MMMoE: Advancing Multimodal Mixture-of-Experts Architecture to Power Next-Generation Multimodal Paradigm

Jianzhu Yao
yjz19@mails.tinghua.edu.cn

Abstract

Large sparsely-activated models have a great performance across different areas. However, such models usually focus on unimodal and uni-task scenarios. We present the first multitask multimodal MoE, MMMoE, a general sparse mixture of experts model with heuristic expert selection technique. MoE is an architecture fit for multimodal machine learning, since expert layers can learn to partition different modality according to token-level features. But modality imbalance problem is a unique problem in this domain. For that, we present a modality level entropy regularisation loss function to solve this problem. We demonstrate the remarkable performance improvement in MultiBench dataset, compared to state-of-art multitask models. We also analyse our model on four domains: Performance, Load Balance, Computing Latency and Few-shot Learning scenarios. The experiments show that MoE is a strong backbone of multimodal machine learning.

Keywords: Mixture of Experts, Multimodal Machine Learning

1 Introduction

Sparsely activated mixture of experts (MoE) models has recently developed rapidly thanks to the development of distributed computing and computing devices. And at the same time, the visual and text MoE are well researched in each specific area. The primary motivation to use MoE for these tasks is to scale model parameters while keeping compute costs under control. What's more, when enlarging the model capacity, the performance of MoE can be better than dense models with equivalent computational cost, which makes MoE even more environmentally friendly.

So it is a smooth transition from unimodal MoE to multimodal MoE. It is very intuitive that different experts can process information of different modalities. And existing works have shown the better performance multimodal MoE can achieve. [1]. Given success in the multimodal modelling domain, and the intuition that sparse models may have much richer model capacity which can handle modality noise or low-resource data problems, we explore the application of MoEs to multimodal multitask and transfer learning. We take the first step in this direction, and develop MoE models that support task generalization for multimodal machine learning. In particular, we train a single general multimodal MoE that can process various modalities for different multimodal tasks, as an alternative to existing task specific models. What's more, considering in some mobile and IoT scenarios, for lower implementation and computation latency, we propose a heuristic Multitask Multimodal MoE (MMMoE) architecture as a universal multimodal processing model with less inference cost.

One intuition of why larger expert capacity can help improve performance is that those extra experts can help correct the representation of the first one when use Top-k ($k \geq 2$) routing. So we fix one expert for each modality, and varying the second expert to correct the representation of the modality-specific expert. Afterward, we add up the output of these two experts to get the final output. The main motivation of this approach is to treat the second expert as an error correction term of the modality-specific expert. We call this technique heuristic expert selection. The details of our model architecture is described in Section 2.

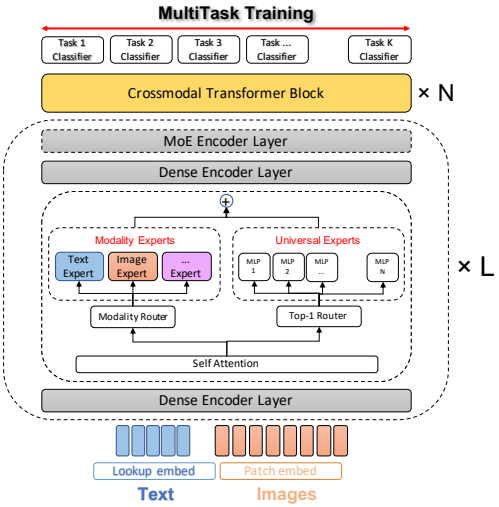

Figure 1: The model architecture of our proposed multitask multimodal MMMoE.

However, to handle the multimodal MoE's unique load balance problem, different from previous approach [1], we present a set of *modality-level entropy regularisation* which can make training stable and solve the imbalance problem.

In summary, our contribution are as follows.

1. We present MMMoE, the first large-scale multitask multimodal mixture of experts models, with modality specific expert selection technique and modality-level entropy regularisation scheme to speed up and stabilise training.
2. We analyze in detail how our approach can facilitate multimodal transfer learning, few-shot, zero-shot learning, even in low-resource scenarios (less training data or missing modalities when inference).

2 Methods

In this section, we describe our model for multitask multimodal machine learning (illustrated in Figure 1). We also describe the definition of our heuristic modality-specific expert selection technique and the modality-level entropy regularisation as the auxiliary loss function.

2.1 Modality-specific embedding

For each distinct modality $m_i \in M$ (which may appear across multiple tasks), where $1 \leq i \leq |M|$ and $|M|$ is the number of modalities, we use a modality-specific embedding network to project each heterogeneous intrinsic data dimension to the desired width for our MoE model.

2.2 Heuristic Unimodal Expert Selection Mixture of Experts Encoder

With unified representation of each modality, then, all tokens are processed by a shared unimodal mixture of experts transformer encoder. To lower the computation latency for our model, inspired by the DeepSpeed-MoE [2], we implement a heuristic expert selection method for faster inference and training.

Specifically, for every MoE layer, we use a fixed modality-specific expert E_{m_i} and another expert according to the routing algorithm presented in GShard [3]. Such that we can achieve the benefit of using 2 experts per layer with the same amount communication volume as Top-1 gating function (because we can assign each modality-specific expert to each host when forwarding, which means there are less all2all communication cost between each host). Related research points out in Top-2 gating algorithm, the second expert tends to correct the output of the first expert, so the implementation of the fixed modality-specific expert will play the leading role in representation. The j-th output hidden state of this $l+1$ MoE layer is

$$h_j^{l+1} = O_{m_i} + g_k O_k \tag{1}$$

, where O_{m_i} is the output of the modality-specific expert E_{m_i}, O_k is the output of the second expert, g_k is the second expert weight calculated by the top-1 algorithm among the remaining experts:

$$k = \text{argmax}_i h_j^l \cdot v_i \tag{2}$$

where v_i is the representation of expert E_i.

2.3 Crossmodal Transformer Block

After we get each modality's representation, we use multiple layers of a general-purpose Crossmodal Transformer block([4], [5]). For example, we use two modalities representation z_1 and z_2 in this task. In those Transformer Blocks' cross attention modules, we set z_1 as query, z_2 as key and value to get representation Z_{12} to learn attention from modality 1 to modality 2. Similarly, we set z_2 as query, z_1 as key and value to get representation Z_{21}. We denote the final output representation of these two modalities as Z_m,

$$Z_m = \boldsymbol{W} f([Z_{12}; Z_{21}]) + \boldsymbol{b} \tag{3}$$

, where $[;]$ is the concatenation operation, $\boldsymbol{W}$ and $\boldsymbol{b}$ are trainable parameters and f is the SiLU activation function [6] followed by layer normalization. For tasks with more than 2 modalities, this approach is applied between every two modality pair and finally concatenate every representation to get Z_m.

2.4 Task-specific Classifier and Model Training

For task-specific prediction, we use a linear classification MLP layer per task applied on Z_m to get the final prediction. To boost the information sharing between different modalities and task knowledge, we train our model across different multimodal tasks. And we optimize a weighted sum of loss over all tasks we select:

$$\mathcal{L}_{Tasks} = \sum_{t=1}^{|T|} w_t \mathcal{L}_t \tag{4}$$

where T is the task set and $|T|$ is the number of tasks, $\mathcal{L}_t$ is the loss function of task t, w_t is a task-specific weight as a hyper parameter.

2.5 Auxiliary Loss Function

Previous work showed that classical auxiliary loss function ($\mathcal{L}_{aux}$) of mixture of experts model can't make multimodal MoE stable and perform well. Because if one modality's tokens are all dropped, while other modality's tokens are equally distributed to all experts, the classical auxiliary loss can also be very low. So inspired by this phenomenon, we propose a modality-level entropy loss for our multitask multimodal MoE.

Definition Suppose for task t, we will use t_m modalities, and for each modality m_k ($1 \leq m_k \leq t_m$), we denote there are n_{m_k} tokens in each batch. For a token x, the probability distribution over E experts are $p_{m_k}(\text{experts}|x)$, and our modality-level entropy loss is defined by:

$$\mathcal{L}^t_{modal} = -\frac{1}{t_m} \sum_{k=1}^{t_m} \mathcal{H}(\tilde{p}_{m_k}(\text{experts})) \tag{5}$$

$$\tilde{p}_{m_k}(\text{experts}) = \frac{1}{n_{m_k}} \sum_{i=1}^{n_{m_k}} p_{m_k}(\text{experts}|x_i) \tag{6}$$

where $\tilde{p}_m(experts)$ is the expert probability distribution averaged over the tokens and $\mathcal{H}$ is the entropy function $\mathcal{H}(p) = -\sum_{e=1}^{E} p_e log(p_e)$. $\mathcal{L}_{modal}$ guarantee a consistent expert distribution among different modalities. So the final loss function of our MMMoE is

$$\mathcal{L} = \mathcal{L}_{Tasks} + \sum_{t=1}^{|T|} w'_t \mathcal{L}^t_{modal}, \tag{7}$$

where w'_t is a task-specific modality-level loss function weight as a hyper parameter.

Research Area	Size	Dataset	Modalities	# Samples	Prediction task
Affective Computing	S	MUSTARD [24]	$\{\ell, v, a\}$	690	sarcasm
	M	CMU-MOSI [181]	$\{\ell, v, a\}$	2,199	sentiment
	L	UR-FUNNY [64]	$\{\ell, v, a\}$	16,514	humor
	L	CMU-MOSEI [183]	$\{\ell, v, a\}$	22,777	sentiment, emotions
Healthcare	L	MIMIC [78]	$\{t, ta\}$	36,212	mortality, ICD-9 codes
Robotics	M	MUJOCO PUSH [90]	$\{i, f, p\}$	37,990	object pose
	L	VISION&TOUCH [92]	$\{i, f, p\}$	147,000	contact, robot pose
Finance	M	STOCKS-F&B	$\{t \times 18\}$	5,218	stock price, volatility
	M	STOCKS-HEALTH	$\{t \times 63\}$	5,218	stock price, volatility
	M	STOCKS-TECH	$\{t \times 100\}$	5,218	stock price, volatility
HCI	S	ENRICO [93]	$\{i, s\}$	1,460	design interface
Multimedia	S	KINETICS400-S [80]	$\{v, a, o\}$	2,624	human action
	M	MM-IMDB [8]	$\{\ell, i\}$	25,959	movie genre
	M	AV-MNIST [161]	$\{i, a\}$	70,000	digit
	L	KINETICS400-L [80]	$\{v, a, o\}$	306,245	human action

Figure 2: The multimodal datasets provided by MultiBench.

3 Experiments

I will design experiments to analyze the multitask capabilities of MMMoE. We use a set

of multimodal datasets provided by MultiBench [1][7] benchmark spanning 15 datasets, 10 modalities, 20 prediction tasks and 6 research areas. The datasets have different modalities, such as language, image, video, audio, time-series, force sensors, set, table and optical flow, and different domains spanning healthcare, robotics, finance, HCI, multimedia and so on. The dataset statistics is shown in Figure 2

For baselines, we will use **the same amount of activated parameters** to set the baseline capacity. The MultiBench provides some unimodal approaches like MLP, GRU, LeNet, CNN, LSTM, Transformer and so on. And the repository also provides different training structures like supervised learning (with early fusion, late fusion, MVAE, etc), gradient blend, architecture search, etc. The benchmark paper [7] provides the performance range and SOTA of different tasks, so we can use those models and scores as baseline to evaluate our MMMoE's performance on each task.

At the same time, we use a multitask multimodal HIGHMMT [8] as another strong baseline. HIGHMMT is also a general multimodal model and the difference from our model just lies in the task-agnostic unimodal encoder. Specifically, our MMMoE enlarged the model capacity by increasing the number of experts. By comparing the performance of two general multitask multimodal models, we can test the ability of mixture of expert.

And we will also replace our heuristic expert selection method with traditional Top-2 gating algorithm in GShard [3], to compare the inference speed of MMMoE, which is important in edge computing efficiency and real-time computing latency. What's more, the comparison between classical MoE architecture and HIGHMMT baselines can also give us insight of the ability and unique features to process multimodal data of mixture of experts architecture.

4 Analysis

We conduct experiment analysis in three different dimensions: performance analysis, load balance analysis, latency analysis on our model MMMoE. However, the low-resource scenarios are also common in multimodal tasks. Especially in the real-time applications, different level noises and missing data often influence the recognition or generation performance. So we also conduct a few-shot learning analysis to test our model's ability.

Performance Analysis To conduct the performance analysis, we compare MMMoE's score with other baselines. Different tasks have different evaluation metrics, so we will use the evaluation metrics provided by MultiBench benchmark itself.

Load Balance Analysis Because the load balance problem only occurs in mixture of experts architecture, so to conduct this load balance analysis, we explore how routing is distributed across different layers, experts, and modalities for each task. We detect and count the number of tokens passing through each expert and the tokens dropped by each layers, in order to evaluate the load balance issues.

First, we will compare the token distribution of each experts of MMMoE and traditional Top-2 gating algorithm in GShard [3] baseline in order to observe the influence of heuristic expert selection technique. Then, we will replace the $\mathcal{L}_{modal}$ with $\mathcal{L}_{aux}$ in MMMoE to observe the load performance of different loss function. Most importantly, we will compare the routing results between different multimodal tasks to check the consistency of loading each modality's tokens.

If our modality level entropy loss function does not work, we will try the following alternatives as ablation study:

1. remove the $\mathcal{L}_{modal}$, use Ω_{local} and Ω_{global} proposed in [1] and compare the performance.
2. remove the heuristic expert selection and use the Ω_{local} and Ω_{global} loss function. [1]
3. use Residual-MoE proposed in [2] to observe the influence of universal MLP on load balance problem.

[1]https://github.com/pliang279/MultiBench

Latency Analysis To conduct the latency analysis, we will compare the training convergence speed of different models/approaches and the speed of inference, which can seriously affect the efficiency of real-time computing. Just as most of MoE papers, we will compare the power consumption during our MMMoE and baselines' training. For real-time computing application, the implement speed and distributed computing technique of MMMoE have a promising advantage.

Few-shot learning Because MMMoE will share different modality information across tasks and between experts, we will test this hypothesis in a few-shot scenario by evaluating model's performance on low-resource target tasks. We will use a MMMoE trained on one single task and a MMMoE trained on multitask to compare their performance on that single task. We can choose any single task in the MultiBench datasets. Because it is always difficult to collect enough data in a target domain, using an expert-shared multimodal model is a great approach for improving performance.

5 Conclusion

We have presented MMMoE, the first multitask multimodal sparse mixture of experts model. We proposed a new modality-level entropy loss function to handle the load balance problem of multimodal MoEs. Thanks to the model capacity provided by MoE and transfer learning features, our model can perform well even in low-resource or modality-noisy scenario, which is also an open problem in multimodal domain. Specially, our heuristic expert selection technique is also the first routing algorithm designed for multimodal scenario, which lowers the computing latency and make the training more stable and environment friendly.

6 Future Work

There are many interesting directions from multimodal mixture of experts.

Unimodal and Multimodal MoE The application and algorithm of MoE for NLP haven't carried over perfectly to Vision, and also other modalities. Previous work observed different behaviour between images and text. Of course, MoE extensions to more modalities like audio, video, time series should also be explored. The study of unimodal features is also instructive for multimodal MoE. What's more, the load balance problem raised from LIMoE also shows there may be amazing interactions between different modalities in the MoE domain and routing algorithms. So the possible research paths are as follows:

1. Conduct unimodal MoE experiment on each modality and observe the unique features, like the expert distribution.
2. Merge different modalities to one MoE and conduct the load balance experiment to find maybe the proper loss function.
3. Study on expert pruning or optimized routing algorithm and do profiling research for lower computing latency.

Multimodal Translation Inspired by the surprising performance MoE has achieved in multilingual neural machine translation, I think the multimodal translation also deserves further study. I suppose the transfer learning between different modalities can take place with scalable expert capacity that will help improve the overall translation quality for different modality translation pairs. The shared parameters and collaborated experts will also perform well in low-resource modality scenario. So I think in multimodal translation (especially for many modality pairs, not limited in just one modality pair), the use of sparsely-activated models can boost the development in this field.

7 Reference paper Introduction

Since there is only one related work on the MoE in the multimodal field, most of my paper survey focus on unimodal MoE and multimodal machine learning.

7.1 Intensively Reading

HIGHMMT For the HIGHMMT paper [8], it gives me insights on multitask machine learning. I learned about the details and advantages

of multitask learning and transfer learning from this paper. What's more, the general-purpose cross-modal transformer block also provides an interesting approach for modality fusion. We use the same dataset as this paper for multitask experiment. The model scalability problem raised by this paper can also be solved with sparsely-activated models which have a larger model capacity for any task.

LIMoE For the LIMoE paper [1], it gives me insight on multimodal load balance problem and the first example of the multimodal mixture of experts. The experiment on two modalities (image and text) show the unique feature of multimodal MoE, so I proposed a novel modality level entropy loss function to handle the imbalance problem.

7.2 Extensive Reading

GShard In this paper, they proposed a classical mixture of experts architecture for text processing [3]. And this approach has already been extended to other modalities, such as Vision. And the Top-2 Routing algorithm also serves as a baseline in our work.

Mixture of Experts series Due to space limitations, I won't list every article and introduce one by one. But I have done a research survey on latest mixture of experts papers, which show different challenging problems and chances we meet in both text and vision MoEs. By understanding the basic characteristics of sparsely-activated model, I get to know the computing latency problem introduced by all2all communication cost. And to reduce the impact of MoE, I try to use an expert selection technique according to each modality.

References

[1] B. Mustafa, C. Riquelme, J. Puigcerver, R. Jenatton, and N. Houlsby, "Multimodal contrastive learning with limoe: the language-image mixture of experts," *arXiv preprint arXiv:2206.02770*, 2022.

[2] S. Rajbhandari, C. Li, Z. Yao, M. Zhang, R. Y. Aminabadi, A. A. Awan, J. Rasley, and Y. He, "Deepspeed-moe: Advancing mixture-of-experts inference and training to power next-generation ai scale," *arXiv preprint arXiv:2201.05596*, 2022.

[3] D. Lepikhin, H. Lee, Y. Xu, D. Chen, O. Firat, Y. Huang, M. Krikun, N. Shazeer, and Z. Chen, "Gshard: Scaling giant models with conditional computation and automatic sharding," *arXiv preprint arXiv:2006.16668*, 2020.

[4] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proceedings of the conference. Association for Computational Linguistics. Meeting*, vol. 2019, p. 6558, NIH Public Access, 2019.

[5] J. Lu, D. Batra, D. Parikh, and S. Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," *Advances in neural information processing systems*, vol. 32, 2019.

[6] S. Elfwing, E. Uchibe, and K. Doya, "Sigmoid-weighted linear units for neural network function approximation in reinforcement learning," *Neural Networks*, vol. 107, pp. 3–11, 2018.

[7] P. P. Liang, Y. Lyu, X. Fan, Z. Wu, Y. Cheng, J. Wu, L. Y. Chen, P. Wu, M. A. Lee, Y. Zhu, *et al.*, "Multibench: Multiscale benchmarks for multimodal representation learning," in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021.

[8] P. P. Liang, Y. Lyu, X. Fan, S. Mo, D. Yogatama, L.-P. Morency, and R. Salakhutdinov, "Highmmt: Towards modality and task generalization for high-modality representation learning," *arXiv preprint arXiv:2203.01311*, 2022.