Deep Learning par la pratique

Jour 3, après midi : texte

Recap

On a vu

Plan

- Le NLP
  -
  - générer du texte avec les RNN
  - Transformers
- Récap des 3 jours

NLP

Le Traitement du langage naturel - NLP

le NLP couvre un très vaste champs d'applications

- Part of speech tagging (POS), analyse sémantique,
- classification : sentiment, sujets, …
- detection : discours de haine,
- NER : reconnaissance d'entités

Puis de façon plus générale

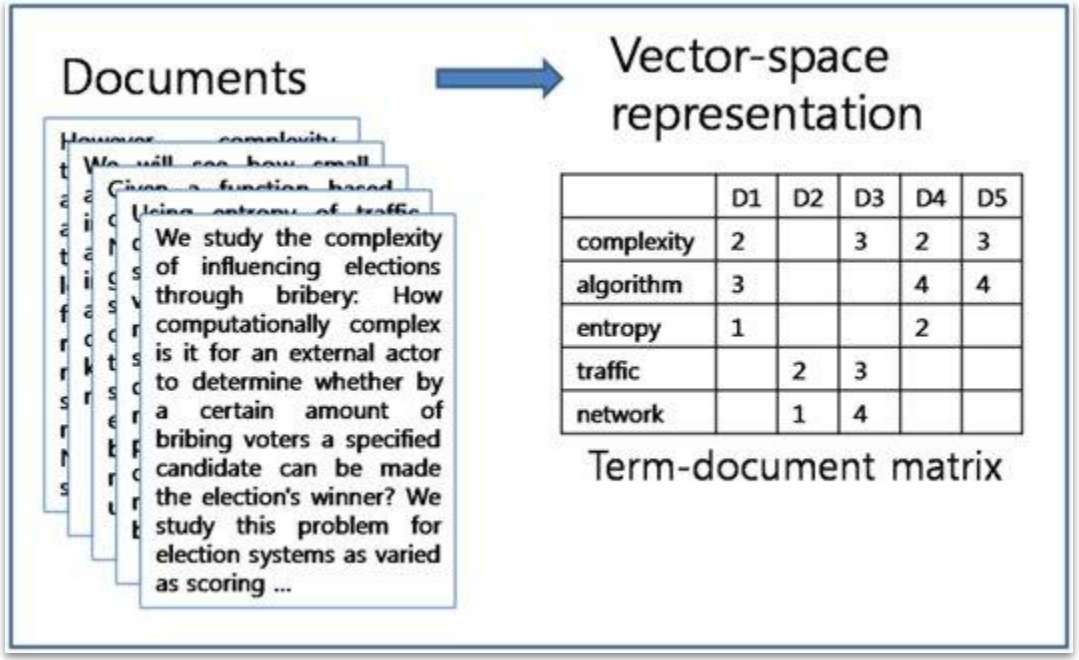
- traduction, résume, compréhension, génération, chatbots

et le texte comme origine ou cible

- transformations : texte - image - speech - audio - video - SQL
- génération :
  - text to image & image to text
  - text to audio & audio to text
  - text to video & video to text

du décompte des mots aux LLMs

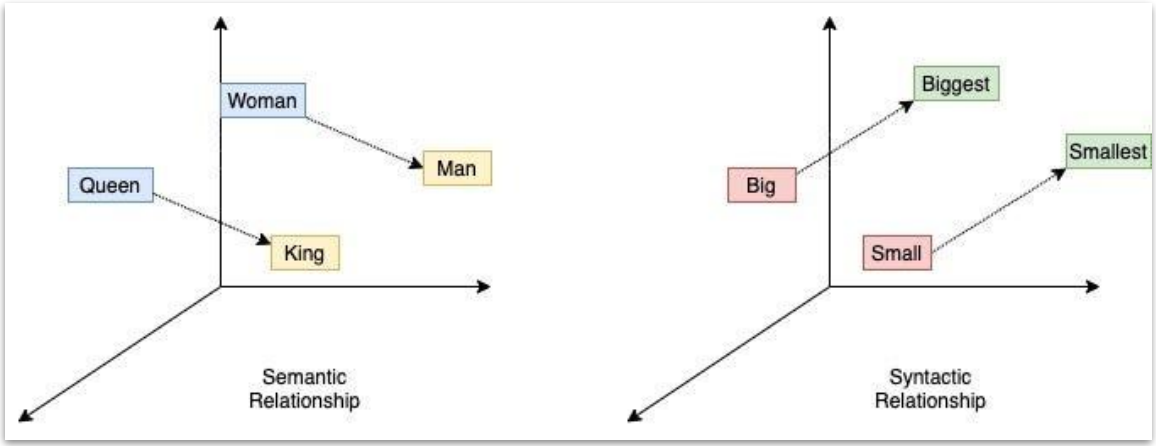
Dans l'ancien temps, on se basait sur la fréquence des mots dans un corpus, l'approche tf-idf



	D1	D2	D3	D4	D5
complexity	2		3	2	3
algorithm	3			4	4
entropy	1			2	
traffic		2	3		
network		1	4		

du décompte des mots aux LLMs
2013 embeddings - word2vec

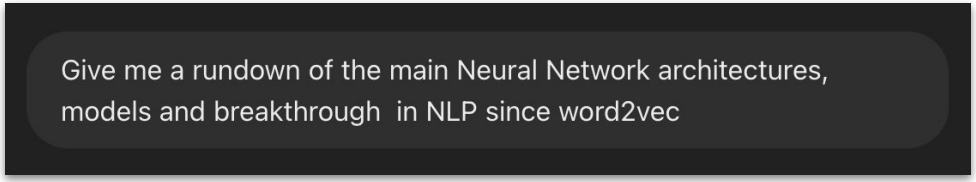
- on entraîne un RNN sur un large corpus pour prédire le mot manquant dans une phrase
- on récupère les poids de la dernière couche comme représentation de chaque mot. Ce vecteur est appelé **embedding.**
- La représentation vectorielle capture la signification du mot

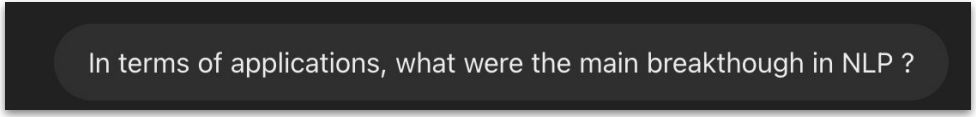


Petit à petit

les architectures s'étoffent, le champ d'application s'élargit et les performances s'accroissent

voir ce résumé par chatGPT4o
https://chatgpt.com/share/8fe90678-8113-48e5-a1eb-943a4e0996af





Principales étapes

- **word2vec** : 2013 capture les relations semantiques. suivi de GloVe, FastText
- **Seq2Seq** : 2014 encoder-decoder architecture with RNNs
- **Attention** : 2015 Allows the model to focus on different parts of the input sequence for each output step.

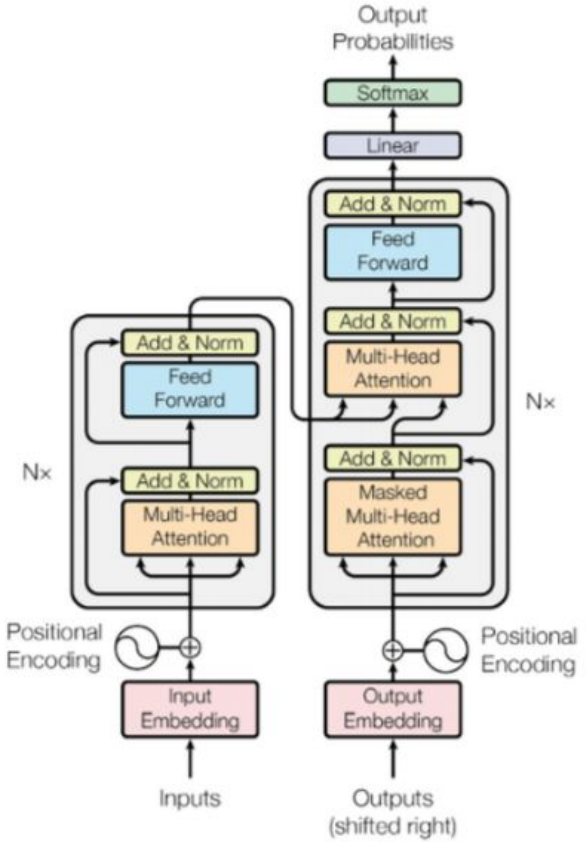
- **Transformers** : 2017 attention mechanisms sans CNN ou RNN
- **BERT** : 2018 evolution de l'architecture Transformers
- **GPT** 1, 2, 3, 4: de 2018 à 2023 : transformers; modèles de plus en plus grand avec des datasets d'entraînement de plus en plus gigantesques
- Les LLMs : GPT4o, Claude Opus, Mistral, Gemini, LLama, ….

Du décompte des mots aux LLMs - Ce qui compte c'est l'attention

2017 - Attention Is All You Need - Transformers

https://arxiv.org/abs/1706.03762

- mélange de CNN et de RNN
- focus sur certain mots : l'attention
  - multi-head attention : focus sur plusieurs mots à la fois
  - self-attention : lien entre les mots



Ressources sur les transformers

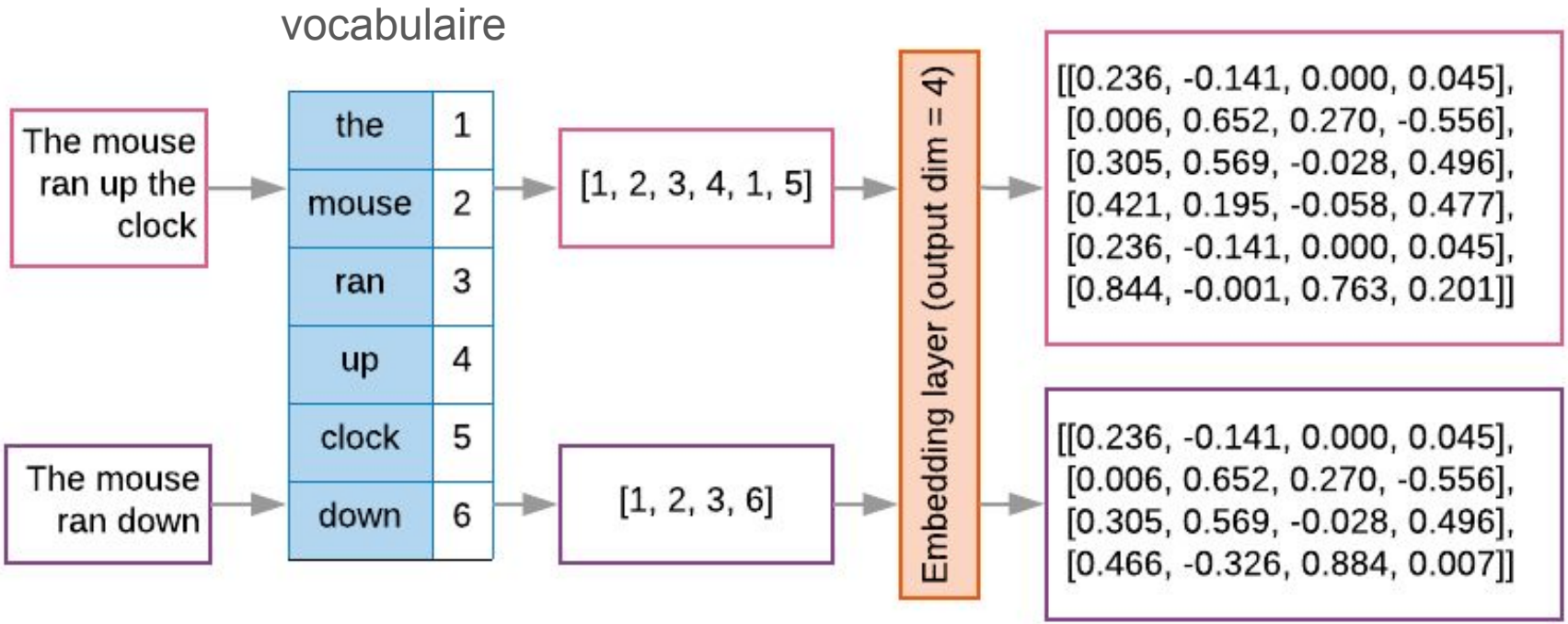
- The Illustrated Transformer : https://jalammar.github.io/illustrated-transformer/
- Analyse de l'architecture : https://kikaben.com/transformers-encoder-decoder/

Preparer les données

Preparer les données

Comment passer d'un texte de dimension variable à une représentation numérique que l'on peut utiliser pour entraîner un modèle ?

1. **Tokenization** : processus de découpage d'un texte en unités élémentaires appelées "tokens". Un token peut être un mot, un signe de ponctuation ou un autre élément atomique.
2. Sequencage (prochaine slide)
3. **Embedding** : remplacer chaque token par son équivalent vectoriel


vocabulaire
The mouse ran up the clock
The mouse ran down
the 1
mouse 2
ran 3
up 4
clock 5
down 6
[1, 2, 3, 4, 1, 5]
[1, 2, 3, 6]
Embedding layer (output dim = 4)
[[0.236, -0.141, 0.000, 0.045],
[0.006, 0.652, 0.270, -0.556],
[0.305, 0.569, -0.028, 0.496],
[0.421, 0.195, -0.058, 0.477],
[0.236, -0.141, 0.000, 0.045],
[0.844, -0.001, 0.763, 0.201]]
[[0.236, -0.141, 0.000, 0.045],
[0.006, 0.652, 0.270, -0.556],
[0.305, 0.569, -0.028, 0.496],
[0.466, -0.326, 0.884, 0.007]]
tokenisation des phrases
représentation vectorielle

Preparer les données

3. Sequencage : découper le texte en séquence de même longueurs avec du padding

"Dynamic Padding"

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	
1	_Eh	_bien	_c	'	_est	_un	_bon	indicateu	[PAD]	[PAD]	[PAD]	[PAD]	[PAD]		
2	_Ouais	_je	_suis	_un	_coureur	[PAD]	[PAD]	[PAD]	[PAD]	[PAD]	[PAD]	[PAD]	[PAD]		Batch Length: **13**
3	_Ils	_ne	_sont	_pas	important	_	.	[PAD]	[PAD]	[PAD]	[PAD]	[PAD]	[PAD]		
4	_Il	_y	_a	_de	nombreus	condition	_qui	_ne	_sont	_pas	_visibles	_	.		
5	_Chaque	_zone	_de	_l	'	_île	_offre	_quelque	_chose	_de	_différent	_	.		
6	_Mais	_tu	_peux	_vivre	_avec	_eux	_	.	[PAD]	[PAD]	[PAD]	[PAD]	[PAD]		Batch Length: **13**
7	_Un	_grand	_homme	_	,	_dit	-	il	_	.	[PAD]	[PAD]	[PAD]		
8	_Elle	_a	_été	_menée	_en	_silence	_	.	[PAD]	[PAD]	[PAD]	[PAD]	[PAD]		
9	_Tu	er	beaucou	_de	_fourmis	_de	_feu	[PAD]	[PAD]	[PAD]	[PAD]	[PAD]	[PAD]	[PAD]	
10	_La	_questior	_est	_de	_savoir	_si	_clin	ton	_a	_le	_cul	ot	_	.	Batch Length: **14**
11	_C	'	_est	_vrai	_	.	[PAD]	[PAD]	[PAD]	[PAD]	[PAD]	[PAD]	[PAD]	[PAD]	
12	_Dans	_ce	_domaine	_	,	_seuls	_les	_sa	ther	i	_le	_savent	_	.	

Total Tokens: **160**

Text classification

https://keras.io/examples/nlp/text_classification_from_scratch/

https://www.geeksforgeeks.org/sentiment-analysis-with-an-recurrent-neural-networks-rnn/?ref=ml_lbp

https://www.geeksforgeeks.org/sentiment-classification-using-bert/?ref=lbp

https://kikaben.com/transformers-encoder-decoder/

https://keras.io/examples/nlp/text_classification_from_scratch/

https://neptune.ai/blog/bert-and-the-transformer-architecture

https://galaxyinferno.com/3-lessons-from-the-paper-attention-is-all-you-need-as-a-beginner/

https://heidloff.net/article/foundation-models-transformers-bert-and-gpt/

https://towardsdatascience.com/attention-is-all-you-need-discovering-the-transformer-paper-73e5ff5e0634

https://github.com/ageron/handson-ml3/blob/main/16_nlp_with_rnns_and_attention.ipynb

with the invention of the transformer architecture are LSTM still relevant ?

https://chatgpt.com/c/e53f96b6-388f-403c-90f3-bd6bb2890b1e
https://claude.ai/chat/d52fa833-e313-441a-ba55-72bd3a76e223

transformers 2022.02 Transformers - Lucas Beyers
https://www.youtube.com/watch?v=UpfcyzoZ644

Attention is all you need
https://arxiv.org/abs/1706.03762