Bridging Feature Spaces: Survey on Multimodal Large Language Model

WashU CSE527A Survey Paper

Jianhong Tu
Washington University in St. Louis
jianhong.t@wustl.edu

Second Author
Washington University in St. Louis
email@domain

Third Author
Washington University in St. Louis
email@domain

Abstract

1 Introduction

1.1 Multimodality

"Multimodal" in the context of machine learning typically refers to models or systems that can process and understand information from multiple types of input, such as text, images, and sound.(Yuzijiang, 2023)

In the early days of computing, the primary mode of interaction was text. Computers were primarily tools for mathematical computations and basic data storage, and the primary focus was on text-based input and output.

With the advent of more powerful computers and graphical user interfaces (GUIs) in the late 20th century, images and sounds began to play a more significant role in computing. This led to the development of models and algorithms that could process not just text but also images and sounds.

In the 21st century, with the proliferation of devices equipped with cameras, microphones, and other sensors, the importance of being able to process multiple modes of input simultaneously became even more pronounced. This era saw the rise of truly multimodal systems capable of understanding and generating content across text, images, sound, and even video.

The evolution of multimodal systems mirrors the evolution of computing itself, from basic text processing to the rich, multimedia experiences we have today. As technology continues to advance, we can expect even more integration of different modalities, leading to more intuitive and powerful computing experiences.(Chu, 2023)

1.2 Large Language Model: Deficiencies

1. **Limited Data Type Handling**: Large language models primarily process textual data and cannot directly manage images, sounds, or other non-textual data. In contrast, multimodal models can analyze and generate content across multiple data types.

2. **Lack of Intuitive Understanding**: When tasks related to images, audio, or other data modalities arise, pure text models might not offer an intuitive understanding or response. For instance, a text-centric model might struggle to interpret or comment on the content within an image, whereas a multimodal model can.

3. **Interaction Limitations**: In interactions with users, large language models can only communicate through text. Multimodal models, on the other hand, can interact with users through images, audio, and other modes, offering a richer and more diverse experience.

4. **Task Restrictiveness**: For tasks that demand simultaneous handling of various data modalities, singular text models may face challenges. For instance, a task might require analyzing an image and its associated description simultaneously, where a multimodal approach would be more appropriate.

5. **Complexity and Diversity**: As the complexity and diversity of AI applications grow, relying solely on text models might not suffice to meet all requirements. Multimodal models, with their capability to process multiple data modalities, thus offer greater adaptability and flexibility.

6. **Completeness of Perception**: Humans typically rely on several senses (e.g., visual, auditory, tactile) to understand the world. Utilizing only text models might not wholly emulate this multimodal perceptual experience.(Wikipedia contributors, 2023)

1.3 Challenges

1. **Data Alignment**: Integrating information from different modalities often requires precise temporal and spatial alignment. Any misalignment can lead to ineffective or even misleading results.

2. **Data Representation**: Choosing an appropriate representation for each modality, as well as a unified representation for all combined modalities, remains a significant challenge.

3. **Scalability**: With the addition of each new modality, the complexity of the model can grow exponentially. Ensuring that the computational and memory requirements do not explode is essential.

4. **Inter-modality Gaps**: Differences in the nature of data from different modalities can make it challenging to create a cohesive model that learns from all modalities equally effectively.

5. **Annotation and Labeling**: Obtaining ground truth labels for multimodal data can be time-consuming and expensive. Semi-supervised or unsupervised approaches might be required.

6. **Transfer Learning and Generalization**: Models trained on specific multimodal datasets may not generalize well to other scenarios or modalities. Transfer learning in a multimodal context is still an open challenge.

7. **Interpretablility**: Understanding the contribution of each modality to the final decision or output of the model is crucial for trust and further model improvement.(Grifoni et al., 2021)

2 Methods

A multimodal large language model (MLLM) leverages the idea of *natural language supervision* (Radford et al., 2021) to enhance the model's awareness of an entity through various input channels by taking advantage of the *emergent capabilities* (Wei et al., 2022) of large language models (LLM), such that a unimodal LLM multi-step can achieve reasoning (Rae et al., 2021) and instruction following (Ouyang et al., 2022). To limit the scope, this survey focuses on **Vision-Language Models** (VLM) that accepts textual tokens and images and details the methods to construct a joint feature space for two modalities via step-by-step feature transforming.

2.1 Datasets

Obtaining a large dataset is crucial to the generalization ability of VLM because the quality of unimodal representation learning is positively proportional to the scale of a dataset. Particularly, in learning image representation, the vision transform, a popular choice in VLM architecture, yields a modest accuracy when trained on a mid-sized dataset but performs better than the SOTA ResNet in a set of common image recognition benchmarks when trained on larger datasets (Dosovitskiy et al., 2021). One method to gather sufficient data points is to integrate benchmark datasets across domains. The PaLM-E group, concerning robotic policy in addition to visuals and texts, gathered the image captioning data "COCO", visual question answer data "VQAv2", text corpus "Wikipedia", and robotic navigation data "SayCan" Driess et al. (2023). Appendix A summarizes the common datasets. However, crowd-labeled datasets are often limited in size or tailored for tasks other than VLM. There is a trend to exploit the richness of information on the internet by compiling a customized dataset (Radford et al., 2021). Sun et al. (2019) utilized YouTube's audio-transcription to gather 176 hours of synchronous video frames and captions through web API, while Radford et al. (2021) collected approximately 400 million pairs of image and descriptions from public sources.

2.2 Joint Representation

LLMs trained on textual data are incompatible with the raw image data as a 3D tensor with height, width, and multiple channels. A common multimodal modeling approach is learning a joint representation of texts, images, and other modalities. Formally, let $X, \ Y$ be the feature spaces, and we intend to find $\phi : X \times Y \rightarrow Z$, where Z is the feature space for the VLM. Due to the cost of training end-to-end multimodel (Yin et al., 2023), it is more feasible to concatenate unimodels into a multimodel, where conventional unimodels are pretrained separately. With the LLM as the backbone, the function ϕ serves as a learnable interface that projects the image space X onto the embedding space $Z = Y$ (Yin et al., 2023), keeping the feature space invariant to pretrained LLMs.

Pretraining refers to the stage where a taskagnostic model is trained to optimize an objective function in order to capture the general features of data. Akin to pretrained embeddings can capture semantic similarities, pretrained image classifiers can encode abstract signals. Consider a trained deep classifier as a composite function $h \odot \theta$ where h is a SoftMax classifier and θ is a feature transformation. The function h may be understood as a mapping function from raw feature space to the transformed feature space where the data is linearly separable. Sun et al. (2019) obtains such feature mapping by pretraining a ConvNet on the Kinetics dataset then detaching the final SoftMax layer, resulting in 1024-dimensional continuous feature vectors. Continuous features are tokenized using hierarchical K-Means to extract 2^{14} centroids as the final image representations. An alternative model is to apply vision transformer, which is an extension of the transformer architecture that conveniently maps images as a sequence of patches into patch embeddings (Dosovitskiy et al., 2021; Driess et al., 2023; Radford et al., 2021). However, patch embeddings are not equivalent to the word embeddings.

2.3 Multimodal Pretraining

Unimodal pretraining learns a useful image prior, and multimodal pretraining further establishes the connection between image and word embeddings. Most simply, the connection can be learnt by treating video embeddings as word embeddings. Sun et al. (2019) appends the new image embeddings to a pretrained BERT's lookup table and optimizes a weighted sum of unimodel and multimodel objectives detailed in Appendix B. More directly, *contrastive objectives* are used to fine-tune the text and image encoders by explicitly training on a image-caption paring task to maximize the cosine-similarity between paired image and text embeddings. (Radford et al., 2021). Another light-weight training scheme as opposed to the end-to-ending training above, is to learn a linking function that transform images to "prompts" while "freezing" the encoders and the LLM (Driess et al., 2023). The multimodel is pretrained on a next-sentence prediction tasks with sequence of visual and linguistic embeddings, updating the affine transformation function only.

2.4 Fine-tuning

Fine-tuning is a technique to adapt a pretrained model for specific downstream tasks by transferring acquired unimodal knowledge into the final multimodoel. The multimodel optimizes some appropriate objective functions used to assess the model's performance on a curated set of benchmarks, such as visual question answering (VQA) (Yin et al., 2023). Fine-tuning enables application of the model in high level tasks, which may be evaluated qualitatively. Figure 2 provides 2 examples: VideoBERT model (Sun et al., 2019) captions the given image by describing the scene and actions, and predicting successive steps while the PaLM-e model (Driess et al., 2023) generates a robotic policy based on the given prompt.

Model	VQAv2 test-dev	test-std	OK-VQA val	COCO Karpathy test
PaLI-3B	81.4	-	52.4	145.4
PaLI-15B	82.9	-	56.5	146.2
PaLI-17B	84.3	84.3	64.5	149.1
PaLM-E-12B	77.7	77.9	60.1	136.0
PaLM-E-66B	-	-	62.9	-
PaLM-E-84B	80.5	-	63.3	138.0

Table 1: Performance of PaLI (Chen et al., 2023) and PaLM-e (Driess et al., 2023) on General Visual-language Benchmarks

3 Results

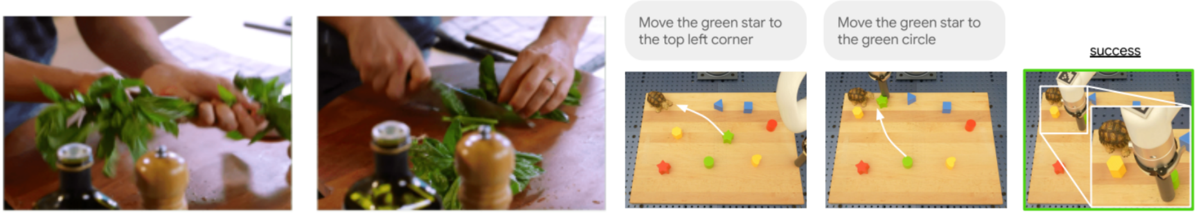
Evaluation of multimodels is data and task specific. We generally discusses 3 perspectives to assess model's performance on downstream tasks and the quality of join representation learning.

3.1 Benchmark scores

Usage of common benchmarks enables cross evaluation of different model. Table 1 provides the reference scores of 2 SOTA MLLM on multimodal tasks. VQA (Goyal et al., 2019) provide image and prompt as the inputs and access the model's ability to retrieve information from the images as the answer to the prompt, while COCO caption (Chen et al., 2015) let the model to generate context-free description of the image. The scales have a positive impact on both models.

3.2 Customized Task Performance

Models fine-tuned to a more restricted or less common domains tend to assessed using customized benchmarks, and so is their training. VideoBERT (Sun et al., 2019) is similarly evaluated on video captioning, but on cooking instructional videos. MLLM is also applied to the robotics field with the model as the "brain". PaLM-e (Driess et al., 2023) is able to interact with the environment, which is


Move the green star to the top left corner
Move the green star to the green circle
success

VideoBERT: “chop the basil and add to the bow”

PaLM-e: “Move the green blocks to the turtle”

Figure 1: Examples of High Level Tasks

evaluated on the success rate to carry out human instrcution, i.e. pick up the sponge and then bring it to the user. In a VQA fashion, researchers can assess the model’s understanding of the affordance, that whether an *action* can be performed on an *entity*. In addition, *zero-shot classification* is of great interest to understand the generalization, *transfer*, of abstract knowledge into unseen examples. Driess et al. (2023), Radford et al. (2021) and Sun et al. (2019) report a better zero-shot image classification accuracy than the baseline unimodel and more efficient learning, requiring less data to achieve the same level of performance.

3.3 Representation Effectiveness

Recall that a representation mapping can be obtained by detaching the output layer. The quality of the learnt representation can be thus accessed by the accuracy of a linear classifier on some classification tasks. Compared against ResNet (He et al., 2015) and BiT-M (Kolesnikov et al., 2020), the features outputted by CLIP (Sun et al., 2019) achieves a >70% accuracy on the 16-shot classification tasks versus 63% and 53% accuracyf for BiT-M and ResNet respectively, meaning a more effective representation learning result.

4 Discussion

5 Conclusion

References

Xi Chen, Xiao Wang, Soravit Changpinyo, A. J. Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish V. Thapliyal, James Bradbury, and Weicheng Kuo. 2023. Pali: A jointly-scaled multilingual language-image model. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. 2015. Microsoft COCO captions: Data collection and evaluation server. *CoRR*, abs/1504.00325.

Mingqian Chu. 2023. Introduction to multimodality. Accessed: 2023-10-23.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. 2023. Palm-e: An embodied multimodal language model. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 8469–8488. PMLR.

Yash Goyal, Tejas Khot, Aishwarya Agrawal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2019. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. *Int. J. Comput. Vis.*, 127(4):398–414.

Patrizia Grifoni, Arianna D’ulizia, and Fernando Ferri. 2021. When language evolution meets multimodality: Current status and challenges toward multimodal computational models. *IEEE Access*, 9:35196–35206.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition. *CoRR*, abs/1512.03385.

Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. 2020. Big transfer (bit): General visual representation learning. In *Computer Vision -*

ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part V, volume 12350 of *Lecture Notes in Computer Science*, pages 491–507. Springer.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. OK-VQA: A visual question answering benchmark requiring external knowledge. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 3195–3204. Computer Vision Foundation / IEEE.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *NeurIPS*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, H. 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 M. 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 J. Johnson, Blake A. Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Edward Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. 2021. Scaling language models: Methods, analysis & insights from training gopher. *CoRR*, abs/2112.11446.

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. Videobert: A joint model for video and language representation learning. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 7463–7472. IEEE.

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. 2022. Emergent abilities of large language models. *Trans. Mach. Learn. Res.*, 2022.

Wikipedia contributors. 2023. Large language model — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Large_language_model&oldid=1181088602. [Online; accessed 22-October-2023].

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023. A survey on multimodal large language models. *CoRR*, abs/2306.13549.

Yuzijiang. 2023. What is multimodality? Accessed: 2023-10-23.

A Example MLLM Datasets

Dataset	Task	Size	
TAMP	Robotic manipulation planning with VQA	96000 scenes	Source
Language Table	Robotic manipulation planning	600000 sequences	Driess et al. (2023)
Mobile Manipulation	Robotic navigation and manipulation planing with VQA	2912 sequences of action	Driess et al. (2023)
VQA2v	Question answering with images and prompts	1.1M questions with images	Goyal et al. (2019)
COCO	Image captioning	123287 images with captions	Chen et al. (2015)
OK-VQA	Visual question answering requiring external knowledge	14031 questions with images	Marino et al. (2019)
YouCook	Instructional video captioning	2000 videos with caption	Sun et al. (2019)
CLIP benchmark	Image captioning	400M image-text pairs	Radford et al. (2021)

Table 2: Summary of MLLP datasets

B Multimodal BERT Objectives

Pretraining of a multimodel is commonly conducted in a step by step fashion: first the encoders of various modalities is trained, then the pretrained models are concatenated into a multimodel and pretrained on visual-linguistic tasks to acquire a general embedding space of visuals and texts. Here we explain the pretraining objectives of VideoBERT (Sun et al., 2019) as a paradigm of multimodel pretraining.

VideoBERT is a masked multimodal large language model. With the images summarised into $2^1 4$ discrete tokens, both word embeddings and image tokens can be accepted as input to the BERT model. For each modality, the sequences of words or video frames are randomly masked so that the model is trained to restore the maksed embeddings. The video-only training objective is analogous to the text-only, or the valina BERT training objective. It helps the model to acquire long-term state dependencies in video frames. Formally, training each modality separately ensures that the model's capability to capture the marginal distribution of image $P(x)$ or text $P(y)$.

The third training objective to establish the joint distribution $P(x, y)$ so that the model understands the interplay of video and texts. To encode image information into sequences of word embeddings, word-based sentences and visual tokens are combined into visual sentences jointed by a special token "[>]".

```
[CLS] orange chicken with [MASK] sauce
 [>] v01 [MASK] v08 v72 [SEP]
```

Researchers proposed a *linguistic-visual alignment* task, where the state of the "[cls]" token is used to determine whether the linguistic sequence and the visual sequence are aligned, which is, to some degree, similar to the *next-sentence prediction* task in the vanila BERT model. This objective essentially establishes the connection between two domains.

After pretraining, the model would learn a joint representation of multi-modalities. Further finetuning allows it to be used into various downstream stasks.