Large Language Model for automobile

DingFei[1,*]

.

Abstract

With the introduction of ChatGPT (OpenAI, 2022) from OpenAI, the power of these models to generate human-like text has captured widespread public attention.

The scale of language models has burgeoned, progressing from modest multimillion-parameter architectures like ELMo (Peters et al., 2018) and GPT-1 (Radford et al., 2018), to behemoths boasting billions, even trillions of parameters, exemplified by the monumental GPT-3 (Brown et al., 2020), Switch Transformers (Fedus et al., 2022) , GPT-4 (OpenAI, 2023), PaLM-2 (Anil et al., 2023), and Claude (Claude, 2023) and Vicuna (Chiang et al., 2023).

The expansion in scale has significantly raised hardware requirements, making it exceedingly challenging to deploy models on mobile devices such as smartphones and tablets.

To deploy on cars , we trained a 7-billion-parameter automobile model, which outperforms GPT-3.5 in the automotive domain. Surpassing all models in areas such as automotive maintenance, navigation queries, and beyond.

Keywords: Large Language Model , AI

Contents

*. email: dingfei@email.ncu.edu.cn

1 Introduction

Evidence suggests that large models exhibit emergent an emergent capability that is absent in smaller models (Wei et al., 2022a). A typical example is few-shot prompting. Few-shot prompting significantly expands the range of tasks supported by models and lowers the barrier to entry for users seeking automation for new language tasks.

After GPT-3, models grew in size to 280 billion (Gopher, Rae et al., 2021), 540 billion (PaLM, Chowdhery et al., 2022), and 1 trillion parameters (Megatron, Korthikanti et al., 2022). But their attention has been almost exclusively focused on general large language models,GPT-4 (OpenAI, 2023), PaLM-2 (Anil et al., 2023), Claude (Claude, 2023) ,LLaMA (Touvron et al., 2023), Alpaca (Taori et al., 2023), Vicuna (Chiang et al., 2023), and others (Wang et al., 2022; Zhu et al., 2023; Anand et al., 2023).

In recent endeavors aimed at training models solely with domain-specific data, the resulting models, albeit significantly smaller, have outperformed general-purpose LLMs in tasks specific to those domains, such as science (Taylor et al., 2022) and medicine (Bolton et al., 2023; Luo et al., 2022; Lehman et al., 2023). These discoveries inspire further advancement in the development of Large Language Model for automobile .

1.1 Objective

We train a 7 billion-parameter language model tailored to cater to a diverse array of tasks within the automobile sector.

1.2 Domain-specific LLMs.

The few existing domain-specific LLMs are trained exclusively on domain-specific data sources (Luo et al., 2022; Bolton et al., 2023; Taylor et al., 2022), or adapt a very large general purpose model to domain-specific tasks (Singhal et al., 2022; Lewkowycz et al., 2022). Our research aims to train LLMs using a combination of domain-specific and general data. The resulting model performs exceptionally well on domain-specific tasks while also maintaining robust performance on general benchmarks.

1.3 Training data.

In terms of data processing, we focus on data deduplication and high-quality data..

1.4 Architecture .

The model architecture of Automobile 7B is based on the prevailing Transformer (Vaswani et al., 2017). Nevertheless, we made several modifications .

1.5 Tokenizer .

We use byte-pair encoding (BPE) (Shibata et al., 1999) from SentencePiece (Kudo and Richardson, 2018) to tokenize the data .

1.6 Positional Embeddings .

We use Rotary Positional Embedding (RoPE) (Su et al., 2021) .

1.7 Activations and Normalizations .

We use SwiGLU (Shazeer, 2020) activation function, a switch-activated variant of GLU (Dauphin et al., 2017) which shows improved results.
We implement Layer Normalization (Ba et al., 2016) at the input of the Transformer block, known for its resilience to the warm-up schedule (Xiong et al., 2020). Furthermore, we integrate the RMSNorm technique proposed by (Zhang and Sennrich, 2019), which exclusively computes the variance of input features, enhancing computational efficiency.

1.8 Optimizations .

We use AdamW (Loshchilov and Hutter, 2017) optimizer for training. β_1 and β_2 are set to 0.9 and 0.95, respectively. We use weight decay with 0.1 and clip the grad norm to 0.5. Following a regimen of 2,000 linear scaling steps, the models are primed, gradually ascending to the peak learning rate before transitioning to a cosine decay, tapering down to the minimum learning rate.
To stabilize training and improve the model performance, we normalize the output embeddings.Because we observed that the norms of the heads tend to be unstable. Additionally, the norm of embeddings for rare tokens decreases during training, which disrupts the training dynamics. Moreover, we have found that semantic information is primarily encoded through the cosine similarity of embeddings rather than L2 distance.

1.9 Data .

We added 30% of automotive specialized knowledge books, comprehensive automotive information, automotive repair textbooks, and maintenance manuals for all vehicles to the data used for training the general LLM.

1.10 Train .

We use deepspeed zero2 .We have conducted performance optimization to enhance GPU computational efficiency and throughput.

2 Conclusion

We have presented Automobile 7B, a best-in-class LLM for automobile NLP. Our model contributes to the ongoing dialog on effective ways to train domain-specific models. Our training strategy of mixing domain-specific and general-purpose data results in a model that balances performance in both domains.
We've obtained impressive outcomes on general LLM benchmarks, surpassing similar models in automobile tasks. We attribute this to the meticulously curated dataset.We will continue to gather larger-scale automobile domain data to further optimize our model.

References

Yuvanesh Anand, Zach Nussbaum, Brandon Duderstadt, Benjamin Schmidt, and Andriy Mulyar. Gpt4all: Training an assistant-style chatbot with large scale data distillation from gpt-3.5-turbo. *GitHub*, 2023.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2023.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

Claude. Conversation with Claude AI assistant, 2023.

Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks. In *International conference on machine learning*, pages 933–941. PMLR, 2017.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *The Journal of Machine Learning Research*, 23(1): 5232–5270, 2022.

Taku Kudo and John Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*, 2018.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

OpenAI. Introducing chatgpt. *Blog post openai.com/blog/chatgpt*, 2022.

OpenAI. Gpt-4 technical report. *ArXiv*, abs/2303.08774, 2023.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. corr abs/1802.05365 (2018). *arXiv preprint arXiv:1802.05365*, 2018.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

Noam Shazeer. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.

Yusuxke Shibata, Takuya Kida, Shuichi Fukamachi, Masayuki Takeda, Ayumi Shinohara, Takeshi Shinohara, and Setsuo Arikawa. Byte pair encoding: A text compression scheme that accelerates pattern matching. 1999.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*, 2021.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html*, 3(6):7, 2023.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aur'elien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*, 2022.

Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tieyan Liu. On layer normalization in the transformer architecture. In *International Conference on Machine Learning*, pages 10524–10533. PMLR, 2020.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. *Advances in Neural Information Processing Systems*, 32, 2019.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. *Transactions on Machine Learning Research (TMLR)*, 2022a. doi: 10.48550/ARXIV.2206.07682. URL `https://arxiv.org/abs/2206.07682`.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas,

Aurelia Guy, Chris Jones, James Bradbury, Matthew Johnson, Blake Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Ed Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. Scaling language models: Methods, analysis & insights from training gopher. *arXiV*, 12 2021. URL http://arxiv.org/abs/2112.11446.

Vijay Korthikanti, Jared Casper, Sangkug Lym, Lawrence McAfee, Michael Andersch, Mohammad Shoeybi, and Bryan Catanzaro. Reducing activation recomputation in large transformer models, 2022. URL https://arxiv.org/abs/2205.05198.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. Galactica: A large language model for science. *arXiV*, 11 2022. URL http://arxiv.org/abs/2211.09085.

Elliot Bolton, David Hall, Michihiro Yasunaga, Tony Lee, Chris Manning, and Percy Liang. BioMedLM. https://github.com/stanford-crfm/BioMedLM, 2023.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6), sep 2022. doi: 10.1093/bib/bbac409. URL https://doi.org/10.1093%2Fbib%2Fbbac409.

Eric Lehman, Evan Hernandez, Diwakar Mahajan, Jonas Wulff, Micah J. Smith, Zachary Ziegler, Daniel Nadler, Peter Szolovits, Alistair Johnson, and Emily Alsentzer. Do we still need clinical language models?, 2023. URL https://arxiv.org/abs/2302.08091.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Scharli, Aakanksha Chowdhery, Philip Mansfield, Blaise Aguera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. Large language models encode clinical knowledge, 2022. URL https://arxiv.org/abs/2212.13138.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. Solving quantitative reasoning problems with language models, 2022. URL https://arxiv.org/abs/2206.14858.