

Attention, Transformers et LLM

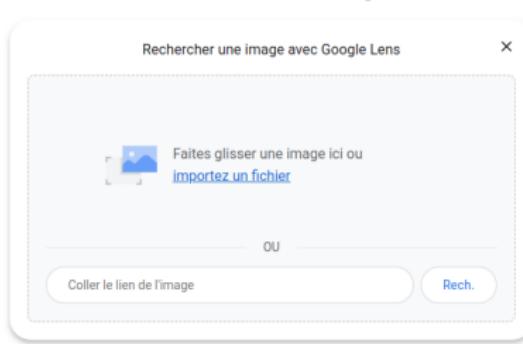
A. Carlier

2025

Plan du cours

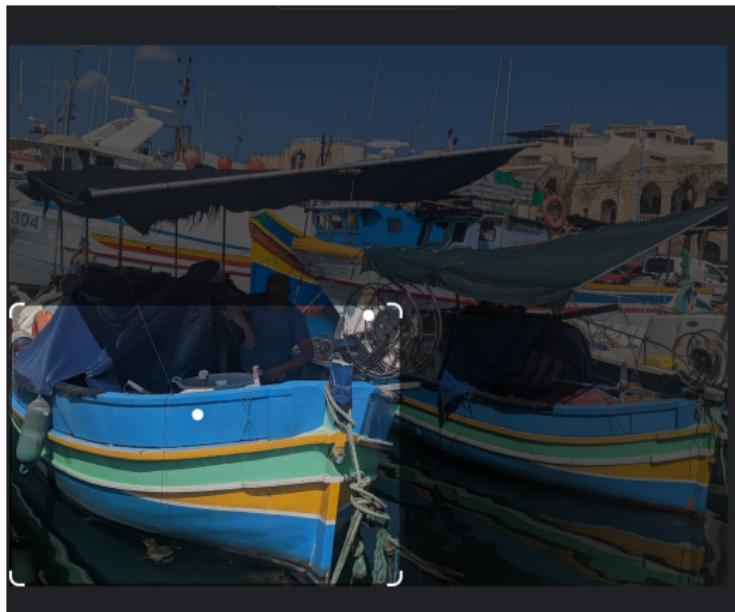
- 1 Mécanisme d'attention
- 2 Transformers
- 3 Modèles de Langage
- 4 De GPT à ChatGPT
- 5 Architectures multimodales
- 6 Limites et défis
- 7 Retrieval Augmented Generation

Recherche d'information



Content-Based Image Retrieval
Recherche d'information basée contenu

Recherche d'information



alamyimages.fr
Bateau de pêche en bois traditionnel maltais



alamyimages.fr
Malta marsaxlokk Banque de...
alamyimages.fr



viator.com
Tours et billets - Marsaxlokk - Réservez...
alamyimages.fr



alamyimages.fr
La Tunisie, le Cap Bon, le port de pêche de...
alamyimages.fr



alamyimages.fr
Un pêcheur maltais dans un port de...



journaldemontréal.ca
Malte, l'île aux 365 clochers | Le Journal d...



pixers.fr
Poster Barques colorées, Malte - ...
alamyimages.fr



alamyimages.fr
Marsaskala marsascala malta Banque de...



flickr.com
Babour (bateau de pêche à moteur) et...



alamyimages.fr
Maltese man Banque de photographies et...



came-true.com
Malte : Marsaxlokk, petit village de...
alamanchelibre.fr



loozap.com
Bateau pêche | Monastir | Tunisie | Loozap



lamanchelibre.fr
Vie du port. Manche :

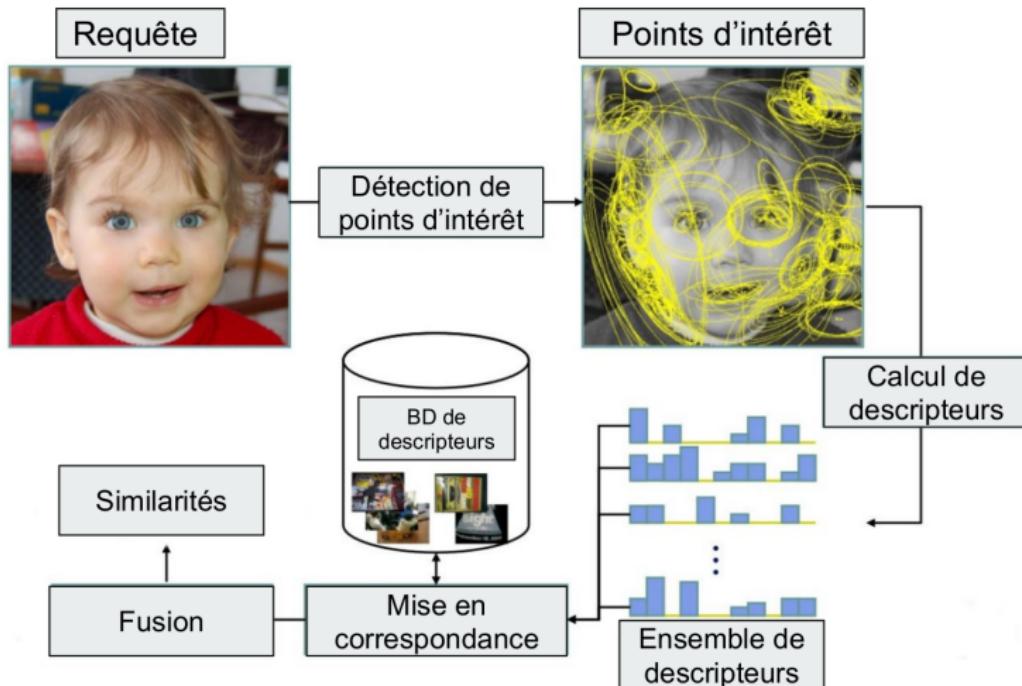


dramstime.com
Marsaxlokk, Malte - Bateau De Pêche...

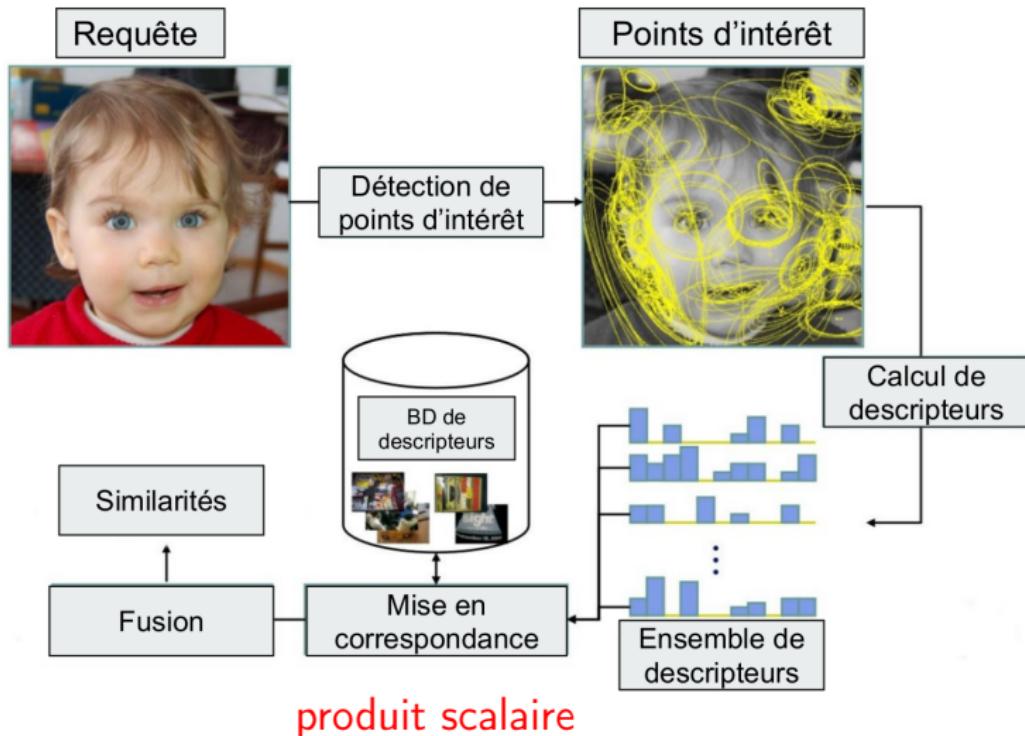


dreamstime.com
Marsaxlokk, Malte - Bateau De Pêche...

Recherche d'information



Recherche d'information

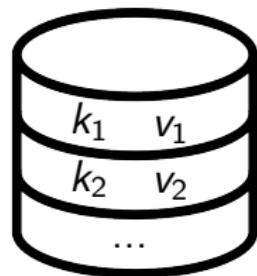


Module d'attention

Recherche d'information dans une base de données :

Soit une requête q , on peut comparer la requête aux clés k_i des différentes valeurs stockées en base de données en calculant le produit scalaire qk_i^T .

La réponse renvoyée à cette requête q sera la valeur correspondant à la clé dont le produit scalaire avec q était maximal.

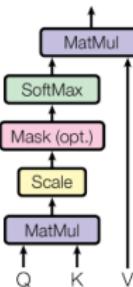


Module d'attention

On appelle **attention** l'implantation de ce mécanisme dans un réseau de neurones. On considère une matrice K contenant les clés, une matrice V contenant les valeurs et une matrice Q contenant les requêtes :

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Les clés et les requêtes sont de même dimension d_k et le facteur de normalisation permet de conserver des valeurs dans une zone où les gradients ne sont pas trop faibles.

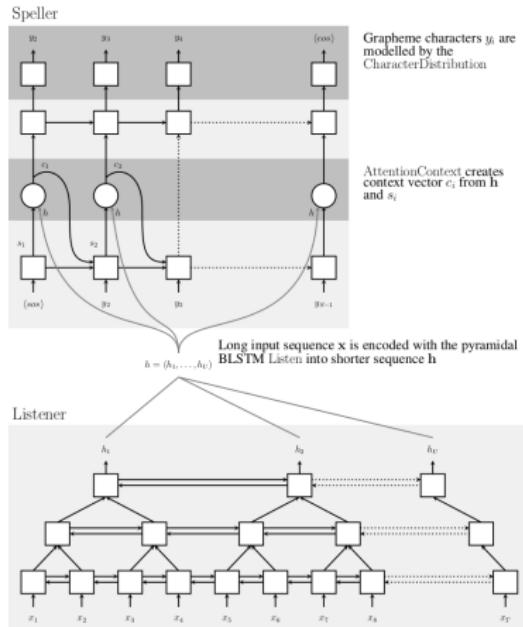


Listen, Attend and Spell

En 2015, l'état de l'art des modèles *Seq2seq* (séquence à séquence) utilise des LSTM et un modèle d'attention.

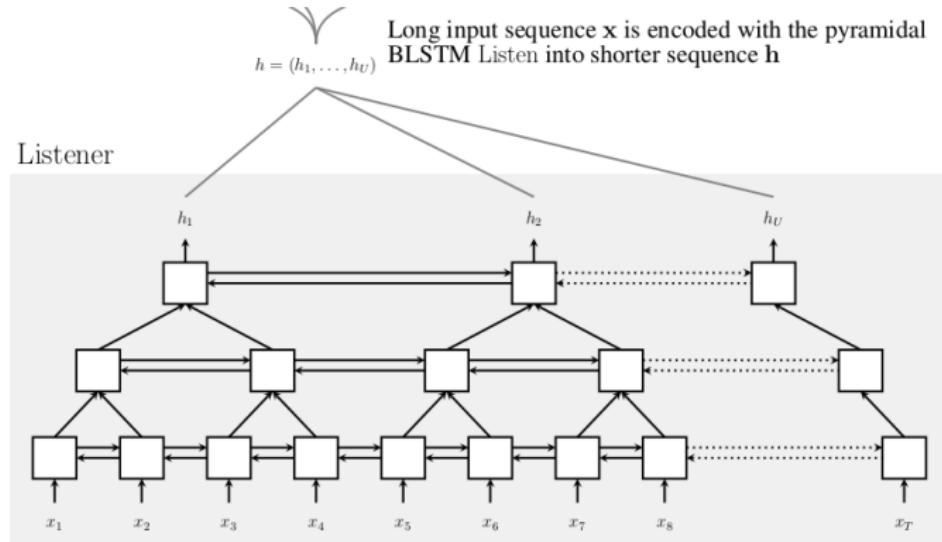
Il s'agit d'un modèle encodeur-décodeur pour la transcription automatique, où l'encodeur “écoute” (Listener) le signal audio et le décodeur “épelle” (spell) sa transcription.

[Chan 2015] Listen, Attend and Spell



Listen

LSTM bidirectionnel pyramidal, avec diminution de la dimension à chaque nouvelle couche pour simplifier le signal d'entrée.



[Chan 2015] Listen, Attend and Spell

Spell

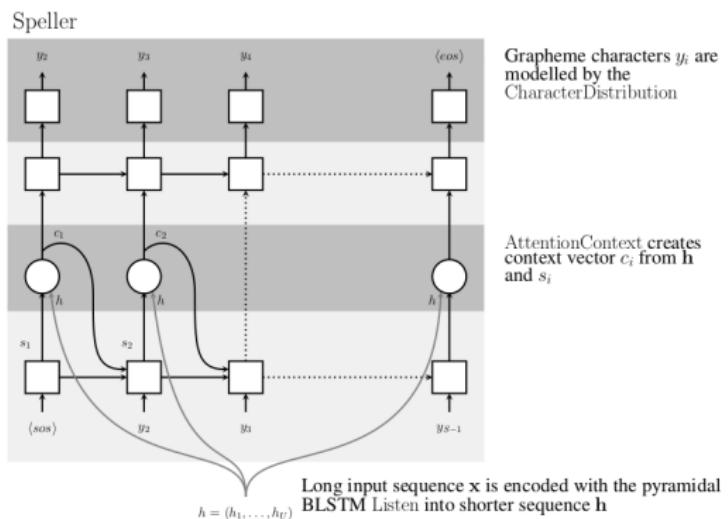
LSTM avec attention qui permet de générer une séquence en portant, à différents moments, un poids (une attention) particulier à différents éléments de la séquence h .

$$e_{i,u} = \langle \phi(s_i), \psi(h_u) \rangle$$

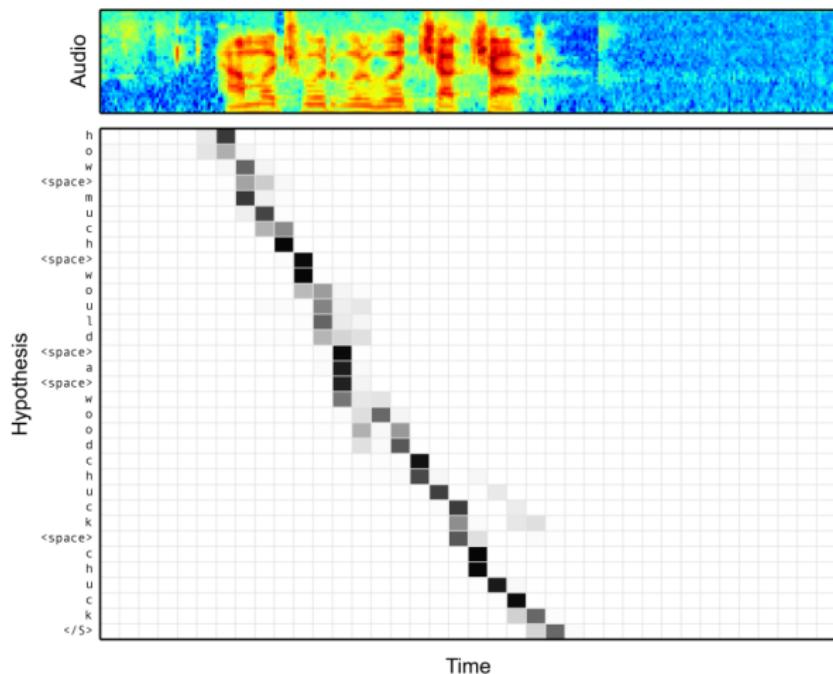
$$\alpha_{i,u} = \frac{\exp(e_{i,u})}{\sum_u \exp(e_{i,u})}$$

$$c_i = \sum_u \alpha_{i,u} h_u$$

[Chan 2015] Listen, Attend and Spell

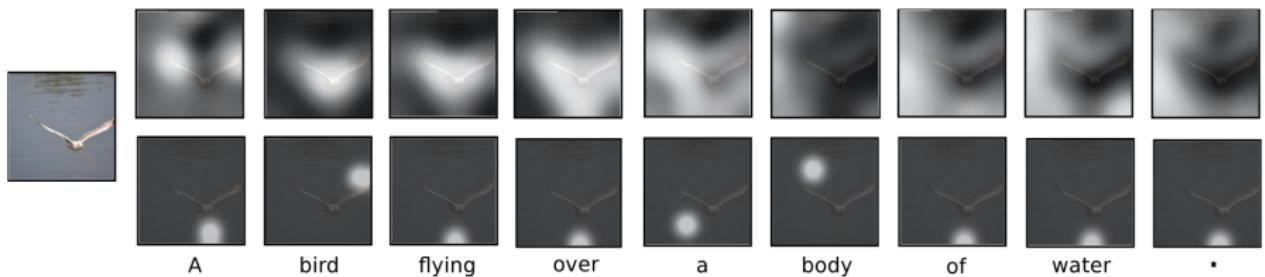
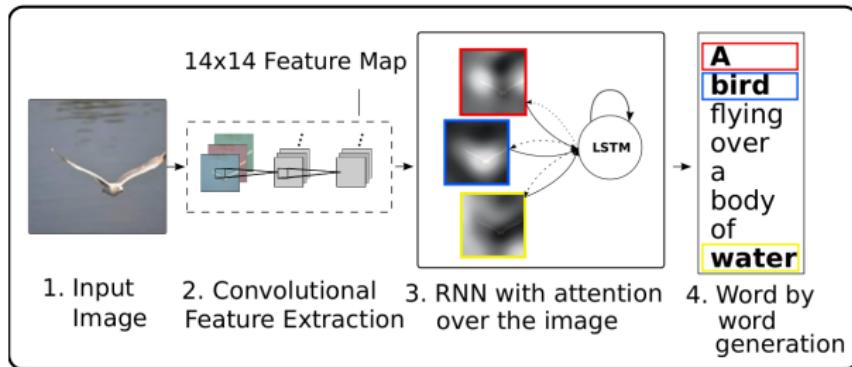


Listen, Attend and Spell



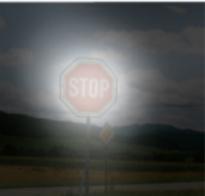
Visualisation de l'attention portée au signal d'entrée pour la génération de chaque caractère en sortie
[Chan 2015] Listen, Attend and Spell

Show, Attend and Tell



[Xu 2016] Show, Attend and Tell : Neural Image Caption Generation with Visual Attention

Show, Attend and Tell



A woman is throwing a frisbee in a park.

A dog is standing on a hardwood floor.

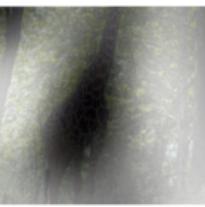
A stop sign is on a road with a mountain in the background.



A little girl sitting on a bed with a teddy bear.



A group of people sitting on a boat in the water.



A giraffe standing in a forest with trees in the background.

[Xu 2016] Show, Attend and Tell : Neural Image Caption Generation with Visual Attention

Plan du cours

- 1 Mécanisme d'attention
- 2 Transformers
- 3 Modèles de Langage
- 4 De GPT à ChatGPT
- 5 Architectures multimodales
- 6 Limites et défis
- 7 Retrieval Augmented Generation

Attention Is All You Need

Ashish Vaswani*

Google Brain

avaswani@google.com

Noam Shazeer*

Google Brain

noam@google.com

Niki Parmar*

Google Research

nikip@google.com

Jakob Uszkoreit*

Google Research

usz@google.com

Llion Jones*

Google Research

llion@google.com

Aidan N. Gomez* †

University of Toronto

aidan@cs.toronto.edu

Lukasz Kaiser*

Google Brain

lukaszkaiser@google.com

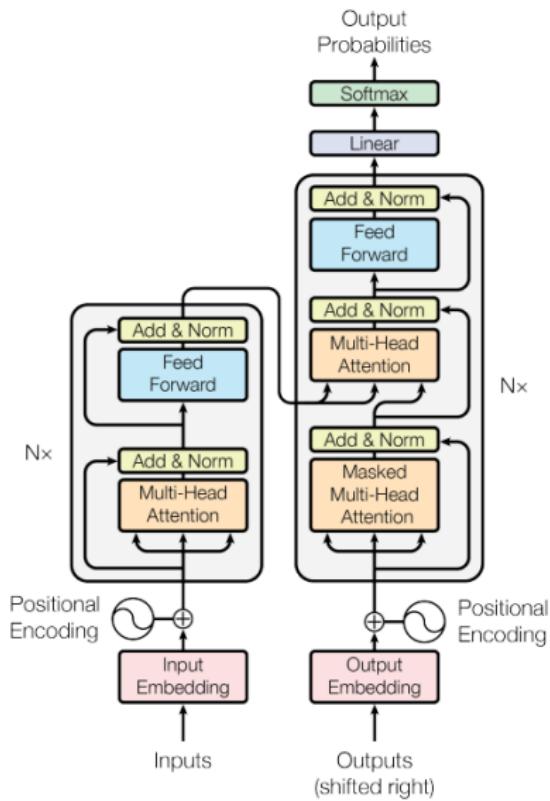
Illia Polosukhin* ‡

illia.polosukhin@gmail.com

Architecture d'un *Transformer*

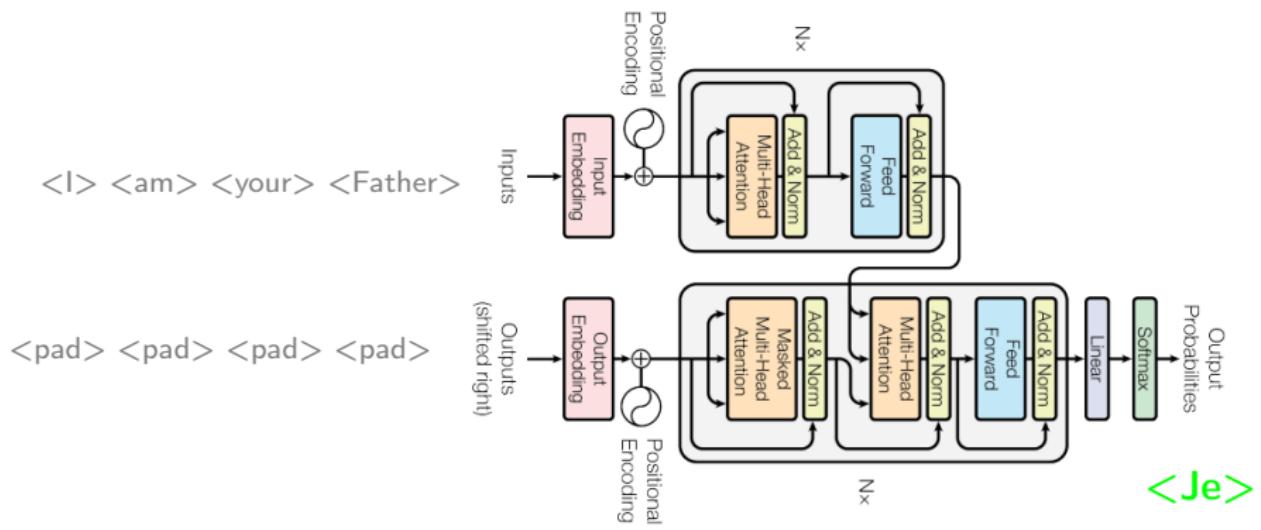
Remplacement des couches récurrentes ou convolutives par un mécanisme d'auto-attention (*self-attention*).

Il s'agit à nouveau d'un modèle encodeur-décodeur pour des problèmes *seq2seq*. L'encodeur prend en entrée un ensemble de *tokens*, le décodeur prend également en entrée un ensemble de *tokens* et prédit un *token* en sortie (via une distribution de probabilités sur les *tokens* possibles).



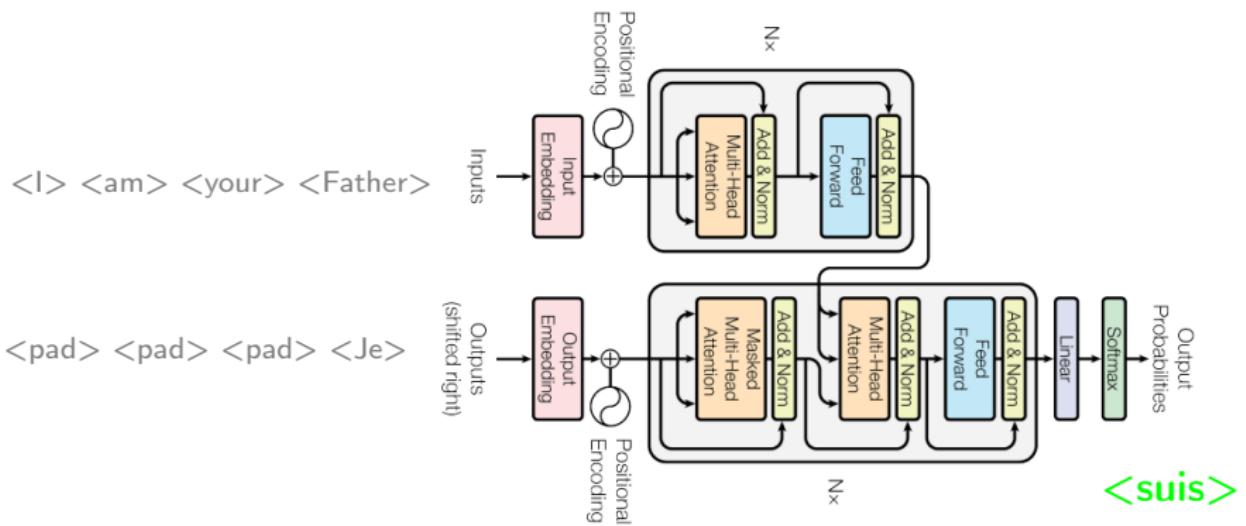
Utilisation du Transformer

Exemple de la traduction automatique :



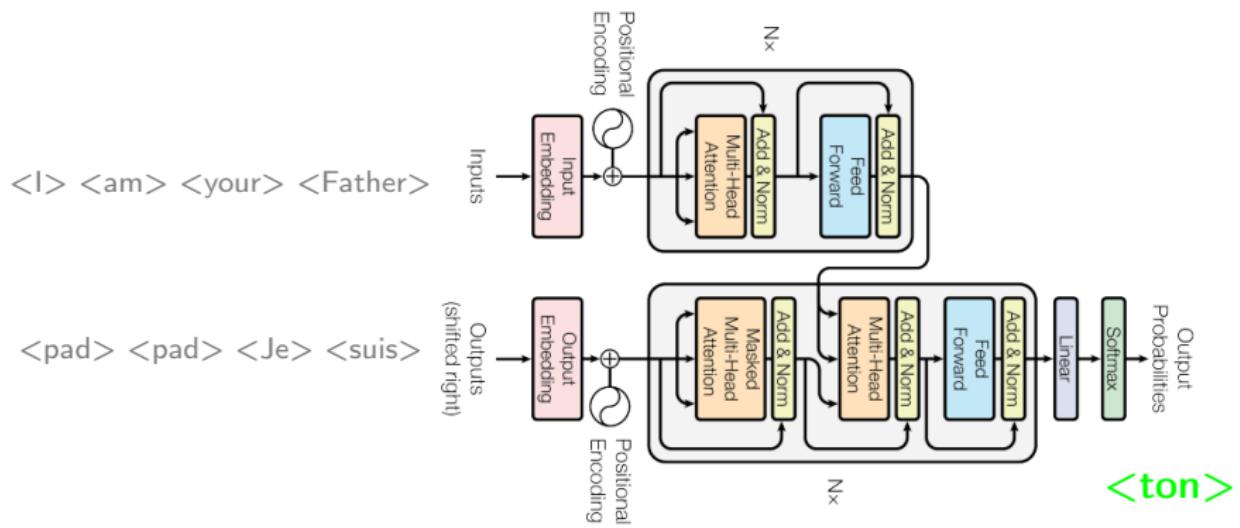
Utilisation du Transformer

Exemple de la traduction automatique :



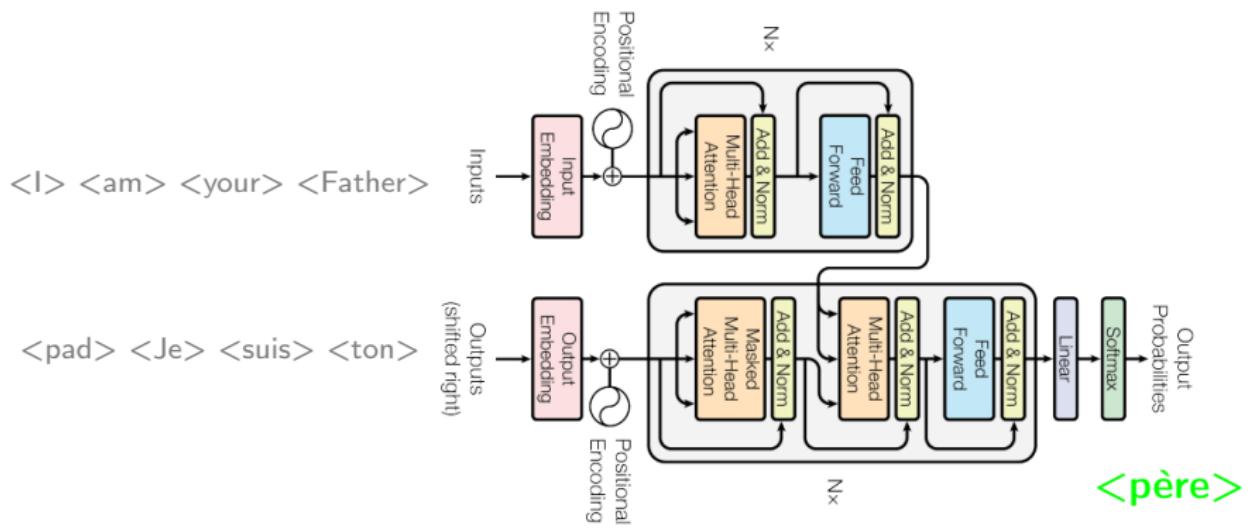
Utilisation du Transformer

Exemple de la traduction automatique :



Utilisation du Transformer

Exemple de la traduction automatique :

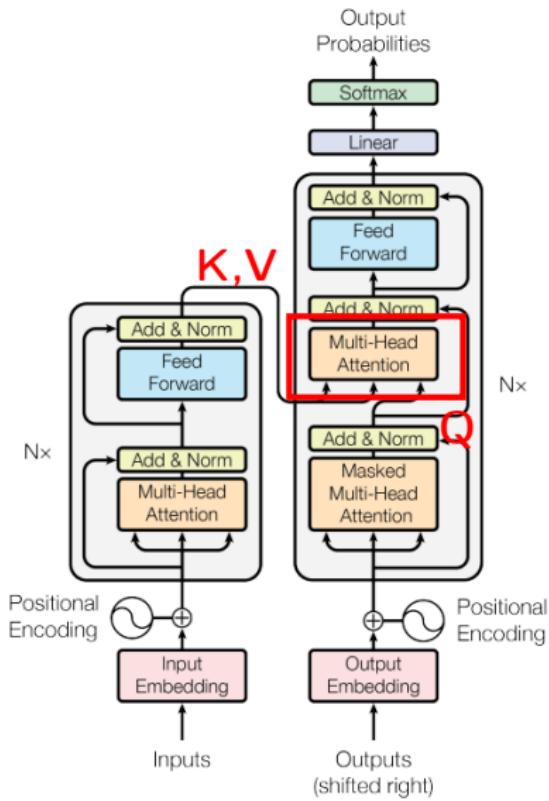


Architecture d'un Transformer

L'encodeur analyse la séquence d'entrée et produit un ensemble de clés et de valeurs.

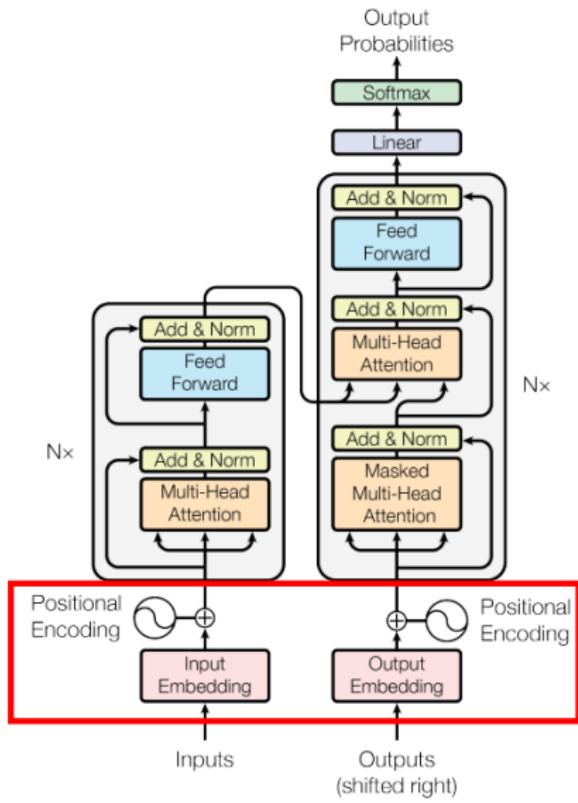
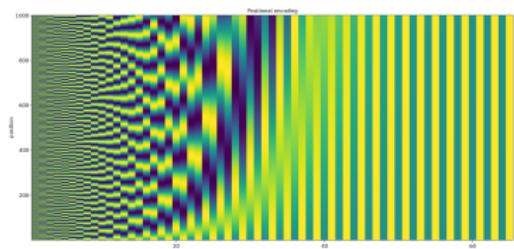
Le décodeur analyse la séquence de sortie et produit une requête qui lui permet de se focaliser sur les parties de l'entrée les plus pertinentes à sa prédiction.

Une séquence de sortie porte attention sur une séquence d'entrée : on parle de *cross-attention*.



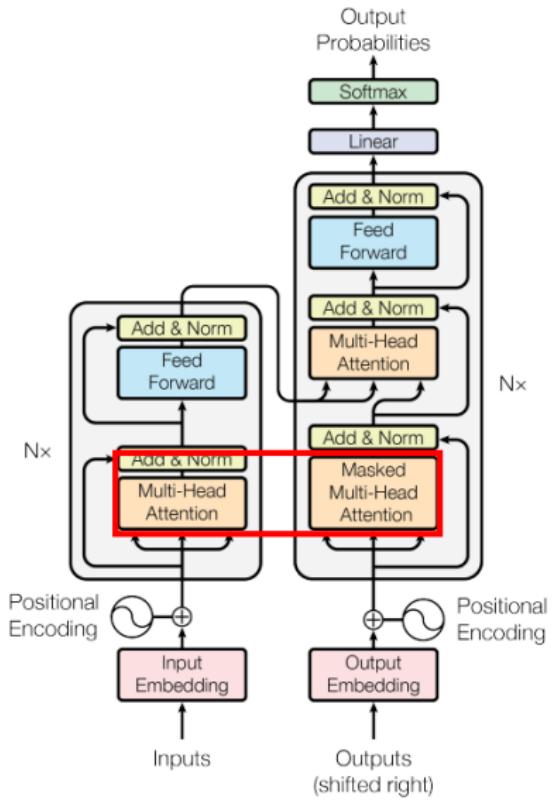
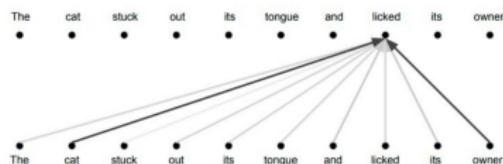
Architecture d'un Transformer

Il n'y a pas de notion d'ordre, ou de séquence, sur les tokens d'entrée. Les descripteurs associés à chaque token sont sommés à un descripteur de position (*positional encoding*), unique, qui peut également être appris (l'est souvent dans les travaux suivants).



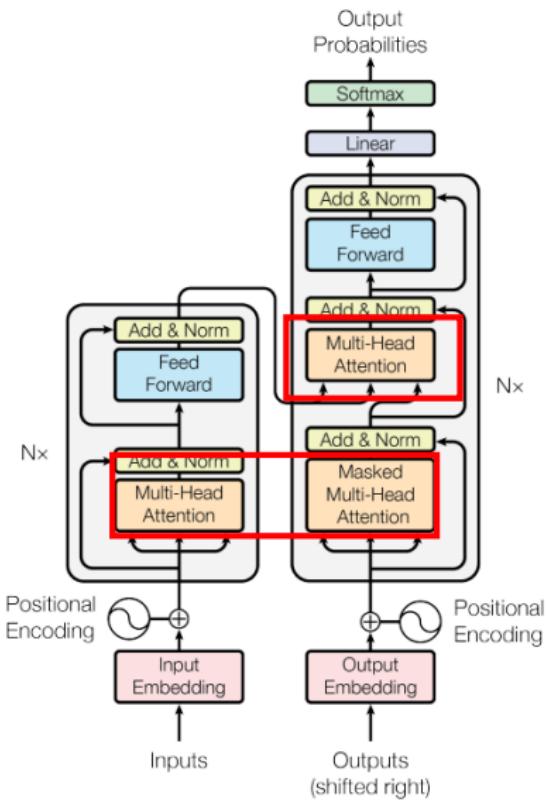
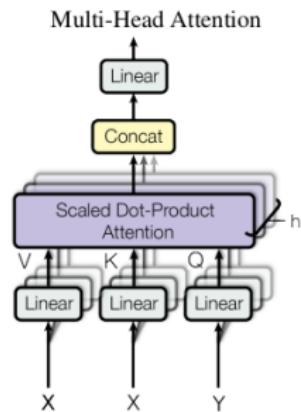
Architecture d'un Transformer

Les séquences d'entrée et de sortie sont analysées via un mécanisme d'auto-attention. Les clés, valeurs, et requêtes sont toutes générées à partir de la même séquence.



Architecture d'un Transformer

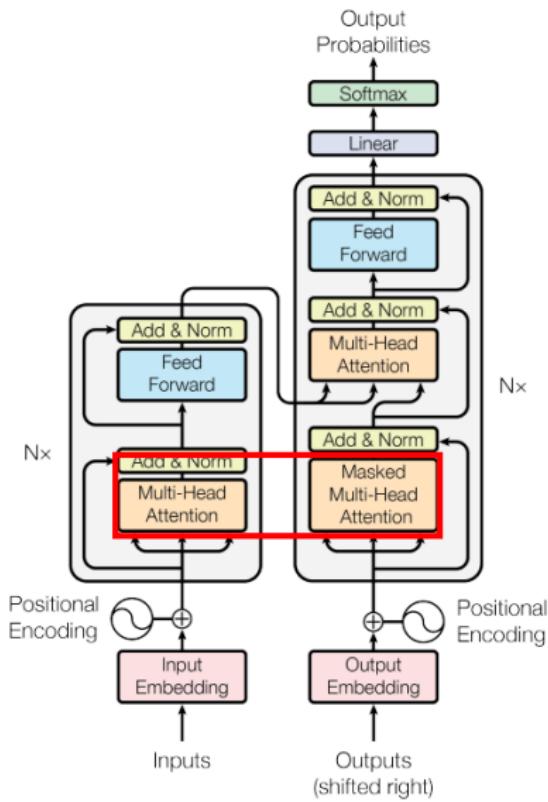
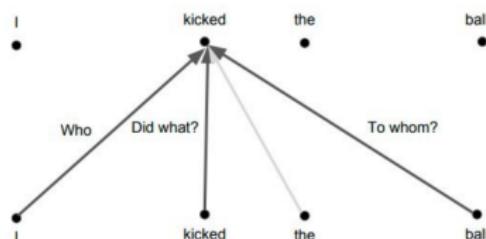
Le mécanisme d'attention est dit à "plusieurs têtes" (multi-head), ce qui permet de porter attention sur plusieurs éléments de la séquence avec des regards différents.



Nombre de paramètres ?

Architecture d'un Transformer

Les mécanismes d'attention et de *self-attention* ont “plusieurs têtes” (*Multi-head*) : plusieurs ensembles différents de clés, valeurs et requêtes permettent de porter attention à différents aspects.

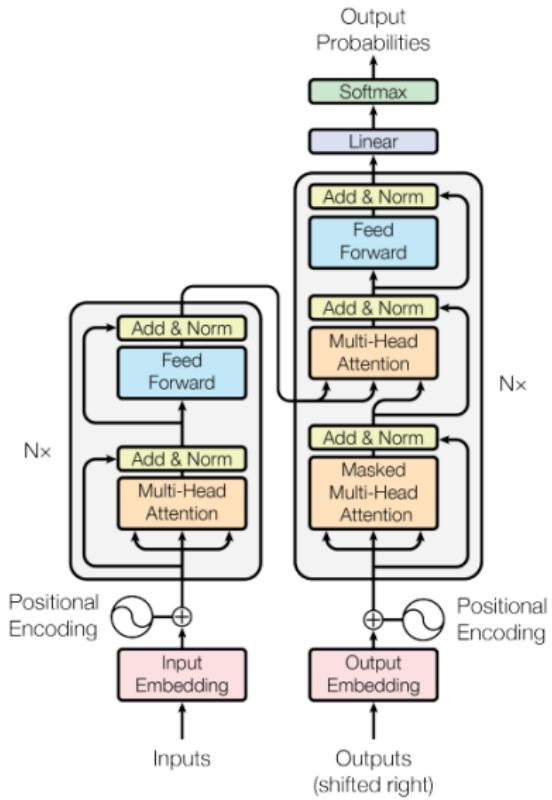


Architecture d'un *Transformer*

Dans l'architecture ci-contre :

- Nombre de répétitions de chaque bloc : $N = 6$
- 8 têtes d'attention en parallèle
- Dimension maintenue pour la représentation interne des *tokens* : $d = 512$

65 millions de paramètres



Intérêt de l'auto-attention

- **Complexité plus faible** que les couches récurrentes dans les cas où la longueur de la séquence est inférieure à la dimension de la représentation maintenue ($n < d$)
- Plus de **parallélisation** possible que pour les couches récurrentes car pas de séquentialité nécessaire.
- Le **chemin minimal** dans le réseau connectant deux éléments de la séquence est beaucoup plus **court**, ce qui favorise l'apprentissage de "dépendances à long terme".

Layer Type	Complexity per Layer	Sequential Operations	Maximum Path Length
Self-Attention	$O(n^2 \cdot d)$	$O(1)$	$O(1)$
Recurrent	$O(n \cdot d^2)$	$O(n)$	$O(n)$
Convolutional	$O(k \cdot n \cdot d^2)$	$O(1)$	$O(\log_k(n))$

où n est la longueur de la séquence, d est la dimension de la représentation, et k la dimension du noyau de convolution.

Transfert d'apprentissage

Les représentations apprises par les transformers sont transférables, et, il est possible de pré-entraîner les transformers de manière non supervisée (cf. BERT, GPT).

Ce n'est pas le cas des réseaux récurrents, pour lesquels le transfert d'apprentissage n'a jamais réellement fonctionné.

LSTM vs. Transformers

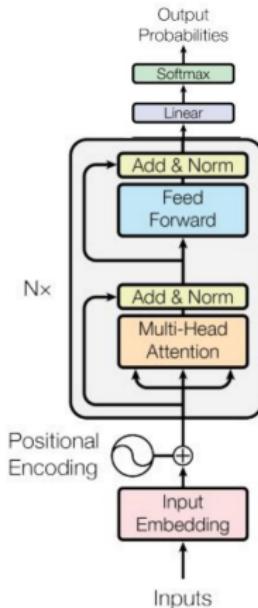
- Les Transformers ont maintenant majoritairement remplacé les LSTM pour les tâches séquentielles (NLP, audio, vidéo).
- Les LSTM sont toujours utilisés dans deux cas de figure :
 - ▶ Séquences très longues (complexité en $O(n^2)$ des Transformers),
 - ▶ Pas de large base de données pour pré-entraîner les Transformers (sur de petits échantillons et sans pré-entraînement, LSTM > Transformers)

Plan du cours

- 1 Mécanisme d'attention
- 2 Transformers
- 3 Modèles de Langage
- 4 De GPT à ChatGPT
- 5 Architectures multimodales
- 6 Limites et défis
- 7 Retrieval Augmented Generation

Google BERT

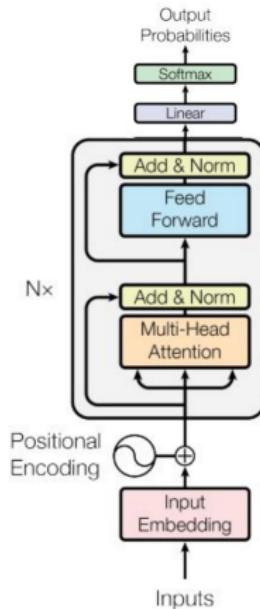
- Architecture ressemblant à l'encodeur du Transformer originel : l'attention peut être portée à tous les *tokens* de la séquence (bidirectionnel).



Google BERT

- Nombre de répétitions de chaque bloc :
 $N = 24$
- 16 têtes d'attention en parallèle
- Dimension maintenue pour la représentation interne des *tokens* :
 $d = 1024$

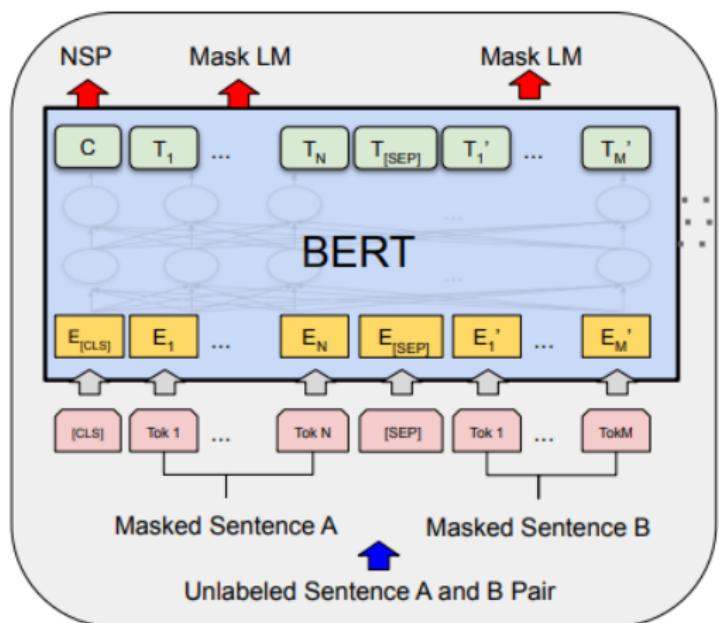
340 millions de paramètres



[Devlin 2018] BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding

Google BERT - pré-entraînement

Pré-entraînement non-supervisé sur de larges corpus de textes, avec masquage de mots dans la séquence et prédiction de la prochaine phrase.

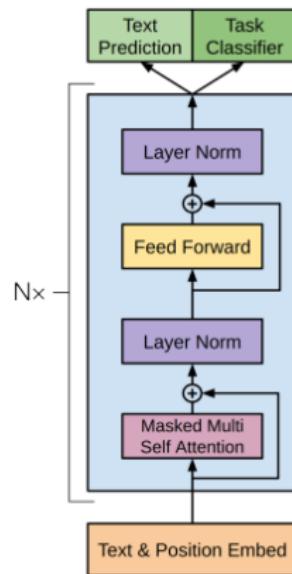


[Devlin 2018] BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding

OpenAI GPT

- Architecture ressemblant au décodeur du Transformer originel : l'attention ne peut être portée qu'aux tokens précédents dans la séquence.
- Pré-entraînement non-supervisé sur un large corpus de textes (BooksCorpus : 11000 livres non publiés, 1 milliard de mots).

A obtenu les meilleures performances de l'état de l'art sur 9 tâches simultanément (Similarité entre phrases, classification, réponse à une question, etc.)

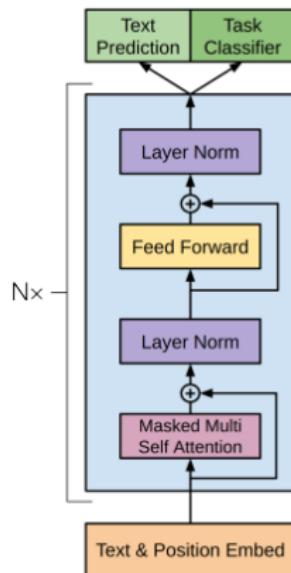


[Radford 2018] Improving Language Understanding by Generative Pre-Training

OpenAI GPT

- Nombre de répétitions de chaque bloc :
 $N = 12$
- 12 têtes d'attention en parallèle
- Dimension maintenue pour la représentation interne des *tokens* :
 $d = 768$

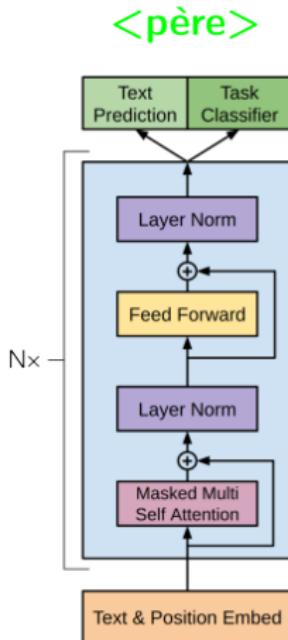
117 millions de paramètres



[Radford 2018] Improving Language Understanding by Generative Pre-Training

OpenAI GPT

- Pré-entraînement non-supervisé sur un large corpus de textes (BooksCorpus : 11000 livres non publiés, 1 milliard de mots).



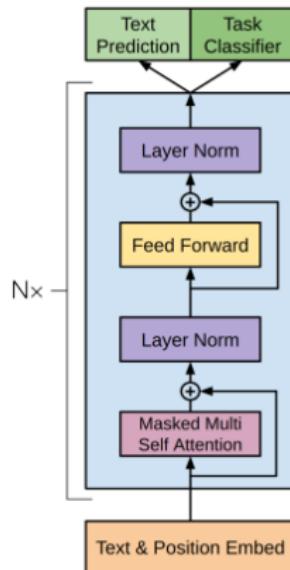
<Je> <suis> <ton> <mask>

[Radford 2018] Improving Language Understanding by Generative Pre-Training

OpenAI GPT2

- Nombre de répétitions de chaque bloc :
 $N = 48$
- 12 têtes d'attention en parallèle
- Dimension maintenue pour la représentation interne des *tokens* :
 $d = 1600$

1,5 milliards de paramètres



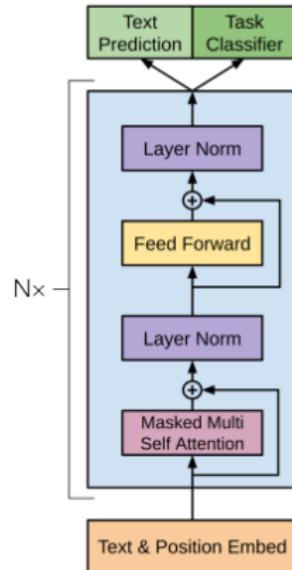
Pré-entraînement sur WebText, 8 millions de pages Web représentant plus de 40 Go de texte

[Radford 2019] Language Models are unsupervised multitask learners

OpenAI GPT3

- Nombre de répétitions de chaque bloc :
 $N = 96$
- 96 têtes d'attention en parallèle
- Dimension maintenue pour la représentation interne des *tokens* :
 $d = 12888$ (!!)

175 milliards de paramètres



Pré-entraînement sur CommonCrawl (45 To de texte), WebText, des bases de données de livres, ainsi que Wikipedia (qui ne représente que 3% du total des données)

[Brown 2020] Language models are few shot learners

Llama

Llama v1 compte 4 tailles de modèles

params	dimension	n heads	n layers	learning rate	batch size	n tokens
6.7B	4096	32	32	$3.0e^{-4}$	4M	1.0T
13.0B	5120	40	40	$3.0e^{-4}$	4M	1.0T
32.5B	6656	52	60	$1.5e^{-4}$	4M	1.4T
65.2B	8192	64	80	$1.5e^{-4}$	4M	1.4T

LLAMA 2		Time (GPU hours)	Power Consumption (W)	Carbon Emitted (tCO ₂ eq)
	7B	184320	400	31.22
	13B	368640	400	62.44
	34B	1038336	350	153.90
	70B	1720320	400	291.42
Total		3311616		539.00

Llama v2 reprend à peu près les mêmes architectures et fournit une estimation du coût énergétique des entraînements.

Une vision d'ensemble

Major Large Language Models (LLMs)

ranked by capabilities, sized by billion parameters used for training

Parameters (Bn) open access

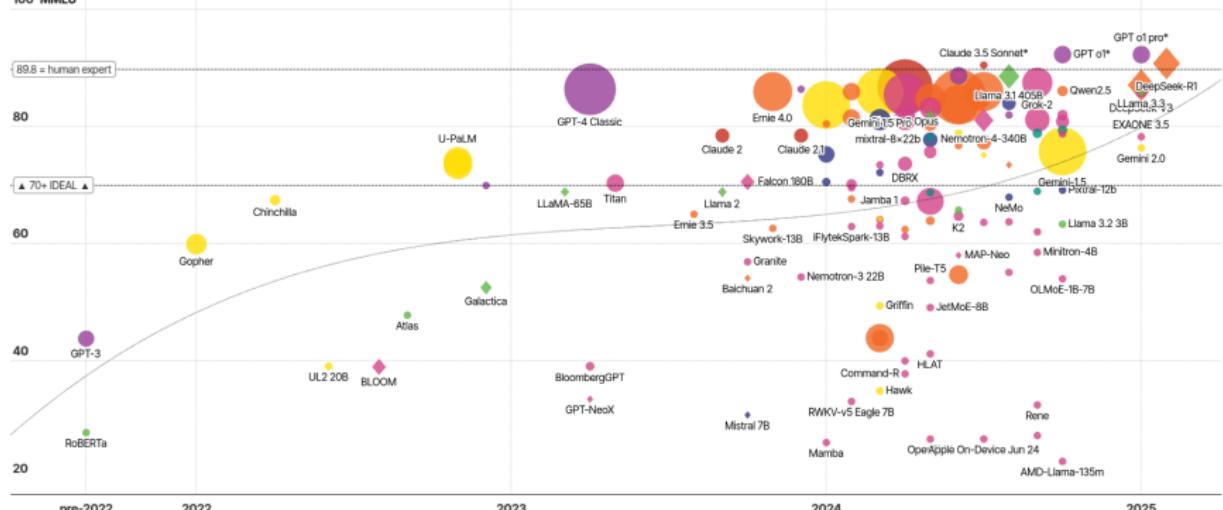
CLICK LEGEND ITEMS TO FILTER

anthropic chinese google meta microsoft mistral openAI other

search...

show only: all

100 MMLU



David McCandless, Tom Evans, Paul Barton
Informationisbeautiful // Jan 2024

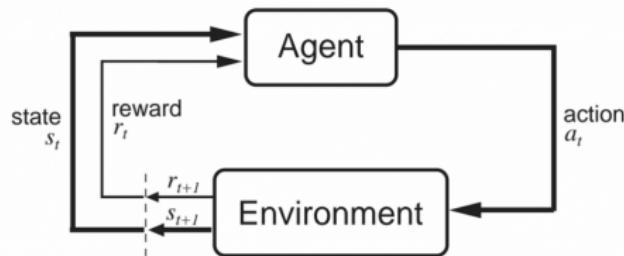
MMLU = benchmark for measuring LLM capabilities
* = parameters undisclosed // source: LifeArchitect // data

Plan du cours

- 1 Mécanisme d'attention
- 2 Transformers
- 3 Modèles de Langage
- 4 De GPT à ChatGPT
- 5 Architectures multimodales
- 6 Limites et défis
- 7 Retrieval Augmented Generation

Apprentissage par renforcement

- Définition d'un **agent** comportemental, qui peut prendre un ensemble de décisions (**actions**) en fonction de l'**état** d'un certain système
- L'agent obtient des **récompenses** pour chacune de ses actions
- L'objectif est d'apprendre une **politique**, c'est-à-dire une fonction pour déterminer l'action optimale à effectuer en fonction du contexte (état)

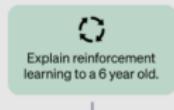


De GPT à ChatGPT

Step 1

Collect demonstration data and train a supervised policy.

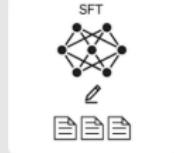
A prompt is sample from our prompt dataset.



A labeler demonstrates the desired output behavior.



This data is used to fine-tune GPT-3.5 with supervised learning.



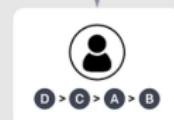
Step 2

Collect comparison data and train a reward model.

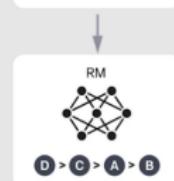
A prompt and several model outputs are sampled.



A labeler ranks the outputs from best to worst.



This data is used to train our reward model.



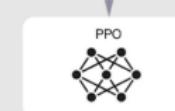
Step 3

Optimize a policy against the reward model using the PPO reinforcement learning algorithm.

A new prompt is sampled from the dataset.



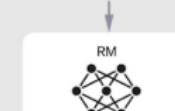
The PPO model is initialized from the supervised policy.



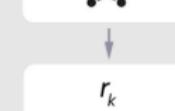
The policy generates an output.



The reward model calculates a reward for the output.



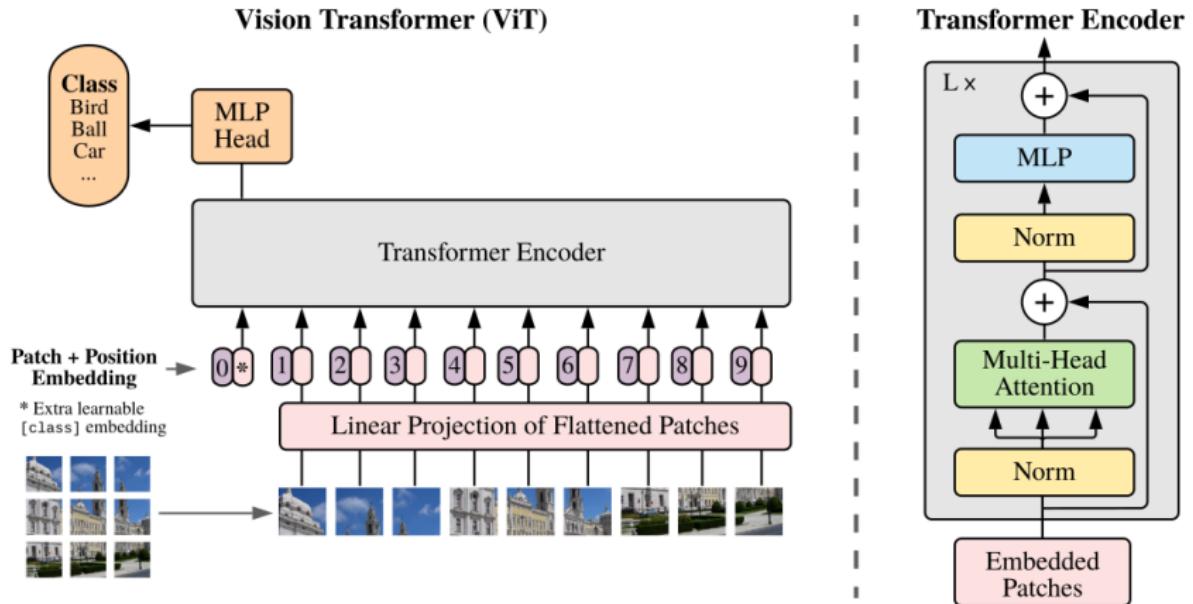
The reward is used to update the policy using PPO.



Plan du cours

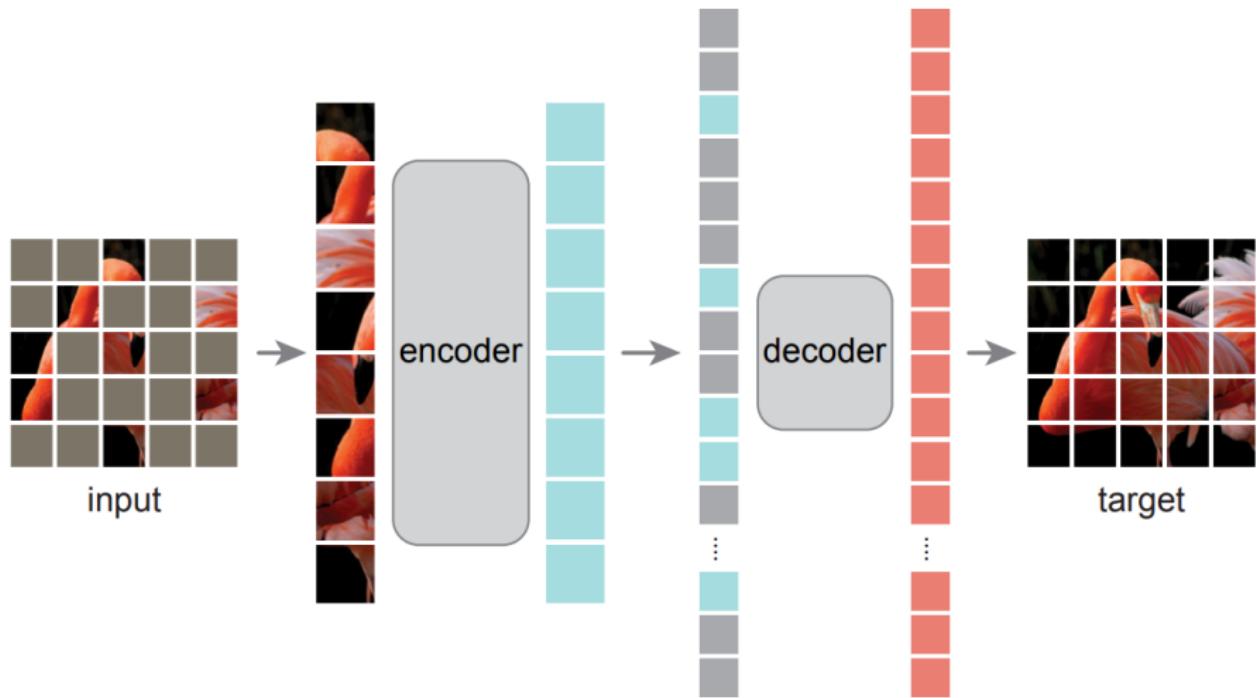
- 1 Mécanisme d'attention
- 2 Transformers
- 3 Modèles de Langage
- 4 De GPT à ChatGPT
- 5 Architectures multimodales
- 6 Limites et défis
- 7 Retrieval Augmented Generation

Vision Transformers (ViT)



[Dosovitskiy 2020] An Image is Worth 16×16 Words : Transformers for Image Recognition at Scale

Vision Transformers : pré-entraînement non supervisé



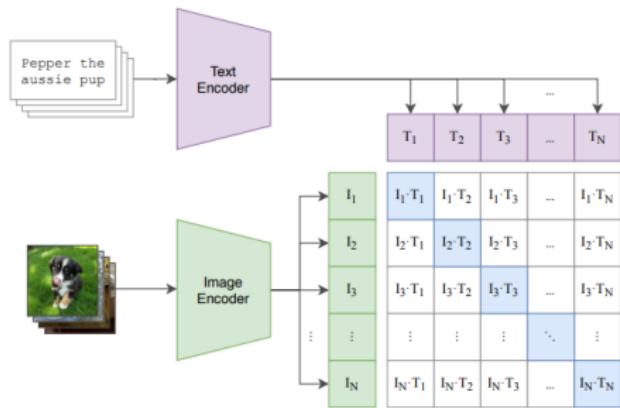
[He 2021] Masked Autoencoders Are Scalable Vision Learners

Le grand remplacement ?

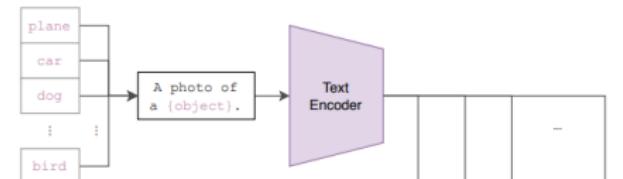
CoCa (finetuned)	91.0%	2100M	x	CoCa: Contrastive Captioners are Image-Text Foundation Models			2022	ALIGN	Transformer	JFT-3B	
Model soups (BASIC-L)	90.98%	2440M	x	Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time			2022	Conv+Transformer	JFT-3B	ALIGN	
Model soups (ViT-G/14)	90.94%	1843M	x	Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time			2022	JFT-3B	Transformer		
ViT-e	90.9%	3900M	x	PaLI: A Jointly-Scaled Multilingual Language-Image Model			2022	Transformer	JFT-3B		
CoAtNet-7	90.88%	2440M	2586	x	CoAtNet: Marrying Convolution and Attention for All			2021	Conv+Transformer	JFT-3B	

Modèles multimodaux : CLIP

(1) Contrastive pre-training



(2) Create dataset classifier from label text



(3) Use for zero-shot prediction

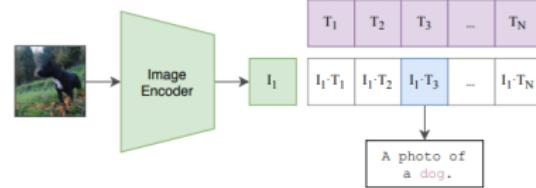


Image de [Radford et al.] Learning Transferable Visual Models From Natural Language Supervision.

Modèles multimodaux : exemple de Pixtral

Le modèle Pixtral de Mistral :

Pixtral Model Architecture

Pixtral also performs very well
as compared to closed
models like Claude-3 Haiku, ...


Multimodal Transformer Decoder



Text

Image

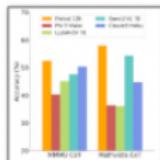
Text

Image

Vision Transformer
Encoder

Vision Transformer
Encoder

Pixtral is trained to understand both natural images and documents. The model shows strong abilities in document question answering, optical character recognition ...



Pixtral outperforms other open source models on multimodal understanding and reasoning benchmarks, such as MMLU and Mathvista ...

Model	MMLU left	Mathvista right	ChatGPT left	DiffVQA right	VQA4T right (mean)
Pixtral 1B	50.9	58.0	51.6	52.2	78.4
Claude-3 Haiku	50.4	44.9	52.6	70.3	65.4
Gemini-1.5B	50.7	56.9	57.0	-	65.5
QFT-4B	55.0	64.0	55.1	94.7	77.8
Claude-3.5 Sonnet	50.0	64.4	57.1	83.8	70.7

Plan du cours

- 1 Mécanisme d'attention
- 2 Transformers
- 3 Modèles de Langage
- 4 De GPT à ChatGPT
- 5 Architectures multimodales
- 6 Limites et défis
- 7 Retrieval Augmented Generation

Limites des LLM

Nous en avons déjà mentionné quelques-unes :

- Limitation de la taille du contexte (et donc de la longueur des échanges)
- Coût matériel **exorbitant** d'un entraînement et des prédictions
- Biais créés par le processus d'évaluation de ce qu'est une "bonne" réponse
- Propriété intellectuelle des données d'entraînement

Les LLM ne raisonnent pas (même s'ils en ont l'air)

GSM-NoOp

Oliver picks 44 kiwis on Friday. Then he picks 58 kiwis on Saturday. On Sunday, he picks double the number of kiwis he did on Friday, but five of them were a bit smaller than average. How many kiwis does Oliver have?

01-mini: Oliver picks kiwis over three days as follows:

- Friday: 44 kiwis
- Saturday: 58 kiwis
- Sunday: Double the number he picked on Friday, which is $2 \times 44 = 88$ kiwis

However, on Sunday, 5 of these kiwis were smaller than average. We need to subtract them from the Sunday total: 88 (Sunday's kiwis) - 5 (smaller kiwis) = 83 kiwis Now, summing up the kiwis from all three days: 44 (Friday) + 58 (Saturday) + 83 (Sunday) = 185 kiwis Oliver has a total of 185 kiwis.

Llama3-8B: Let's break it down step by step:

Oliver picks 44 kiwis on Friday. He picks 58 kiwis on Saturday. On Sunday, he picks double the number of kiwis he did on Friday, which is $2 \times 44 = 88$ kiwis.

Five of the kiwis picked on Sunday are a bit smaller than average, so we subtract 5 from the total number of kiwis picked on Sunday: $88 - 5 = 83$ kiwis. Now, let's add up the total number of kiwis Oliver has:

44 (Friday) + 58 (Saturday) + 83 (Sunday) = 185 kiwis

So, Oliver has 185 kiwis in total.

[Mirzadeh et al. 2024] GSM-Symbolic : Understanding the Limitations of Mathematical Reasoning in Large

Language Models

Les LLM peuvent inventer des réponses plausibles



Who was the first person to walk on the moon?



Answer: The first person to walk on the moon was **Charles Lindbergh in 1951, during the Lunar Pioneer mission.** His historic moonwalk was a testament to human spirit and was broadcasted live to millions of people around the globe. 



Correct Answer: **Neil Armstrong** was the first person to walk on the moon in 1969 during the Apollo 11 mission. 

(a) Factuality Hallucination



Please summarize the following news article:



Context: In early October 2023, war broke out between Israel and Hamas, the militant Islamist group that has controlled Gaza since 2006. Hamas fighters fired rockets ... civilians and taking dozens of hostages.

Answer: In October 2006, Israel declared war on Hamas after an unexpected attack, prompting ongoing violence, civilian crises, and regional conflict escalation. 

(b) Faithfulness Hallucination

Plan du cours

- 1 Mécanisme d'attention
- 2 Transformers
- 3 Modèles de Langage
- 4 De GPT à ChatGPT
- 5 Architectures multimodales
- 6 Limites et défis
- 7 Retrieval Augmented Generation

Depuis quelques semaines

Qui a été élu président des Etats-Unis en 2024 ?

Donald Trump a été élu président des États-Unis lors de l'élection présidentielle du 5 novembre 2024, battant la vice-présidente sortante, Kamala Harris. Il a obtenu 312 voix au collège électoral contre 226 pour Harris, ainsi que 49,91 % du vote populaire contre 48,44 % pour sa concurrente.

[WIKIPEDIA](#)

Cette victoire fait de Donald Trump le 47^e président des États-Unis et marque son retour à la Maison-Blanche après un premier mandat de 2017 à 2021. Il devient ainsi le premier président depuis Grover Cleveland en 1893 à effectuer deux mandats non consécutifs.

[WIKIPEDIA](#)

La réélection de Trump a suscité des interrogations et des inquiétudes sur la scène internationale, notamment concernant la politique étrangère des États-Unis et leur soutien à l'Ukraine dans le contexte du conflit avec la Russie.

[LEMONDE.FR](#)

Parallèlement, le Parti républicain a obtenu la majorité absolue à la Chambre des représentants et au Sénat, assurant à Donald Trump un soutien significatif du Congrès pour son mandat.

[LEMONDE.FR](#)

Réactions internationales à la réélection de Donald Trump

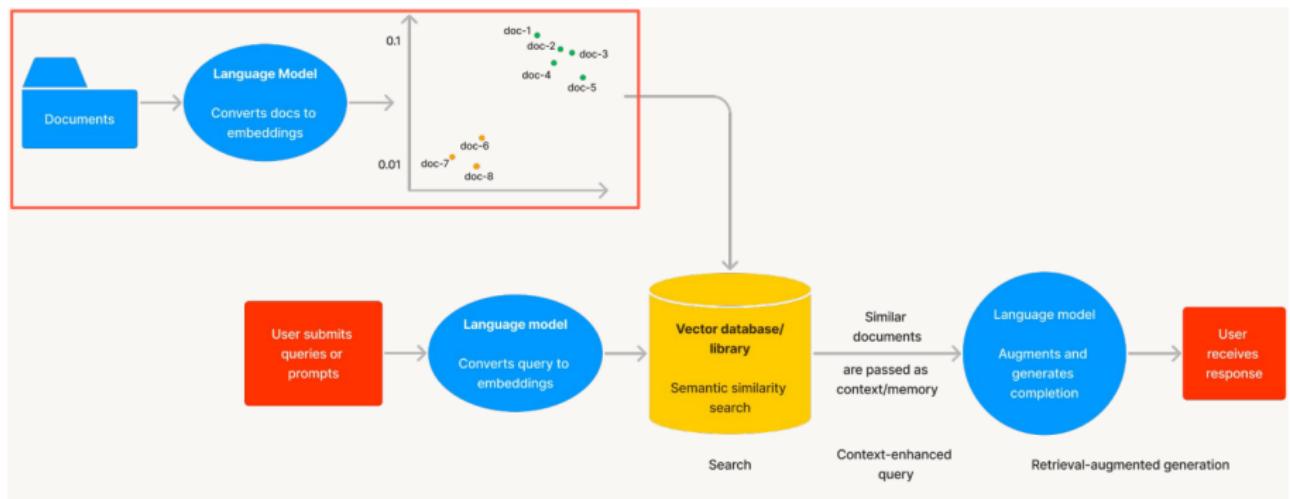
 [lemonde.fr](#)

Présidentielle américaine 2024 : beaucoup d'interrogations dans le monde après l'élection de Donald Trump

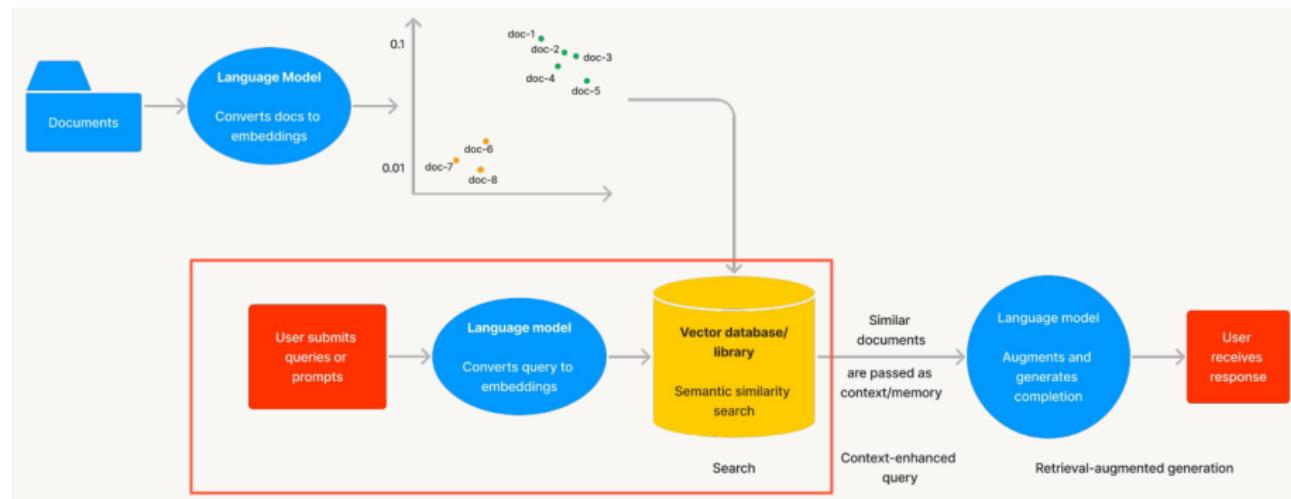
il y a 125 jours



Indexation



Recherche d'information



Génération "augmentée"

