Definition:

1. However, a variant of RNN, the Long Short-Term Memory  
   (LSTM), is paramount in mitigating some of these limitations. The LSTM provided a  
   fair solution to the long-range sequence dependency problem but disregarded parallel  
   computations and was even slower than the RNN (Sundermeyer et al. 2012)
2. The original Transformer architecture, introduced by Vaswani et al. in 2017, marks a pivotal departure from the prevailing paradigm of Long Short-Term Memory (LSTM) and other Recurrent Neural Network (RNN) variations, which dominated natural language processing (NLP) until then. The Transformer emerged as a breakthrough in NLP, addressing the limitations of LSTMs, especially in handling long sequences and transfer learning reliably. Unlike its predecessors, The transformer presented a novel way to incorporate an attention mechanism into deep feedforward networks, enabling it to capture long-range sequence dependencies with a broader context window, akin to LSTMs but with enhanced efficiency for longer sequences. Notably, the Transformer architecture demonstrated superior performance, scalability, and parallelizability, allowing for training on unprecedentedly large corpora, as evidenced by studies such as those by Raffel (2020) and Brown (2020). In light of these advancements, language models leveraging the Transformer architecture are referred to as Transformer Language Models (TLMs), a term coined to underscore their remarkable superiority over LSTM-based models and their exceptional scalability, as highlighted in research by Kaplan (2020).
3. The Transformer in NLP is a novel architecture that aims to solve sequence-to-sequence tasks while handling long-range dependencies with ease. It relies entirely on self-attention to compute representations of its input and output WITHOUT using sequence-aligned RNNs or convolution.
4. The transformer model, introduced in 2017 by Google AI researchers, revolutionized natural language processing (NLP) and artificial intelligence (AI) as a whole. Before transformers, recurrent neural networks (RNNs) and convolutional neural networks (CNNs) were the dominant AI architectures for language and computer vision tasks, respectively. However, transformers have quickly become the cornerstone of modern AI, achieving state-of-the-art results on a wide range of NLP tasks, including machine translation, text summarization, and question answering. Transformers are also now used in many other AI domains, such as processing multiple types of input simultaneously or planning tasks.
5. Un transformeur, ou modèle auto-attentif, est un modèle d'apprentissage profond introduit en 2017, utilisé principalement dans le domaine du traitement automatique des langues (TAL).
6. Les transformers sont des réseaux de neurones qui apprennent le contexte et la signification en suivant les relations dans les données séquentielles, comme les mots dans une phrase 1. Ils appliquent une série de techniques mathématiques évolutives, appelées attention ou auto-attention, pour détecter les subtilités des éléments de données distants dans une série qui influencent et dépendent les uns des autres 1. Les Transformers sont rapidement devenus le modèle de choix pour les problèmes de TAL, remplaçant les anciens modèles de réseaux de neurones récurrents tels que le LSTM (Long Short-Term Memory)
7. Transformer ([Vaswani et al., 2017](https://www.sciencedirect.com/science/article/pii/S2666651022000146" \l "b137)) is a prominent [deep learning model](https://www.sciencedirect.com/topics/computer-science/deep-learning-model) that has been widely adopted in various fields, such as [natural language processing](https://www.sciencedirect.com/topics/computer-science/natural-language-processing) (NLP), computer vision (CV) and speech processing. Transformer was originally proposed as a sequence-to-sequence model ([Sutskever et al., 2014](https://www.sciencedirect.com/science/article/pii/S2666651022000146" \l "b130)) for machine translation. Later works show that Transformer-based pre-trained models (PTMs) ([Qiu et al., 2020](https://www.sciencedirect.com/science/article/pii/S2666651022000146" \l "b100)) can achieve state-of-the-art performances on various tasks.
8. The transformer model made its debut in 2017 through the paper "Attention is all you need" by Google researchers. Before the advent of transformers, distinct AI architectures dominated various applications, with recurrent neural networks for language and convolutional neural networks for computer vision being prevalent. However, transformers have evolved into the cornerstone of modern artificial intelligence, revolutionizing the AI landscape over the past five years. Virtually every advanced AI system now relies on transformers, and researchers have become accustomed to working with them. The optimization efforts of thousands of researchers over several years have endowed transformers with a formidable incumbency advantage that makes them challenging to replace. The widespread success of transformers has spurred exploration into their application in other AI domains, such as processing multiple types of input simultaneously or planning tasks. The popularity of the transformer model stems from its potential universality, prompting researchers to investigate the possibility of converging disparate subfields of AI.

Le Transformer, introduit en 2017, est une architecture de réseau neuronal auto-attentif qui a révolutionné le traitement automatique du langage naturel (TAL). En s'appuyant sur des techniques mathématiques avancées, notamment l'attention, les Transformers sont capables d'apprendre le contexte et la signification des données séquentielles, telles que les mots dans une phrase. Cela leur permet de surpasser les modèles précédents, notamment les réseaux neuronaux récurrents (RNN), en particulier dans les tâches de traitement de séquences longues.

Le Transformer, également connu sous le nom de modèle auto-attentif, constitue une innovation majeure dans le domaine de l'apprentissage profond, ayant été introduit en 2017 et largement adopté dans le traitement automatique des langues (TAL). Il représente une évolution cruciale par rapport aux modèles traditionnels encodeur-décodeur, tels que les mémoires à court terme (LSTM) et autres variantes de réseaux neuronaux récurrents (RNN), qui prédominaient dans le traitement du langage naturel (NLP) jusqu'à cette époque.

Les transformers, en tant que réseaux de neurones, excellent dans l'apprentissage du contexte et de la signification en analysant les relations séquentielles dans les données, telles que les mots dans une phrase. Cette capacité est rendue possible par l'application de techniques mathématiques évoluées, notamment l'attention ou auto-attention. Ces mécanismes permettent de détecter les subtilités des éléments de données distants dans une séquence, établissant des relations d'influence et de dépendance entre eux.

L'architecture du Transformer surmonte les limitations des modèles antérieurs, notamment en ce qui concerne la gestion efficace de séquences longues et l'adaptation réussie à l'apprentissage par transfert. En intégrant un mécanisme d'attention dans des réseaux feedforward profonds, le Transformer capture des dépendances de séquences à longue portée avec une fenêtre contextuelle élargie, surpassant ainsi les performances des LSTMs. Sa capacité à traiter des corpus de données d'une ampleur sans précédent, démontrée par des études telles que celles de Raffel (2020) et Brown (2020), a élevé les modèles linguistiques exploitant l'architecture du Transformer, également désignés sous le terme de Modèles Linguistiques Transformer (TLMs), au rang de choix privilégié pour les problèmes de TAL. Cette désignation souligne leur supériorité remarquable par rapport aux modèles basés sur les LSTMs, ainsi que leur extensibilité exceptionnelle, comme le met en évidence la recherche de Kaplan (2020).

Historique :

Au début des années 1990, Jürgen Schmidhuber a introduit le premier modèle de transformateur, le "fast weight controller", qui utilisait un mécanisme d'auto-attention pour apprendre les relations entre les mots d'une phrase. Cependant, il était inefficace et n'a pas été largement adopté. En 2017, Vaswani et al. ont publié l'article "Attention is All You Need", qui a introduit un modèle de transformateur beaucoup plus efficace. Ce modèle est rapidement devenu l'état de l'art pour de nombreuses tâches de traitement du langage naturel (NLP), notamment la traduction automatique, la synthèse de texte et la réponse aux questions.

Les transformateurs sont capables d'apprendre les dépendances à longue séquence entre les mots d'une phrase, ce qui est essentiel pour de nombreuses tâches de NLP. Ceci est possible grâce au mécanisme d'auto-attention, qui permet au modèle de se concentrer sur les mots les plus pertinents d'une phrase lorsqu'il décode les jetons de sortie.

Le transformer a eu un impact majeur dans le domaine du TAL, ils sont désormais l'approche privilégiée pour de nombreuses tâches de NLP, et sont constamment améliorés.

Étapes clés dans l'histoire des transformateurs dans les réseaux de neurones :

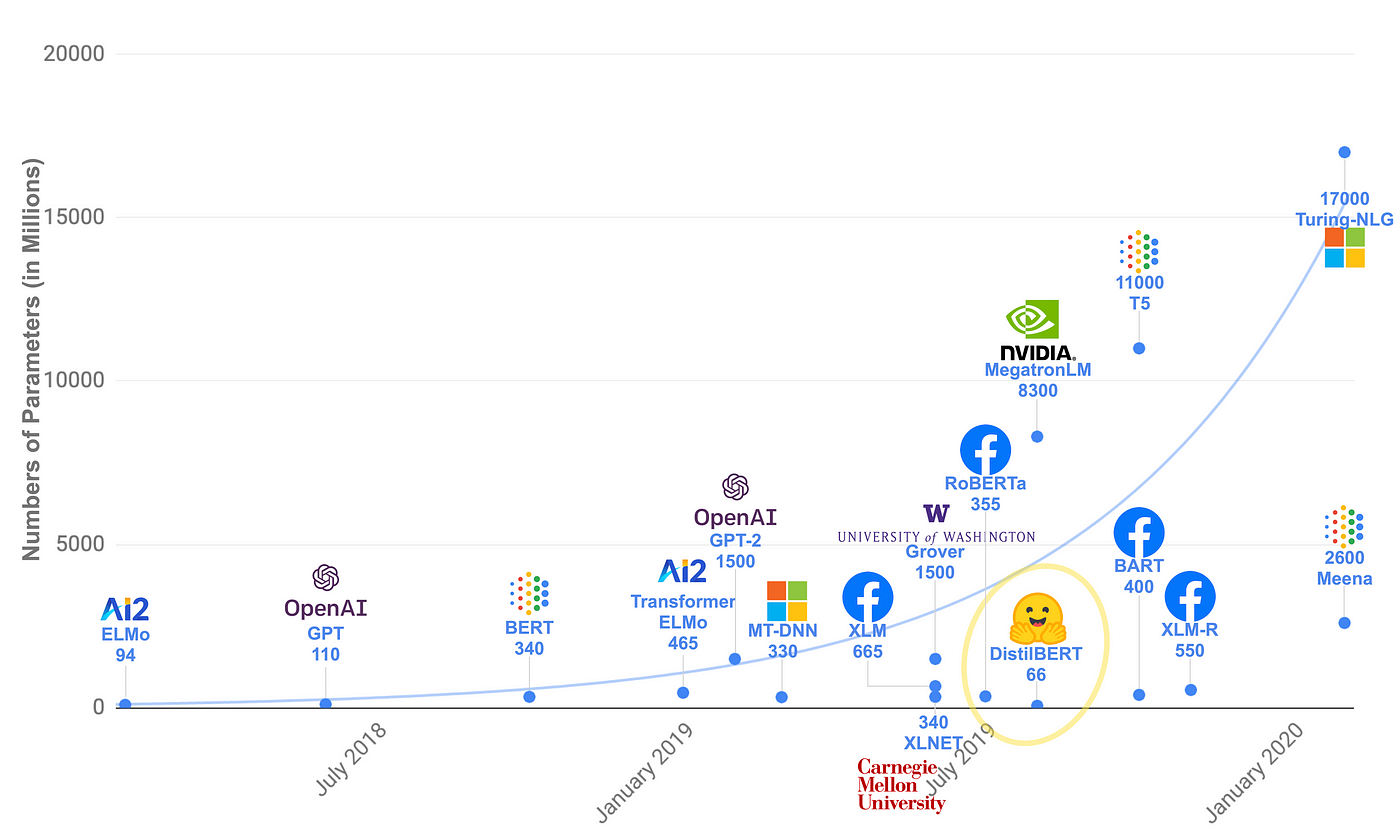
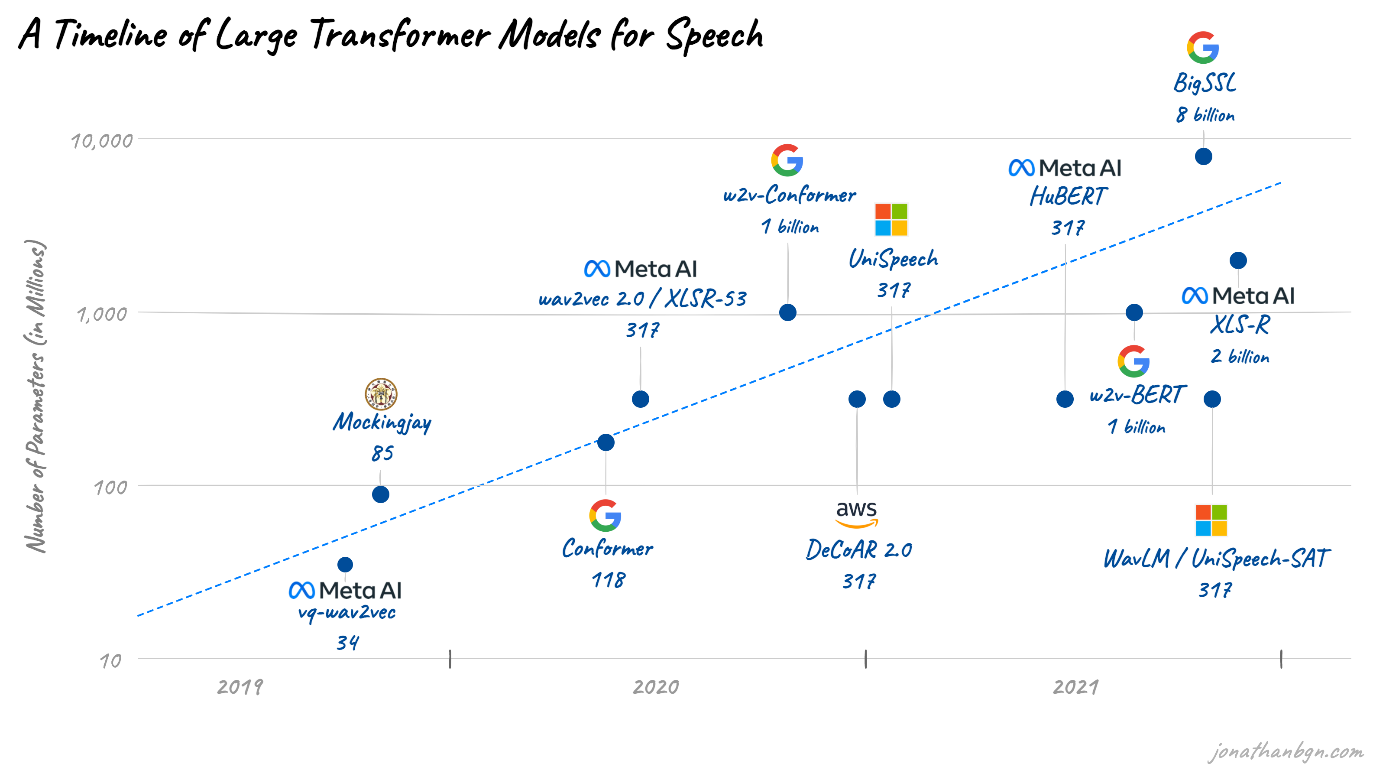
- 1990 : Jürgen Schmidhuber propose le premier modèle de transformer, le "contrôleur de poids rapide".

- 2017 : Vaswani et al. publient l'article "Attention is All You Need", introduisant le modèle de transformer.

- 2018 : Les modèles de transformer atteignent des résultats de pointe dans une large gamme de tâches en TAL, dont la traduction automatique, la résumé de texte et la réponse aux questions.

- 2019 : Les transformers sont utilisés pour créer des modèles de langage volumineux (MLV) tels que BERT et GPT-2.

- 2020 : Les MLV sont utilisés pour créer des modèles encore plus puissants tels que GPT-3.



The transformer model, providing a substantive solution to the long-standing problems  
faced in sequential manipulations, breaks grounds at a breath-taking pace in the NLP  
research space. It has yielded many SOTA results in some NLP applications since its birth  
in 2017, particularly due to its constituents. Figure 3 shows the architecture of the transformer model.  
The model consists of the encoder and decoder blocks with a softmax activation function for normalizing output probabilities. The input to the model is a sequence of data. The  
input words are embedded and passed through the positional encoders, which assign vectors to words based on their positioning in a sentence; thus, extracting input words’ contextual meaning. The encoder blocks consisting of the multi-head attention and a feed-forward  
network receive the embeddings. The multi-head attention layers compute attention vectors  
for each input to represent how each word is related to other words in the same sentence,  
i.e., further capturing the contextual relationship between the words in the sentence. The  
attention vectors are passed through a feed-forward network one vector at a time to the  
decoder block. It is worth stating that parallelization is reached in the multi-head attention  
layer because the attention networks are independent.  
The decoder has the positional encoders and masked multi-head attention layers that  
work similarly to the encoder block. The attention vectors from its masked multi-head  
attention layers and that from the encoder block are passed into another multi-head attention block. Each vector represents the relation with other words in the entire document.  
The vectors are then passed to a feed-forward network and then to the linear layer and  
finally to a softmax activation function that converts it into a probability distribution for  
the output. For the detailed operations of the transformer model, readers are encouraged to  
refer to the paper (Vaswani et al. 2017)