (3-5 pages),
  - Code source : code source documenté, sur base du dernier notebook de départ, ce qui nous permettra d'effectuer des tests sur des données de test que nous gardons secrètes à ce usage,
  - Données additionnelles pour l'apprentissage, et une documentation.

- L'évaluation de vos travaux se fera selon cette pondération:
  - qualité du rapport (20%),
  - qualité du code (20%),
  - exploitation de différentes approches parmi les façons de faire vues au cours des séances (30%)
  - données d'apprentissage additionnelles (10%)
  - taux de classification sur nos données cachées (20%)

Critical Viewpoints

The Submerged Part of the AI-Ceberg

This article discusses the contradiction between the exploding energy demand of artificial intelligence (AI) and the information and communication (ICT) industry as a whole and the parallel strong request for energy sobriety imposed by the need to mitigate the impact of climate change and the anticipated collapse of civilization as we know it. Under the form of an open reflection on the goods and evils of AI, the article raises the suggestion of a drastic change in the AI paradigm, more in phase with the vital obligation to design a more resilient society.

world with a few clicks, to name only a few [1], [2].

Deep neural network learning is at the forefront of this development and has spread rapidly, far beyond the confidential fields of its beginnings. In a matter of 10 years, this specific computer science tool—theorized as early as the 1980s [3]—has reached all levels of society: in companies, institutions, research laboratories, in virtually all engineering disciplines as well as life sciences. Easy to use as a black box thanks to an important software development effort—multiple "plug-and-play" solutions have been

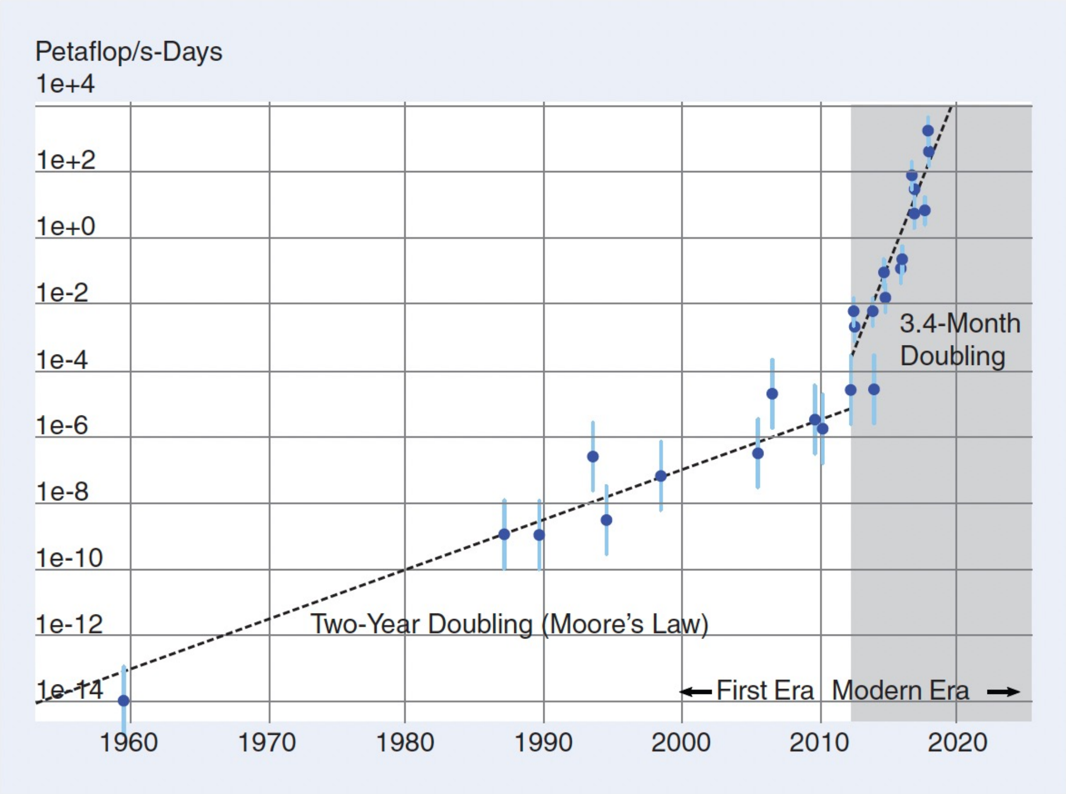


FIGURE 2. The amount of computation required by neural networks, from Rosenblatt's perceptron (far left) to the latest deep networks (far right) [21]. Each blue dot corresponds to an instance of a popular neural network. The growth rate of AI ($\simeq \times 11/\text{year}$) is a decade higher than the average growth rate of other industrial goods and services, which is already too high to avoid ecological collapse (see Figure 3).

Energy Consumtion in MWh => $500,000

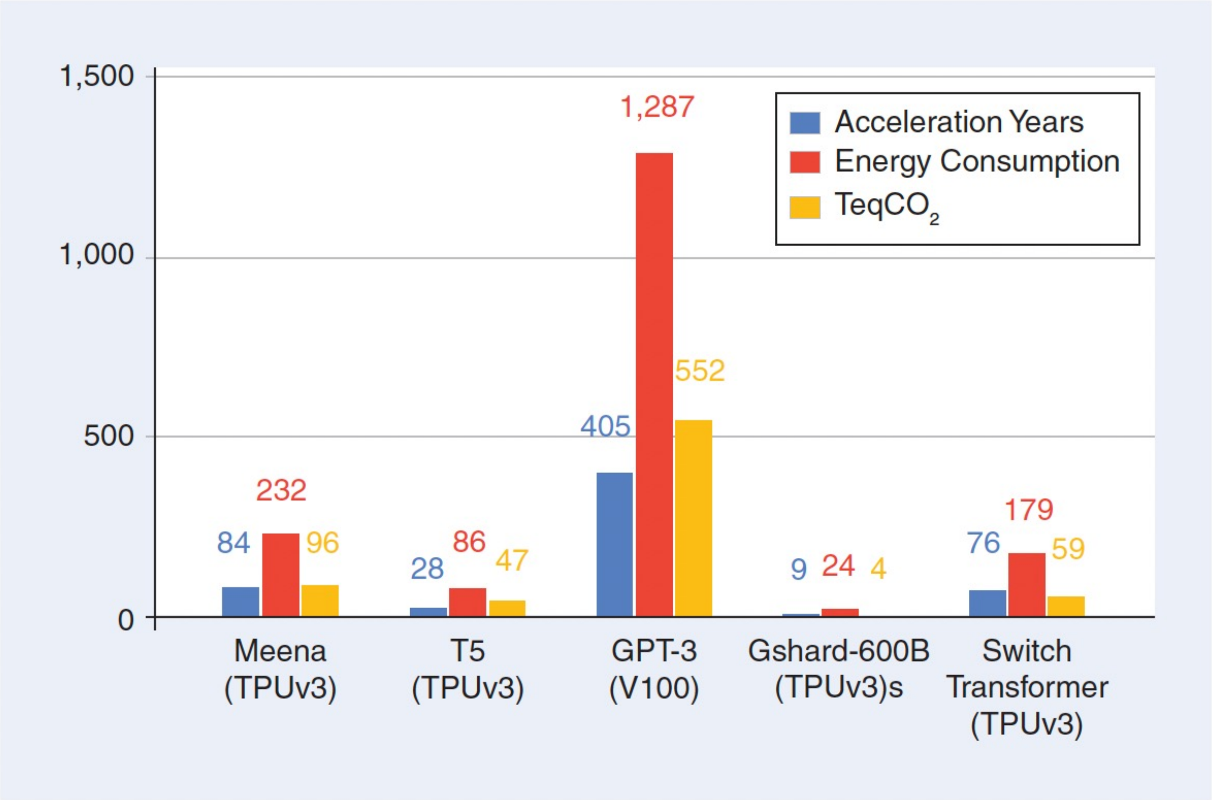


FIGURE 1. Some factors of energy used to train several modern neural networks (here, for language processing); see [15] for details. The graph shows the number of GPU-equivalent acceleration years; electricity consumption, in megawatt hours (1 MWh = €100); and CO_2-equivalent tonnage ($TeqCO_2$) (1 $TeqCO_2$ = one Paris–New York round-trip flight). The mere cost of training a neural network on a targeted application is 250× higher than the annual maximum allotted to each European (2 $TeqCO_2$) to reach the carbon balance in 2050 [17]. TPU: tensor processing unit.

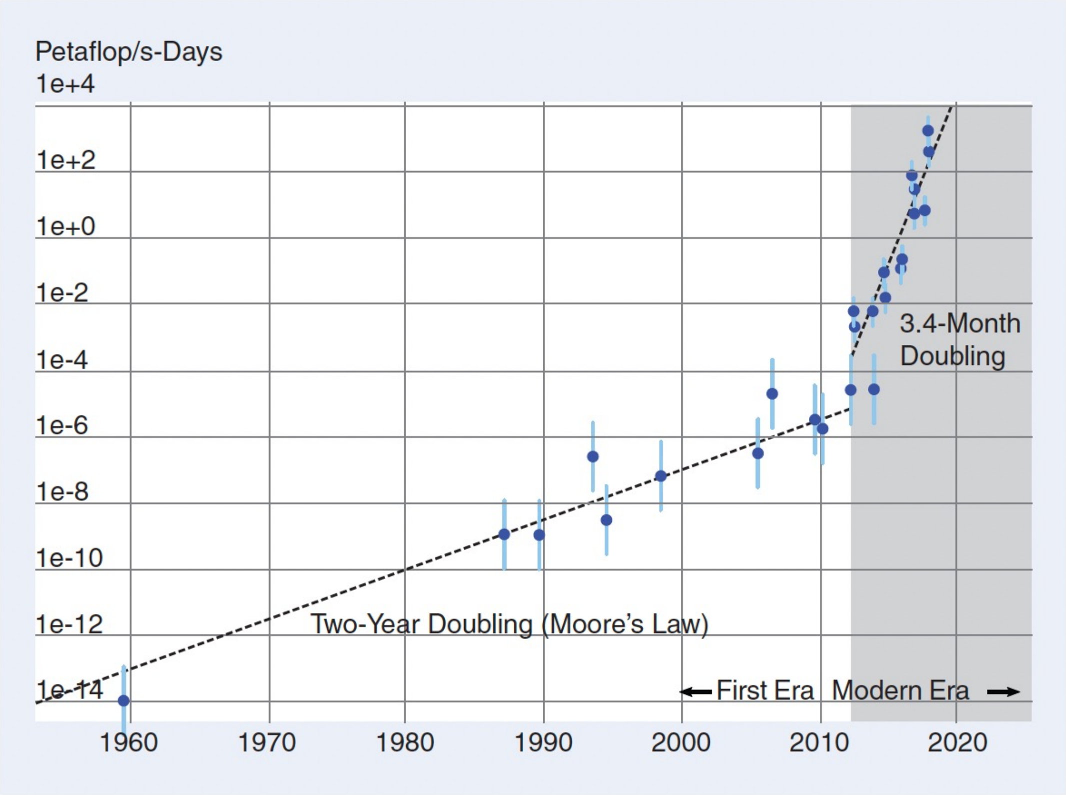


FIGURE 2. The amount of computation required by neural networks, from Rosenblatt's perceptron (far left) to the latest deep networks (far right) [21]. Each blue dot corresponds to an instance of a popular neural network. The growth rate of AI ($\simeq \times 11$/year) is a decade higher than the average growth rate of other industrial goods and services, which is already too high to avoid ecological collapse (see Figure 3).

Why it's totally unsurprising that Amazon's recruitment AI was biased against women

Isobel Asher Hamilton

DALL·E 2 Fails to Reliably Capture Common Syntactic Processes

Evelina Leivada[1*], Elliot Murphy[2], Gary Marcus[3]

1. Universitat Rovira i Virgili, Tarragona, Spain
2. Vivian L. Smith Department of Neurosurgery, University of Texas Health Science Center at Houston, Texas, USA
3. New York University, New York, USA

*Correspondence should be addressed to evelina.leivada@urv.cat

Abstract: Machine intelligence is increasingly being linked to claims about sentience, language processing, and an ability to comprehend and transform natural language into a range of stimuli. We systematically analyze the ability of DALL·E 2 to capture 8 grammatical phenomena pertaining to compositionality that are widely discussed in linguistics and pervasive in human language: binding principles and coreference, passives, word order, coordination, comparatives, negation, ellipsis, and structural ambiguity. Whereas young children routinely master these phenomena, learning systematic mappings between syntax and semantics, DALL·E 2 is unable to reliably infer meanings that are consistent with the syntax of the prompts. These results challenge recent claims concerning the capacity of such systems to understand human language. We make available the full set of test materials as a benchmark for future testing.

Keywords: DALL·E; syntax; semantics; linguistics; compositionality; large language models; neural networks

Questions?
Merci!