Technical Report

MedAIPaca

Desi Arti

Master Degree of Mathematicss
Department Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Indonesia

Abstract

This technical report presents the innovative instruction-following model, Alpaca, and its collaborative integration with MEDAIPACA, an open-source collection of medical conversational AI models and training data. Alpaca, designed by Rohan Taori, et.all represents a paradigm shift in machine learning by combining a robust neural network architecture with transfer learning methodologies. Its proficiency is showcased across diverse applications, including robotics, natural language processing, and human-machine interaction.

In parallel, MEDAIPACA, focuses on advancing medical conversational AI. The initiative aims to foster collaboration and innovation in the medical community by providing open-source models and annotated training data.

The collaboration between Alpaca and MedAIPaca represents a synergistic approach to advancing instruction-following capabilities and medical conversational AI. This integration holds promise for transformative applications in healthcare, ranging from enhanced medical assistance to improved diagnostics and communication.

Keywords: Alpaca, MedAIPaca, instruction-following model, neural network, transfer learning, medical conversational Al

1. Introduction

The intricate labyrinth of medical knowledge often leaves both patients and healthcare professionals adrift, yearning for clear navigation. Navigating complex diagnoses, treatment options, and preventive measures often requires deciphering dense jargon and traversing mountains of dense literature. This information asymmetry can hinder patient comprehension, impede informed decision-making, and ultimately impact health outcomes.

MedAIPAca emerges as a beacon of hope, an open-source project aiming to revolutionize healthcare communication through the magic of Conversational AI (CAI). By offering a readily accessible library of medical CAI models and training data, MedAIPAca empowers both patients and healthcare professionals to engage in natural language dialogues about health, transforming their journey from one of frustration to understanding.

Alpaca, a powerful instruction-following model, is complemented by MEDAIPACA, an open-source collection focused on advancing medical conversational AI. This technical report provides insights into Alpaca's architecture, training methodologies, and applications, followed by an overview of MedAIPaca 's contributions to medical conversational AI.

2. Fundamental

2.1 AlPaca

Alpaca is a large language model (LLM) specifically designed to excel at following instructions. It was developed by researchers at Stanford University to address limitations in existing LLMs for instruction-following tasks, such as high resource requirements and susceptibility to hallucinations. Large Language Models, or LLMs, are powerful neural networks with billions of parameters. They've been trained on massive amounts of text data using semi-supervised learning. These models can perform tasks like mathematical reasoning and sentiment analysis, demonstrating their understanding of the structure and meaning of human language.LLMs have been trained on data spanning hundreds of Terabytes, which gives them a deep contextual understanding. This understanding extends across various applications, making them highly effective at responding to different prompts.

2.2 Neural Network Architecture

Alpaca leverages a sophisticated neural network architecture designed for sequential instruction processing.Attention mechanisms and memory modules enhance the model's ability to handle complex instructions.

This simplified model captures the essence of how a single neuron in a neural network processes information. By combining multiple neurons with different connections and weights, neural networks can learn complex relationships and perform a variety of tasks, making them powerful tools for machine learning.

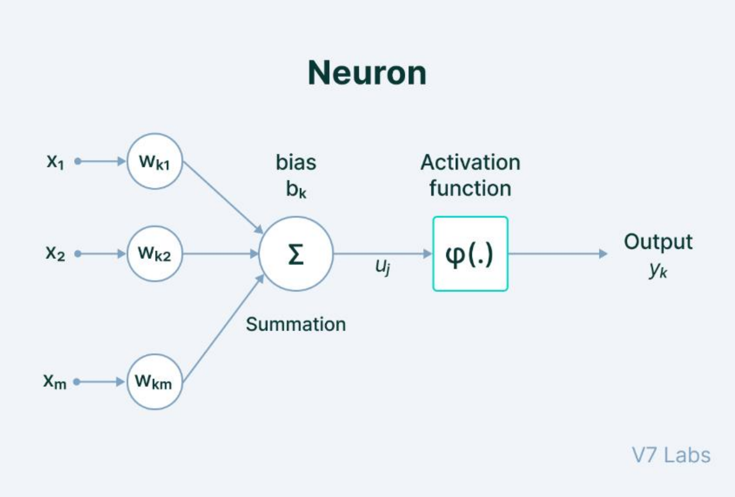


Figure 1. Neural Network Architecture

Source:

https://assets-global.website-files.com/5d7b77b063a9066d83e1209c/614fc05e2486109794ed3bdc_neuron.png

Components:

- Input Layer: Represented by the arrows on the left, it carries information from other neurons or external sources.
- Soma: The central circle, also called the cell body, processes incoming signals and performs calculations.
- Weights: Associated with each input arrow, these values determine the strength of the signal received from that specific input.
- Activation Function: Represented by the curved line within the soma, it applies a mathematical transformation to the weighted sum of the inputs.
- Output: Shown as the arrow on the right, it carries the activated output signal to other neurons in the network.

Function:

1. Signal reception: Incoming signals from other neurons travel through the input arrows.
2. Weighted sum: Each input signal is multiplied by its corresponding weight, representing the influence it has on the neuron's output.
3. Activation: The weighted sum is then passed through the activation function, which determines the overall output signal strength.
4. Output transmission: The activated output signal is then sent to other neurons through the output arrow, potentially influencing their activation as well.

The image likely illustrates a simplified perceptron model, a basic type of neuron with a linear activation function. Real neural networks typically have multiple layers of interconnected neurons, creating complex networks capable of higher-level processing.The specific weights and activation functions used in a neural network are determined during the training process, allowing the network to learn from data and improve its performance.

2.3 Natural Language Processing (NLP)

Natural language processing (NLP) is a machine learning technique from computer science that uses algorithms to analyze textual data.
NLP studies comprise theories and methods that enable effective communication between humans and computers in natural language. As a scientific field of study, NLP assimilates computer science, linguistics, and mathematics with a primary goal of translating human (or natural) language into commands that can be executed by computers.

Large language models (LLMs) have recently transformed the landscape of natural language processing, marking significant advancements in applications like conversational bots and text generation. Progress in deep-learning algorithms, increased computational resources, and substantial engineering endeavors have made it possible to train language models with billions of parameters, utilizing extensive datasets, exemplified by the expansive 800GB collection known

2.4 Transformer Architecture

The Transformer architecture has two main components.Encoder-decoder architecture basically allows the model to process a sequence of inputs and produce a sequence of outputs.

Encoder: The encoder takes an input sequence of symbol representations, denoted as $x_1, x_2 \ldots . . x_n$ These symbols can represent words, characters, or subword units, depending on the application. The main task of the encoder is to map each symbol xi to the corresponding continuous representation z_i

Decoder.The continuous representation $z = (z_1, z_2, \ldots . z_n)$ generated by the encoder is then passed as input to the decoder. The decoder produces a sequence of outputs $(y_1, y_2, \ldots . y_n)$

Attention : The attention function on a set of queries simultaneously, packed together into a matrix Q. The keys and values are also packed together into matrices K and V . We compute the matrix of outputs as:

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

Multi Head Attention : Multi-head attention allows the model to jointly attend to information from different representation subspaces at different positions.

$$\begin{aligned}\mathrm{MultiHead}(Q, K, V) &= \mathrm{Concat}(\mathrm{head_1}, \ldots, \mathrm{head_h})W^O \\ \text{where } \mathrm{head_i} &= \mathrm{Attention}(QW_i^Q, KW_i^K, VW_i^V)\end{aligned}$$

Feed-Forward Network (FFN) : Each layer in the encoder and decoder has a feed-forward network applied to each position separately and identically.

The FFN is described by the following formula

$$\mathrm{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2$$

where x is the input at each position, W_1 and W_2 are linear transformation matrices, b_1 and b_2 are bias vectors, and max(0,·) max represents the ReLU (Rectified Linear Unit) function.

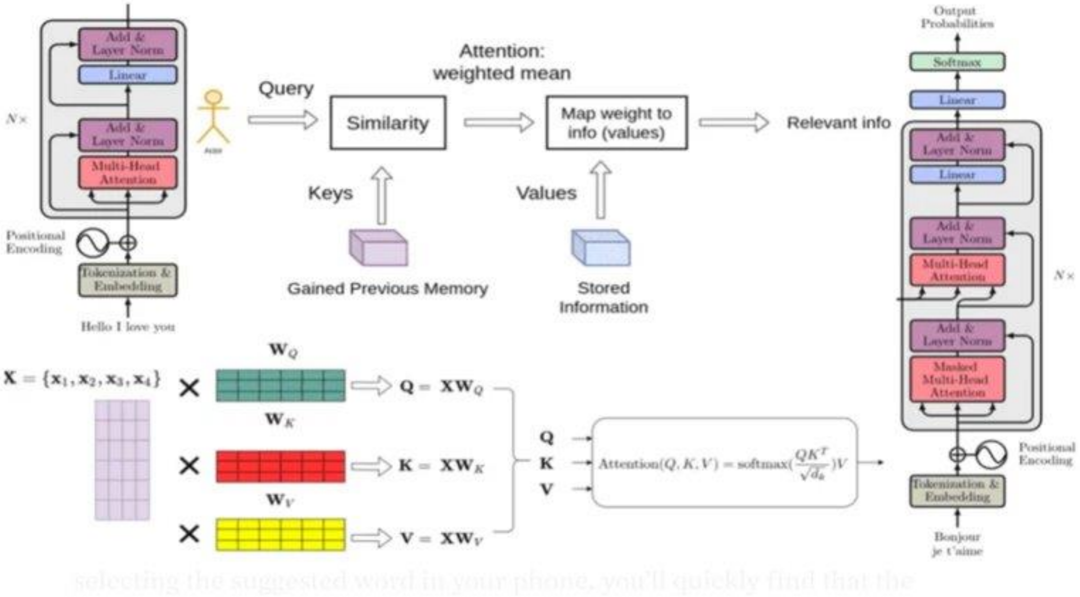


Figure 1. Trasformer Architecture

Source : https://medium.com/@amanatulla1606/transformer-architecture-explained-2c49e2257b4c

4. MedAIPaca - Medical Conversational AI

MedAIPaca aims to advance medical conversational AI by providing an open-source collection of models and training data.

4.1 Model Architecture:

MedAIPaca includes a diverse set of pre-trained models based on advanced natural language processing (NLP) architectures. These models are specifically designed for understanding and generating medical conversations, taking into account the complex and domain-specific nature of healthcare dialogue.

The dataset provided with MedAIPaca is a carefully curated collection of medical conversations, covering a wide range of scenarios and medical specialties. This high-quality training data is intended to facilitate the fine-tuning of models for specific healthcare contexts.

All components of MedAIPaca, including models and training data, are released under an open-source license. This encourages collaboration, transparency, and innovation in the development of medical conversational AI systems.

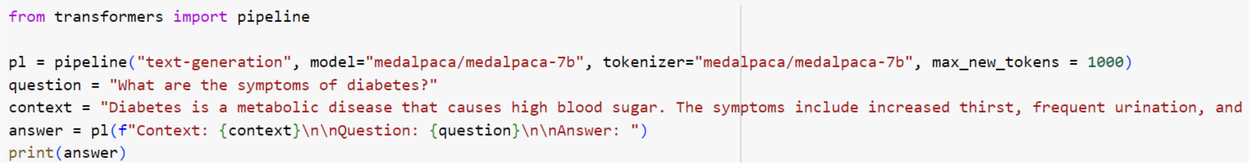
4.2. Implementation Details

Model Training: The models included in MedAIPaca are trained using a combination of publicly available medical dialogue datasets and domain-specific corpora. The training process employs transfer learning techniques to leverage pre-trained language models and adapt them to medical conversational tasks. Models are constructed based on the foundational LLaMA (Large Language Model Meta AI) models. LLaMA, an advanced large language model introduced by Meta, showcases their dedication to open science. It is accessible in multiple configurations, such as 7 billion, 13 billion, 38 billion, and 65 billion parameters. In our research, we conducted fine-tuning on the 7 billion and 13 billion parameter variants of the LLaMA model.Model

Evaluation: To ensure the effectiveness of the models, a comprehensive evaluation framework is provided within MedAIPaca. This framework includes standard metrics for assessing the performance of medical conversational agents, such as accuracy, fluency, and domain relevance.

4.3 Use Cases and Applications:

The versatility of MedAIPaca makes it suitable for a range of applications, including but not limited to virtual health assistants, medical education tools, and patient engagement platforms. Developers and researchers can leverage the models and training data to create AI-driven solutions that improve healthcare accessibility and communication. Here an example model of medalpaca



```
from transformers import pipeline

pl = pipeline("text-generation", model="medalpaca/medalpaca-7b", tokenizer="medalpaca/medalpaca-7b", max_new_tokens = 1000)
question = "What are the symptoms of diabetes?"
context = "Diabetes is a metabolic disease that causes high blood sugar. The symptoms include increased thirst, frequent urination, and
answer = pl(f"Context: {context}\n\nQuestion: {question}\n\nAnswer: ")
print(answer)
```

5. Conclusion

The combined strength of Alpaca's instruction-following capabilities and MedAIPaca 's focus on medical conversational AI presents a formidable synergy. This integration opens avenues for enhanced medical assistance, diagnostics, and communication, showcasing the potential for AI-driven advancements in healthcare. MedAIPaca represents a significant step forward in the domain of medical conversational AI. By providing open access to advanced models and curated training data, MedAIPaca aims to catalyze innovation in healthcare applications, ultimately contributing to improved patient care and communication in the digital age. The project is available for exploration and collaboration.

6. References

Han, T. et al. (2023). MedAlpaca – An Open-Source Collection of Medical Conversational AI Models and Training Data. https://arxiv.org/abs/2304.08247v2.H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A.

Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al., “Llama: Open and efficient foundation language models,” arXiv preprint arXiv:2302.13971, 2023.

L. Yunxiang, L. Zihan, Z. Kai, D. Ruilong, and Z. You, “Chatdoctor: A medical chat model fine-tuned on llama model using medical domain knowledge,” arXiv preprint arXiv:2303.14070, 2023.

https://crfm.stanford.edu/2023/03/13/alpaca.html

https://arxiv.org/pdf/2304.08247.pdf

https://huggingface.co/medalpaca