MoME: Mixture-of-Masked-Experts for Efficient Multi-Task Recommendation

Jiahui Xu*
xujh2022@shanghaitech.edu.cn
ShanghaiTech University
Shanghai, China

Lu Sun
sunlu1@shanghaitech.edu.cn
ShanghaiTech University
Shanghai, China

Dengji Zhao
zhaodj@shanghaitech.edu.cn
ShanghaiTech University
Shanghai, China

ABSTRACT

Multi-task learning techniques have attracted great attention in recommendation systems because they can meet the needs of modeling multiple perspectives simultaneously and improve recommendation performance. As promising multi-task recommendation system models, Mixture-of-Experts (MoE) and related methods use an ensemble of expert sub-networks to improve generalization and have achieved significant success in practical applications. However, they still face key challenges in efficient parameter sharing and resource utilization, especially when they are applied to real-world datasets and resource-constrained devices. In this paper, we propose a novel framework called Mixture-of-Masked-Experts (MoME) to address the challenges. Unlike MoE, expert sub-networks in MoME are extracted from an identical over-parameterized base network by learning binary masks. It utilizes a binary mask learning mechanism composed of neuron-level model masking and weight-level expert masking to achieve coarse-grained base model pruning and fine-grained expert pruning, respectively. Compared to existing MoE-based models, MoME achieves efficient parameter sharing and requires significantly less sub-network storage since it actually only trains a base network and a mixture of partially overlapped binary expert masks. Experimental results on real-world datasets demonstrate the superior performance of MoME in terms of recommendation accuracy and computational efficiency. Our code is available at https://https://github.com/Xjh0327/MoME.

CCS CONCEPTS

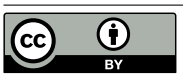
• **Computing methodologies** → **Multi-task learning**; • **Information systems** → **Recommender systems**.

KEYWORDS

Multi-Task Recommendation; Mixture-of-Experts; Binary Masks

ACM Reference Format:
Jiahui Xu, Lu Sun, and Dengji Zhao. 2024. MoME: Mixture-of-Masked-Experts for Efficient Multi-Task Recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24), July 14–18, 2024, Washington, DC, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3626772.3657922

*Corresponding author.



SIGIR '24, July 14–18, 2024, Washington, DC, USA

ACM ISBN 979-8-4007-0431-4/24/07
https://doi.org/10.1145/3626772.3657922

1 INTRODUCTION

The rapid development of the Internet and mobile applications highlights the critical importance of providing personalized information to users. For instance, e-commerce platforms need to recommend products that are more in line with user preferences to enhance users' purchasing intentions and encourage them to purchase. In recommendation systems, users often perform multiple behaviors on items, such as clicking, collecting, sharing, and purchasing. It is challenging to rely on a single metric to evaluate the user's level of interest in items and make recommendations that better match their preferences. Multi-task learning techniques have attracted great attention in recommendation systems because they can model multiple tasks simultaneously, which helps enhance recommendation performance by leveraging shared knowledge among tasks.

Current Multi-Task Recommendation System (MTRS) models mainly focus on different parameter sharing mechanisms, which can be roughly divided into two categories: hard sharing and soft sharing. Hard sharing [1, 8, 17] forces all tasks to share the same bottom network and therefore is highly efficient yet does not work well when tasks are weakly or even negatively correlated, while soft sharing [3, 14, 16] avoids negative transfer by building task-specific networks and therefore has more parameters and is less computationally efficient. To better control the trade-off between performance and efficiency, expert sharing models based on Mixture-of-Experts (MoE) [6], such as MMoE [11], PLE [19] and AESM^2 [23], have received widespread attention recently. They treat multi-layer networks as experts and use the gating network to combine the ensembles of experts. Although they provide greater flexibility than hard or soft sharing models, multiple experts with the same network structure may lead to inefficient utilization of model parameters and resources. It becomes especially severe when dealing with large-scale datasets and complex models. To solve this problem, SNR [10] learns binary coding variables to achieve sparse connections between sub-networks, while MSSM [2] uses two different types of sparse masks applied to feature fields and connections between model sub-networks to achieve feature selection and cell-level sparse sharing. However, they only sparsify the connections between sub-networks, which is coarse-grained and still requires many parameters in learning. Furthermore, their mask learning algorithms suffer from high gradient variance [22], which often leads to unstable training and low convergence rate. Dselect-k [4] selects k experts through binary encoding by a continuously differentiable and sparse gate. Compared to these methods, MoME is based on a stable gradient-based masking learning algorithm and applies fine-grained multi-level masks to extract a mixture of experts, enabling to learn more flexible model sharing structure among experts using fewer parameters.

To address these challenges, in this paper, we propose a novel and efficient MTRS method, namely **M**ixture-**o**f-**M**asked-**E**xperts (**MoME**), based on a multi-level binary mask learning mechanism consisting of coarse-grained neuron-level model masking and fine-grained weight-level expert masking. To improve computational efficiency, MoME starts from an over-parameterized base network and applies neuron-level model masking to roughly filter out unimportant neurons. Then, weight-level expert masking is employed on the masked base network to dynamically learn expert-specific binary masks. as relevant experts, allowing personalized and fine-grained sparse sharing among experts. Unlike existing MoE-based methods that train different sub-networks for different experts, MoME treats the learned binary masks as experts, enabling efficient parameter sharing with low computational overhead. To avoid combinatorial optimization and high gradient variance in mask learning, we use a probabilistic formulation based on approximate Bernoulli distribution to enforce sparsity on masks by L_0 regularization, which can be solved by scalable and stable gradient-based algorithm. Experimental results on multiple datasets show that MoME outperforms state-of-the-art models with significantly reduced number of model parameters. Fig. 1 illustrates the framework of the proposed MoME model. The contributions of this work can be summarized as follows:

- We propose a novel MTRS method, namely MoME, which learns a mixture of experts by applying different fine-grained binary masks on the same base network, enabling efficient and flexible parameter sharing across multiple tasks with low computational overhead.
- Based on approximate Bernoulli gate and probabilistic L_0 regularization, we develop a stable and scalable gradient-based optimizer for MoME.
- Comprehensive experiments on real-world datasets demonstrate that MoME has superior performance in recommendation accuracy and computational efficiency compared to cutting-edge methods.

2 MIXTURE-OF-MASKED-EXPERTS (MOME)

Existing MoE-based methods for MTRS train a separate network for each expert, resulting in computational inefficiency and high memory consumption. To improve model efficiency, some methods [2, 10, 23] enforce sparsity in model parameters by binary encoding, but the encouraged sparsity is typically coarse-grained, which poses challenges in effectively leveraging shared knowledge and flexibly exploiting task dependencies. Furthermore, the mask learning algorithms used in prior works [12, 13] suffer from the high variance issue, leading to unstable model training. To overcome these limitations, we propose a novel MTRS method named Mixture-of-Masked-Experts (MoME), which employs a single backbone network in conjunction with a set of binary expert masks, to achieve computational efficiency and low memory consumption.

2.1 Framework

MoME is constructed based on the multi-gate MoE model [11]. Given T tasks and E experts, it combines expert outputs $\{f_e(\cdot)\}_{e=1}^{E}$ on the data matrix $\mathbf{X} \in \mathbb{R}^{N \times D}$ through the gating network $g_t(\cdot)$ and uses task-specific tower network $h_t(\cdot)$ to estimate the ground truth output $\mathbf{y}_t \in \mathbb{R}^N$ of the t-th task. Thus, the basic framework can be formulated as:

$$\min_{\mathbf{W}} \frac{1}{T} \sum_{t=1}^{T} \mathcal{L}\left(h_t(f_t(\mathbf{X})), \mathbf{y}_t\right), \text{ s.t. } f_t(\mathbf{X}) = \sum_{e=1}^{E} g_{t,e}(\mathbf{X}) f_e(\mathbf{X}), \forall t, \tag{1}$$

where f_e, g_t and h_t are parameterized by $\mathbf{W}_e$, $\mathbf{W}_{g,t}$ and $\mathbf{W}_{h,t}$, respectively, g_t is the softmax function with outputs $\{g_{t,e}\}_{e=1}^{E}$ and $\mathcal{L}$ is the loss function.

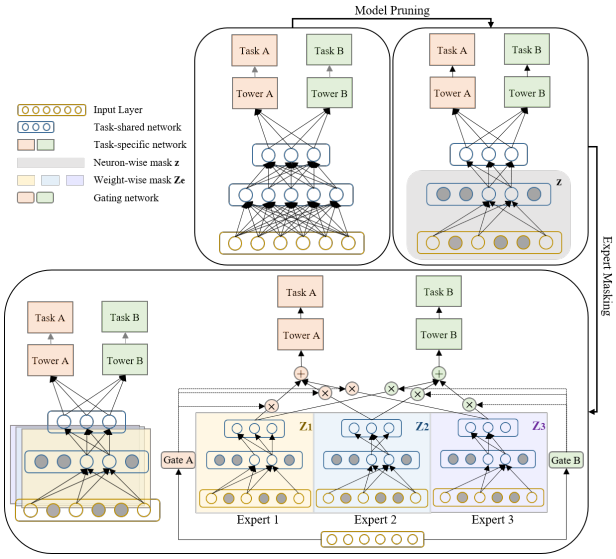


Figure 1: Illustration of the proposed MoME model. Based on an over-parameterized base neural network, a mixture of experts is learned through model pruning (z) and expert masking ($\{Z_e\}_{e=1}^{E}$). The prediction is made by an ensemble of experts weighted by task-specific gating networks.

2.2 Expert Masking and Model Pruning

Inspired by previous works [12, 13, 18] that build task-specific models by covering separate binary masks on a shared model, we propose to achieve efficient and flexible sparse parameter sharing by learning an ensemble of binary masks of the base network. Instead of using E independent sub-networks f_e's as experts in (1), MoME learns E weight-level binary masks $\{\mathbf{Z}_e\}$ acting on the same backbone network[1] parameterized by $\mathbf{W} \in \mathbb{R}^{d_i \times d_o}$, leading to $\mathbf{W}_e = \mathbf{W} \odot \mathbf{Z}_e$, and treats $f(\mathbf{X}, \mathbf{W} \odot \mathbf{Z}_e)$ as the e-th expert. Here we use $\odot$ to denote the element-wise product. These binary masks can be stored using only 1 bit per parameter; in contrast, each expert in the MoE-based models typically requires 32 bits per parameter. Therefore, it allows MoME to significantly save storage space (32×) while improving the inference speed of the model.

To further eliminate unnecessary model parameters of the over-parameterized backbone network f, coarse-grained pruning is introduced by applying neuron-level binary mask $\mathbf{z} \in \mathbb{R}^{d_i}$ on $\mathbf{W}$ before performing the weight-level expert masking, leading to

[1]For brevity, we consider a single-layer base network with d_i input neurons and d_o output neurons. It can be easily extended to handle multi-layer network.

$\mathbf{W}_e = \mathbf{W} \odot (\mathbf{z} \otimes \mathbf{1}_{d_o}) \odot \mathbf{Z}_e$, where $\otimes$ denotes the outer product and $\mathbf{1}_{d_o}$ is an all-one vector in size of d_o. Note that $\mathbf{z}$ is shared across E experts and is aimed at eliminating useless neurons for all the experts. Therefore, if z_i is zero, it will prune the weights belonging to the i-th neuron for all the E experts. In this way, neuron-level model pruning further improves computational efficiency and reduces memory consumption by removing unnecessary neurons before fine-grained weight-level masking.

Motivated by L_0-regularized sparse learning [9, 22], we propose to learn the binary masks by penalizing the L_0-norm (the number of non-zero elements in an arbitrary vector or matrix) of $\mathbf{z}$ and $\{\mathbf{Z}_e\}$, and then develop the following optimization problem:

$$\begin{aligned} &\min_{\mathbf{W},\mathbf{z},\{\mathbf{Z}_e\}} \frac{1}{T}\sum_{t=1}^{T} \mathcal{L}\left(h_t(f_t(\mathbf{X})), \mathbf{y}_t\right) + \lambda_1 \|\mathbf{z}\|_0 + \lambda_2 \sum_{e=1}^{E} \|\mathbf{Z}_e\|_0, \\ &\text{s.t. } f_t(\mathbf{X}) = \sum_{e=1}^{E} g_{t,e}(\mathbf{X}) f(\mathbf{X}, \mathbf{W}_e),\ \mathbf{W}_e = \mathbf{W} \odot (\mathbf{z} \otimes \mathbf{1}_{d_o}) \odot \mathbf{Z}_e. \end{aligned} \tag{2}$$

Compared with (1), the proposed model in (2) essentially trains a pruned base network $\mathbf{W} \odot (\mathbf{z} \otimes \mathbf{1}_{d_o})$ and a set of binary masks $\{\mathbf{Z}_e\}$, making task dependencies efficiently shared through $\mathbf{z}$ and overlapped $\mathbf{Z}_e$'s with significantly reduced memory consumption (32×) in inference.

2.3 MoME with Approximate Binary Mask

The proposed model in (2) learns a mixture of binary masks rather than a mixture of sub-networks in MoE-related methods, in order to maintain the strong generalization ability of MoE with significantly reduced computational resources. However, the L_0 regularization used in (2) is non-convex and non-differentiable, making the optimization problem NP-hard. To overcome these limitations, we introduce a probabilistic formulation by treating each entry of $\mathbf{z}$ or $\mathbf{Z}_e$ as an independent Bernoulli variable, i.e.,

$$z_i \sim \text{Bern}(\pi_i),\ \ z_{e,ij} \sim \text{Bern}(\pi_{e,ij}),\ \ \forall e, i, j, \tag{3}$$

where $\text{Bern}(\pi)$ is the Bernoulli distribution that takes the value 1 with probability π. In this sense, the L_0 regularization in (2) can be represented as:

$$\mathbb{E}\left[\|\mathbf{z}\|_0\right] = \sum_{i=1}^{d_i} \pi_i,\ \ \mathbb{E}\left[\|\mathbf{Z}_e\|_0\right] = \sum_{i=1}^{d_i}\sum_{j=1}^{d_o} \pi_{e,ij}, \tag{4}$$

where $\mathbb{E}[X]$ denotes the expectation of a random variable X. Note that the optimization of the probabilistically formulated problem is still difficult to solve since its loss function involves discrete Bernoulli random variable $\mathbf{z}$ and $\{\mathbf{Z}_e\}$. Although some optimization algorithms such as REINFORCE [20] and approximation techniques [9] can be used to solve the problem, they usually suffer from high variance [9, 15]. Inspired by [22], we introduce Gaussian based continuous relaxation to approximate Bernoulli masks $\mathbf{z}$ and $\{\mathbf{Z}_e\}$:

$$\begin{aligned} z_i &= \max(0, \min(1, \mu_i + \epsilon_i)),\ \ \epsilon_i \sim \mathcal{N}(0, \sigma^2), \\ z_{e,ij} &= \max(0, \min(1, \mu_{e,ij} + \epsilon_{e,ij})),\ \ \epsilon_{e,ij} \sim \mathcal{N}(0, \sigma_e^2), \end{aligned} \tag{5}$$

where $\mathcal{N}(\mu, \sigma^2)$ denotes the Gaussian distribution with expectation μ and standard deviation σ. In experiments, σ and σ_e are fixed as 0.5 throughout training. The approximate Bernoulli distribution for binary masking in (5) makes the gradient w.r.t. μ and $\mu_{\mathbf{e}}$ computable using standard back-propagation, and thus the L_0 regularization in (4) can be reformulated by:

$$\begin{aligned} \mathbb{E}\left[\|\mathbf{z}\|_0\right] &= \sum_{i=1}^{d_i} \mathbb{P}(z_i \geq 0) = \Phi\left(\frac{\mu_i}{\sigma}\right), \\ \mathbb{E}\left[\|\mathbf{Z}_e\|_0\right] &= \sum_{i=1}^{d_i}\sum_{j=1}^{d_o} \mathbb{P}(z_{e,ij} \geq 0) = \sum_{i=1}^{d_i}\sum_{j=1}^{d_o} \Phi\left(\frac{\mu_{e,ij}}{\sigma_e}\right), \end{aligned} \tag{6}$$

where Φ denotes the CDF of the standard Gaussian distribution. Based on (2) and (6), we have the optimization problem of **MoME**:

$$\begin{aligned} &\min_{\mathbf{W},\boldsymbol{\mu},\{\boldsymbol{\mu}_e\}} \mathbb{E}_{\mathbf{z},\{\mathbf{Z}_e\}}\left[\frac{1}{T}\sum_{t=1}^{T} \mathcal{L}\left(h_t(f_t(\mathbf{X})), \mathbf{y}_t\right)\right] + \lambda_1 \sum_{i=1}^{d_i} \Phi\left(\frac{\mu_i}{\sigma}\right) + \\ &\qquad \lambda_2 \sum_{e=1}^{E}\sum_{i=1}^{d_i}\sum_{j=1}^{d_o} \Phi\left(\frac{\mu_{e,ij}}{\sigma_e}\right), \\ &\text{s.t. } f_t(\mathbf{X}) = \sum_{e=1}^{E} g_{t,e}(\mathbf{X}) f(\mathbf{X}, \mathbf{W}_e),\ \mathbf{W}_e = \mathbf{W} \odot (\mathbf{z} \otimes \mathbf{1}_{d_o}) \odot \mathbf{Z}_e, \end{aligned} \tag{7}$$

where $\mathbf{z}$ and $\mathbf{Z}_e$ are calculated according to (5). In (7), the expectation $\mathbb{E}[\cdot]$ over $\mathbf{z}$ and $\{\mathbf{Z}_e\}$ can be approximated by Monte Carlo (MC) sampling. By introducing fine-grained sparsity through multi-level binary masks, MoME enables personalized and efficient parameter sharing across multiple tasks with a significantly reduced number of parameters in inference. Moreover, we employ a stable surrogate of probabilistic L_0-norm for mask learning, thereby mitigating the high variance problem and ensuring its robustness.

2.4 Implementation

MoME uses Kaiming initialization [5] to set up an over-parameterized backbone network and tower networks. Neuron-level model pruning on the backbone network is initialized with $\boldsymbol{\mu} = 0.5$ and applied for M epochs. After M epochs, the stochasticity part of $\mathbf{z}$ is removed by $\hat{z}_i = \max(0, \min(1, \mu_i))$ and keep $\hat{\mathbf{z}}$ unchanged throughout the following training, implying that only the pruned network parameterized by $\mathbf{W} \odot (\hat{\mathbf{z}} \otimes \mathbf{1}_{d_o})$ is adopted in expert mask learning. Besides, the weights belonging to neuron i if $\hat{z}_i = 0$ are also removed to reduce the number of trainable parameters. In expert masking, we use E masks $\{\mathbf{Z}_e\}$ and adopt the gating network g_t to flexibly combine the outputs of masked experts. After training, we abandon the stochasticity of $\mathbf{Z}_e$ through $\hat{z}_{e,ij} = \max(0, \min(1, \mu_{e,ij}))$ and remove the corresponding parameter once $\hat{z}_{e,ij} = 0, \forall e, i, j$. In these two phases, we update masks $\boldsymbol{z}$ and $\boldsymbol{Z}_e$ using gradient descent by calculating the gradient w.r.t μ_i and $\boldsymbol{\mu}_e$, respectively.

3 EXPERIMENTS

3.1 Experimental Setting

We conduct experiments on four real-world datasets to evaluate the performance of MoME. The used four datasets are extracted from the AliExpress dataset[2] collected from traffic logs on AliExpress[3] running in more than 200 countries. We choose the subsets from four countries: Spain (ES), French (FR), Netherlands (NL), and America (US). All of them contain 80 features and their number of

[2]https://tianchi.aliyun.com/dataset/74690
[3]https://www.aliexpress.com/

Table 1: Experimental results on four real-world datasets.

Model	AliExpress_NL				AliExpress_FR			
	AUC1	AUC2	Logloss	Size(MB)	AUC1	AUC2	Logloss	Size(MB)
Single Task	72.51	85.75	0.05709	10.18	72.60	87.48	0.05238	10.18
Shared-Bottom	72.39	86.01	0.05713	5.26	72.50	87.96	0.05250	**5.26**
MoE	72.53	85.77	0.05715	39.32	72.50	86.67	0.05248	39.32
MMoE	72.54	85.49	0.05710	39.39	72.35	87.89	0.05266	39.39
CGC	72.33	85.83	0.05710	39.35	72.46	87.61	0.05251	39.35
PLE	72.37	85.78	0.05726	47.90	72.62	87.56	0.05238	47.90
AITM	72.58	86.03	0.05695	11.17	72.58	87.88	0.05243	11.17
MoME	**72.66**	**86.80**	**0.05656**	**0.64**	**74.04**	**88.53**	**0.05218**	7.38
Model	**AliExpress_US**				**AliExpress_ES**			
	AUC1	AUC2	Logloss	Size(MB)	AUC1	AUC2	Logloss	Size(MB)
Single Task	70.51	86.63	0.05211	10.18	73.07	88.76	0.06238	10.18
Shared-Bottom	70.50	87.55	0.05214	5.26	73.08	88.67	**0.06234**	**5.26**
MoE	70.40	86.90	0.05228	39.32	72.97	88.93	0.06238	39.32
MMoE	70.74	86.97	0.05219	39.39	72.77	88.54	0.06258	39.39
CGC	70.63	87.16	0.05209	39.35	72.98	88.78	0.06243	39.35
PLE	70.73	87.06	**0.05201**	47.90	73.07	88.70	0.06243	47.90
AITM	70.56	87.17	0.05235	11.17	73.02	88.78	0.06239	11.17
MoME	**70.87**	**87.82**	0.05209	**0.64**	**74.35**	**89.54**	0.06355	6.91

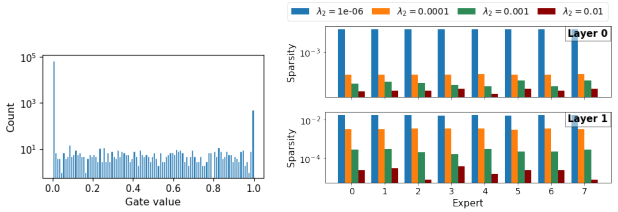
training & testing samples are: 22.3M & 9.3M (ES), 18.2M & 9.2M (FR), 12.2M & 5.6M (NL), and 20.0M & 7.4M (US), respectively. They have two prediction tasks: CTR (click-through rate) and CVR (conversion rate). In experiments, MoME is compared with seven models including **Single Task**, **Shared-Bottom** [1], **MoE** [6], **MMoE** [11], **CGC** [19], **PLE** [19], and **AITM** [21]. We conduct experiments of these comparison models using a public MTRS library[4].

We evaluate the models using average binary cross-entropy loss (Logloss) and the Area Under the Curve (AUC) score. For fair comparisons, all the models are trained for 50 epochs using Adam optimizer [7] with learning rate 0.001. We tune hyper-parameters of MoME according to $\lambda_1, \lambda_2 \in [10^{-2}, 10^{-3}, 10^{-4}, 10^{-6}]$, and set hyperparameters of comparing models as recommended by corresponding papers. The best performance in average AUC is reported in experiments. We set the number of experts as 8 for the MoE-based models and adopt MLP with two hidden layers and ReLU activation as the experts. The input embedding dimension is fixed to 128 for all models. We set the hidden layer sizes to [512, 256] for the bottom network and expert network, and set the size to [128, 64] for the tower network. Batch normalization and dropout with rate 0.2 are employed, and the batch size is set as 2,048.

3.2 Experimental Results

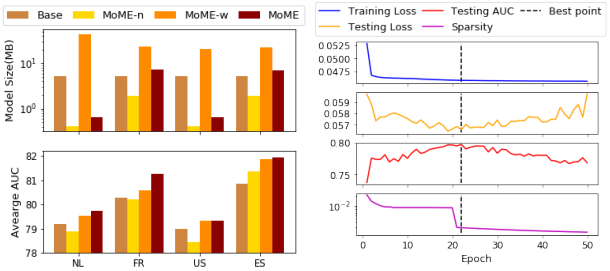
Table 1 shows the experimental results of comparison models in terms of AUC, Logloss and model size (MB) on four datasets, where the best results are highlighted in boldface. We can see that MoME achieves the best result in AUC on all the four datasets with significantly reduced model size compared to other MoE-based models and AITM, and works the best or the second best in Logloss on all datasets except ES. The performance advantage of MoME may come from its ability to mimic a mixture of experts through an ensemble of fine-grained binary masks, thereby mitigating negative transfer by enabling flexible parameter sharing at low computational overload. Fig. 2 illustrates the sparsity in expert masks $\{\mathbf{Z}_e\}$ on NL. We can see that only a small subset of mask elements are

[4]https://github.com/easezyc/Multitask-Recommendation-Library



(a) Distribution of mask elements (b) Mask sparsity across eight experts

Figure 2: Illustration of mask sparsity on AliExpress_NL.



(a) Ablation study (b) Convergence analysis on NL

Figure 3: Quantitative analysis of MoME.

in the range of $(0, 1)$ (Fig. 2(a)), and therefore the mask we use can effectively approximate binary mask. Besides, expert masks exhibit strong sparsity once a proper value of λ is selected (Fig. 2(b)). In most cases, MoE-based models slightly outperform the single-task model, but at the cost of much larger model sizes. Shared-Bottom learns smaller model sizes than MoME on the FR and ES datasets, but it performs worse than MoME in prediction accuracy.

An ablation study is conducted by comparing MoME with its three degraded variants: Base (without masking), MoME-n (neuron-level pruning), and MoME-w (weight-level masking). Fig.3(a) shows the results on four datasets in average AUC. Compared to Base, MoME-n prunes the backbone network to a large extent while still achieving comparable performance. MoME-w performs better than Base and MoME-n in AUC, revealing the efficacy of learning an ensemble of binary expert masks. By combining two types of masking, MoME consistently outperforms its variants. In addition, convergence analysis of MoME ($\lambda_1 = 0.01, \lambda_2 = 10^{-6}$) on NL is shown in Fig. 3(b). It shows that the objective can stably converge to the local minimum while the sparsity gradually increases.

4 CONCLUSION

In this paper, we propose a novel MoME model for efficient MTRS. MoME uses an ensemble of binary marks and an over-parameterized base network to improve generalization and computational efficiency. Experiments show that MoME outperforms state-of-the-art MTRS models with significantly reduced model size. For future work, it would be interesting to combine MoME with transformer and evaluate its efficacy on large-scale multi-modal datasets.

ACKNOWLEDGMENTS

This work is partially supported by Science and Technology Commission of Shanghai Municipality (No. 23010503000).

REFERENCES

[1] Rich Caruana. 1997. Multitask learning. *Machine learning* 28 (1997), 41–75.

[2] Ke Ding, Xin Dong, Yong He, Lei Cheng, Chilin Fu, Zhaoxin Huan, Hai Li, Tan Yan, Liang Zhang, Xiaolu Zhang, et al. 2021. MSSM: a multiple-level sparse sharing model for efficient multi-task learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2237–2241.

[3] Long Duong, Trevor Cohn, Steven Bird, and Paul Cook. 2015. Low resource dependency parsing: Cross-lingual parameter sharing in a neural network parser. In *Proceedings of the 53rd annual meeting of the Association for Computational Linguistics and the 7th international joint conference on natural language processing (volume 2: short papers)*. 845–850.

[4] Hussein Hazimeh, Zhe Zhao, Aakanksha Chowdhery, Maheswaran Sathiamoorthy, Yihua Chen, Rahul Mazumder, Lichan Hong, and Ed Chi. 2021. Dselect-k: Differentiable selection in the mixture of experts with applications to multi-task learning. *Advances in Neural Information Processing Systems* 34 (2021), 29335–29347.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*. 1026–1034.

[6] Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. 1991. Adaptive mixtures of local experts. *Neural computation* 3, 1 (1991), 79–87.

[7] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[8] Iasonas Kokkinos. 2017. Ubernet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6129–6138.

[9] Christos Louizos, Max Welling, and Diederik P Kingma. 2018. Learning Sparse Neural Networks through L_0 Regularization. In *International Conference on Learning Representations*.

[10] Jiaqi Ma, Zhe Zhao, Jilin Chen, Ang Li, Lichan Hong, and Ed H Chi. 2019. Snr: Sub-network routing for flexible parameter sharing in multi-task learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 216–223.

[11] Jiaqi Ma, Zhe Zhao, Xinyang Yi, Jilin Chen, Lichan Hong, and Ed H Chi. 2018. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*. 1930–1939.

[12] Arun Mallya, Dillon Davis, and Svetlana Lazebnik. 2018. Piggyback: Adapting a single network to multiple tasks by learning to mask weights. In *Proceedings of the European conference on computer vision (ECCV)*. 67–82.

[13] Arun Mallya and Svetlana Lazebnik. 2018. Packnet: Adding multiple tasks to a single network by iterative pruning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 7765–7773.

[14] Ishan Misra, Abhinav Shrivastava, Abhinav Gupta, and Martial Hebert. 2016. Cross-stitch networks for multi-task learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3994–4003.

[15] Andriy Mnih and Danilo Rezende. 2016. Variational inference for monte carlo objectives. In *International Conference on Machine Learning*. PMLR, 2188–2196.

[16] Sebastian Ruder, Joachim Bingel, Isabelle Augenstein, and Anders Søgaard. 2019. Latent multi-task architecture learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 4822–4829.

[17] Ozan Sener and Vladlen Koltun. 2018. Multi-task learning as multi-objective optimization. *Advances in neural information processing systems* 31 (2018).

[18] Tianxiang Sun, Yunfan Shao, Xiaonan Li, Pengfei Liu, Hang Yan, Xipeng Qiu, and Xuanjing Huang. 2020. Learning sparse sharing architectures for multiple tasks. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 8936–8943.

[19] Hongyan Tang, Junning Liu, Ming Zhao, and Xudong Gong. 2020. Progressive layered extraction (ple): A novel multi-task learning (mtl) model for personalized recommendations. In *Proceedings of the 14th ACM Conference on Recommender Systems*. 269–278.

[20] Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning* 8 (1992), 229–256.

[21] Dongbo Xi, Zhen Chen, Peng Yan, Yinger Zhang, Yongchun Zhu, Fuzhen Zhuang, and Yu Chen. 2021. Modeling the sequential dependence among audience multi-step conversions with multi-task learning in targeted display advertising. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 3745–3755.

[22] Yutaro Yamada, Ofir Lindenbaum, Sahand Negahban, and Yuval Kluger. 2020. Feature selection using stochastic gates. In *International Conference on Machine Learning*. PMLR, 10648–10659.

[23] Xinyu Zou, Zhi Hu, Yiming Zhao, Xuchu Ding, Zhongyi Liu, Chenliang Li, and Aixin Sun. 2022. Automatic expert selection for multi-scenario and multi-task search. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1535–1544.