MASTER 02 IREN
Industries de Réseau et Économie Numérique

From W2V to LLAMA2:

10 years of history on the diffusion of LLMs and the emergence of regulatory requirements

Author: Vincent TO

Supervising professor: Christophe BENAVENT

Master in Network Industries and Digital Economy (IREN)

School year : 2023-2024

Contents

Abstract

This thesis examines the trajectory and impacts of Large Language Models (LLMs) over a decade, focusing on key innovations from Word2Vec to LLAMA2. By analyzing the evolution of LLMs, this work highlights their transformative role in Artificial Intelligence (AI) and Natural Language Processing (NLP). It explores how advancements such as word embeddings, contextual models, Transformer architectures, and generative approaches have redefined the understanding and generation of human language by machines.

The thesis also addresses the regulatory and ethical implications of these technologies, emphasizing the challenges associated with their integration into practical and societal applications. Through this exploration, the study aims to provide a holistic perspective on LLMs, illustrating their potential and limitations, and prompting reflection on the future of AI in natural language processing.

Glossary

This glossary lists in alphabetical order both the list of abbreviations, acronyms, as well as initialisms used throughout this document.

Activation Function: An activation function used in some natural language processing models, such as LLAMA. It is designed to improve model efficiency by allowing better control over information passing through network layers.

AI (Artificial Intelligence): Refers to the simulation of human intelligence by machines, particularly computer systems. These processes include learning (acquiring information and rules for using information), reasoning (using rules to reach approximate or definitive conclusions), and self-correction.

Attention Mechanism: A technique used in natural language processing models to enhance focus on specific aspects of input while ignoring others. It enables models to better process relevant information, especially in long sequences.

BERT (Bidirectional Encoder Representations from Transformers): A language model based on the Transformer architecture, designed to understand the context of words in a sentence in both left-to-right and right-to-left directions. BERT is used to enhance natural language understanding in various NLP tasks.

BPE (Byte Pair Encoding): A tokenization method used in natural language processing to break down words into smaller subunits, reducing the size of the vocabulary and effectively handling rare or unknown words.

CBOW (Continuous Bag of Words): A natural language processing model that predicts a target word based on the words surrounding it. It is part of the Word2Vec approach for creating word embeddings.

Contextual Models: Natural language processing models that consider the context in which words appear to determine their meaning. They generate word representations that vary based on their usage context, providing a deeper understanding of the text.

Deep Learning: A machine learning process that utilizes neural networks with multiple layers of hidden neurons. These algorithms have a large number of parameters, requiring a substantial amount of data for training.

Distillation: A machine learning technique where a smaller model (student) is trained to mimic the behavior of a larger and more complex model (teacher). This helps reduce the model size while retaining much of its performance.

ELMo (Embeddings from Language Models): A pre-trained language model that generates word embeddings by considering their context within a sentence. ELMo uses recurrent neural networks to better capture contextual nuances.

Embeddings: Vector representations of words that capture their meaning and semantic and syntactic relationships. These dense vectors assist NLP models in processing language more effectively by capturing linguistic nuances.

Fine-tuning: The process of adjusting or optimizing a pre-trained language model on a specific dataset or for a specific task. This allows the model to be adapted to specific needs and improve its performance on targeted tasks.

Few-shot Learning: An approach in machine learning where a model can learn or adapt to a new task with a very small number of examples or training data, in contrast to traditional methods that require large amounts of data.

GPT (Generative Pretrained Transformer): A series of language models based on the Transformer architecture, designed for text generation. These models are pre-trained on large text corpora and can be used for various NLP tasks, including text generation.

Generative Models: Machine learning models designed to generate new data that resembles a training dataset. In NLP, they are used to generate text that mimics the style and content of human language.

Hyperparameters: Parameters used to control the learning process in machine learning models. Unlike model parameters, which are learned during training, hyperparameters are predefined and include things like learning rate, batch size, number of epochs, etc.

LLAMA: A natural language processing model developed by Meta AI, using the Transformer architecture. It is designed to efficiently process natural language with improved contextual understanding and is trained on a diverse dataset for enhanced performance in various NLP tasks.

LSTM (Long Short-Term Memory): A type of recurrent neural network (RNN) used in natural language processing and other domains. It is designed to address the vanishing gradient problem, allowing the model to retain information over long sequences.

Neural Networks: Computational structures inspired by the human brain, used in machine learning to model complex relationships between inputs and outputs. They consist of layers of neurons that transmit and transform data.

NLP (Natural Language Processing): A branch of AI that helps computers understand, interpret, and manipulate human language.

Out-of-Vocabulary (OOV) Words: Terms or words that do not appear in the known vocabulary of a language processing model. Models must find ways to handle these unknown words, often by breaking them down into smaller subunits or using special tokens.

Pruning: A machine learning technique involving the reduction of a neural network's size by removing the least important weights. This simplifies the model and reduces computational and memory requirements.

RNN (Recurrent Neural Network): A type of neural network specialized in processing sequences of data, where outputs from one step are reused as inputs for the next. They are particularly useful for natural language processing and other sequential tasks.

Skip-gram: A model in the Word2Vec approach for creating word embeddings, where the model uses one word to predict its context. It is useful for capturing relationships between words in large text corpora.

SwiGLU: An activation function used in some natural language processing models, such as LLAMA. It is designed to improve model efficiency by allowing better control over information passing through network layers.

Tokenization: The process of dividing a text into smaller pieces, called tokens. In NLP, this can mean splitting text into words, syllables, or sub-words, facilitating processing and analysis by language models.

Transformer: An architecture for natural language processing models that uses attention mechanisms to enhance understanding of word relationships in a sentence. It is known for its ability to efficiently handle large data sequences.

Word2Vec: A natural language processing model for creating word embeddings, where words are represented by dense vectors. It captures semantic and syntactic relationships between words from large text corpora.

WordPiece: An algorithm for tokenization used to break down words into smaller subunits. It allows better handling of rare or unknown words in natural language processing, improving model quality.

Introduction

Large Language Models (LLMs), such as GPT-3, BERT, and their successors, represent a significant advancement in the fields of Artificial Intelligence (AI) and Natural Language Processing (NLP). Built on deep learning techniques and trained on vast text corpora, these models have the remarkable ability to generate, understand, and interpret human language with great sophistication. They play an essential role in various applications, from language translation to content generation, and are a cornerstone of modern NLP research.

However, this technological advancement also raises crucial questions. How do these models influence the development of machine language processing capabilities? What are the ethical, practical, and regulatory challenges associated with their use and development? These questions form the central theme of this thesis, aiming to provide a comprehensive account of the evolution of LLMs, highlighting not only their growing significance in AI and NLP but also exploring the implications of their increasing adoption in a constantly evolving social and regulatory context.

The genesis of modern NLP: Word2Vec (2013)

Brief History of NLP and AI before Word2Vec

Before the emergence of Word2Vec, the field of Natural Language Processing (NLP) and Artificial Intelligence (AI) went through several crucial phases. The origins of NLP and AI date back to the 1950s and 1960s, a period marked by pioneering work in artificial intelligence research. During this time, the emphasis was primarily on creating systems capable of understanding and manipulating natural language in a very rudimentary manner, often based on programmed rules and logic.

In the 1970s and 1980s, NLP underwent significant evolution through the adoption of rule-based models and computational linguistics. This era was dominated by symbolic approaches, where language understanding relied on explicit rules and parse trees. These methods, while innovative, were limited by their inability to handle the flexibility and ambiguity inherent in human language.

With the advent of the 1990s, NLP began to incorporate statistical approaches, marking a major shift in the way natural language was processed. Statistical models, such as Hidden Markov Models and probability-based methods, started replacing purely rule-based approaches. This period also saw the development of the first automatic translation systems and speech recognition, which used statistical techniques to improve their performance.

In the early 2000s, with improved computing capabilities and the availability of large amounts of textual data, NLP made a leap forward with the use of machine learning-based approaches. This period was characterized by the use of machine learning models for various NLP tasks, such as document classification, sentiment analysis, and named entity recognition. Systems during this time could learn from data and adapt to new tasks, offering significantly greater flexibility and performance compared to earlier methods.

It was in this historical context that Word2Vec was introduced in 2013, marking another major evolution in the field of NLP. Word2Vec revolutionized the way words are represented in AI models by introducing the concept of word embeddings, where words are transformed into dense vectors capturing complex semantic and syntactic relationships. This innovation paved the way for more sophisticated advances in natural language processing, thus laying the foundation for future developments in the fields of NLP and AI.

Detailed analysis of Word2Vec, its innovative approach and its methodology

The Word2Vec model, introduced by Tomas Mikolov and his team at Google, marked a revolutionary milestone in the field of Natural Language Processing (NLP). This model is renowned for its ability to transform words into dense vectors, thereby providing a rich and meaningful numerical representation of natural language.

Their work, detailed in the paper *"Efficient Estimation of Word Representations in Vector Space*", introduced innovative model architectures: CBOW (Continuous Bag of Words) and Skip-gram. Word2Vec, through these models, improved the efficiency and effectiveness of word embeddings learning. The CBOW model predicts a target word from context words, while Skip-gram predicts context words from a target word. This advancement was fundamental in the development of more advanced word embedding techniques, deeply influencing the field of NLP.

Word2Vec is innovative primarily for two reasons:

- Vector Representation of Words: Unlike previous methods that treated words as discrete and isolated entities, Word2Vec represents each word as a vector in a continuous space. This approach captures word semantics based on their usage context, allowing similar words to be close in the vector space.
- Context and Word Models: Word2Vec uses two main architectures - CBOW (Continuous Bag of Words) and Skip-gram. The CBOW model predicts a target word from a set of contextual words, while the Skip-gram model operates in reverse, predicting contextual words from a target word. This flexibility enables the model to adapt to various language processing requirements.

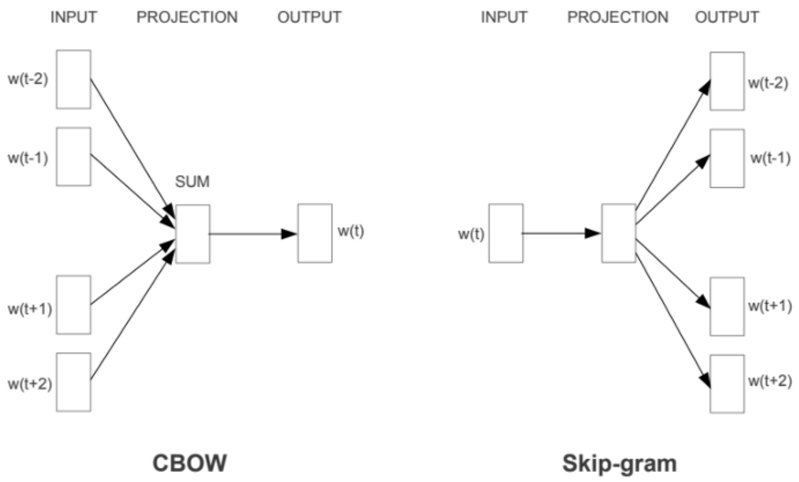

Figure 1: Word2Vec, CBOW & Skip-gram, https://datascientest.com/nlp-word-embedding-word2vec

The Word2Vec methodology is based on unsupervised learning and focuses on the efficient construction of word vectors:

- Training and Optimization: Word2Vec employs deep learning techniques, particularly neural networks, for training its models. The training process involves adjusting the neural network weights to minimize the prediction error between contextual words and target words.
- High-Quality Word Embeddings: The resulting vectors capture complex semantic and syntactic relationships. For example, simple vector operations can reveal relationships between words (such as "king" - "man" + "woman" ≈ "queen").
- Computational Efficiency: One key to Word2Vec's success is its ability to be trained efficiently on large text corpora, making word embeddings both high-quality and practical for real-time NLP applications.

Word2Vec not only introduced a revolutionary method for word representation but also set a new standard in the field of NLP. Its word vectors have become a fundamental component of many modern NLP systems, greatly influencing subsequent research and developments in this field.

Before Word2Vec, word representation in NLP models was primarily based on methods like "bag of words," where words were treated as discrete and isolated entities. Word2Vec introduced the idea of representing words as vectors in a continuous space, capturing semantic and contextual nuances. This approach allowed models to better understand language subtleties, such as synonyms, antonyms, and contextual relationships.

Thanks to Word2Vec, it became possible to extract complex semantic and syntactic relationships directly from large text corpora. The vectors generated by Word2Vec demonstrated their ability to capture relationships like word analogies, revolutionizing how machines understand language. This advancement opened the door to new NLP applications, such as automatic translation, sentiment analysis, and named entity recognition.

Word2Vec sparked a wave of research in the NLP field. Many researchers began exploring different variants and improvements of word embedding models. It also encouraged the scientific community to develop more complex and efficient models, such as recurrent neural networks and attention-based models, leading to the creation of transformer models like BERT.

In the realm of practical applications, Word2Vec greatly facilitated the development of more accurate and reliable NLP systems. For example, in question-answering systems and virtual assistants, the use of Word2Vec enabled a better understanding of user queries and more precise responses.

Word2Vec was a step forward in reducing the computational complexity of NLP models. Its ability to work with large datasets efficiently made it possible to use sophisticated language models in environments where computational resources are limited.

Evolution of contextual language models

From word integration to contextual models

The transition from Word2Vec and its static embeddings to dynamic contextual models like ELMo and BERT marked a significant advancement in the field of NLP. This section examines this evolution and its implications.

Word2Vec and similar models generate static word embeddings, where a word is represented by a unique vector, regardless of its context of use. This means that the word "bank" would have the same vector whether it refers to a financial institution or the edge of a river. Although these models were revolutionary, they were limited in their ability to capture contextual nuances in language.

ELMo (Embeddings from Language Models), introduced by AllenNLP in 2018, was one of the first models to utilize contextual embeddings. In ELMo, the meaning of a word is determined based on its context, allowing the same word to have different representations depending on its usage in a sentence. ELMo uses a bidirectional Long Short-Term Memory (LSTM) model to generate these embeddings, enabling it to consider the complete context of a word in a text.

Developed by Google in 2018, BERT (Bidirectional Encoder Representations from Transformers) took the idea of contextual embeddings even further. Unlike ELMo, BERT uses the Transformer architecture, which allows for a comprehensive bidirectional understanding of a word's context. The paper "*BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*" introduced BERT as an innovative method for pre-training linguistic representations. Its main innovation lies in its bidirectional training, a approach that significantly differs from previous unidirectional methods. BERT is designed to pre-train deeply bidirectional representations by conditioning on both left and right context. This approach allowed BERT to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, with just an additional output layer fine-tuning.

BERT is pre-trained on a large text corpus using two tasks: masked word prediction and understanding relationships between sentences. This enables BERT to better understand the structure and meaning of natural language.

Contextual models like ELMo and BERT have revolutionized the understanding of natural language, allowing for finer and more accurate analyses of word and phrase meanings. These models have enhanced tasks such as text comprehension, machine translation, and question answering by providing a more nuanced analysis of language.

The advent of BERT has led to a wave of new Transformer-based models, each seeking to improve or adapt the technology for specific use cases. These developments continue to push the boundaries of what is possible in NLP, paving the way for even more sophisticated and accurate applications.

The Transformers Model Revolution

The revolution of Transformer models in the field of Natural Language Processing (NLP) represents a major breakthrough that has redefined modern approaches to natural language understanding and generation.

Introduced for the first time in the paper "Attention is All You Need" by Vaswani et al. in 2017, Transformer models introduced a new mechanism called "Attention," which allows the model to focus on different parts of a sentence to better understand its context and meaning.

This paper was foundational for the Transformer architecture, as it demonstrated that attention mechanisms alone, without recurrent or convolutional layers, were sufficient to achieve state-of-the-art results in machine translation tasks. The Transformer architecture, composed of stacks of encoders and decoders and using multi-head attention mechanisms, enabled more efficient handling of long-range dependencies, thus revolutionizing natural language processing applications.

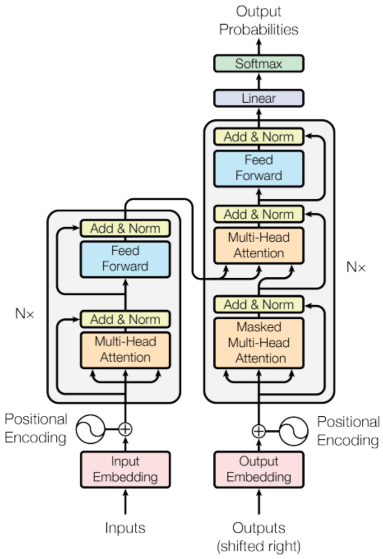

Figure 2: The Transformer model architecture, from "Attention Is All You Need", A. Vaswani and al., 2017

Unlike previous architectures based on recurrent neural networks (RNNs) or LSTMs, Transformers do not need to process text sequentially. This allows them to handle all parts of a sentence simultaneously, offering increased computational efficiency.

The central aspect of Transformers is the attention mechanism, which allows the model to weigh the importance of different words in a sentence. This results in a better ability to understand complex relationships and nuances in language. Attention can be "self-directed," allowing the model to refer to itself to improve its understanding of context.

Transformers outperform previous models in nearly all NLP tasks, including machine translation, text comprehension, and content generation. They are particularly effective in processing long sequences of data, where RNNs and LSTMs tended to lose performance.

Transformer-based models, such as BERT, GPT (Generative Pretrained Transformer), and their variants, are now used in a variety of NLP applications, redefining possibilities in fields such as dialogue systems, intelligent personal assistants, and sentiment analysis. Their ability to understand and generate language has profound implications not only for NLP but also for AI in general, paving the way for more natural and intelligent interactions between humans and machines.

The revolution of Transformer models marked a turning point in the field of NLP. Their innovative approach to parallel processing and attention mechanisms has not only improved the efficiency and accuracy of language understanding but has also opened new avenues for research and development in AI. These models continue to evolve and adapt, promising even more significant advancements in natural language processing in the future.

The article "*A Comprehensive Overview of Large Language Models*" provides a detailed analysis of recent developments in LLMs. It covers a wide range of topics, including architectural innovations, training strategies, context length improvements, fine-tuning, multi-modal LLMs, and benchmarking efficiency. This comprehensive and concise review of advances in LLMs enriches the understanding of current progress and future directions in Transformer and large-scale language model research.

The rise of generative models

Introduction to Generative Models

Generative models in the field of Natural Language Processing (NLP) have radically transformed the way machines can create content that mimics human language.

The document "*Pretrained Language Models for Text Generation: A Survey*" provides a comprehensive overview of advances in pretrained language models for text generation. It discusses various architectures of these models and their applications in different text generation domains. The review also explores methods for adapting pretrained language models to various types of input data and strategies to ensure that the generated text meets specific required properties. This overview aims to synthesize major developments and guide researchers in related areas of text generation using pretrained language models.

A generative model in NLP is designed to produce text that appears natural and coherent. Unlike discriminative models used to classify or predict existing data, generative models can generate new data, specifically text, ranging from simple sentences to entire paragraphs. These models learn to mimic the style and structure of human language by training on vast text corpora, thereby capturing language nuances and patterns.

Early generative models in NLP often relied on simple statistical methods, such as Hidden Markov Models. However, with the advent of deep neural networks, the capacity of generative models to produce high-quality text has significantly improved. Models like Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) were among the first to show promising results in text generation.

The introduction of the Transformer architecture ushered in a new era for generative models. OpenAI's GPT (Generative Pretrained Transformer) model, in particular, set new standards by generating text that can often be indistinguishable from human writing. GPT and its subsequent versions (GPT-2, GPT-3) have demonstrated remarkable ability to create coherent and contextually relevant content in various styles and formats, ranging from literary prose to poetry and informative articles.

A notable evolution lies in tokenization techniques, particularly in addressing the out-of-vocabulary (OOV) word problem. Recent research favors subword tokenization, offering a flexible solution to the OOV challenge by decomposing words into subword units when they are not in the dictionary. This approach strikes a balance between word-level and character-level models, with techniques like Byte Pair Encoding (BPE) and WordPiece used in advanced models like GPT-2 and RoBERTa. Beyond individual words or subwords, there is growing interest in sentence and phrase-level language models, replacing common word sequences with phrases and directly modeling sentence probabilities, rather than relying on conditional probabilities of smaller linguistic units. This approach is particularly suitable for applications like Automatic Speech Recognition (ASR), where it offers robustness against recognition errors.

Generative models are used in a multitude of applications, such as automated content creation, chatbots, machine translation, and even artistic creation. Their ability to generate realistic and engaging responses makes them valuable tools for human-machine interaction, paving the way for more sophisticated and interactive applications. Despite their success, generative models pose challenges, including issues related to bias inherent in training data and the difficulty of controlling and guiding content generation. Ongoing research in this field aims to improve the accuracy, relevance, and ethics of generative models, with a growing focus on responsible text generation and bias mitigation.

The introduction and evolution of generative models in NLP have opened new frontiers in machines' ability to understand and produce human language. These models continue to evolve, offering exciting prospects for the future of human-machine interaction and automated linguistic content creation.

The GPT model: from GPT to GPT-4

The series of Generative Pretrained Transformer (GPT) models developed by OpenAI represents a major advancement in the field of NLP. These models have not only revolutionized text generation but have also set new standards in natural language understanding.

Launched in 2018, the first GPT model introduced the idea of pretraining on a large text corpus followed by task-specific fine-tuning. This model used a Transformer as a language model capable of predicting the next word in a sentence. Although relatively simple compared to its successors, GPT-1 laid the foundation for future advancements in Transformer-based language models.

Introduced in 2019, GPT-2 marked a significant leap forward with 1.5 billion parameters. This model was capable of generating text of astonishing quality, often indistinguishable from human writing. The paper "*Language Models are Unsupervised Multitask Learners*" explores GPT-2's capabilities as a multitask language model. It demonstrates that GPT-2, trained on a diverse dataset, can perform a variety of language tasks without specific task-specific training. This includes translation, question answering, and reading comprehension. This research highlights GPT-2's ability to generate coherent and contextually relevant text, emphasizing its potential for multitasking in natural language processing. GPT-2 stood out for its ability to perform various NLP tasks without the need for specific fine-tuning, showcasing remarkable contextual understanding of language.

Introduced in 2020, GPT-3 revolutionized the industry with its 175 billion parameters, becoming the largest language model to date. Its ability to generate text, answer questions, translate, and perform a multitude of other NLP tasks, often with minimal or no fine-tuning, was widely acclaimed. In "GPT-3: What's it good for?", Robert Dale provides a critical evaluation of GPT-3. He highlights its impressive ability to generate human-like text while also shedding light on the model's limitations, particularly in producing coherent long-form content and its tendency to reproduce biases from training data. Dale recommends caution in interpreting GPT-3 outputs, especially for tasks requiring precision and accuracy, offering a balanced view of GPT-3's technological advancements while acknowledging the associated challenges and ethical considerations. One of the most notable features of GPT-3 is its ability to perform what is known as "few-shot learning" or learning with few examples, where the model can perform specific tasks based on just a few examples.

The GPT series of models has played a pivotal role in advancing NLP, pushing the boundaries of what is possible in text generation and natural language understanding. GPT-4, the latest iteration, continues in this lineage, aiming to further refine these capabilities and explore new potential applications in various domains.

Specialized and domain-specific models

Adaptation of models to specific needs

The adaptation of natural language processing (NLP) models to specific domains such as healthcare, law, or finance is a growing field. These specialized models are designed to meet the unique requirements and specific nuances of each industry.

In the healthcare domain, NLP models are used to interpret electronic health records, aid in diagnosis, and support clinical decision-making. These models need to understand medical terminology and be capable of processing sensitive and complex data. For example, models like BioBERT, a variant of BERT trained on biomedical texts, can extract relevant information from medical records, facilitating clinical research and personalized medicine.

In the legal sector, NLP models are applied to analyze legal documents, legislation, and judicial precedents. These models must navigate highly technical and often archaic language. They are used for predicting legal outcomes, contract analysis, and assisting in the drafting of legal documents. This not only saves time but also increases the accuracy and consistency of legal interpretations. In the financial sector, NLP models are adapted to analyze market reports, economic news, and assist in investment decision-making. In the media industry, they are used for automatic article generation, sentiment analysis on social media, and content personalization.

Each sector requires specific adaptations to enable NLP models to understand and effectively process specialized language and terminologies. The customization of NLP models for specific domains represents a major advancement, enabling more accurate and efficient application of these technologies in various industries.

By understanding and processing specialized language, these models provide significant assistance in analysis, decision-making, and information management in each specific domain.

Advances in personalization and efficiency

The recent evolution of natural language processing (NLP) models is characterized by a notable trend towards greater efficiency and increased customization. This section explores how these advancements are transforming the application of NLP models.

One of the major trends is the customization of NLP models to meet the specific needs of different industries or applications. This customization often involves training or fine-tuning with domain-specific datasets, enabling models to better understand and process specialized language and terminologies. Customization also allows models to adapt to regional linguistic variations or specific communication styles, offering broader and more precise applicability.

With the increasing size of NLP models, efficiency has become a crucial concern. Efforts are made to optimize model performance while reducing their computational footprint. This includes techniques like model distillation, where a larger and more complex model is used to train a smaller and faster model.

Approaches such as neural network pruning and the use of efficient learning techniques also contribute to making NLP models faster and less resource-intensive. The trend towards more efficient models makes NLP more accessible, especially for organizations with limited computational resources. This paves the way for wider adoption of NLP technologies across various sectors, including small and medium-sized enterprises. The increased efficiency of models also enables the processing of larger volumes of data in real-time, which is essential for applications such as social media monitoring, real-time sentiment analysis, or interactive virtual assistants.

Advancements in terms of customization and efficiency of NLP models reflect a shift towards smarter, faster, and user-specific applications. These advancements are crucial to ensure that NLP remains at the forefront of technological innovation while being applicable and useful in a variety of contexts and industries.

Recent developments: LLAMA and LLAMA2

The path to LLAMA and LLAMA2

LLAMA, a natural language processing (NLP) model developed by Meta AI, represents a significant milestone in the evolution of large-scale language models. Its development is part of a series of key technological advancements that have shaped the current landscape of NLP.

Advancements in the field of language models, particularly with the introduction of models like BERT and GPT, laid the foundation for the creation of more advanced and specialized models like LLAMA. These models brought significant improvements in text understanding and generation. LLAMA and LLAMA2 use the Transformer architecture, with minor architectural differences compared to models like GPT-3. For example, LLAMA uses the SwiGLU activation function instead of ReLU and rotational positional embeddings instead of absolute positional embeddings. LLAMA2 extends the context length from 2K tokens (LLAMA1) to 4K tokens. These models are trained on massive datasets, with LLAMA1 using 1.4 trillion tokens and LLAMA2 approximately 2 trillion tokens.

LLAMA has been trained on a variety of data sources, including CCNet, C4, GitHub, Wikipedia, and books, contributing to a comprehensive and diverse understanding of linguistic styles and domains. This diversity in data sources has enhanced LLAMA's versatility and broad applicability. LLAMA's performance has been evaluated through a range of metrics, testing its abilities to understand and infer information from texts. LLAMA has demonstrated competitive performance, especially in scientific domains and for tasks like question answering.

The configuration of LLAMA's model hyperparameters plays a crucial role in optimizing its performance for specific use cases. Selecting and fine-tuning these hyperparameters allows LLAMA to be tailored to specific needs, ensuring optimal performance and efficient resource utilization. LLAMA represents a significant step in the development of large-scale language models, illustrating the ongoing evolution and innovation in the field of NLP. Its advanced capabilities in natural language understanding and generation make it a valuable tool for researchers and developers looking to push the boundaries of artificial intelligence and natural language processing.

LLAMA models have shown competitive performance in various NLP tasks, including solving mathematical theorems, predicting protein structures, and comprehension reading. Their smaller size compared to other large-scale AI models makes them more accessible for research and experimentation, contributing to democratizing access in this rapidly evolving field.

In July 2023, Meta introduced LLAMA2, the next generation of its large-scale language model. LLAMA2, available for both research and commercial use, marks another step in the partnership between Meta and Microsoft. By making LLAMA2 accessible, Meta and Microsoft encourage an open approach to AI model development, enabling a wider range of developers and researchers to test and improve these tools and explore their potential applications in various contexts.

LLAMA and LLAMA2 illustrate the ongoing evolution and innovation in the field of NLP. Their versatility and accessibility make them valuable for researchers and developers looking to push the boundaries of artificial intelligence and natural language processing.

The emergence of regulatory requirements: the AI Act of the European Union

The European Union's "AI Act" represents a major step in the regulation of Large Language Models (LLMs) and AI technologies, addressing the rapid advancement of these technologies and the challenges they bring. This pioneering legislation imposes stricter rules for "base models," including prominent LLMs like OpenAI's GPT or Google's Bard.

Economic and Global Impact

Highlighted by "*What's next for AI regulation in 2024?*" by MIT Technology Review, these powerful AI models will be subject to rigorous security checks and data governance measures before entering the market. This includes verifying that training data complies with copyright laws and assessing risks to fundamental rights, health, safety, the environment, democracy, and the rule of law. The AI Act's regulations could significantly impact the EU's AI industry, potentially creating barriers for smaller companies while also setting a global regulatory standard akin to GDPR's influence on data privacy.

Challenges and Opportunities for AI Development

Ensuring compliance, particularly with unbiased training data and assessing AI's impact on fundamental rights, poses technical hurdles. This necessitates a balance between innovation and regulatory adherence. The Act potentially spurs innovation in AI governance and compliance technologies, presenting new opportunities for the industry.

Broader Implications for AI Ethics and Governance

Under the "AI Act," base models are required to have improved documentation and comply with EU copyright laws. This mandates comprehensive documentation and strict adherence to EU copyright laws for training data, as emphasized in "*The imperative for regulatory oversight of large language models (or generative AI) in healthcare*" by npj Digital Medicine.

The Act is pivotal in integrating AI responsibly into society, emphasizing safety, fairness, and the protection of fundamental rights. It reflects a growing awareness of the importance of accountability in AI development.

Impact on Advanced AI Models

For the most advanced AI models, additional requirements are imposed, including the disclosure of their security and energy efficiency measures. This ensures that these models are not only high-performing but also secure and sustainable. Models must disclose information about the data used for their training, crucial for assessing objectivity, reliability, and potential biases.

To enforce these rules, the "AI Act" will establish a new European AI Office, making the EU a significant global regulator in the field of AI. This body will coordinate compliance, implementation, and enforcement of the Act, with fines for non-compliance ranging from 1.5% to 7% of a company's global turnover.

The AI Act marks a significant stride in AI regulation, focusing on responsible and ethical use of AI. This regulatory evolution aims to ensure safety and fairness in the use of AI technologies while promoting their responsible integration into society. The Act's impact extends beyond the EU, influencing global AI governance and setting a standard for future regulatory efforts.

Conclusion

This first part of the thesis on the topic "From W2V to LLAMA2: 10 Years of History on the Diffusion of LLMs and the Emergence of Regulatory Requirements" traces a decade of innovations and developments in the field of Large Language Models (LLMs). It begins with the advent of Word2Vec, which revolutionized word representation, laying the foundation for subsequent developments in NLP. Word2Vec introduced rich vector representations, enabling machines to grasp complex semantic and syntactic relationships.

The emergence of contextual models, such as ELMo and BERT, marked a significant advancement by moving beyond Word2Vec's static embeddings to understand language in diverse contexts. These developments were crucial for handling the subtleties and complexity of natural language, making machines more competent in tasks such as text comprehension and machine translation.

The revolution of Transformers models, embodied by breakthroughs like GPT and its successive iterations, redefined the standards for language generation and comprehension. Transformers, with their innovative attention mechanism, allowed better handling of long-distance dependencies and increased capacity to process longer data sequences.

This first part of the thesis also highlights the evolution of generative models, with a particular focus on the GPT series, which pushed the boundaries of automatic text generation, making it possible to create content that is often indistinguishable from that produced by humans.

The adaptation of NLP models to the specific needs of different domains, such as healthcare, law, and finance, represents a major advancement, enabling more precise and efficient application of these technologies in various sectors.

Finally, the development of LLAMA and LLAMA2 by Meta AI represents the culmination of these developments, combining the versatility of previous models with increased accessibility and efficiency, thereby opening new prospects for research and practical application of LLMs.

This journey illustrates not only the rapid evolution and technological advancements in the field of LLMs but also underscores the emergence of regulatory requirements, reflecting the ethical and societal challenges posed by these advanced technologies. Understanding this history is crucial for anticipating the future directions of research in NLP and navigating the ever-changing landscape of regulatory and ethical requirements in artificial intelligence.

Bibliography

Tomas Mikolov et al. (2013), *“Efficient Estimation of Word Representations in Vector Space”*

Ashish Vaswani et al. (2017), *“Attention Is All You Need”*

Alec Radford et al. (2018), *“Language Models are Unsupervised Multitask Learners”*

Jinhyuk Lee et al. (2019), *“BERT: a pre-trained biomedical language representation model for biomedical text mining”*

Robert Dale (2021), *“GPT-3: What’s it good for?”*

Junyi Li et al. (2021), *“Pretrained Language Models for Text Generation: A Survey”*

Humza Naveed et al. (2023), *“A Comprehensive Overview of Large Language Models”*

Bertalan Meskó et al. (2023), *“The imperative for regulatory oversight of large language models (or generative AI) in healthcare”*

Hugo Touvron et al. (2023), *“LLaMA: Open and Efficient Foundation Language Models”*

Council of the European Union (2023), *“Législation sur l'intelligence artificielle: le Conseil et le Parlement parviennent à un accord sur les premières règles au monde en matière d'IA”*

Tate Ryan-Mosley et al. (2024), MIT Technology Review, *“What’s next for AI regulation in 2024?”*