LLM + Multi-Modality

1 Introduction

让模型能够接受并处理多种不同性质的信息是实现通用人工智能（AGI）的必由之路。近年来，随着 scaling law [5] 的潜力被不断发掘，我们已经拥有了能力强悍的大语言模型，比如 Bert [2]，GPT4 [8] 等等。这些端到端的大语言模型拥有处理文本信息并给予自然语言回复的能力，然而他们一开始被设计成只能处理本文编码成的 token，因此并不具有能直接泛化到其他模态数据的能力。

直觉上，文本、图像、音频以及其他各种模态的输入都是信息的载体，一定有某种方式把这些信息的编码结合起来，从而让模型拥有处理不同数据的能力，这也就是所谓多模态模型的出发点与目标。一个经典的例子就是图像描述：给模型一张图片，让它描述这个图里有什么东西。对于这种任务，早年间的 RNN 都可以做出还行的效果。但是我们的目标并不止于此，对于一个理想的多模态模型，它一定是能完全掌握各种模态上的信息，并处理各种各样的任务。比方说，图片描述、图文问答等等。并且我们希望它在大部分数据上都有很好的结果。这对我们的多模态模型提出了两点要求。(1) 它一定是 “one for all” 的，也就是能够处理各种东西；(2) 它一定是泛化的，并不局限于某一个小数据集。对于处理文本的模型，我们其实也有同样的要求，基于这两点目标，我们设计出了大语言模型（Large Language Model, LLM）。因此，考虑到现成的大语言模型已经有很强的泛化能力，也接受了非常大量的数据输入，它对这个世界已经存在一些先验的信息和认知，一个简单的想法就是：我们是否能够利用 LLM 来处理多模态数据?

现在我们看到，已经有 GPT4V [8] 这样的强大文本-视觉模型，能够同时处理文本和图像两种模态的信息。这证明了 LLM-base 的多模态大模型（Multimodal Large Language Model, MLLM）是有前途的。图 1显示了近年来 MLLM 的发展，可以见得这是一个非常蓬勃发展的领域。这张图来自这篇有关 MLLM 的综述 [14]，本文很大程度上受益于这篇文章，之后便不再指出明确的引用。我把视角聚焦于文本图像对齐的多模态大模型，一方面，图文理解是最常见的人工智能问题；另一方面，无论是图片、语音，还是视频乃至三维体素等等，他们虽然需要做特化的处理，但本质上的思路是一样的。在这篇报告中，我将概括性地介绍图文对齐 MLLM 的技术以及发展，并不着重于强调理论或工程细节。

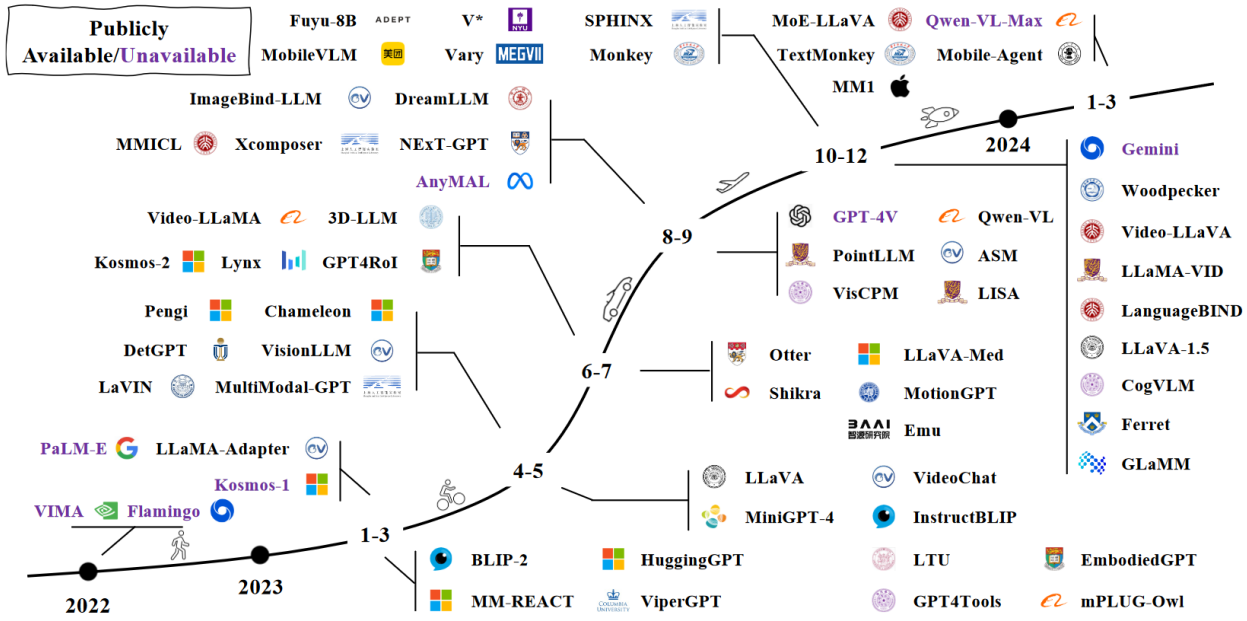

图 1: 近年来的各种多模态大模型

2 preliminary

在我们着手来谈 MLLM 的基本架构之前，我们必须先搞清楚现在的人怎么处理单模态信息的。直觉上来说，只要把图文各自编码成一个向量，然后再做一个 cross-attention，就可以把图文信息对接起来了。当然，事实上并不会有这么简单，但我们就先从注意力开始。

2.1 注意力机制

简单来说，注意力 [13] 就是一个加强版的多层感知机。对于一个很长的向量，我们希望模型能学到哪里是重要的，哪里是不重要的，也就是向量每个位置的权重，然后把“强化”版本的向量再传给下一层。对于一个向量，我们就用他自己的信息来编码这个权重向量，然后把权重向量和原向量做逐元素乘法，这就是自注意力机制；对于两个向量，我们想通过一个向量来判断另一个向量哪里重要，于是我们就用第一个向量来编码权重向量，再和第二个向量乘，这就是跨注意力机制。

2.2 图像模型

对图像的编码已经是一个很成熟的技术了。大体上可以分成两种，像素级别或者 token 级别。传统的卷积神经网络，比如 Resnet [4] 以及它的各种变体，可以从像素层面上提取图像的特征。后来的 ViT [3] 基于 Transformer [13] 把一张图片拆成若干个 token，然后提取信息。这些都已经成为许多任务的基本 backbone。

2.3 文本模型

Bert [2] 通过在文本中引入 `[MASK]`，然后让模型去预测每个位置的单词，从而学到文本的特征表示。GPT [10] 则用的是自回归的方法，让模型基于先验来预测下一个单词。基于这两种模式的模型还有很多，通过利用这些预训练的语言模型，我们能很好地编码各种文本。

2.4 多模态对齐

一些工作已经探索了结合文、图输入来同时编码两种模态的领域，在经过对齐的编码后，特征向量能够被传递进各种下游任务进行运用，比如赫赫有名的 CLIP [9]。它用的是一种叫做对比学习的策略，这种方法非常朴素，也就是最大化两个正确文本对所编码成的两个特征向量的余弦相似度。CLIP 所生成的特征有很强的迁移能力。

3 Basic Architecture

CLIP 这样的工作虽然探索了图文对齐的可能性，但是它毕竟只是一个编码器，离我们在节 1中提出的目标还很远。实际上，形如 CLIP 的编码器只是 MLLM 的一环。通常来说，MLLM 的架构如同图 2所示。简单来说可以分为两个部分，(1) 编码图文；(2) 对接图文。

3.1 编码多模态输入

在常见的 MLLM 中，编码不同模态输入的方法其实和我们在节 2中谈到的架构差别不大。对于图编码器，有使用基于 CLIP 的 ViT 模型 [17]，也有使用预训练卷积网络的 [16]，甚至也有抛弃预训练的图编码器，转而重新训练一个自己的编码器（当然也逃不开卷积或者 ViT。另一方面，MLLM 本身的重要特征就是 LLM，因而文本编码器都是使用现成的大语言模型，比如 T5 [11]，LLaMa [12] 等等。

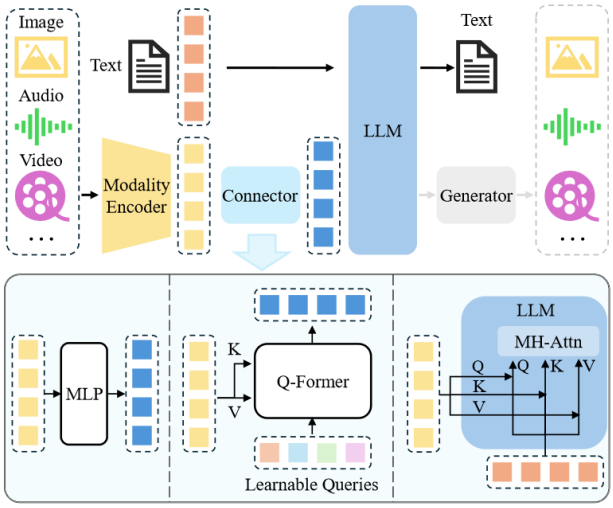

图 2: MLLM 的基本架构

3.2 对齐

现在我们已经有了两种模态输入的编码向量，剩下的事情就是把他们对接起来。目前的方法主要分为如下四种：

- 直接再编码图像的输入 token，过线性层。从理论上来说，模型是可以学到从图像编码到文本编码的一个映射 [7]。
- 用可学习的 token 来把图片编码成 LLM 能读懂的文本 token，这种方法被称为 Q-Former，被用在 BLIP-2 中 [6]。
- 直接在图文向量之间做 cross-attention 来达到对接的目的 [1]。
- 用一个专家模型把图像编码先转换成文本，再传到 LLM 中 [15]。

4 Challenge

虽然现在已经有一些很强大的 MLLM，但是它和其他的大模型一样都存在一些问题。

- **幻觉**。如普通的 LLM 一样，MLLM 的幻觉问题其实是最为主要的问题。在 LLM 中，我们只有一种模态的输入（文本），而当模态变多，无论是从训练数据、训练手段、模

型复杂度来看，都给研究者们除了很大的难题。现在的模型还是有很大概率输出与画面/文本相悖的产物。

- **长上下文**。现在的 MLLM 往往只局限于处理一张或者几张图片，然而更理想的是，我们给 MLLM 输入一段长视频，它也能告诉我们视频里面的具体信息。这其实是一个非常困难的任务。直觉上，一个视频就是成千上万帧的图片，这无疑对模型的处理能力提出了很高的要求。
- **具身 MLLM**。更为理想地，我们希望多模态大模型能够真实地理解自然数据，并且用机械臂等等来和外界交互。

References/参考文献

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[5] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020.

[6] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[7] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

[8] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning,

Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024.

[9] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[10] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever,

et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[11] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

[12] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[14] Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and Philip S. Yu. Multimodal large language models: A survey. In *2023 IEEE International Conference on Big Data (BigData)*, pages 2247–2256, 2023.

[15] Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. Woodpecker: Hallucination correction for multimodal large language models. *arXiv preprint arXiv:2310.16045*, 2023.

[16] Yuqian Yuan, Wentong Li, Jian Liu, Dongqi Tang, Xinjie Luo, Chi Qin, Lei Zhang, and Jianke Zhu. Osprey: Pixel understanding with visual instruction tuning. *arXiv preprint arXiv:2312.10032*, 2023.

[17] 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.