

# A Comprehensive Survey of Mixture-of-Experts: Algorithms, Theory, and Applications

Siyan Mu and Sen Lin

**Abstract**—Artificial intelligence (AI) has achieved astonishing successes in many domains, especially with the recent breakthroughs in the development of foundational large models. These large models, leveraging their extensive training data, provide versatile solutions for a wide range of downstream tasks. However, as modern datasets become increasingly diverse and complex, the development of large AI models faces two major challenges: (1) the enormous consumption of computational resources and deployment difficulties, and (2) the difficulty in fitting heterogeneous and complex data, which limits the usability of the models. Mixture of Experts (MoE) models have recently attracted much attention in addressing these challenges, by dynamically selecting and activating the most relevant sub-models to process input data. It has been shown that MoEs can significantly improve model performance and efficiency with fewer resources, particularly excelling in handling large-scale, multimodal data. Given the tremendous potential MoE has demonstrated across various domains, it is urgent to provide a comprehensive summary of recent advancements of MoEs in many important fields. Existing surveys on MoE have their limitations, e.g., being outdated or lacking discussion on certain key areas, and we aim to address these gaps. In this paper, we first introduce the basic design of MoE, including gating functions, expert networks, routing mechanisms, training strategies, and system design. We then explore the algorithm design of MoE in important machine learning paradigms such as continual learning, meta-learning, multi-task learning, and reinforcement learning. Additionally, we summarize theoretical studies aimed at understanding MoE and review its applications in computer vision and natural language processing. Finally, we discuss promising future research directions.

**Index Terms**—Mixture-of-Experts, Continual Learning, Meta-Learning, Multi-Task Learning, Reinforcement Learning, Computer Vision, Natural Language Processing

## I. INTRODUCTION

Artificial Intelligence (AI) has emerged as a transformative force in modern technologies, demonstrating remarkable successes across multiple domains from computer vision [1], [2], [3] and speech recognition [4], [5] to healthcare [6] and autonomous systems [7]. A key driver behind this is the development of foundation models, such as BERT [8], CLIP [9], GPT-4 [10] which are large-scale neural networks pre-trained on massive datasets and provide versatile solutions for a wide range of downstream tasks. Building upon the pre-trained knowledge, these models can be further fine-tuned with minimal additional training to adapt to specific applications,

making them indispensable tools for handling increasingly complex and diverse tasks in many real-world scenarios.

However, as AI applications expand, modern datasets are becoming more diverse and complex. They often contain multimodal data (e.g., text, images, and audio) and exhibit intricate structures (e.g., graphs or hierarchical relationships). This diversity and complexity introduce two significant challenges: (1) the computational cost of training and deploying large models grows exponentially, making it unsustainable for many applications [11], and (2) integrating conflicting or heterogeneous knowledge within a single model becomes increasingly difficult, often leading to unstable training dynamics and suboptimal performance [12]. These challenges underscore the need for more efficient and scalable architectures capable of meeting the growing demands of modern AI tasks.

One promising approach to addressing these challenges is the Mixture of Experts (MoE) architecture, which has attracted much attention recently. Originally proposed in [13], [14], MoE adopts a “divide and conquer” strategy that fundamentally differs from traditional dense models. While conventional models activate all parameters for every input, MoE models dynamically select and activate only the most relevant subset of parameters based on the characteristics of the input data. This approach not only enhances the specialization of individual experts but also mitigates the difficulties associated with training on diverse and conflicting tasks. Furthermore, the selective activation mechanism allows MoE models to significantly expand their capacity and handle diverse knowledge domains without proportionally increasing computational costs, thereby achieving an optimal balance between performance and efficiency [15], [16]. By leveraging the strengths of specialized “experts” for different tasks or data types, MoE provides a scalable and flexible framework for tackling the challenges posed by complex, multifaceted datasets.

Notably, the MoE model architecture has demonstrated its unique advantages and tremendous potentials especially in large language models (LLMs) [171], [172], [173], [21], [15], [20], [52], [174], [175], [176], [177], [178], [179], [180], [181], [182]. For instance, the Switch Transformer [15] achieves 7 times faster in pre-training speed compared to the T5-Base [183] model, and shows improved performance across all 101 languages in a multilingual setting. It has also successfully scaled the number of parameters to the trillion level, achieving a pre-training speed 4 times faster than the T5-XXL [183] model at this scale. GLaM [20] has also expanded its parameters to the trillion level, enhancing the model’s ability to utilize contextual information. DeepSpeed-moe [52] leverages MoE and model compression techniques to reduce

The work Siyan Mu was done when she was an intern at University of Houston. She is also with the College of Science, Sichuan Agricultural University, Ya’an, Sichuan, China (e-mail: 202203863@stu.sicau.edu.cn).

Sen Lin is with the Department of Computer Science, University of Houston, Houston, TX, USA (e-mail: slin50@central.uh.edu).

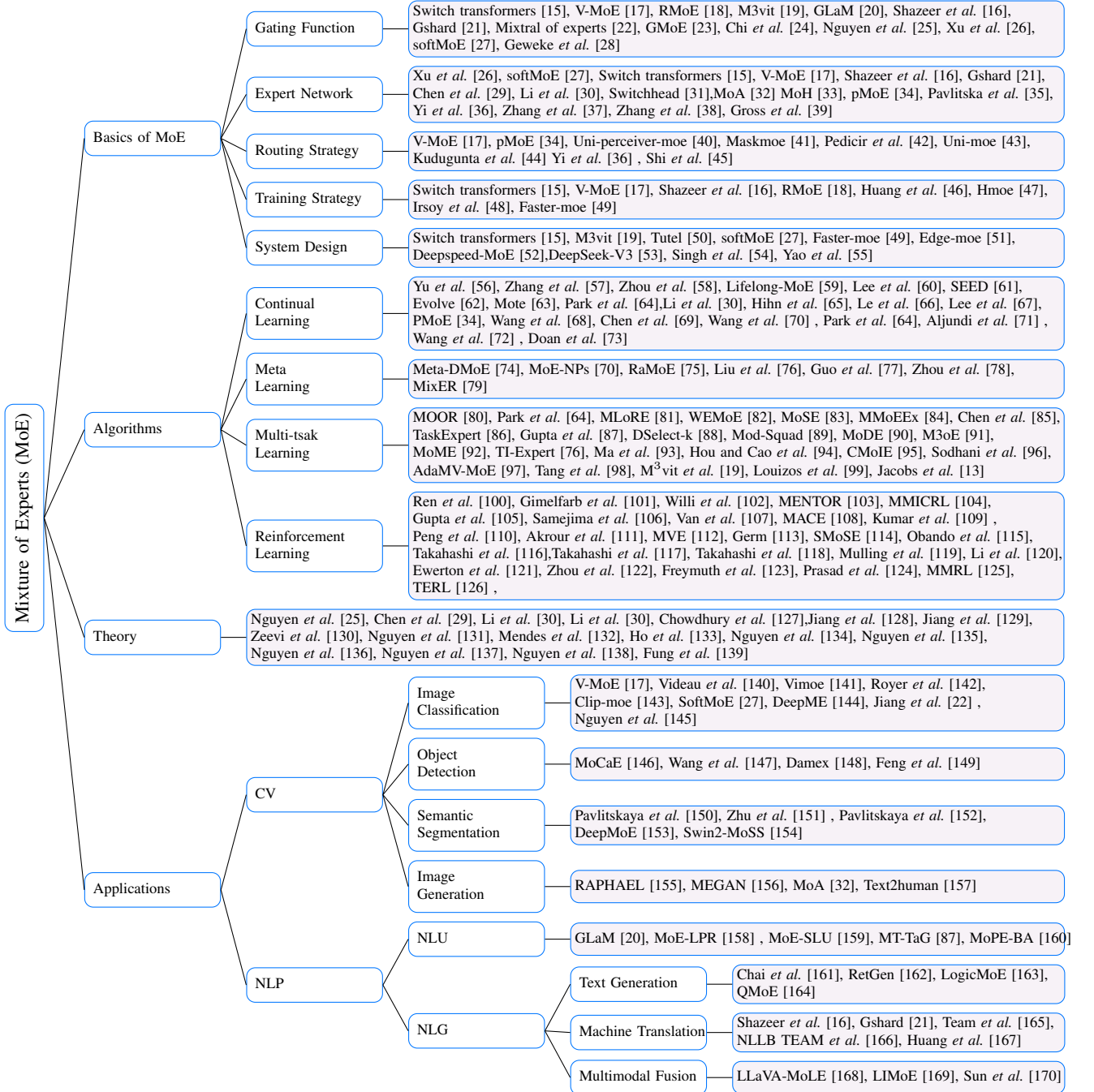


Fig. 1. The roadmap of Mixture of Experts (MoE) covered in this paper.

the size of MoE models by 3.7 times. ST-MoE [174] has further investigated the stability and transferability of MoE models, proposing various methods such as Router z-loss to ensure training stability. Ultimately, on the SuperGLUE [184] benchmark, the ST-MoE-32B model surpassed the previous state-of-the-art models. OpenMoE [178] has experimented with decoder-only MoE and provided an in-depth discussion on the routing mechanisms of MoE, making significant contributions to the open-source community's understanding of MoE architectures. Mixtral 8x7B [22], despite processing each token with only 13 billion active parameters, is able to access a total of 47 billion parameters thanks to its distinctive

architecture. This allows Mixtral 8x7B to achieve higher parameter efficiency and effectively control computational costs. The impressive performance of the DeepSeek series [185], [186], [53], [187] has also garnered significant attention. These models leverage MoE architectures to achieve state-of-the-art results on a variety of benchmarks while maintaining manageable computational requirements.

Beyond efficiency, MoE models also offer opportunities to improve model interpretability [101], [188], [35], [189]. By learning intrinsic allocation mechanisms, researchers can gain insights into how different experts specialize in handling specific types of data or tasks. This interpretability not only

enhances our understanding of the model behaviors but also opens new avenues for designing more robust and transparent AI systems.

Given the rapidly growing research interests in MoE and the huge potentials of MoE in various application domains, there is an urgent need of a comprehensive survey for summarizing and disseminating the recent advances of MoE, in order to elicit escalating attentions and inspire further research ideas in this field. Nevertheless, the early surveys of MoE [190], [191] were published ten years ago and have not incorporated the new development in the area. There is only one new survey [192] that emerged last year, which however 1) emphasized more on the basic designs of MoE and 2) did not provide a broad discussion on the use of MoEs in important machine learning paradigms and application fields such as computer vision. The theory development in MoE has also not been covered.

And although some other new concurrent surveys have been uploaded online very recently during the preparation of our paper, these surveys still either focus on a special aspect of MoE, such as applications in big data [193], inference optimization [194] and LLMs [195], or cover multiple topics very briefly [196]. To the best of our knowledge, our paper is the first survey that comprehensively summarizes the latest advancement in MoE, which includes four key components: basic design strategies of MoE, MoE-based algorithm designs in mainstream machine learning directions, theory towards understanding MoE in various scenarios, and an in-depth review of applications of MoE in both computer vision and natural language processing.

This paper will be organized as follows (as illustrated in Figure 1). In Section II, we will provide a comprehensive introduction to the basic designs of MoE, ranging from basic architecture design to training strategies to system optimization strategies. In Section III, we will introduce the recent MoE-based algorithm design in four important machine learning paradigms, including continual learning, meta-learning, multi-task learning, and reinforcement learning, which seeks to provide readers a general idea about how MoE can be leveraged to improve the algorithm design in these domains. In Section IV, we will summarize the efforts towards building the theoretical understanding of MoE, which covers a variety of aspects in MoE, such as different gating functions, different expert models, and different learning scenarios. In Section V, we will present the recent applications of MoE in two major domains, i.e., computer vision and natural language processing, where we look into multiple important subproblems in each domain. Future research directions will be discussed in Section VI, followed by the conclusion in Section VII.

## II. BASIC DESIGNS OF MOE

In this section, we will focus on the basic framework and design considerations in MoE to help readers understand the general workflow of MoE, as shown in Figure 2. More specifically, we will first introduce the key components in the design of MoE, including the gating function, the expert networks, and the routing mechanism. The first two components essentially form a basic framework of MoE, whereas

the last component characterizes how the gating function handles the input data. Next, we will introduce the strategies, i.e., loss function and pipeline design, so as to guarantee proper training of MoEs. In the end, system designs will be discussed to optimize the system efficiency of MoE from multiple perspectives.

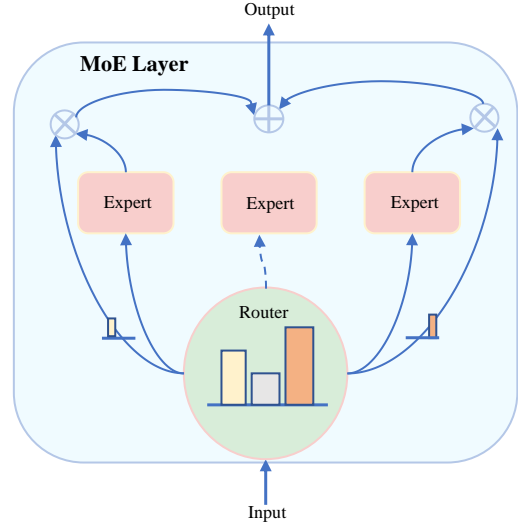


Fig. 2. The simple schematic of a standard MoE architecture.

### A. Gating Function

The gating function serves as the mathematical implementation of a router, determining how input data is allocated to designated experts. Clearly, the effectiveness of MoE models is intrinsically tied to this allocation strategy, making the design of the gating function a critical consideration. Below, we provide a detailed discussion on the principles and considerations for designing an effective gating function.

When selecting a gating function, the following criteria should be prioritized: (1) The function must accurately discern the characteristics of both input data and experts. This enables the assignment of similar data to the same expert or group of experts, ensuring that each expert receives a sufficiently large and coherent training set. Such specialization allows experts to develop expertise in specific knowledge domains. (2) The input data should be distributed as evenly as possible among the predefined experts. An uneven distribution can lead to model collapse, which reduces the efficiency of the MoE framework by underutilizing experts and impairs performance due to insufficient separation of conflicting knowledge in the training data.

**1) Linear Gating.** Based on these considerations, most existing MoE models [15], [17], [18], [19], [20] use a linear function with softmax as their gating functions due to its simplicity and effectiveness, which is also referred as softmax gating:

$$G(x)_i = \text{softmax}(\text{TopK}(g(x) + \mathcal{R}_{\text{noise}}, k))_i, \quad (1)$$

$$\text{TopK}(v) = \begin{cases} v_i & \text{if } v_i \text{ is one of the top } k \text{ elements,} \\ -\infty & \text{otherwise.} \end{cases} \quad (2)$$

Here  $G$  represents the gating function, the TopK function retains the  $k$  highest scoring inputs out of  $N$  experts and sets the rest to negative infinity,  $g$  represents the gating value calculated with a linear function prior to the softmax operation,  $x$  denotes the input,  $\mathcal{R}_{\text{noise}}$  represents the noise to encourage expert exploration,  $w$  denotes the model parameters, and  $k$  is a hyperparameter that can either be learned or manually set.

The order of the TopK and softmax operations in the above expression is flexible and can be designed based on specific requirements. One approach is to perform the TopK operation before softmax [16], [22], [20]. This method quickly filters out the most relevant experts, eliminating the need to compute softmax for all experts and thereby reducing computational overhead. However, the scores obtained after TopK may not conform to a probability distribution, necessitating an additional normalization step. Moreover, the TopK operation is inherently “hard,” potentially excluding experts with low scores that could still contribute meaningfully. Alternatively, the TopK function can be applied after softmax [21], [17], [15], [18]. In this case, the softmax function first normalizes the scores into a probability distribution, providing statistically meaningful activation weights for each expert. This approach offers more clear guidance on determining the value of  $k$  in the TopK operation. However, it requires computing softmax for all experts, resulting in higher computational costs.

When the MoE is incorporated into existing models, usually a Transformer [197], we typically replace the Feed-Forward Network (FFN) layer within a transformer block by using an MoE layer. There are several reasons [16] behind this design choice: First, the computational cost of the FFN layer typically constitutes a significant portion of the total computational cost in a Transformer, especially in deep models. By replacing the FFN layer with an MoE layer, the computational cost can be significantly reduced while maintaining the model’s expressive power. Additionally, replacing the Self-Attention layer would disrupt the core mechanism of the Transformer, which is not aligned with our objectives.

**2) Non-linear Gating.** The gating functions can also be non-linear. Specifically, a gating function design based on cosine distance is proposed in GMoE [23] for domain generalization tasks, which is shown as follows:

$$G(x) = \text{TopK} \left( \text{softmax} \left( \frac{E^T W_{\text{linear}} x}{\tau \|W_{\text{linear}} x\| \|E\|} \right) \right) \quad (3)$$

where  $x \in \mathbb{R}^{d_e}$  is the input of MoE module,  $W_{\text{linear}} \in \mathbb{R}^{d_e}$  is a learnable linear transformation for  $x$ , responsible for projecting  $x$  into a hypersphere space. The number of expert networks is  $N$ , and  $E \in \mathbb{R}^{d_e \times N}$  represents the features of the  $N$  expert networks. In the gating function, the input  $x$  is projected into a new space and compared with the expert embeddings  $E$  using cosine similarity to determine which experts each input  $x$  should be assigned to. In this way, the expert embeddings  $E$  can capture the feature representations of different experts and are gradually optimized during training to better adapt to the task requirements. Additionally, the temperature parameter

$\tau$  controls the sharpness of the gating distribution. A smaller  $\tau$  makes the distribution sharper, favoring the selection of a few experts, while a larger  $\tau$  makes the distribution smoother, allowing more experts to participate.

Comprehensive theoretical and experimental evidence has been provided in [23] to demonstrate the superiority of using cosine routers over linear routers in domain generalization tasks. Specifically, the cosine router excels at handling cross-domain data, capturing visual attributes, and enhancing the model’s generalization capability and efficiency. There are also more studies that have employed cosine routers for better performance [24], [25], which encourages us to explore more on the design of gating functions beyond the cosine routers.

In addition to the cosine gating function, there are many other designs of nonlinear functions, such as the gating function based on exponential Family of Distributions [26] as shown in Equation (4),

$$G_j(x, \nu) = \frac{\beta_j D(x|\nu_j)}{\sum_i \beta_i D(x|\nu_i)}, \quad (4)$$

Here  $\beta_j$  is the prior probability of the  $j$ th expert network, satisfying  $\sum_j \beta_j = 1$  and  $\beta_j \geq 0$ .  $D(x|\nu_j)$  is the conditional probability density function of input  $x$  under the  $j$ th expert network, belonging to the exponential family distribution. This approach analytically updates the gating function during iterations, and overcomes the nonlinear relationships introduced by softmax and the additional iterative optimization steps that follow (such as IRLS, Iteratively Reweighted Least Squares). The most common example is the Gaussian distribution.

There is also a special gating function in Soft MoE [27] that no longer employs a discrete token allocation mechanism. Specifically, Soft MoE calculates the weights between each token and each expert, then uses these weights for weighted averaging to generate input slots. Consequently, the slots processed by each expert are a weighted combination of all tokens, rather than individual tokens. This approach avoids potential token dropping issues due to load imbalances and optimization difficulties caused by discrete operations such as TopK.

Gating functions based on the Student-t distribution [198] are particularly suitable for data groups exhibiting long-tail phenomena or outliers, because they offer more robust fitting capabilities. [28] utilizes a multinomial probit model [199], [200] to determine the expert model that each observation belongs to.

## B. Expert Networks

The expert network is a core component of the MoE architecture. By dynamically selecting the most suitable expert network through the gating function, MoE efficiently allocates input data, enabling different experts to specialize in distinct knowledge domains. This design enhances the model’s overall performance, generalization ability, and computational efficiency, particularly when handling complex and diverse tasks.

In principle, each expert network in MoE can function as an independent network model, similar to a single network model (e.g., [29], [30]). However, in practice, to ensure efficiency

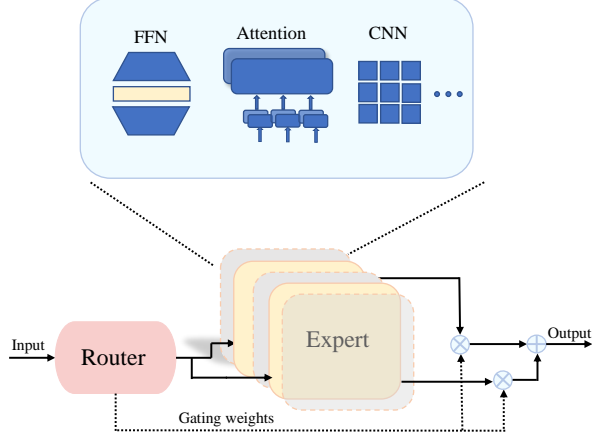


Fig. 3. Various expert networks for the MoE layer schematic.

and scalability, expert networks are often integrated into a single network model, with specific layers replaced by MoE layers [16], [15], [21]. Notably, MoE layers can be introduced at specific levels of existing neural network models without altering the overall structure, demonstrating high flexibility and broad applicability. Currently, the most widely used MoE layers include the following types (Figure 3):

1) **Replace the FFN layer in Transformer with an MoE layer.** The MoE layer can be characterized as follows:

$$\text{MoE}(x) = \sum_{i \in ID} w_i M_i(x) \quad (5)$$

where  $ID$  represents the set of indices of experts selected by the gating function,  $w_i$  is the weight assigned to the  $i$ th expert, and  $M_i(x)$  defines a two-layer FFN expert model. One key reason for introducing the MoE mechanism into the FFN layer is due to the higher sparsity and domain specificity of the FFN layer in Transformer. It has been shown [17] that the FFN layer exhibits higher sparsity compared to self-attention layers, mainly because not all units in the fully-connected FFN layer will be activated simultaneously in order to prevent overfitting. This will make the FNN layer more suitable to incorporate sparse activation mechanisms. Furthermore, [17] reveals the Emergent Modularity phenomenon in pre-trained Transformer models, indicating a significant association between neuron activations and specific tasks. This further motivates the idea of leveraging the MoE structure to reflect the modular nature of pre-trained Transformers. Therefore, due to its inherent sparsity and task relevance, the FFN layer in Transformer becomes an ideal choice for introducing the MoE mechanism.

2) **Apply MoE to the attention module in Transformer.** This will lead to a special type of MoE, namely mixture-of-attention (MoA) [32]. Specifically, MoA consists of a set of attention heads with different parameters, where each attention head can be regarded as an expert  $f_1, \dots, f_N$ . For a given input, the gating function selects the TopK attention heads based on the importance scores of the input, and the final

output is generated as a weighted sum of only these selected heads' outputs. This design allows the model to focus on the most relevant attention heads, enhancing performance without significantly increasing computational costs, and is highly scalable to support larger parameter sizes.

Compared to the standard multi-head attention mechanism [197], the key advantage of MoA lies in its use of a gating network to dynamically select the most relevant attention heads for each input token and compute a weighted sum of their outputs. This approach eliminates the requirement for all attention heads to process every input, enabling MoA to achieve superior model performance with reduced computational overhead. Consequently, MoA exhibits enhanced scalability, and particularly this model can attain higher BLEU scores than Transformer with 213 million parameters while requiring significantly less computation.

Inspired by MoA, numerous studies have been proposed to improve upon this design. SwitchHead [31] attempts to independently introduce the MoE mechanism for each head's key, query, value, and output projection, and employs a non-competitive selection activation function. This will therefore reduce the number of attention matrices that need to be computed, significantly lowering computational and memory requirements. MoH [33] enhances the standard MoE method by sharing heads and a two-stage routing strategy, and can further fine-tune pre-trained multi-head attention models into MoH models, enhancing their applicability. In addition, this work integrates the proposed MoH framework with various model frameworks including the encoder-decoder architecture, such as ViT [1], DiT [201], and decoder-only LLMs [202],

which are applied in various tasks and demonstrate great potentials of MoH.

3) **Apply MoE to CNN layers.** Integrating CNN into the MoE architecture can fully leverage CNN's strengths in local feature extraction, achieving more refined task allocation and improved computational efficiency. As a result, many studies have attempted to effectively combine MoE with CNN. For example, in [127], a simple pMoE structure is adopted, where the MoE layer consists of a gating kernel (a trainable weight network) and multiple expert networks composed of two-layer CNN networks.

Image data possesses high local correlation and spatial hierarchical structure, which aligns well with the design philosophy of CNNs. As a result, Convolutional Mixture of Experts (CMoE) is often applied to computer vision tasks. For instance, [38] utilizes the MoE framework to decompose fine-grained classification problems into subspace problems, allowing each expert CNN network to focus on different subproblems and improving classification performance. Building on [38], [37] proposes an attention-based MoE framework to enhance the model's generalization ability and classification performance under limited data conditions. [39] addresses the issue of parallel training of independent expert networks in MoE models, alleviating the need for multiple GPUs in large-scale supervised visual tasks. Meanwhile, [35] focuses on exploring how to enhance the interpretability of CNNs while maintaining model performance in computer vision tasks. Additionally, a small number of works have explored

the potential of CMoE in other domains, such as [36].

### C. Routing Strategy

A critical consideration in the design of MoE is determining the level at which experts should specialize in knowledge. For instance, in computer vision, certain tasks may emphasize more on the global features, while others demand precise local perception. Consequently, selecting an appropriate routing level—defined by the frequency of routing decisions—is essential to ensure the MoE model effectively meets the requirements of specific tasks. The following discussion extends the classification strategy proposed in Uni-Perceiver-MoE [40] and outlines several common routing strategies: token-level routing, modality-level routing, task-level routing, and other-level routing.

**1) Token-Level Routing Strategy** determines routing decisions based on token representations, making it the most classic routing approach. Tokens also come in various types, among which the most common ones we encounter are text tokens and image tokens (patch tokens).

In terms of text tokens, to mitigate underfitting and preserve representation diversity, MaskMoE [41] introduces a novel token-level routing strategy. This approach generates a mask vector for each token in the vocabulary and adjusts the number of visible experts for tokens of varying frequencies before training. For infrequent tokens, MaskMoE restricts routing to a single expert via routing masks, ensuring consistent training for these tokens. For frequent tokens, it permits multiple visible experts to capture contextual nuances. Additionally, [42] proposes a token-level recurring routing strategy that employs recurrent neural networks to dynamically adjust expert selection based on input sequence context. This strategy improves the model’s accuracy in processing complex linguistic structures, optimizes computational resource utilization, and effectively addresses long-range dependencies. Consequently, it achieves enhanced efficiency and performance across diverse natural language processing tasks.

The usage of patch tokens, as discussed in [127], is often built upon the ViT architecture. This strategy divides an image into multiple patches and dynamically routes each patch to the most suitable expert for processing. By adaptively selecting experts based on the local features of the image, it optimizes computational resource utilization and enhances the model’s performance in handling complex visual tasks. Additionally, V-MoE [17] introduces a Batch Priority Routing (BPR) strategy, which 1) calculates the importance of each image patch, 2) sorts them in descending order of importance, and 3) assigns experts to patches accordingly. Experimental results demonstrate that BPR can maintain performance comparable to a dense model while processing only 15%-30% of the image patches, showcasing significant advantages by standard Routing.

Apart from these two common types, there are also audio tokens [36], time series token [45], etc., all of which can be freely combined with MoE to help the model achieve better learning performance.

**2) Modality-Level Routing Strategy** routes tokens based on their modality, emulating the practice of employing special-

ized encoders for distinct modalities to minimize interference between them. The Uni-MoE model [43] is designed to effectively process and understand multi-modal data through a structured approach. First, it introduces connectors trained using cross-modal data to achieve data alignment, mapping diverse modalities into a unified language representation space. Simultaneously, it trains modality-specific expert systems on cross-modal instruction datasets, so as to enhance each expert’s specialized capabilities and activate their task-specific preferences. Subsequently, the model utilizes LoRA [203]

for fine-tuning, optimizing performance with mixed multi-modal instruction data while maintaining computational efficiency. Following instruction tuning, Uni-MoE is evaluated on a comprehensive and diverse multi-modal dataset. The results indicate that the model significantly reduces performance discrepancies when processing mixed data, improves collaboration efficiency among multiple experts, and enhances generalization capabilities, highlighting the potential of the MoE architecture in developing a unified multi-modal large language model.

**3) Task-Level Routing Strategy** determines routing based on task IDs, ensuring that different tasks are routed to distinct experts, thereby effectively minimizing interference between tasks. This approach is particularly advantageous for multi-task learning. For instance, [44] explores multilingual machine translation tasks, where task boundaries are defined by target languages, language pairs, or unique task IDs. And the experts are selected at the task level according to these boundaries. Compared to token-level routing strategies, this approach only requires loading a subset of experts relevant to the current task during inference, rather than loading all experts of the entire model. This results in reduced communication costs between devices and lower memory usage.

**4) Other-Level Routing Strategies** encompass a variety of approaches, such as context-level and attribute-level routing strategies [40]. The context-level routing strategy employs global pooling to provide global context information to the router, enabling it to generate reliable routing strategies based on the contextual information of the current token representation. In contrast, attribute-level routing incorporates an 8-dimensional binary embedding, which includes attributes such as the modalities of the current task and token (indices 0-5), the causation type of the model (index 6), and the token source (index 7). These rich information enables models to handle more complex and different tasks.

### D. Training Strategies

Due to the sparse activations inherent in MoE-based architectures, directly applying the functions and procedures used for training dense models to MoE-based sparse models is not straightforward. To improve training effectiveness, achieve faster convergence, and enhance performance on target tasks, numerous studies like [15], [17], [16], [18], [46]

have explored the methods for training these sparse models. In the following section, we summarize and discuss some representative techniques.

**1) Auxiliary Loss Function Design.** In addition to standard training losses, auxiliary loss functions are often introduced



during MoE training to enhance learning performance, improve model efficiency, and stabilize the training process. One type of auxiliary losses widely used in practice is the load balancing loss, which aims to balance the allocation of inputs across all experts to fully utilize the capacity of the MoE.

A well-known challenge in MoE training is the risk of model collapse, where a small subset of experts receives the majority of inputs, while others receive almost none, leading to an imbalance in expert utilization. To mitigate this issue, [16] proposes Importance and Load loss functions, which are defined as follows:

$$\text{Load}(X)_i = \sum_{x \in X} P(x, i) \quad (6)$$

$$\text{Importance}(X)_i = \sum_{x \in X} G(x)_i \quad (7)$$

$$L_{\text{load}}(X) = w_{\text{load}} \cdot (\text{CoV}(\text{Load}(X)))^2 \quad (8)$$

$$L_{\text{importance}}(X) = w_{\text{importance}} \cdot (\text{CoV}(\text{Importance}(X)))^2 \quad (9)$$

where  $w_{\text{importance}}$  and  $w_{\text{load}}$  are hyperparameters controlling the influence of expert importance balance and load balancing, respectively, and  $P(x, i)$  denotes the probability that  $G(x)_i$  is non-zero given new random noise selections. Here, CoV stands for the coefficient of variation. These two loss functions are incorporated into the overall model loss function to promote balanced usage and load distribution among experts, thereby enhancing model performance and stability.

The Switch Transformer [15] (as shown in Equation (12)) simplifies the approach above based on [16], trading a slight reduction in accuracy for improved efficiency:

$$f_i = \frac{1}{T} \sum_{x \in \mathcal{B}} \mathbb{1}\{\arg\max G(x) = i\} \quad (10)$$

$$Q_i = \frac{1}{T} \sum_{x \in \mathcal{B}} G(x)_i, \quad (11)$$

$$\text{loss} = \alpha \cdot N \cdot \sum_{i=1}^N f_i \cdot Q_i. \quad (12)$$

Here  $T$  is the total number of tokens in the batch,  $\mathcal{B}$  is the set of tokens in the batch,  $f_i$  is the proportion of tokens assigned to expert  $i$ ,  $Q_i$  is the routing probability ratio assigned to expert  $i$ , and  $\alpha$  is a hyperparameter used to control the weight of the auxiliary loss. Ideally, all experts should receive tasks evenly, with each processing an equal amount of input data, such that the expected values of  $f$  and  $Q$  are both  $1/N$ . When the distribution is uniform, this auxiliary loss function reaches its minimum. Expert capacity [16] refers to the maximum number of data points each expert can process. Insufficient expert capacity may result in skipped input data during training, leading to under-trained experts; conversely, increasing capacity reduces dropped tokens but may also waste computational and memory resources. V-MoE [17] normalizes  $L_{\text{Importance}}$  and  $L_{\text{Load}}$  when applying the auxiliary loss function to visual tasks, achieving a more stable training process.

**2) Expert Selection.** It is clear that the performance and efficiency of MoE models heavily depend on the expert selection for data inputs. While the gating function primarily determines expert selection, additional design choices can further enhance MoE models. The TopK strategy, which selects the best  $K$  experts among all experts to process the data, significantly increases model capacity without proportionally escalating computational costs. For example, [16] has shown that the MoE model can scale to over 1000 times of the size of traditional neural networks while maintaining computational efficiency. On the other hand, the TopK strategy can be removed in order to further enhance the efficiency. For instance, V-MoE [17] replaces TopK with a Top-1 strategy to simplify computation and reduce resource wastage. Moreover, randomness is usually introduced for expert selection inside the TopK function (as shown in Equation (1)), by adding Gaussian noise to expert activations during the selection process [16]. This mechanism encourages expert exploration with random selections and prevents over-reliance on specific experts, which mitigates load imbalance and optimizes resource utilization.

Dynamic expert selection mechanisms have also been explored in the literature. Notably, [46] proposes a Top-P routing strategy, which determines the number of activated experts at each time by setting a probability threshold  $P$ . For each input data, the gating network calculates a score for each expert, corresponding to a probability distribution of an expert being selected, and the experts are sorted in a descending order of their scores. The strategy then progressively accumulates the scores of the experts until the cumulative value exceeds  $P$ , activating all accumulated experts to process the token. This allows the model to adaptively adjust the number of selected experts based on input complexity, enabling flexible and efficient resource allocation. To reduce overfitting and improving generalization, [48] proposes a dropout regularization method tailored for the tree structure of Hierarchical Mixture of Experts (HMoE) [47], which randomly drops expert networks of different branches according to the hierarchical structure. Experimental results demonstrate that this method enhances model performance, providing smoother fitting outcomes, better generalization, and lower validation error rates compared to models without dropout.

**3) Pipeline Design.** Given the sparsity nature and need of dynamic data allocation in MoE architectures, a well-designed pipeline plays a crucial role in the training of MoE models, in order to maximize the utilization of computational resources, reduce time and costs, and ensure stable convergence of the model. In general, the pipeline design for MoE seeks to optimize resource allocation and efficiently distribute data among experts.

As mentioned before, BPR [17] achieves sparsity reuse by prioritizing important samples for processing. The advantages of BPR lie in its efficient resource utilization and reduction of unnecessary computations.

RMoE [18] designs a pipeline particularly for downstream tasks such as segmentation and detection. More specifically, RMoE decomposes the weights of MoE experts into core and residual weights, where the core weights inherit knowledge from pre-trained non-MoE models and only the residual

weights will be finetuned. This will significantly shorten the training time and reduce the training costs. During inference, both sets of weights will be combined to ensure the good performance of the MoE model.

### E. System Design

The MoE architecture is gaining popularity in the edge deployment and capacity scaling of large models [15], [51], [52],

offering significant flexibility and scalability. However, its nature of inherent non-uniform workload distribution and dynamic expert selection imposes strict requirements on the system's resource allocation, communication efficiency, and overall performance optimization. This will introduce new system challenges such as increased synchronization overhead, hampering the efficiency and stability of model execution. In this section, we outline some common methods currently available to enhance the system designs for MoE from three dimensions: computation, communication, and memory.

**1) Computation.** In the application of MoE models, due to their inherent dynamic characteristics, there may be significant disparities in the number of tasks processed by different experts, making the issue of non-uniform workload distribution particularly pronounced. Particularly, in large-scale distributed system environments, standard MoE architectures require synchronization among expert operations [49], i.e., all experts must complete their current tasks before moving on to the next computation step. However, some experts may finish work earlier due to smaller assigned task volumes, leading to idle timeslots and wasted computational resources. Moreover, as the MoE model capacity expands, computational complexity significantly increases [15], intensifying the demand for better computation strategy design.

Various attempts have been made in the literature to mitigate this problem. The paper [15] proposes and discusses several parallel strategies, including data parallelism, expert parallelism, and model parallelism, but different parallel methods have varying requirements for the distribution of data and model parameters. To enable the system to quickly adapt to dynamic load changes, [50] designs a unified distribution layout to manage the model's parameters and input data. This can accommodate the distribution requirements of various parallel strategies, achieving zero-cost switching of parallel strategies. Based on this distribution, [50] also further develops multiple optimization techniques, such as the ALL-to-ALL optimization algorithm mentioned later. Moreover, Soft MoE [27] introduces an improved scheduling algorithm that uses gradient-based routing selection methods, making the computation process smoother and more efficient. This approach reduces unnecessary synchronization wait times, lowers overall synchronization overhead, and boosts system throughput and efficiency. DeepSeek-V3 [53] introduces several enhancements to the communication process. Firstly, it concurrently processes two micro-batches with comparable computational loads, overlapping the attention mechanism and MoE component of one micro-batch with the dispatch and combine operations of another. This technique not only

boosts throughput but also masks the overhead associated with all-to-all and tensor parallel communications, thereby substantially diminishing reliance on communication bandwidth. Furthermore, an FP8 format is employed for low-precision communication; activations are quantized to FP8 prior to up-projection in MoE using integral power-of-two scaling factors to mitigate quantization errors, a method similarly applied to activation gradients before down-projection in MoE. Lastly, to alleviate all-to-all communication bottlenecks, particularly among GPUs, a more efficient processing strategy has been devised. The specific implementation is available as open-source on GitHub at <https://github.com/deepseek-ai/DeepEP>.

**2) Communication.** In the training of MoE models, input data must be distributed to the experts via an All-to-All communication pattern, which usually results in a significant demand on the communication bandwidth. On the other hand, due to the dynamic nature of expert selection, the amount of communication in the system can substantially fluctuate, such that the communication demands between iterations can also be very different. Clearly, this communication pattern is difficult to predict and optimize, thereby increasing the risk of network congestion.

Considering that only a portion of MoE models are activated at any given time, on-demand loading strategies [19] can be adopted when scheduling the model, i.e., weights are loaded into the on-chip memory only when the corresponding expert is selected. This method reduces frequent access to external DRAM, effectively alleviating the pressure on memory bandwidth. Moreover, different experts within the MoE can be assigned to distinct compute units or cores, achieving task-level parallel processing [15]. This strategy not only increases the system throughput but also fully exploits the parallelism provided by Multi-Chip Modules (MCMs). In addition to expert-level parallelism, other studies have proposed various parallelization strategies such as data parallelism, model parallelism or tensor parallelism [15], allowing more flexible design and the use of different parallel methods according to task characteristics [50], [54], [55]. To further optimize performance, systems such as TUTEL [50] and FasterMOE [49] simultaneously run the communication operations with other computational tasks to hide communication latency and reduce the impact of communication on overall runtime. Moreover, hardware features have also been utilized to motivate various communication patterns to reduce the need of bandwidth, e.g., EdgeMOE [51] and M3ViT [19].

**3) Memory.** The MoE model typically contains a large number of parameters due to the usage of multiple experts. This will need a large amount of memory, which sometimes may even exceed the storage capacity of a single device, thereby limiting the scalability of the MoE model. Moreover, due to the randomness and dynamics in expert selection, the memory access patterns of MoE models can become extremely complex.

To address these issues, Switch-Transformer [15] proposes a parameter migration strategy that leverage the memory space of multiple devices in a collective manner, to store a large volume of parameters through parallel communication mechanisms. This method effectively distributes the pressure



of parameter storage, enhancing the system’s overall storage capacity and computational efficiency. DeepSpeed-MoE [52] adopts a hierarchical storage management scheme, integrating local high-speed caches with remote slower storage to construct a multi-level storage architecture. This design ensures that frequently used parameters can be accessed quickly while also accommodating less frequently used but necessary parameters, thus achieving more efficient memory utilization without compromising performance.

### III. ALGORITHMS

Due to the effectiveness and the strength in handling diverse knowledge in data, there have been rapidly growing interests in applying the MoE architecture into various machine learning paradigms, leading to improved algorithm designs. To showcase this and the great potentials of MoE in solving more complex learning scenarios, we will delve into the MoE-based algorithm designs for four widely investigated domains, i.e., continual learning, meta-learning, multi-task learning, and reinforcement learning.

#### A. Continual Learning

Continual learning [204] seeks to build an agent that can continuously learn a sequence of different tasks, corresponding to a non-stationary environment with data distribution shifts, to mimic the extraordinary lifelong learning capability of human beings. In particular, in continual learning a model is continuously adapted based on the new data without access to previous data. This will lead to a key challenge in continual learning, the so-called “Catastrophic Forgetting” [205], which refers to the phenomenon that neural networks can easily forget the previously acquired knowledge when learning the new task. Seeing that the MoE framework dynamically selects the most appropriate sub-model for each input, enabling the model to handle diverse tasks without a substantial increase in computational overhead, multiple efforts have been made in the literature to leverage MoE to facilitate better continual learning algorithm designs.

The idea of applying MoE in continual learning is first explored [60] which builds a model with a set of experts for task-free continual learning [206], i.e., a more general continual learning setup where the explicit task definition and task boundary information are unavailable. In particular, by following the expansion-based continual learning approaches, a Continual Neural Dirichlet Process Mixture (CN-DPM) model is introduced, where each expert model in CN-DPM handles a subset of data and the number of expert models dynamically expands based on the Bayesian nonparametric framework. However, the gating mechanism to determine the expert selection for current input data in [60] relies on a generative model for each expert. [65] improves over the design by showing that a gating mechanism based only on the input data is feasible. In particular, [65] considers a hierarchical variational continual learning (HVCL) setup with multiple priors, and shows that leveraging multiple experts can alleviate problems in HVCL such as known task boundaries and increased computation cost. To this end, a standard MoE

model is built based on [16] where two different strategies are proposed to encourage expert diversity, which, as the authors believe, may mitigate forgetting since experts specialize in different tasks.

Recently, pre-trained model based continual learning [207] has attracted much attention by leveraging the powerful pre-trained models to facilitate better continual learning. This further spurs a new category of continual learning approaches, i.e., the prompt-based continual learning approaches [70], which have demonstrated superior performance compared to standard continual learning approaches. Thus motivated, there have emerged several attempts in leveraging MoE to design better prompt-based approaches for continual learning with pretrained models. Specifically, by shedding lights on the connection between self-attention in Vision Transformer [1] and a mixture of experts, [66] interprets applying prefix tuning in pre-trained models as adding new experts, which can be leveraged to facilitate efficient model adaptation for new tasks. Based on this, a novel gating mechanism named Non-linear Residual Gates (NoRGa) is proposed to improve the parameter efficiency in prefix tuning, where non-linear activation functions and residual connections are introduced within the gating function. [63] develops a mixture of task-specific experts (MoTE) framework, where a task-specific adapter (i.e., expert) is trained for each new task. To enhance the feature robustness and reduce cross-task feature overlap, an expert evaluation strategy and an Expert Peer-Voting Mechanism are further introduced, where the most trustworthy experts will be selected for feature fusion at inference.

Considering the recent breakthroughs in large language models, several studies have explored the usage of MoE in the continual adaptation of language models. [59] investigates the continual adaptation of pretrained language models, where an MoE-based architecture, i.e., Lifelong-MoE, is proposed to address the forgetting problem. In particular, new experts will be introduced for new data distributions, while regularization-based strategies on both experts and gating functions are used to retain the knowledge for old tasks. A new architecture PMoE is proposed in [34] to address the forgetting for continual learning in large language models. By using shallow layers for general knowledge and deep layers for new knowledge, new experts are progressively added to deep layers, where a router efficiently allocates new knowledge to appropriate experts based on deep features.

The application of MoE in continual adaptation based on vision-language models has also been studied very recently. [56] introduces the parameter-efficient MoE-Adapters framework for continual adaptation with the CLIP model [9]. This framework leverages Adapters as experts and employs a task-specific router to select the most suitable experts for the current task, thereby accelerating adaptation and fostering expert collaboration. Additionally, the authors propose a Distribution Discriminative Auto-Selector (DDAS) to enhance the model’s zero-shot recognition capabilities, which automatically allocates test data to either the MoE-Adapters or the original CLIP model, ensuring accurate predictions for both seen and unseen data. [64] observes that the usage of a shared layer in MoE-based models can lead to unsatisfying performance for

continual learning. To address this, an expert merging strategy is proposed to merge the two most frequently selected experts based on the tracking of expert usage. The merged experts will be used to update the least frequently selected expert. This helps to prevent the same feature being learned as different features by multiple experts. Additionally, the least important parts of the model are regularly removed, saving computational resources.

Moreover, the MoE-based models have been leveraged in various specific applications of continual learning. For example, [57] studies the problem of continual test time adaptation and introduces the Mixture-of-Activation-Sparsity-Experts (MoASE) adapter to address the forgetting problem therein. By decomposing the neuron activation into high and low activation components, MoASE develops a Domain-Aware Gate to utilize domain information for expert combination and an Activation Sparsity Gate for more accurate feature decomposition. [67] studies the continual traffic forecasting problem and introduces a novel MoE-based model, i.e., TFMoE, to efficiently retain past knowledge. The traffic flow is divided into homogeneous groups, where each group is handled by an expert model. [69] applies MoE to continual medical image segmentation, updating only relevant experts for new data while fixing other network parameters. [30] exploits MoE in mobile edge computing (MEC) networks, dynamically selecting experts for task offloading while optimizing gating network parameters to minimize computational delays.

There have also been some studies in the literature that leverage multiple expert models, instead of a single model, for continual learning. For example, [71] proposes a framework where an expert model is trained for each task and a gating mechanism to select the relevant expert at test time. [72] uses a fixed number of small networks to learn all tasks in continual learning in parallel and shows that the performance can be better than a single big model. [73] investigates the impact of different ensemble models on the performance of continual learning. The SEED method proposed in [61] optimizes the model ensemble training by selecting a single expert for new task learning based on Gaussian distribution similarity, ensuring effective learning of new tasks while preserving prior knowledge. EVOLVE [62] leverages multiple pre-trained models as experts to enhance self-supervised continual learning on local clients. In particular, EVOLVE employs an expert aggregation loss to distill guidance information and dynamically adjusts expert weights based on new task data. While these studies do not exactly leverage the MoE architecture, they still demonstrate the great potentials of leveraging multiple experts to facilitate better continual learning.

## B. Meta-Learning

The core of meta-learning [208], [209], [210], [211], [212] lies in learning to learn, focusing on enabling models to quickly learn new tasks from a small amount of data, thereby improving the generalization and adaptability of machine learning systems. MoE can be used to enhance the rapid learning capability of meta-learning methods because it leverages multiple expert models to learn the distinct core features of

each domain. By capturing the differences and connections between various tasks in a better way, MoE has great potentials in promoting more efficient learning and adaptation of new tasks by the model. In what follows, we will go over the recent development in the field of meta-learning by leveraging MoE models.

While the most popular meta-learning approach in the past decade is the optimization-based approach [213], [214] due to its simplicity and efficiency, neural processes (NPs) based approaches [215] recently gain increasing attention for meta-learning, which directly learn meta-representations and do not require additional gradient updates during fast adaptation. However, the standard NPs consider use a single global Gaussian latent variable to model tasks, leading to a lack of expressive power when dealing with complex tasks generated from a mixture of stochastic processes. To alleviate this issue, [70] proposes the Mixture of Expert Neural Processes (MoE-NPs), with multiple latent variables and discrete assignment latent variables. The multiple latent variables are used to construct a mixture of NPs experts to define various functional priors, and the discrete assignment variable controls the expert selection for each data point in prediction. An evidence lower bound is derived to optimize the proposed model, which demonstrates promising performance in both few-shot supervised learning and meta-reinforcement learning.

As meta-learning has demonstrated the superior capability in fast adaptation for novel scenarios, it has been recently applied to generalizable dynamical system reconstruction (DSR) [216] for learning across varying environments with a small amount of data, which is particular important for scientific discovery using scientific data. While promising, meta-learning typically relies on the assumption that all tasks (or environments) follow the same task distribution, limiting its applications in scenarios with less similar environments. To address this, [79] proposes the Mixture of Experts Reconstructors (MixER) by augmenting existing contextual meta-learning approaches [217] with the Top-1 MoE architecture. In particular, MixER consists of multiple experts, each corresponding to a meta-model to capture the shared knowledge, and a linear gating function which takes the context information, instead of state vectors, as input. An unsupervised routing mechanism is further designed to optimize the MoE-based models.

Besides these studies, MoE has also been leveraged together with meta-learning for various applications. Domain adaptation (DA) [218] is a subfield of transfer learning which seeks to adapt the model trained for the source domain to perform well on a related but different target domain. The standard formation of DA considers the transfer from a single domain, whereas multiple sources can be available in practice. However, most existing methods use a single model to learn all source tasks. To address this, [77] leverages MoE models to explicitly characterize the relationships between different source domains and the target domain, where the prediction for a target example will be based on a weighted combination of all experts' predictions. Here the weights capture the similarity of the target example with each source domain, which are learned through metric-based meta-learning approaches [219] by constructing a meta-training procedure with multiple multi-

source-single-target pairs. [74] investigates a slightly different problem where some unlabeled data from the target domain will be exploited, and propose Meta-Distillation of MoE where a student model is built from each multi-source-single-target pair. Here, a MoE model is used as a teacher student, and each expert is separately trained using supervised learning on each source domain to extract its discriminative features. For domain adaptation, the unlabeled data from the target domain will be passed through all experts and the knowledge from all experts will be aggregated by an aggregator to guide the distillation from the teacher to the student. The student model and the aggregator will be jointly trained using optimization based meta-learning.

Other studies, which also leverage MoE together with meta-learning, investigate some specific applications of domain adaptation. For instance, [75] studies the domain generalizable person re-identification task (DG ReID), where each expert in the MoE model learns the knowledge from one source domain and all the features from these experts are integrated based on an adaptive voting process for the unseen target domain. As the specific target domain is generally unknown during training, an optimization-based meta-learning approach is proposed to train the voting network, which is jointly trained with the expert networks given their couplings in the algorithm design. [78] shares the same design philosophy as [75] but studies the domain generalization of face anti-spoofing, where the domain expert models are meta-trained using optimization-based methods. [76] investigates the generalization of multi-access control in heterogeneous wireless networks. Traditional meta-reinforcement learning (meta-RL) methods may not fully capture the subtle differences between different tasks, leading to insufficient adaptability in new environments [220]. When designing a General Multiple Access (GMA) protocol, the challenges posed by varying network configurations in different testing environments need to be considered. [76] thus introduces the MoE model to enhance the model's representational capacity and reduce overfitting to specific tasks [190]. Specifically, an MoE-enhanced encoder network is introduced in designing the GMA protocol, where each expert model in the linear gating MoE inside the encoder independently encodes each state transition into latent representations. These latent representations are then combined to generate a final comprehensive task representation, which will be leveraged by SAC [221] for individual protocol learning within meta-RL of the generalizable GMA protocol learning.

### C. Multi-task Learning

Multi-task learning (MTL) [222], [223], [224] is a machine learning paradigm whose core idea is to simultaneously learn multiple related tasks to improve the generalization ability and performance of the model. Essentially, it provides the model with rich prior knowledge to enhance its performance. This determines that multi-task learning algorithms need to leverage shared information among tasks, typically achieved by sharing part of the model's parameters or representations. However, this approach has exposed some shortcomings when dealing with complex data [93]. The MoE model, on the

other hand, can intelligently select the most suitable sub-model for processing based on the characteristics of the input data, naturally decoupling different tasks and effectively handling the complex relationships between them. This gives MoE a greater advantage when dealing with complex tasks [87], and as a result, it has gradually gained more attention and been applied to many important sub-tasks in multi-task learning.

To more efficiently handle the relationships among multiple tasks in multi-task learning, [93] proposes an important variant of MoE called Multi-gate Mixture-of-Experts (MMoE), which sets up a separate gating network for each task. The gating network can further select different expert networks based on the specific characteristics of the current task, decoupling each task into different gate functions. This avoids the mutual interference of tasks with large differences, and alleviates the significant performance degradation of traditional methods when processing multi-task data with large differences due to shared underlying networks. [95] finds that MMoE still cannot fully resolve the issue of Negative Transfer, and in some tasks, its performance is even worse than that of single-task models [95]. As a result, they design several optimization strategies to further mitigate the Negative Transfer problem in MMoE by penalizing overly similar experts, enhancing the fine-grained interaction among experts, and balancing the importance of tasks with varying amounts of data during model training.

Multi-task reinforcement learning is an important branch of multi-task learning [225]. It aims to promote the learning of diverse skills by agents to enhance their ability to handle complex tasks. However, existing methods lack guarantees on the diversity of learned representations, which may lead to representation collapse into similar representations, thus limiting their generalization ability [96], [226], [96], [227], [228], [229]. To address this issue, [80] introduces a method called Mixture of Orthogonal Experts (MOOR), which utilizes multiple expert models responsible for generating different representations. In this case, each expert is responsible for different aspects of the task, encouraging the model to learn more diverse skills. To ensure efficient differentiations among experts, MOOR introduces the Gram-Schmidt orthogonalization method into MoE to force each expert to generate mutually orthogonal features, further limiting the possibility of experts generating redundant representations. [96] investigates contextual multi-task RL and proposed a novel approach to encode an input observation into multiple representations using a mixture of encoders. These representations correspond to different skills or objects, and can be selected for any given task based on the context information. This approach gives the learning agent a more fine-grained control on the information shared across tasks and alleviates negative interference. Instead of naively sharing parameters across all tasks in traditional approaches, [85] proposes an attention-based MoE approach for multi-task reinforcement learning, which seeks to learn a compositional policy for each task. In particular, each expert network learns task-specific skills and specializes in different parts of the multi-task representation space. An attention module was introduced to integrate the expert outputs, we propose an attention module to generate connections between tasks and experts to achieve the best performance automati-

cally.

Multi-task learning techniques have also garnered significant attention in recommendation systems because they can simultaneously meet the modeling needs of multiple perspectives, improving recommendation performance. However, many existing multi-task recommendation systems struggle to balance parameter sharing and resource utilization. Some MoE-based methods, while more flexible, also face issues such as unstable training or resource waste [93], [13], [99]. [92] improves the standard MoE framework by proposing a new framework called Mixture-of-Masked-Experts (MoME). First, MoME no longer trains independent sub-networks for each expert but extracts expert sub-networks from an over-parameterized base network, generating diverse expert networks by learning different binary masks. Thus, during training, only one base network and a set of binary masks need to be trained, effectively saving resources. Additionally, MoME designs multi-level masks: neuron-level masks are used to filter out unimportant neurons in the base network, while weight-level masks are used to generate diverse expert networks. To further enhance the robustness of the mask learning process, MoME uses a probability formula based on an approximate Bernoulli distribution to determine mask elements and achieves model sparsity through  $L_0$  regularization, enabling efficient parameter sharing. [76] also focuses on recommendation system design, noting that existing multi-task learning models [93], [98] typically use MLP as the expert model. However, using MLP models may not generate distinctive features for different tasks, leading to gradient conflicts and thus affecting model performance. Therefore, this work proposes a new expert network called the Task-Intensive Expert Model, whose parameters are specifically generated by a hyper-network based on different task embeddings, making it more distinctive and alleviating gradient conflict issues.

[91] further investigates multi-domain multi-task recommendation tasks, where multi-domain scenarios introduce more complex data dependencies that existing methods cannot handle well [230]. [91] decouples the originally complex tasks and introduces shared expert modules, domain expert modules, and task expert modules using MoE. These modules learn cross-domain and cross-task common knowledge, domain-specific user preferences, and task-specific user preferences, respectively, enabling the model to capture user preferences from multiple perspectives and better handle complex dependencies between domains and tasks.

Tasks in machine vision encompass various types, each with its corresponding deep models. However, many of these tasks often follow similar pipeline designs [81]. Aggregating the models of these similar tasks into a multi-task model can significantly improve the efficiency of training and inference while maintaining the performance of individual tasks. Considering that MoE not only allows for flexible model expansion but also handles dependencies between different tasks, it is frequently used in the design of visual multi-task models. [89], [19], [97] primarily apply MoE to improve the encoder. [89] introduces MoE layers into the attention modules and MLP blocks of ViT [1] and designs a new loss function to encourage strong yet sparse dependencies between tasks

and experts, ensuring that each task is only related to a few experts. This can reduce the likelihood of gradient conflicts between tasks. [19] also incorporates MoE layers into ViT and discusses the advantages and disadvantages of two gating functions in application scenarios: one configures a different router for each task, while the other uses a shared router to achieve dynamic routing by concatenating task embeddings with input tokens. It was found that the first gating method is more effective for handling multiple tasks. [97] inserts sparsely activated MoE layers into ViT and computes the model's loss on the validation set during each iteration. By using the loss as feedback to dynamically increase or decrease the number of experts, the cumbersome process of manually adjusting model size can be avoided.

A few works have also attempted to introduce MoE into the decoder. [86] finds that different task decoders, due to shared decoding parameters, result in a static feature decoding process, leading to insufficient distinction in task-specific features. Therefore, MoE was introduced as a decoder to decompose backbone features into task-agnostic features from different aspects, making the decoding process of task-specific features more fine-grained. This work also utilizes convolutional operations to expand the receptive field of the gating network, which enriches the contextual information in the decision-making process. [81] argues that in traditional MoE structures, experts can only interact through routers, thus only establishing connections between some tasks. By adding a  $3 \times 3$  convolutional layer after the encoder, which is optimized by all tasks, it can learn common features across all tasks, providing the model with more global information to aid decision-making.

The combination of multi-task learning and MoE has further been explored in other application scenarios. [83] replaces the expert networks in MoE with LSTMs [231], leveraging the advantages of LSTMs in handling sequential data to improve performance in processing user activity data streams. The work in [94] applies MMoe to Digital Rock Physics Analysis tasks, addressing the issue of significant computational and memory resource consumption encountered by traditional numerical simulation methods for rock analysis [232], [233]. By utilizing MMoe, [94] transforms serial operations into parallel operations, reducing interference between multiple analysis tasks. [84] focuses on Heterogeneous Multi-task Learning tasks. In particular, this study only employed the MMoe framework but also introduced some stochastic mechanisms that restrict certain experts to handle specific tasks or disconnect certain experts from the current task, thus enhancing expert exploration.

#### D. Reinforcement learning

Reinforcement Learning (RL) plays a pivotal role in decision making in many real-world applications, such as robotics, games, autonomous driving, and healthcare [234], [235], [236], [237], which seeks to learn a policy by interacting with the environments. However, its practical development suffers from several significant challenges, including computational inefficiency and limited adaptability within high-dimensional

state spaces and complex dynamic environments [102]. By dynamically selecting the most appropriate expert based on input data, the MoE mechanism enhances RL agents’ flexibility in responding to diverse demand patterns and environmental changes, thus improving overall performance. In what follows, we summarize the recent advance in leveraging MoE in RL and showcase its potential in various applications.

The application of MoE in RL has a long history. One notable direction is the so-called modular RL, which breaks down complex RL problems into smaller modules and builds multiple sub-agents to handle a specific aspect of the task. Notably, [125] considers nonlinear and nonstationary control tasks, and proposes a multiple model-based RL (MMRL) system with multiple modules (i.e., experts). Each module has a dynamics model and a controller, where the state predictions and the action outputs are the weighted combination of all module outputs, with the weights calculated using a gaussian softmax function of the module state prediction error. Multiple studies have been proposed by following up this work. [106] introduces a concept of ‘modular reward’ which combines both the actual reward and the imaginary reward of appropriate module selections for the task. This will promote independent learning of different modules, which further leads to accurate estimation of global value functions of the composite policy. [107] constructs multiple experts to specialize in different regions of the overall state space, where standard RL approaches can be used for each expert to capture different information of the underlying state space. This is particularly useful for RL problems with large state spaces.

On the other hand, the idea of modular RL has also been investigated in the design of value-based approaches. Specifically, [108] introduces a novel mixture of actor-critic experts (MACE) architecture with multiple actor-critic pairs for learning terrain-adaptive dynamic locomotion skills, where each pair will specialize in particular aspects of the motions. MACE is shown to be able to learn more quickly than the standard single actor-critic pair approach. [109] follows this network architecture and learning algorithm to solve MDP problems with a mixture of discrete and continuous actions, in the context of policy learning for the safe falling problem. [126] considers the skill knowledge transfer in multi-task RL and proposes a transfer expert RL architecture (TERL), where both actor and critic consist of a hierarchical structure with the experts and a gating network. The gating network will thus determine the expert selection of different tasks for either policy learning in the actor or value function learning in the critic. Several key features are designed in both gating network and experts to increase the capability of solving new tasks. [110] also considers a multi-task RL framework for transfer learning, which seeks to leverage the pretraining of multiple source tasks to facilitate the learning of new target tasks. Different from the traditional case where only one expert is activated at a particular timestep, this work proposes a model to enable the activation of multiple expert simultaneously. Each expert can specialize in a distribution of actions, whereas the composite policy can be obtained through a composition of these distributions.

More recently, [111] constructs the policy in policy iteration

by leveraging the MoE architecture, where the action of each expert is determined based on the distance to a prototypical state, in order to ensure the policy is interpretable. [100] introduces Probabilistic Mixture-of-Experts (PMOE), which employs Gaussian mixture models for multi-modal policy representation and frequency approximation gradients to address non-differentiability during optimization. This approach overcomes the limitations of unimodal policies in traditional deep RL, enabling algorithms to learn varied strategies in tasks with multiple optimal solutions. More importantly, PMOE can be potentially applied in generic off-policy and on-policy deep RL algorithms using stochastic policies, e.g., Soft Actor-Critic (SAC) [221] and Proximal Policy Optimization (PPO) [238]. To improve the sample efficiency of [108] and accelerate model-free RL for terrain-adaptive locomotion skills learning, [112] proposes to generalize model-based value expansion (MVE) [239], a technique to obtain better state-action values by using more reliable targets, to a mixture of actor-critic experts. [101] proposes an MoE framework incorporating Bayesian functions that autonomously identify and combine promising sub-regions from multiple source tasks across different regions of the state space. By utilizing state-dependent Dirichlet prior distributions to learn similarities between dynamics in source and target tasks, and updating these priors using state transition data from the target environment, this method improves robustness against sparse or delayed rewards, even when there are errors in dynamic estimation.

The attention of leveraging MoE in RL has gained increasing attention last year due to the success of MoE in LLMs. [115] investigates the impact of MoE layers in the DNNs used in value-based deep RL approaches, and shows that the performance of various deep RL algorithms can be significantly improved by incorporating soft MoEs [27] to replay the penultimate layers. Similar phenomena have been observed in different RL setups, such as online RL, offline RL [240], and RL with a small amount of interactions. [113] considers multi-task robot learning and seeks to leverage offline RL to learn from both good demonstrations and sub-optimal data. In particular, the policy network is constructed as an MoE-based Transformer encoder to generate action tokens. [241] proposes a Diverse Skill Learning (Di-Skill) method which aims to enable agents to acquire more skills in RL. In particular, each expert is modeled as a contextual motion primitive, adjusting behavioral strategies according to the current context and optimizing related distributions to focus on optimal performance sub-regions. Di-Skill effectively addresses hard discontinuities and multimodality in unknown environments by representing each expert’s contextual distribution using energy-based models. [114] proposes a novel method named SMOSE for continuous control tasks. The controller is built upon a Top-1 MoE architecture, where each expert is trained to learn different basic skills and the router is trained to learn to appropriately assign tasks to experts. [102] investigates MoEs’ capability to manage non-stationarity — handling and adapting to changes in the environment or data distribution over time — in deep RL settings characterized by “amplified” non-stationarity via multi-task training. It finds that MoE supports network plasticity and is particularly effective in

highly non-stationary conditions, providing fresh insights into addressing non-stationary training environments in RL. [103] explores visual deep RL to enable robots to acquire skills from visual inputs and proposes a method named MENTOR, which replaces the traditional MLP layers in visual RL agents by using the MoE architecture for action generations.

The application of MoE in other domains of RL, such as multi-agent RL and imitation learning, has also been explored. For instance, [116] applies modular RL to soccer robot. [117] proposes a method to assign multiple modules to different situations, in order to learn purposive behaviors for specified situations related to other agents. Similar to [125], in [117] each module consists of a prediction model and a planner, and the module that provides the best estimation of state transitions will be chosen. [118] uses modular RL with MoE to learn competitive behaviors in multi-agent systems. For imitation learning, [119] considers the problem of robot table tennis by allowing robot to learn through interactions with human, where an approach with mixtures of motor experts is proposed to generalize basic movements to various situations. Here each expert corresponds to a motor policy. [121] constructs a mixture of interaction experts to learn various interaction patterns from human demonstrations, based on Gaussian mixture models. [122] leverages Gaussian mixture models to handle the model collapse problem in learning generalized movement representations from human demonstrations. [123] leverages a mixture of movement primitives to imitate versatile human demonstrations. [124] trains a variational autoencoder with a mixture density network to capture the complexity and variability in Human-Robot Interaction. Given the diversity of constraints under which demonstration data may be collected, [104] presents Multi-Modal Inverse Constrained Reinforcement Learning (MMICRL), which uses flow-based density estimation for unsupervised expert identification and infers specific agent constraints, thus enhancing the adaptability and reliability of inverse constrained reinforcement learning algorithms. The strategy of ensemble methods that carries a very similar idea with MoE has also been widely used in RL, in terms of model ensembles [242], value function ensembles [243], and policy ensembles [244], [245].

Finally, we summarize several attempts on applying MoE in specific applications of RL. For example, [105] confronts the challenge of driving chatbots with RL by developing several RL algorithms tailored for Dialogue Management. Leveraging the hierarchical structure of MoE-based language models enables offline RL methods to operate within a significantly reduced action space, making RL problems more tractable while generating discourse that reflects diverse intentions. Similarly, in video processing [246] and investment portfolio management [247], MoE-based approaches dynamically activate specific expert models based on current data, achieving higher accuracy, lower computational costs, and enhanced adaptability to changing environments. [120] leverage MoE models to scale up the parameters of RL models in optimal execution tasks.

## IV. THEORY

Although the MoE model has a long history and achieved great empirical success in practice during the past few years [248], [16], the theory behind it is largely under-explored, especially in deep neural networks. Recently, some studies have initiated the attempt to build theoretical understandings of MoE. In this section, we will delve into these studies, aiming to provide the readers a basic idea of the theory development in this area.

Most of the theory development for MoE so far mainly focus on simple models. Early studies in this area seek to understand the approximation capacity of MoE models. For instance, [128] investigates the convergence rate for hierarchical MoE models where the experts are exponential family regression models. [129] characterizes the convergence rate for hierarchical MoE models with generalized linear models as the experts. [130] provides the error bounds for functional approximation using MoE models and shows the convergence to any sufficiently differentiable target function in the Sobolev space. [131] provides an alternative result of this through a universal approximation theorem for MoE models with softmax gating functions and linear experts, without imposing assumptions on the differentiability of the target function. [132] considers a standard MoE model with softmax gating functions where the experts are polynomial regression models, and characterizes the convergence rate of the maximum likelihood estimator (MLE) depending on the number of experts and the order of the polynomial model.

Recently, some studies have emerged and aimed to characterize the convergence of MLE for MoE models with different gating functions for regression problems. For instance, [133] characterizes the convergence rate of the MLE for Gaussian MoE models with covariate-free gating functions, by leveraging techniques from optimal transport. [134] attempts to understand the parameter estimation of the MLE for Gaussian MoE models with softmax gating functions. By leveraging novel Voronoi loss functions to capture the interactions between the gating and experts, the convergence rate of MLE is established. The convergence of MLE for Gaussian MoE models with other types of gating functions, such as Gaussian density gating [135] and top-K sparse softmax gating [136], has also been investigated. [137] studies a more general setup of MoE models with softmax gating for least square estimation, and characterizes the convergence rates for different types of expert model. The theoretical understanding behind the recently emerged cosine gating function is developed in [25] for least square estimation.

In contrast, the theoretical understanding of MoE models for classification problems is less explored. [138] investigates the classification problem with the MoE models including softmax gating and multinomial logistic expert models. In particular, the convergence rates for density estimation and parameter estimation are established. A novel class of modified softmax gating functions is further proposed to improve the convergence rate for parameter estimation when part of expert parameters vanish. [138] also studies the inefficiency issues encountered by the gating function in MoE during



training. Specifically, when some expert parameters vanish (i.e., the parameters of certain expert models approach zero or become insignificant), the interaction between the standard softmax gating function and the expert functions through partial differential equations (PDEs) significantly slows down the parameter estimation rate [133], [136], [134], [135]. This paper thoroughly investigated the problem under an improved softmax gating function setting. Specifically, they used a bounded function  $M(X)$  to transform the input  $X$  before passing it to the softmax gating function. Although the expert parameters still depend on the input  $X$ , the transformed  $X$  no longer has a linear relationship with the original  $M(X)$ , thereby eliminating the interaction between gating parameters and expert parameters and significantly improving the model's convergence speed and stability.

Most of the successes of MoE models nowadays happen with more complex expert models, such as deep neural networks. However, the theoretical understanding of MoEs in deep learning remains elusive compared to the theoretical studies of MoE in simple models, largely due to the limited understanding of deep neural networks. [29] seeks to fill this gap and makes the first attempt to understand the behaviors of MoEs in deep learning. More specifically, they consider a standard MoE architecture with softmax gating, where each expert model is a two-layer CNN, and investigate a binary classification problem where the dataset contains multiple clusters. The authors show that any single-expert model with a two-layer CNN cannot achieve more than 87.5% accuracy on this specific dataset, while a linear MoE model performs slightly better than a single-expert model but still falls short of a nonlinear MoE model. With sufficient expert exploration during the training, it can be shown that the router can automatically learn the cluster structure of the data and dynamically route the data to the most suitable expert. Following this work, [127] investigates the patch-level routing for MoE models (pMoE) with linear gating in solving a supervised binary classification task, where each expert is a two-layer CNN. They show that the pMoE model can achieve similar generalization performance with standard CNN but with a reduced sample complexity. One of the key reasons behind this, also theoretically and empirically justified, is that an appropriately trained patch-level router can route the class-discriminative patches of one class to the same expert and drop some class-irrelevant patches.

With the increasingly broad applications of MoE in many problems, the theoretical studies of MoE start to emerge beyond the general machine learning setup. [139] explores the usage of MoE models in handling multilevel data while building the theoretical understanding therein. Notably, multilevel data is very common in practical applications [249], [250], such as hierarchical data with a nested structure where different levels are correlated. Ignoring the dependencies within the data can lead to spurious, misleading, or biased clustering and prediction outcomes [251]. To better capture the dependencies, [139] proposes a Mixed MoE (MMoE) architecture to handle multi-level data. This approach not only introduces random effects into the gating function, making the computed gating values depend not only on input variables but also on random

effects. This helps the model better capture random effects in multi-level data when dynamically adjusting expert weights. At the same time, to simplify the model, the expert functions are separated from the random effects, and the probability distribution of the expert model no longer depends on input variables but only on output variables. Under certain regularity conditions, this paper proves that the MMoE is dense in the sense of weak convergence for any continuous mixed-effects model. In the special case of hierarchical multi-level data, this paper further proves that MMoE can approximate the complex dependency structures of random effects between different factor levels in multi-level data.

[30] investigates how integrating the MoE framework into models can effectively mitigate generalization and forgetting issues in continual learning [252], [253], under the setup of over-parameterized linear regression tasks and a linear gating MoE model. The authors prove that a sufficiently trained MoE model can diversify experts to specialize in different tasks, which minimize the interference across tasks and reduce the forgetting in continual learning. To ensure the stability of the gating function, the update of the gating network must be terminated in a timely manner for new task learning. Explicit expressions for expected forgetting and overall generalization error are also derived, quantitatively evaluating the impact of MoE in continual learning. [30] theoretically investigates the usage of the MoE framework in mobile edge computing (MEC) on handling continuous task streams therein. Specifically, each MEC service station is treated as an expert model, updating its local model based on the data distribution of tasks. An adaptive gating network was proposed to dynamically allocate tasks to different experts, ensuring that each expert converges and specializes in specific task types. Additionally, the paper derived the minimum number of experts required to ensure system convergence and proved that the MoE model can control the overall generalization error within a small constant range, significantly outperforming traditional MEC offloading strategies.

## V. APPLICATIONS

In this section, we investigate the applications of MoE in two broad application domains, i.e., computer vision (CV) and natural language processing (NLP), among which various specific applications will be explored.

### A. Computer vision

With the rapid development and maturation of the CV field, the focus of both academia and industry has gradually shifted towards how CV technologies can be effectively implemented in specific application scenarios. However, this process has encountered several challenges such as resource constraints in edge deployment, difficulties in model scalability, and computational efficiency. These issues have pointed to the need of the redesign of existing models, especially when dealing with large volumes of data and complex tasks, where traditional model architectures often struggle to find an ideal balance between performance, flexibility, and stability. Since MoE has attracted renewed attention due to its performance

advantages demonstrated in Mixtral 8x7B [22], an increasing number of studies have attempted to apply it in the CV field, to boost the state-of-the-art performance. Since the MoE architecture introduces multiple specialized experts to handle specific types of inputs and dynamically adjusts the activation levels of different experts based on the input data, it not only helps alleviate hardware resource bottlenecks but also enhances the model’s ability to handle diverse visual tasks, thereby optimizing both performance and stability. Therefore, in this subsection we will focus on the application of MoE in the field of CV by reviewing recent advancements from four bedrock aspects, i.e., classification, detection, segmentation, and generation. The aim is to provide valuable references for future explorations to address the challenges in the CV field.

**1) Image Classification.** Image classification, as one of the most fundamental machine learning tasks, is a key component of many CV tasks. Therefore, understanding the performance and limitations of MoE in image classification can, to some extent, reflect its potentials for application in other CV tasks.

Notably, [17] proposes and validates a new approach, namely V-MoE, to cope with visual tasks, by replacing the MLP layers in Vision-Transformer [1] with sparse MoE layers. V-MoE can scale the model to 15 billion parameters and demonstrates the efficiency gains from this simple combination in image classification tasks. Further explorations within the traditional MoE framework have been conducted. [140] dissects the MoE architecture, providing a detailed exploration of its training and optimization. The paper discusses how the number of expert networks, the number of MOE layers, and their placement affect the model’s performance, efficiency, and stability.

As expert networks are a key component of MoE, many efforts have been made to optimize them. ViMOE [141] introduces the concept of shared experts, where a shared expert handles common knowledge required for classification, while other specialized experts focus on specific knowledge. This approach mitigates the difficulty in exploring optimal configurations due to the sensitivity of MoE layers to expert setups and enhances training stability. The paper also systematically analyzes routing behaviors, examining routing strategies, the number of experts, and the placement of MoE layers. It visualizes routing behaviors with heatmaps, finding that MoE layers closer to the output handle more semantically rich feature maps, leading to more specialized tasks for experts. Consequently, the last  $L$  layers of the model should be replaced with sparse MoE layers. The analysis suggests that both the number of experts and the size of  $L$  are related to the degree of routing, and more does not necessarily mean better performance. To some degree, the base model introduced in [142] shares similarities with the shared expert concept proposed in ViMOE [141]. In CLIP-MoE [143], the MoE architecture is applied to enhance the capabilities of the CLIP model through a method called Diversified Multiplet Upcycling. The authors fine-tune a series of pre-trained CLIP models using multi-stage contrastive learning, extracting expert networks that focus on different aspects of the input data, and integrating these experts into the MoE architecture.

Other papers delve deeper into the design of gating func-

tions. [27] introduces a novel technique called Soft MoE, which differs from traditional hard assignment of input tokens to experts by using a weighted average of all tokens for soft assignment. Therefore, each expert processes different parts of these weighted combinations. This method retains the advantages of the MoE architecture in scaling model capacity. Besides, it also improves training stability, reduces inference time, and effectively addresses the token dropping issue common in MoE training, creating a new paradigm for gating functions. [142] addresses the overfitting problem encountered when using MOE on small datasets by designing an architecture that incorporates early exit mechanisms and pre-training. Specifically, the model pre-trains a base model during the training phase and pre-determines the number of experts using k-means clustering. During inference, if the base model has high confidence in a sample’s prediction, it can opt for an early exit, bypassing the subsequent MoE layers for specialized learning. Otherwise, the sample is passed to the MoE layers for specialized learning, and the outputs from both the MoE layers and the base model are combined by Ensemblers to produce the final output.

[145] focuses on improving the design of gating mechanisms in Hierarchical Mixture of Experts (HMoE) models to address performance bottlenecks in handling complex inputs and executing specific tasks. Specifically, this work proposes a method that goes beyond the traditional Softmax gating function by using Laplace gating strategies for both hierarchical and task allocation. By customizing gating functions for each expert group, the HMoE model can allocate resources more effectively without increasing computational burden and improve performance on complex datasets. Additionally, it explores the intrinsic interactions between first-level and second-level gating parameters, which significantly affect the convergence speed of model parameter estimation, providing a thoughtful direction for allocating complex or hierarchical datasets. [144] designs a deep mixture algorithm called DeepME (Deep Mixture Experts) for efficiently handling large-scale image processing tasks. It groups similar image categories based on semantic relevance, allowing for some overlap between categories. This ensures that each task group contains image categories with similar learning complexities, and specific base deep networks are trained for these groups. Additionally, a gating network is trained to combine all base deep networks, generating a mixed network with larger outputs to effectively handle large-scale datasets containing thousands of categories.

**2) Object Detection.** Object detection is also one of the important tasks in the field of CV. Building on image classification, it requires the model to not only detect objects but also output the location of each classified object, where ensuring the computational efficiency of the model is more challenging compared to that in image classification.

Similarly, many studies have explored the impact of simply integrating MoE with base models for object detection. MoCaE [146] has shown that directly using simple methods from deep ensemble (DEs) [254] to integrate different object detectors with the MoE architecture does not improve model performance and may even have adverse effects. The under-

lying reason is that simply adding different detectors leads to unfair competition, thereby affecting the final detection results. To address this, this work introduces Early Calibration and Late Calibration to adjust the confidence levels of different detectors, better reflecting their true detection performance. Experimental results have shown that MoCaE yields significant gains over single models and DEs on several real-world challenging detection tasks.

[147] replaces the FFN layer in the traditional transformer with the MoE-HCO block, proposing an event-stream-based object detection framework called MvHeat-DET. The MoE-HCO block selects the most suitable transformation branch for the current features through a policy network, providing various signal transformation expert networks. Therefore, the input events can have more diverse and suitable processing modules. It also uses frequency embedding to predict thermal diffusion coefficients to simulate the heat conduction process. This will enable the model to better capture spatiotemporal dynamic features in event streams while maintaining efficient computation, demonstrating high practical values. To help the model better understand and process multi-source datasets, DAMEX [148] proposes a data-aware MoE architecture, by replacing the FFN layer in the transformer with an MoE layer. Different experts are trained to learn data from different sources, enhancing the model’s generalization capability. Traditional two-stage methods [255], [256] require multiple classification heads to detect data from different datasets, leading to an increase in parameters with mixed datasets. In contrast, the data-aware MoE layer can achieve excellent performance without significantly increasing the number of parameters. [149] addresses the diversity and ambiguity of the task itself, by transforming the original single-mask prediction task into a multi-mask prediction task and predicting human preference scores for each possible salient object mask. To enable the model to handle multiple tasks more efficiently, this paper leverages the MoE architecture to solve the integrated training problem of the two tasks, and overcomes the inconsistency in input and output formats between the two subtasks. In particular, this work replaces the FFN layer in DaViT [257] with an MoE layer, which contains two experts: the P-FFN layer for handling the PSOD (Pluralistic Salient Object Detection) task and the Q-FFN for handling the MQP (Mask Quality Predictor) task. This allows the model to dynamically select the most suitable expert network for the current task, improving model performance without significantly increasing the number of parameters.

**3) Semantic Segmentation.** Semantic segmentation requires classifying each pixel in the input image. As a result, processing high-resolution images may significantly increase the demand on the hardware, hindering the model’s practical applications. Compared to traditional segmentation models, the MoE-based model can provide the following benefits: 1) The MoE architecture alleviates hardware pressure through its sparse activation and divide-and-conquer strategy, enhancing the model’s practicality. 2) Some studies [150], [151] have also found that using MoE can achieve excellent performance in semantic segmentation tasks that require greater model stability and generalization. 3) MoE also enjoys a better

interpretability in terms of knowledge specialization.

[153] seeks to leverage MoE to achieve a good balance between computational complexity and representational capacity in deep neural networks. In particular, traditional convolutional networks are combined with shallow embedding networks and multi-head sparse gating networks, dynamically selecting and executing partial networks at each convolutional layer. This can improve the accuracy while reducing prediction costs without significantly increasing the network width. Furthermore, this work proposes two DeepMoE variants including wide-DeepMoE and narrow-DeepMoE. The first one is suitable for applications requiring high accuracy, whereas the second one is designed for computational resource-limited scenarios. Swin2-MoSE [154] demonstrates that MoE can also be applied to improve semantic segmentation models, especially on remote sensing images. The authors design an MoE layer called MoE-SM, which includes an SM (Smart Merger) module to merge the outputs of various experts and adopts a new per-example strategy instead of the commonly used per-token one. This can ensure that all tokens of each example are processed by the same expert, enhancing the model’s competitiveness in semantic segmentation tasks.

Some studies have also investigated the usage of MoE in the segmentation for autonomous driving related scenarios. For example, [152] aims to leverage MoE to better understand autonomous driving scenarios which can further lead to improved algorithm design. Due to its architecture design, MoE inherently enjoys better transparency. Compared to other general models [254], [258], the MoE models can not only provide the final overall model output, but also enable the analysis of the outputs of individual expert models and their consistency or divergence with the overall output during the decision-making process. This mechanism provides more detailed information about how the models work, offering better interpretability while maintaining performance close to that of a single baseline model. A common issue in segmenting urban and highway traffic scenes is vulnerability to adversarial attacks. To address this, [150] attempts to use MoE given its properties of more dynamic model selections compared to traditional methods [259], [260], [261]. This work shows that the MoE models exhibits higher robustness against instance-specific attacks, universal white-box adversarial attacks, and cross-model transfer attacks, maintaining relatively high accuracy under these attacks. Additionally, the MoE models with additional convolutional layers show even stronger resistance to attacks. These results indicate that MoE not only improves model performance but also enhances model robustness and stability.

**4) Image Generation.** Image generation can produce realistic and diverse images, which is very useful in many applications [262], [263], [264], [265]. However, current image generation technologies still face various challenges such as low generation quality, limited diversity, and poor adaptability to complex tasks. By decomposing the complex problem of image generation into multiple simpler and core tasks, MoE leverages the collaboration among the experts to enhance the quality of generated images, improve the diversity of outputs, and further enrich detail representations.

RAPHAEL [155] introduces two types of MoE models for temporal control and spatial settings during generation, respectively. On one hand, a spatial MoE layer is responsible for depicting different text concepts in specific image regions, allowing each text token to learn specific visual features through specialized experts. This can enhance the representation of different concepts. On the other hand, a temporal MoE layer focuses on processing these concepts at different time steps of the diffusion process to handle varying degrees of noise impact. This configuration results in billions of diffusion paths from the network input to the output, each path acting as a “painter” for specific concepts and image regions, achieving finer text-to-image alignment and improving the quality and aesthetic appeal of generated images. The application of MoE enables RAPHAEL to flexibly switch between multiple styles and closely adhere to text prompts.

GANs [266]

are known to struggle to learn multimodal data distributions when processing complex datasets, leading to generated images with poor-quality. To address this, [156] proposes MEGAN with multiple generator networks, where each network focuses on learning specific modal distributions in the dataset. This design allows each generator to focus on different data subsets, generating more diverse and higher-quality images.

A novel design of MoA is introduced in [32] by creating two attention paths, i.e., a personalized branch and a non-personalized prior branch, allowing the model to retain its powerful generation capabilities while minimizing interference with the personalized part. This specially designed MoA can intelligently adjust the generation process based on the input text and subject image. By optimizing the fusion of personalized and generic content, high-fidelity personalized image generation can be achieved without sacrificing image diversity or quality. Additionally, the learning routing mechanism in [32] dynamically manages pixel distribution at each level, further improving the performance for image generation tasks that involve multiple subjects and complex interactions.

Text2Human [157] applies MoE to the transformer encoder based on the diffusion model [267], [268], [269], enabling conditional generation of clothing with different texture types. Specifically, the MoE model routes input features to different expert heads based on the texture attributes mentioned in the text prompt, with each expert head responsible for predicting tokens of specific textures. Based on the synthesis and control of more complex textures, this method increases the model’s ability to handle details while avoiding the computational burden of training separate samplers for each texture.

## B. Natural Language Processing

As a core research area of artificial intelligence, NLP focuses on enabling computers to understand, generate, and process human language [270], [271]. It plays a significant role in promoting human-computer interaction, information extraction, knowledge mining, cross-lingual communication, and automation. However, the field of NLP still faces challenges such as the trade-off between model capacity and

computational cost, poor adaptability to multi-task and multi-domain scenarios, sparse data and long-tail problems, insufficient reasoning and logical capabilities, and the need for personalization and dynamic adaptability.

The MoE architecture effectively alleviates these challenges through mechanisms such as sparse activation and the allocation of different expert models for different tasks or domains. This not only reduces computational costs but also enhances model performance in multi-task, low-resource scenarios, and complex reasoning tasks, while supporting personalized services. In light of the application of MoE in the NLP field, this subsection will summarize and review relevant work from the perspectives of Natural Language Understanding (NLU) and Natural Language Generation (NLG).

**1) Natural Language Understanding.** The advancement of NLU technology enables machines to understand and interpret human language, by bridging the gap between human communication and machine processing [272], [273]. This allows systems to perform tasks such as intent recognition, entity extraction, and semantic analysis [8], [274], which are crucial for applications such as virtual assistants, chatbots [275], sentiment analysis [276], and information extraction [277]. However, the further development of NLU is hindered by the complexity, ambiguity, and diversity of human language. MoE allows systems to dynamically leverage expert networks for different linguistic tasks or domains. For example, one expert may specialize in syntactic parsing, while another focuses on sentiment analysis, thereby enhancing the model’s ability to effectively handle various complex linguistic patterns. This specialization and adaptability make MoE a powerful method for extending and refining NLU systems, enabling more accurate and context-aware language understanding. Since the use of specially designed MoE architectures to enhance model performance in NLG tasks has not yet reached a scale sufficient for sub-domain classification, this section will only summarize some representative work in recent years.

GLaM [20] introduces an MoE architecture to make the training and inference processes more efficient. It significantly reduces the required computational resources and energy consumption while maintaining or even improving model performance. This encouraging result highlights the great potentials in leveraging MoE to address the issue of high computational resource consumption in large-scale, intensive language models for NLU tasks.

MoE-LPR [158] designs a two-phase training strategy for the MoE-based model. In the first phase, the original model is converted into an MoE architecture, and new expert modules are added, which improves the model’s capability for new languages without using original language data. In the second phase, a small amount of original language data is used for review, and a language prior routing mechanism is employed to restore and maintain the performance of the original language. Experimental results demonstrate the excellent scalability and stability of MoE-LPR, providing a new perspective for multi-lingual natural language understanding tasks.

When speech is transcribed into text, errors from Automatic Speech Recognition (ASR) systems [278]

often affect the accuracy of subsequent NLU components.

To address this, [159] proposes an MoE-based framework, namely MoE-SLU, aiming to mitigate the impact of ASR errors on spoken language understanding performance. MoE-SLU employs three strategies to generate additional transcripts and uses MoE to weight and average these transcripts. By enhancing the model’s ability to capture keywords, MoE-SLU is more robust to ASR errors, achieving state-of-the-art performance on three benchmark SLU datasets. The performance of MoE-SLU can be further improved through regularized predictions.

To address the issue of interference in multi-task learning, particularly among NLU tasks, MT-TaG [87] introduces a sparsely activated MoE architecture and designs a task-aware gating mechanism to route inputs to specific expert networks. Compared to traditional dense models, MT-TaG demonstrates superior performance in multi-task learning, especially in low-resource task transfer, efficient generalization, and handling unrelated tasks. MoPE-BAF [160] designs various soft prompt experts, including text prompts, image prompts, and unified prompts, based on a unified vision-language model to enrich unimodal representations and promote multimodal interaction. By introducing a block-aware prompt fusion mechanism, the model achieves cross-modal prompt attention across Transformer layers, smoothly transitioning from unimodal representations to multimodal fusion. Experimental results show that MoPE-BAF significantly improves performance in few-shot settings for multimodal sarcasm detection and sentiment analysis tasks, demonstrating the effectiveness of MoE in deep multimodal semantic understanding tasks.

**2) Natural Language Generation.** NLG can transform structured data or concepts into natural and fluent text, widely applied in machine translation [279], dialogue systems [280], content creation [281], and information summarization [282], among other fields

. To achieve diverse and precise language expression, NLG systems require strong model capacity to capture complex linguistic structures and contextual dependencies, while effectively handling interference in multi-task learning to maintain high performance across different scenarios. MoE allows models to dynamically select the most suitable expert networks and flexibly adjust model capacity, reducing task interference and enhancing model specialization, thereby contributing to the construction of more powerful and flexible NLG systems.

*(a) Text Generation.* Text generation has numerous applications in NLP. Through effective text processing, NLP systems can accurately understand, generate, and transform human language, which is crucial for achieving more advanced NLP applications. However, text processing still faces challenges such as insufficient generation diversity and uncontrollable generated content. MoE dynamically selects the most suitable sub-network for specific tasks, making models more efficient and precise when processing complex or diverse text data.

In [161], MoE is introduced into the generator of a language GAN, leveraging multiple experts to collaboratively generate high-quality sentences, with each expert acting as a recurrent neural network that autoregressively produces the current token representation. Additionally, the Feature Statistics Alignment (FSA) paradigm is incorporated to further

optimize the learning signals during generator training, making the generated text more closely aligned with the real data distribution. In RetGen [162], during inference, a retriever first obtains the top K most relevant documents and their corresponding probabilities. Then, K independently trained transformer-based generation models process each document along with the same context, combining the partially generated results with the current consensus to produce their respective output distributions. The final output distribution is a weighted combination of these independent generation results, with weights determined by the document relevance probabilities. This approach overcomes issues such as ignoring document relevance information when simply concatenating multiple documents as input, allowing the model to more effectively utilize document relevance information to guide the text generation process. LogicMoE [163] focuses on solving the problem of table-to-text generation. In the designed LogicMoE architecture, each expert acts as a specialized generator for a specific logical type, responsible for generating sentences that meet the requirements of that logical type. This design not only improves the quality of generated sentences but also enriches the diversity of generated content at the semantic and logical levels. Experimental results show that LogicMoE achieves absolute improvements of 0.8 and 2.2 points in BLEU-3 scores [283], demonstrating the inherent advantages of LogicMoE in generation diversity and controllability. The optimization of MoE system design has also been explored [164] to enhance model efficiency in text processing tasks.

*(b) Machine Translation.* Machine translation is the application of automatically translating text from one language to another, greatly facilitating cross-lingual communication, information access, and global collaboration [284], [285]. However, machine translation faces challenges such as linguistic diversity, grammatical complexity, and cultural differences. MoE uses conditional computation to dynamically allocate tasks, enabling the system to select the most suitable expert based on the input content. This mechanism not only improves translation accuracy and fluency but also better handles complex scenarios involving multiple languages and domains, thereby enhancing the performance and adaptability of machine translation.

[16] innovatively introduces Sparsely-Gated MoE technology for conditional computation. Applying MoE to language modeling and machine translation tasks not only significantly improves model performance on large datasets but also achieves efficient utilization of computational resources. Gshard [21] replaces traditional feedforward network layers with MoE layers and adopts conditional computation to achieve efficient utilization of computational resources and flexible expansion of model capacity. MoE-based Transformer models at different scales are designed and trained to meet translation needs from high-resource to low-resource languages. Experimental results show that the MoE models trained using GShard significantly improve translation quality, while achieving higher training efficiency and lower costs compared to dense models under the same hardware conditions. [165] also introduces a selective activation mechanism by replacing some dense model layers with MoE layers,

significantly enhancing the model’s representational capacity with similar inference and training efficiency. This approach is particularly beneficial for high-resource languages due to the increased model capacity, while also helping low-resource languages reduce interference from unrelated languages. Additionally, to address the issues in large-scale MoE models, such as quick overfitting in low-resource directions, regularization and curriculum learning strategies are applied to optimize complex training dynamics. A good balance is achieved between cross-lingual transfer and interference in multilingual machine translation. [166] also leverages sparse activation mechanisms to significantly enhance the model’s representational capacity without increasing computational costs.

In order to address the inefficiency issues encountered during the deployment and inference stages of MoE models, [167] proposes three optimization techniques, including dynamic gating, expert caching, and load balancing methods, to improve the inference efficiency of MoE models in language modeling and machine translation tasks.

**3) Multimodal Fusion.** Multimodal fusion integrates information from text and other modalities, providing richer and more comprehensive data representations. In particular, multimodal fusion can capture contextual clues and semantic details that single modalities cannot provide, thereby enhancing the model’s understanding and predictive capabilities. In multimodal fusion scenarios, using MoE allows for dynamic adjustment of the importance weights of different modalities based on their characteristics, enabling effective integration and utilization of information from different sources while avoiding the high complexity and optimization difficulties faced by traditional single models. This not only improves the overall performance of the model but also enhances its flexibility and scalability, allowing for better performance in complex multimodal tasks.

LIMoE [169], as the first large-scale multimodal mixture of experts model, can simultaneously process image and text data and align their representations through contrastive learning. Additionally, this study demonstrates the organic emergence of modality-specific experts in LIMoE, where experts spontaneously specialize in handling specific types of data or tasks. This indicates that the model not only effectively allocates resources for different tasks but also promotes cross-modal and cross-task knowledge transfer. LLaVA-MoLE [168] addresses the issue of data conflicts in multimodal large language models when fine-tuning with mixed-domain instruction data, by proposing the use of sparse LoRA Mixture of Experts (MoLE). The proposed method selects the most suitable LoRA expert for each input token based on its embedding features in the attention layers, allowing the model to adapt to inputs from different domains. This effectively mitigates performance degradation caused by mixing different types of instruction data. The Hunyuan model [170] is the first Chinese MoE multimodal large model, achieving a score of 71.95 on SuperClue-V, surpassing some international top models and demonstrating its unique advantages in the context of Chinese culture.

## VI. FUTURE DIRECTIONS

MoE models present significant opportunities for advancing machine learning in real-world applications, but they also face several challenges that must be addressed to unlock their full potentials. In this section, we discuss some of the key future research directions for MoE models.

### A. Training stability and load balancing

One of the most critical challenges in MoE models is ensuring training stability and load balancing among experts. Due to the dynamic nature of expert selection, some experts may receive significantly more data than others, leading to imbalanced training and potential model collapse. Future work should focus on developing more robust training strategies that ensure balanced utilization of experts. Techniques such as adaptive load balancing, dynamic expert capacity adjustment, and regularization methods that penalize over-reliance on specific experts could be explored. In some cases, theoretical studies on the convergence properties of MoE models under different load balancing strategies would provide valuable insights into designing more stable training algorithms. Additionally, the sparsity of MoE during training also poses challenges for stable model training, requiring more work to explore and optimize existing training strategies.

### B. Training and system efficiency

While MoE models offer the advantage of conditional computation, their training and inference efficiency remain a concern, especially in large-scale applications. Hardware advancements have reduced computational costs, but high latency remains a persistent issue. Future research should focus on optimizing hardware-software co-design, particularly for conditional computation. Techniques such as efficient memory management, reduced communication overhead, and parallel processing strategies could significantly improve the scalability of MoE models. Combining MoE with conditional computation techniques may also open new avenues for dynamic resource allocation and task-specific adaptations, further enhancing efficiency.

### C. Architecture design

The design of MoE architectures, particularly the determination of the number of experts and their specialization, is another area that requires further exploration. Current approaches often rely on heuristic methods or static setting to decide the number of experts, which may not be optimal for all tasks. Future work should investigate more principled approaches to determining the number of experts, such as using meta-learning or reinforcement learning to dynamically adjust the architecture based on task complexity and data distribution. Additionally, novel architecture designs that integrate MoE with other neural network components, such as attention mechanisms or graph neural networks, could lead to more powerful and flexible models.



## D. Theory development

Improving the interpretability of MoE models remains a critical area for future work. While MoE architectures have demonstrated remarkable scalability and performance, the theoretical underpinnings of their behavior—such as expert routing decisions and clustering mechanisms—are not yet fully understood, especially in modern deep neural networks. More rigorous theoretical studies are needed to explain these characteristics, which could lead to more robust and reliable models. This would also shed light on the design of better gating function and expert network as well.

## E. Tailored algorithm design

MoE models have shown great promise in various machine learning paradigms, but their potential in combination with other learning paradigms remains underexplored. Future work should investigate the integration of MoE with other learning frameworks, such as contrast learning, transfer learning, and self-supervised learning. For example, in federated learning, MoE models could be used to handle heterogeneous data distributions across clients by assigning different experts to different clients. Exploring these combinations could lead to more versatile and powerful learning systems.

## F. New applications

Although MoE has been extensively explored in NLP, its potential in other domains remains underexplored. For instance, in computer vision, recent work has shown that MoE models can achieve performance that continues to improve with training, suggesting untapped potential in areas like image segmentation, object detection, and multimodal learning. Further exploration in these domains could yield significant breakthroughs. Additionally, MoE models could be applied to emerging fields such as healthcare, robotics, autonomous systems, education, finance, as well as recommendation systems, where the ability to handle diverse and complex tasks is crucial. Developing MoE models tailored to these specific applications could lead to more effective and efficient solutions.

## VII. CONCLUSION

This comprehensive survey delves into the integration of the MoE architecture with various domains. Starting from various basic designs of the MoE architecture and training strategies, we highlight its synergy with important machine learning algorithms, alongside the recent theoretical advancements for understanding MoE. Furthermore, we provide a systematic summary of MoE in two critical application domains, computer vision and natural language processing, and shed light on the important future designs for improving the design and impact of MoE. We hope that this work will serve as a valuable reference for researchers in the field and beyond, elicit escalating attentions, and inspire further research ideas in MoE.

## REFERENCES

- [1] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, “An image is worth 16x16 words: Transformers for image recognition at scale,” *arXiv preprint arXiv:2010.11929*, 2020.
- [2] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, *et al.*, “Segment anything,” in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4015–4026, 2023.
- [3] Y. Shi, N. Wang, and X. Guo, “Yolov: Making still image object detectors great at video object detection,” in *Proceedings of the AAAI conference on artificial intelligence*, vol. 37, pp. 2254–2262, 2023.
- [4] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust speech recognition via large-scale weak supervision,” in *International conference on machine learning*, pp. 28492–28518, PMLR, 2023.
- [5] K. An, Q. Chen, C. Deng, Z. Du, C. Gao, Z. Gao, Y. Gu, T. He, H. Hu, K. Hu, *et al.*, “Funaudiollm: Voice understanding and generation foundation models for natural interaction between humans and llms,” *arXiv preprint arXiv:2407.04051*, 2024.
- [6] P. Rajpurkar, E. Chen, O. Banerjee, and E. J. Topol, “Ai in health and medicine,” *Nature medicine*, vol. 28, no. 1, pp. 31–38, 2022.
- [7] L. Cao, “Autoai: Autonomous ai,” *IEEE Intelligent Systems*, vol. 37, no. 5, pp. 3–5, 2022.
- [8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.
- [9] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, “Learning transferable visual models from natural language supervision,” in *International conference on machine learning*, pp. 8748–8763, PMLR, 2021.
- [10] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altschmidt, S. Altman, S. Anadkat, *et al.*, “Gpt-4 technical report,” *arXiv preprint arXiv:2303.08774*, 2023.
- [11] C. Thompson Neil, G. Kristjan, L. Keeheon, and F. Manso Gabriel, “The computational limits of deep learning,” *ArXiv, Cornell University, juillet*, 2020.
- [12] C. Sun, A. Shrivastava, S. Singh, and A. Gupta, “Revisiting unreasonable effectiveness of data in deep learning era,” in *Proceedings of the IEEE international conference on computer vision*, pp. 843–852, 2017.
- [13] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, “Adaptive mixtures of local experts,” *Neural computation*, vol. 3, no. 1, pp. 79–87, 1991.
- [14] M. I. Jordan and R. A. Jacobs, “Hierarchical mixtures of experts and the em algorithm,” *Neural computation*, vol. 6, no. 2, pp. 181–214, 1994.
- [15] W. Fedus, B. Zoph, and N. Shazeer, “Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity,” *Journal of Machine Learning Research*, vol. 23, no. 120, pp. 1–39, 2022.
- [16] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean, “Outrageously large neural networks: The sparsely-gated mixture-of-experts layer,” *arXiv preprint arXiv:1701.06538*, 2017.
- [17] C. Riquelme, J. Puigcerver, B. Mustafa, M. Neumann, R. Jenatton, A. Susano Pinto, D. Keysers, and N. Houlsby, “Scaling vision with sparse mixture of experts,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 8583–8595, 2021.
- [18] L. Wu, M. Liu, Y. Chen, D. Chen, X. Dai, and L. Yuan, “Residual mixture of experts,” *arXiv preprint arXiv:2204.09636*, 2022.
- [19] Z. Fan, R. Sarkar, Z. Jiang, T. Chen, K. Zou, Y. Cheng, C. Hao, Z. Wang, *et al.*, “M<sup>3</sup>vit: Mixture-of-experts vision transformer for efficient multi-task learning with model-accelerator co-design,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 28441–28457, 2022.
- [20] N. Du, Y. Huang, A. M. Dai, S. Tong, D. Lepikhin, Y. Xu, M. Krikun, Y. Zhou, A. W. Yu, O. Firat, *et al.*, “Glam: Efficient scaling of language models with mixture-of-experts,” in *International conference on machine learning*, pp. 5547–5569, PMLR, 2022.
- [21] 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.
- [22] A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, D. d. I. Casas, E. B. Hanna, F. Bressand, *et al.*, “Mixtral of experts,” *arXiv preprint arXiv:2401.04088*, 2024.

- [23] B. Li, Y. Shen, J. Yang, Y. Wang, J. Ren, T. Che, J. Zhang, and Z. Liu, "Sparse mixture-of-experts are domain generalizable learners," *arXiv preprint arXiv:2206.04046*, 2022.
- [24] Z. Chi, L. Dong, S. Huang, D. Dai, S. Ma, B. Patra, S. Singhal, P. Bajaj, X. Song, X.-L. Mao, *et al.*, "On the representation collapse of sparse mixture of experts," *Advances in Neural Information Processing Systems*, vol. 35, pp. 34600–34613, 2022.
- [25] H. Nguyen, P. Akbarian, T. Pham, T. Nguyen, S. Zhang, and N. Ho, "Statistical advantages of perturbing cosine router in sparse mixture of experts," *arXiv preprint arXiv:2405.14131*, 2024.
- [26] L. Xu, M. Jordan, and G. E. Hinton, "An alternative model for mixtures of experts," *Advances in neural information processing systems*, vol. 7, 1994.
- [27] J. Puigcerver, C. Riquelme, B. Mustafa, and N. Houlsby, "From sparse to soft mixtures of experts," *arXiv preprint arXiv:2308.00951*, 2023.
- [28] J. Geweke and M. Keane, "Smoothly mixing regressions," *Journal of Econometrics*, vol. 138, no. 1, pp. 252–290, 2007.
- [29] Z. Chen, Y. Deng, Y. Wu, Q. Gu, and Y. Li, "Towards understanding mixture of experts in deep learning," *arXiv preprint arXiv:2208.02813*, 2022.
- [30] H. Li, S. Lin, L. Duan, Y. Liang, and N. B. Shroff, "Theory on mixture-of-experts in continual learning," *arXiv preprint arXiv:2406.16437*, 2024.
- [31] R. Csordas, P. Piekos, K. Irie, and J. Schmidhuber, "Switchhead: Accelerating transformers with mixture-of-experts attention," *arXiv preprint arXiv:2312.07987*, 2023.
- [32] K.-C. Wang, D. Ostashev, Y. Fang, S. Tulyakov, and K. Aberman, "Moa: Mixture-of-attention for subject-context disentanglement in personalized image generation," in *SIGGRAPH Asia 2024 Conference Papers*, SA '24, (New York, NY, USA), Association for Computing Machinery, 2024.
- [33] P. Jin, B. Zhu, L. Yuan, and S. Yan, "Moh: Multi-head attention as mixture-of-head attention," *arXiv preprint arXiv:2410.11842*, 2024.
- [34] M. J. Jung and J. Kim, "Pmoe: Progressive mixture of experts with asymmetric transformer for continual learning," *arXiv preprint arXiv:2407.21571*, 2024.
- [35] S. Pavlitska, C. Hubschneider, L. Struppek, and J. M. Zöllner, "Sparsely-gated mixture-of-expert layers for CNN interpretability," in *2023 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–10, IEEE.
- [36] Y. Yi, K.-Y. Chen, and H.-Y. Gu, "Mixture of CNN experts from multiple acoustic feature domain for music genre classification," in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 1250–1255, IEEE.
- [37] L. Zhang, S. Huang, and W. Liu, "Enhancing mixture-of-experts by leveraging attention for fine-grained recognition," vol. 24, pp. 4409–4421.
- [38] L. Zhang, S. Huang, W. Liu, and D. Tao, "Learning a mixture of granularity-specific experts for fine-grained categorization," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8330–8339, IEEE.
- [39] S. Gross, M. Ranzato, and A. Szlam, "Hard mixtures of experts for large scale weakly supervised vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6865–6873, 2017.
- [40] J. Zhu, X. Zhu, W. Wang, X. Wang, H. Li, X. Wang, and J. Dai, "Uni-perceiver-moe: Learning sparse generalist models with conditional moes," *Advances in Neural Information Processing Systems*, vol. 35, pp. 2664–2678, 2022.
- [41] Z. Su, Z. Lin, X. Bai, X. Wu, Y. Xiong, H. Lian, G. Ma, H. Chen, G. Ding, W. Zhou, *et al.*, "Maskmoe: Boosting token-level learning via routing mask in mixture-of-experts," *arXiv preprint arXiv:2407.09816*, 2024.
- [42] E. Pedicir, L. Miller, and L. Robinson, "Novel token-level recurrent routing for enhanced mixture-of-experts performance," *Authorea Preprints*, 2024.
- [43] Y. Li, S. Jiang, B. Hu, L. Wang, W. Zhong, W. Luo, L. Ma, and M. Zhang, "Uni-moe: Scaling unified multimodal llms with mixture of experts," *arXiv preprint arXiv:2405.11273*, 2024.
- [44] S. Kudugunta, Y. Huang, A. Bapna, M. Krikun, D. Lepikhin, M.-T. Luong, and O. Firat, "Beyond distillation: Task-level mixture-of-experts for efficient inference," *arXiv preprint arXiv:2110.03742*, 2021.
- [45] X. Shi, S. Wang, Y. Nie, D. Li, Z. Ye, Q. Wen, and M. Jin, "Time-moe: Billion-scale time series foundation models with mixture of experts," *arXiv preprint arXiv:2409.16040*, 2024.
- [46] Q. Huang, Z. An, N. Zhuang, M. Tao, C. Zhang, Y. Jin, K. Xu, L. Chen, S. Huang, and Y. Feng, "Harder tasks need more experts: Dynamic routing in moe models," *arXiv preprint arXiv:2403.07652*, 2024.
- [47] A. Wang, X. Sun, R. Xie, S. Li, J. Zhu, Z. Yang, P. Zhao, J. Han, Z. Kang, D. Wang, *et al.*, "Hmoe: Heterogeneous mixture of experts for language modeling," *arXiv preprint arXiv:2408.10681*, 2024.
- [48] O. Irsoy and E. Alpaydin, "Dropout regularization in hierarchical mixture of experts," *Neurocomputing*, vol. 419, pp. 148–156, 2021.
- [49] J. He, J. Zhai, T. Antunes, H. Wang, F. Luo, S. Shi, and Q. Li, "Faster-moe: modeling and optimizing training of large-scale dynamic pre-trained models," in *Proceedings of the 27th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming*, pp. 120–134, 2022.
- [50] C. Hwang, W. Cui, Y. Xiong, Z. Yang, Z. Liu, H. Hu, Z. Wang, R. Salas, J. Jose, P. Ram, *et al.*, "Tutel: Adaptive mixture-of-experts at scale," *Proceedings of Machine Learning and Systems*, vol. 5, pp. 269–287, 2023.
- [51] R. Sarkar, H. Liang, Z. Fan, Z. Wang, and C. Hao, "Edge-moe: Memory-efficient multi-task vision transformer architecture with task-level sparsity via mixture-of-experts," in *2023 IEEE/ACM International Conference on Computer Aided Design (ICCAD)*, pp. 01–09, IEEE, 2023.
- [52] 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," in *International conference on machine learning*, pp. 18332–18346, PMLR, 2022.
- [53] A. Liu, B. Feng, B. Xue, B. Wang, B. Wu, C. Lu, C. Zhao, C. Deng, C. Zhang, C. Ruan, *et al.*, "Deepseek-v3 technical report," *arXiv preprint arXiv:2412.19437*, 2024.
- [54] S. Singh, O. Ruwase, A. A. Awan, S. Rajbhandari, Y. He, and A. Bhatle, "A hybrid tensor-expert-data parallelism approach to optimize mixture-of-experts training," in *Proceedings of the 37th International Conference on Supercomputing*, pp. 203–214, 2023.
- [55] J. Yao, Q. Anthony, A. Shafi, H. Subramoni, and D. K. D. Panda, "Exploiting inter-layer expert affinity for accelerating mixture-of-experts model inference," in *2024 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, pp. 915–925, IEEE, 2024.
- [56] J. Yu, Y. Zhuge, L. Zhang, P. Hu, D. Wang, H. Lu, and Y. He, "Boosting continual learning of vision-language models via mixture-of-experts adapters," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23219–23230, June 2024.
- [57] R. Zhang, A. Cheng, Y. Luo, G. Dai, H. Yang, J. Liu, R. Xu, L. Du, Y. Du, Y. Jiang, *et al.*, "Decomposing the neurons: Activation sparsity via mixture of experts for continual test time adaptation," *arXiv preprint arXiv:2405.16486*, 2024.
- [58] Y. Zhou, T. Lei, H. Liu, N. Du, Y. Huang, V. Zhao, A. M. Dai, Q. V. Le, J. Laudon, *et al.*, "Mixture-of-experts with expert choice routing," *Advances in Neural Information Processing Systems*, vol. 35, pp. 7103–7114, 2022.
- [59] W. Chen, Y. Zhou, N. Du, Y. Huang, J. Laudon, Z. Chen, and C. Cui, "Lifelong language pretraining with distribution-specialized experts," in *International Conference on Machine Learning*, pp. 5383–5395, PMLR, 2023.
- [60] S. Lee, J. Ha, D. Zhang, and G. Kim, "A neural dirichlet process mixture model for task-free continual learning," *arXiv preprint arXiv:2001.00689*, 2020.
- [61] G. Rypešć, S. Cygert, V. Khan, T. Trzciński, B. Zieliński, and B. Twardowski, "Divide and not forget: Ensemble of selectively trained experts in continual learning," *arXiv preprint arXiv:2401.10191*, 2024.
- [62] X. Yu, T. Rosing, and Y. Guo, "Evolve: Enhancing unsupervised continual learning with multiple experts," in *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2355–2366, IEEE.
- [63] L. Li, Z. Wu, and Y. Ji, "Mote: Mixture of task-specific experts for pre-trained model-based continual learning," *Available at SSRN 5035279*, 2024.
- [64] S. Park, "Learning more generalized experts by merging experts in mixture-of-experts," *arXiv preprint arXiv:2405.11530*, 2024.
- [65] H. Hihn and D. A. Braun, "Mixture-of-variational-experts for continual learning," *arXiv preprint arXiv:2110.12667*, 2021.
- [66] M. Le, H. Nguyen, T. Nguyen, T. Pham, L. Ngo, N. Ho, *et al.*, "Mixture of experts meets prompt-based continual learning," *Advances in Neural Information Processing Systems*, vol. 37, pp. 119025–119062, 2024.
- [67] S. Lee and C. Park, "Continual traffic forecasting via mixture of experts," *arXiv preprint arXiv:2406.03140*, 2024.

- [68] M. Wang, H. Su, S. Wang, S. Wang, N. Yin, L. Shen, L. Lan, L. Yang, and X. Cao, "Graph convolutional mixture-of-experts learner network for long-tailed domain generalization," pp. 1–1.
- [69] Q. Chen, L. Zhu, H. He, X. Zhang, S. Zeng, Q. Ren, and Y. Lu, "Low-rank mixture-of-experts for continual medical image segmentation," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024* (M. G. Linguraru, Q. Dou, A. Feragen, S. Giannarou, B. Glocker, K. Lekadir, and J. A. Schnabel, eds.), vol. 15008, pp. 382–392, Springer Nature Switzerland. Series Title: Lecture Notes in Computer Science.
- [70] Z. Wang, Z. Zhang, C.-Y. Lee, H. Zhang, R. Sun, X. Ren, G. Su, V. Perot, J. Dy, and T. Pfister, "Learning to prompt for continual learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 139–149, 2022.
- [71] R. Aljundi, P. Chakravarty, and T. Tuytelaars, "Expert gate: Lifelong learning with a network of experts," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3366–3375, 2017.
- [72] L. Wang, X. Zhang, Q. Li, J. Zhu, and Y. Zhong, "Coscl: Cooperation of small continual learners is stronger than a big one," in *European Conference on Computer Vision*, pp. 254–271, Springer, 2022.
- [73] T. Doan, S. I. Mirzadeh, and M. Farajtabar, "Continual learning beyond a single model," in *Conference on Lifelong Learning Agents*, pp. 961–991, PMLR, 2023.
- [74] T. Zhong, Z. Chi, L. Gu, Y. Wang, Y. Yu, and J. Tang, "Meta-dmoe: Adapting to domain shift by meta-distillation from mixture-of-experts," *Advances in Neural Information Processing Systems*, vol. 35, pp. 22243–22257, 2022.
- [75] Y. Dai, X. Li, J. Liu, Z. Tong, and L.-Y. Duan, "Generalizable person re-identification with relevance-aware mixture of experts," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16140–16149, IEEE.
- [76] Z. Liu, X. Wang, C. Feng, X. Sun, W. Zhan, and X. Chen, "Meta-reinforcement learning with mixture of experts for generalizable multi access in heterogeneous wireless networks," *arXiv preprint arXiv:2412.03850*, 2024.
- [77] J. Guo, D. J. Shah, and R. Barzilay, "Multi-source domain adaptation with mixture of experts," *arXiv preprint arXiv:1809.02256*, 2018.
- [78] Q. Zhou, K.-Y. Zhang, T. Yao, R. Yi, S. Ding, and L. Ma, "Adaptive mixture of experts learning for generalizable face anti-spoofing," in *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 6009–6018.
- [79] R. D. Nzoyem, D. A. Barton, and T. Deakin, "Towards foundational models for dynamical system reconstruction: Hierarchical meta-learning via mixture of experts," *arXiv preprint arXiv:2502.05335*, 2025.
- [80] A. Hendawy, J. Peters, and C. D'Eramo, "Multi-task reinforcement learning with mixture of orthogonal experts," *arXiv preprint arXiv:2311.11385*, 2023.
- [81] Y. Yang, P.-T. Jiang, Q. Hou, H. Zhang, J. Chen, and B. Li, "Multi-task dense prediction via mixture of low-rank experts," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 27927–27937, 2024.
- [82] L. Shen, A. Tang, E. Yang, G. Guo, Y. Luo, L. Zhang, X. Cao, B. Du, and D. Tao, "Efficient and effective weight-ensembling mixture of experts for multi-task model merging," *arXiv preprint arXiv:2410.21804*, 2024.
- [83] Z. Qin, Y. Cheng, Z. Zhao, Z. Chen, D. Metzler, and J. Qin, "Multitask mixture of sequential experts for user activity streams," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 3083–3091, ACM.
- [84] R. Aoki, F. Tung, and G. L. Oliveira, "Heterogeneous multi-task learning with expert diversity," vol. 19, no. 6, pp. 3093–3102.
- [85] G. Cheng, L. Dong, W. Cai, and C. Sun, "Multi-task reinforcement learning with attention-based mixture of experts," *IEEE Robotics and Automation Letters*, vol. 8, no. 6, pp. 3812–3819, 2023.
- [86] H. Ye and D. Xu, "Taskexpert: Dynamically assembling multi-task representations with memorial mixture-of-experts," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21828–21837, October 2023.
- [87] S. Gupta, S. Mukherjee, K. Subudhi, E. Gonzalez, D. Jose, A. H. Awadallah, and J. Gao, "Sparsely activated mixture-of-experts are robust multi-task learners," *arXiv preprint arXiv:2204.07689*, 2022.
- [88] H. Hazimeh, Z. Zhao, A. Chowdhery, M. Sathiamoorthy, Y. Chen, R. Mazumder, L. Hong, and E. Chi, "Dselect-k: Differentiable selection in the mixture of experts with applications to multi-task learning," *Advances in Neural Information Processing Systems*, vol. 34, pp. 29335–29347, 2021.
- [89] Z. Chen, Y. Shen, M. Ding, Z. Chen, H. Zhao, E. Learned-Miller, and C. Gan, "Mod-squad: Designing mixtures of experts as modular multi-task learners," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11828–11837, IEEE.
- [90] M. Reuss, J. Pari, P. Agrawal, and R. Lioutikov, "Efficient diffusion transformer policies with mixture of expert denoisers for multitask learning," *arXiv preprint arXiv:2412.12953*, 2024.
- [91] Z. Zhang, S. Liu, J. Yu, Q. Cai, X. Zhao, C. Zhang, Z. Liu, Q. Liu, H. Zhao, L. Hu, P. Jiang, and K. Gai, "M3oe: Multi-domain multi-task mixture-of experts recommendation framework," in *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 893–902.
- [92] J. Xu, L. Sun, and D. Zhao, "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*, pp. 2527–2531, ACM.
- [93] J. Ma, Z. Zhao, X. Yi, J. Chen, L. Hong, and E. H. Chi, "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*, pp. 1930–1939, ACM.
- [94] Z. Hou and D. Cao, "Estimating elastic parameters from digital rock images based on multi-task learning with multi-gate mixture-of-experts," vol. 213, p. 110310.
- [95] S. Wang, Y. Li, H. Li, T. Zhu, Z. Li, and W. Ou, "Multi-task learning with calibrated mixture of insightful experts," in *2022 IEEE 38th International Conference on Data Engineering (ICDE)*, pp. 3307–3319, IEEE.
- [96] S. Sodhani, A. Zhang, and J. Pineau, "Multi-task reinforcement learning with context-based representations," in *International Conference on Machine Learning*, pp. 9767–9779, PMLR, 2021.
- [97] T. Chen, X. Chen, X. Du, A. Rashwan, F. Yang, H. Chen, Z. Wang, and Y. Li, "AdaMV-MoE: Adaptive multi-task vision mixture-of-experts," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17300–17311, IEEE.
- [98] H. Tang, J. Liu, M. Zhao, and X. Gong, "Progressive layered extraction (ple): A novel multi-task learning (mtl) model for personalized recommendations," in *Proceedings of the 14th ACM conference on recommender systems*, pp. 269–278, 2020.
- [99] C. Louizos, M. Welling, and D. P. Kingma, "Learning sparse neural networks through  $l_0$  regularization," *arXiv preprint arXiv:1712.01312*, 2017.
- [100] J. Ren, Y. Li, Z. Ding, W. Pan, and H. Dong, "Probabilistic mixture-of-experts for efficient deep reinforcement learning," *arXiv preprint arXiv:2104.09122*, 2021.
- [101] M. Gimelfarb, S. Sanner, and C.-G. Lee, "Contextual policy transfer in reinforcement learning domains via deep mixtures-of-experts," in *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence* (C. de Campos and M. H. Maathuis, eds.), vol. 161 of *Proceedings of Machine Learning Research*, pp. 1787–1797, PMLR, 27–30 Jul 2021.
- [102] T. Willi, J. Obando-Ceron, J. Foerster, K. Dziugaite, and P. S. Castro, "Mixture of experts in a mixture of rl settings," *arXiv preprint arXiv:2406.18420*, 2024.
- [103] S. Huang, Z. Zhang, T. Liang, Y. Xu, Z. Kou, C. Lu, G. Xu, Z. Xue, and H. Xu, "Mentor: Mixture-of-experts network with task-oriented perturbation for visual reinforcement learning," *arXiv preprint arXiv:2410.14972*, 2024.
- [104] G. Qiao, G. Liu, P. Poupart, and Z. Xu, "Multi-modal inverse constrained reinforcement learning from a mixture of demonstrations," *Advances in Neural Information Processing Systems*, vol. 36, pp. 60384–60396, 2023.
- [105] D. Gupta, Y. Chow, A. Tulepbergenov, M. Ghavamzadeh, and C. Boutilier, "Offline reinforcement learning for mixture-of-expert dialogue management," *Advances in Neural Information Processing Systems*, vol. 36, pp. 5912–5935, 2023.
- [106] K. Samejima, K. Doya, and M. Kawato, "Inter-module credit assignment in modular reinforcement learning," *Neural Networks*, vol. 16, no. 7, pp. 985–994, 2003.
- [107] H. Van Seijen, B. Bakker, L. Kester, et al., "Switching between different state representations in reinforcement learning," in *Proceedings of the 26th IASTED International Conference on Artificial Intelligence and Applications*, pp. 226–231, 2008.
- [108] X. B. Peng, G. Berseth, and M. Van de Panne, "Terrain-adaptive locomotion skills using deep reinforcement learning," *ACM Transactions on Graphics (TOG)*, vol. 35, no. 4, pp. 1–12, 2016.

- [109] V. C. Kumar, S. Ha, and C. K. Liu, "Learning a unified control policy for safe falling," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3940–3947, IEEE, 2017.
- [110] X. B. Peng, M. Chang, G. Zhang, P. Abbeel, and S. Levine, "Mcp: Learning composable hierarchical control with multiplicative compositional policies," *Advances in neural information processing systems*, vol. 32, 2019.
- [111] R. Akrou, D. Tateo, and J. Peters, "Continuous action reinforcement learning from a mixture of interpretable experts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, pp. 6795–6806, 2021.
- [112] K. S. Refaat and K. Ding, "Accelerated deep reinforcement learning of terrain-adaptive locomotion skills," in *Deep RL Workshop NeurIPS 2021*, 2021.
- [113] W. Song, H. Zhao, P. Ding, C. Cui, S. Lyu, Y. Fan, and D. Wang, "Germ: A generalist robotic model with mixture-of-experts for quadruped robot," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 11879–11886, IEEE, 2024.
- [114] M. Vincze, L. Ferrarotti, L. L. Custode, B. Lepri, and G. Iacca, "Smose: Sparse mixture of shallow experts for interpretable reinforcement learning in continuous control tasks," *arXiv preprint arXiv:2412.13053*, 2024.
- [115] J. Obando-Ceron, G. Sokar, T. Willi, C. Lyle, J. Farebrother, J. Foerster, G. K. Dziugaite, D. Precup, and P. S. Castro, "Mixtures of experts unlock parameter scaling for deep rl," *arXiv preprint arXiv:2402.08609*, 2024.
- [116] Y. Takahashi and M. Asada, "Modular learning systems for soccer robot," in *Proceedings of the Fourth International Symposium on Human and Artificial Intelligence Systems*, pp. 370–375, Citeseer, 2004.
- [117] Y. Takahashi, K. Edazawa, and M. Asada, "Modular learning system and scheduling for behavior acquisition in multi-agent environment," in *RoboCup 2004: Robot Soccer World Cup VIII* 8, pp. 548–555, Springer, 2005.
- [118] Y. Takahashi, K. Edazawa, K. Noma, and M. Asada, "Simultaneous learning to acquire competitive behaviors in multi-agent system based on a modular learning system," in *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 2016–2022, IEEE, 2005.
- [119] K. Mülling, J. Kober, O. Kroemer, and J. Peters, "Learning to select and generalize striking movements in robot table tennis," *The International Journal of Robotics Research*, vol. 32, no. 3, pp. 263–279, 2013.
- [120] K. Li, M. Cucuringu, L. Sánchez-Betancourt, and T. Willi, "Mixtures of experts for scaling up neural networks in order execution," in *Proceedings of the 5th ACM International Conference on AI in Finance*, pp. 669–676, 2024.
- [121] M. Ewerton, G. Neumann, R. Lioutikov, H. B. Amor, J. Peters, and G. Maeda, "Learning multiple collaborative tasks with a mixture of interaction primitives," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1535–1542, IEEE, 2015.
- [122] Y. Zhou, J. Gao, and T. Asfour, "Movement primitive learning and generalization: Using mixture density networks," *IEEE Robotics & Automation Magazine*, vol. 27, no. 2, pp. 22–32, 2020.
- [123] N. Freymuth, N. Schreiber, P. Becker, A. Taranovic, and G. Neumann, "Inferring versatile behavior from demonstrations by matching geometric descriptors," *arXiv preprint arXiv:2210.08121*, 2022.
- [124] V. Prasad, A. Kshirsagar, D. K. R. Stock-Homburg, J. Peters, and G. Chalkatzaki, "Moveint: Mixture of variational experts for learning human-robot interactions from demonstrations," *IEEE Robotics and Automation Letters*, 2024.
- [125] K. Doya, K. Samejima, K.-i. Katagiri, and M. Kawato, "Multiple model-based reinforcement learning," *Neural computation*, vol. 14, no. 6, pp. 1347–1369, 2002.
- [126] P. Tommasino, D. Caligiore, M. Mirolli, and G. Baldassarre, "A reinforcement learning architecture that transfers knowledge between skills when solving multiple tasks," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 11, no. 2, pp. 292–317, 2016.
- [127] M. N. R. Chowdhury, S. Zhang, M. Wang, S. Liu, and P.-Y. Chen, "Patch-level routing in mixture-of-experts is provably sample-efficient for convolutional neural networks," in *International Conference on Machine Learning*, pp. 6074–6114, PMLR, 2023.
- [128] W. Jiang and M. A. Tanner, "Hierarchical mixtures-of-experts for exponential family regression models: approximation and maximum likelihood estimation," *Annals of Statistics*, pp. 987–1011, 1999.
- [129] W. Jiang and M. A. Tanner, "On the approximation rate of hierarchical mixtures-of-experts for generalized linear models," *Neural computation*, vol. 11, no. 5, pp. 1183–1198, 1999.
- [130] A. J. Zeevi, R. Meir, and V. Maierov, "Error bounds for functional approximation and estimation using mixtures of experts," *IEEE Transactions on Information Theory*, vol. 44, no. 3, pp. 1010–1025, 1998.
- [131] H. D. Nguyen, L. R. Lloyd-Jones, and G. J. McLachlan, "A universal approximation theorem for mixture-of-experts models," *Neural computation*, vol. 28, no. 12, pp. 2585–2593, 2016.
- [132] E. F. Mendes and W. Jiang, "On convergence rates of mixtures of polynomial experts," *Neural computation*, vol. 24, no. 11, pp. 3025–3051, 2012.
- [133] N. Ho, C.-Y. Yang, and M. I. Jordan, "Convergence rates for gaussian mixtures of experts," *Journal of Machine Learning Research*, vol. 23, no. 323, pp. 1–81, 2022.
- [134] H. Nguyen, T. Nguyen, and N. Ho, "Demystifying softmax gating function in gaussian mixture of experts," *Advances in Neural Information Processing Systems*, vol. 36, pp. 4624–4652, 2023.
- [135] H. Nguyen, P. Akbarian, F. Yan, and N. Ho, "Statistical perspective of top-k sparse softmax gating mixture of experts," *arXiv preprint arXiv:2309.13850*, 2023.
- [136] H. Nguyen, T. Nguyen, K. Nguyen, and N. Ho, "Towards convergence rates for parameter estimation in gaussian-gated mixture of experts," in *International Conference on Artificial Intelligence and Statistics*, pp. 2683–2691, PMLR, 2024.
- [137] H. Nguyen, N. Ho, and A. Rinaldo, "On least square estimation in softmax gating mixture of experts," *arXiv preprint arXiv:2402.02952*, 2024.
- [138] H. Nguyen, P. Akbarian, T. Nguyen, and N. Ho, "A general theory for softmax gating multinomial logistic mixture of experts," *arXiv preprint arXiv:2310.14188*, 2023.
- [139] T. C. Fung and S. C. Tseung, "Mixture of experts models for multilevel data: Modelling framework and approximation theory," *Neurocomputing*, p. 129357, 2025.
- [140] M. Videau, A. Leite, M. Schoenauer, and O. Teytaud, "Mixture of experts in image classification: What's the sweet spot?," *arXiv preprint arXiv:2411.18322*, 2024.
- [141] X. Han, L. Wei, Z. Dou, Z. Wang, C. Qiang, X. He, Y. Sun, Z. Han, and Q. Tian, "Vimoe: An empirical study of designing vision mixture-of-experts," *arXiv preprint arXiv:2410.15732*, 2024.
- [142] A. Royer, I. Karmanov, A. Skliar, B. E. Bejnordi, and T. Blankevoort, "Revisiting single-gated mixtures of experts," *arXiv preprint arXiv:2304.05497*, 2023.
- [143] J. Zhang, X. Qu, T. Zhu, and Y. Cheng, "Clip-moe: Towards building mixture of experts for clip with diversified multiplet upcycling," *arXiv preprint arXiv:2409.19291*, 2024.
- [144] M. He, G. Lv, W. He, J. Fan, and G. Zeng, "DeepME: Deep mixture experts for large-scale image classification," in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pp. 722–728, International Joint Conferences on Artificial Intelligence Organization, 2024.
- [145] H. Nguyen, X. Han, C. W. Harris, S. Saria, and N. Ho, "On expert estimation in hierarchical mixture of experts: Beyond softmax gating functions," *arXiv preprint arXiv:2410.02935*, 2024.
- [146] K. Oksuz, S. Kuzucu, T. Joy, and P. K. Dokania, "Mocae: Mixture of calibrated experts significantly improves object detection," *arXiv preprint arXiv:2309.14976*, 2023.
- [147] X. Wang, Y. Jin, W. Wu, W. Zhang, L. Zhu, B. Jiang, and Y. Tian, "Object detection using event camera: A moe heat conduction based detector and a new benchmark dataset," *arXiv preprint arXiv:2412.06647*, 2024.
- [148] Y. Jain, H. Behl, Z. Kira, and V. Vineet, "Damex: Dataset-aware mixture-of-experts for visual understanding of mixture-of-datasets," *Advances in Neural Information Processing Systems*, vol. 36, 2024.
- [149] X. Feng, Y. Li, D. Chen, C. Qiao, J. Yuan, L. Yuan, and G. Hua, "Pluralistic salient object detection," *arXiv preprint arXiv:2409.02368*, 2024.
- [150] S. Pavlitska, E. Eisen, and J. M. Zöllner, "Towards adversarial robustness of model-level mixture-of-experts architectures for semantic segmentation," *arXiv preprint arXiv:2412.11608*, 2024.
- [151] C. Zhu, B. Xiao, L. Shi, S. Xu, and X. Zheng, "Customize segment anything model for multi-modal semantic segmentation with mixture of lora experts," *arXiv preprint arXiv:2412.04220*, 2024.
- [152] S. Pavlitskaya, C. Hubschneider, M. Weber, R. Moritz, F. Huger, P. Schlicht, and J. M. Zollner, "Using mixture of expert models to gain insights into semantic segmentation," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1399–1406, IEEE.

- [153] X. Wang, F. Yu, L. Dunlap, Y.-A. Ma, R. Wang, A. Mirhoseini, T. Darrell, and J. E. Gonzalez, "Deep mixture of experts via shallow embedding," in *Uncertainty in artificial intelligence*, pp. 552–562, PMLR, 2020.
- [154] L. Rossi, V. Bernuzzi, T. Fontanini, M. Bertozzi, and A. Prati, "Swin2-moe: A new single image supersolution model for remote sensing," *IET Image Processing*, vol. 19, no. 1, p. e13303, 2025.
- [155] Z. Xue, G. Song, Q. Guo, B. Liu, Z. Zong, Y. Liu, and P. Luo, "Raphael: Text-to-image generation via large mixture of diffusion paths," *Advances in Neural Information Processing Systems*, vol. 36, pp. 41693–41706, 2023.
- [156] D. K. Park, S. Yoo, H. Bahng, J. Choo, and N. Park, "Megan: Mixture of experts of generative adversarial networks for multimodal image generation," *arXiv preprint arXiv:1805.02481*, 2018.
- [157] Y. Jiang, S. Yang, H. Qiu, W. Wu, C. C. Loy, and Z. Liu, "Text2human: Text-driven controllable human image generation," *ACM Transactions on Graphics (TOG)*, vol. 41, no. 4, pp. 1–11, 2022.
- [158] H. Zhou, Z. Wang, S. Huang, X. Huang, X. Han, J. Feng, C. Deng, W. Luo, and J. Chen, "Moe-lpr: Multilingual extension of large language models through mixture-of-experts with language priors routing," *arXiv preprint arXiv:2408.11396*, 2024.
- [159] X. Cheng, Z. Zhu, X. Zhuang, Z. Chen, Z. Huang, and Y. Zou, "MoE-SLU: Towards ASR-robust spoken language understanding via mixture-of-experts," in *Findings of the Association for Computational Linguistics ACL 2024*, pp. 14868–14879, Association for Computational Linguistics.
- [160] Z. Wu, H.-Y. Huang, F. Qu, and Y. Wu, "Mixture-of-prompt-experts for multi-modal semantic understanding," *arXiv preprint arXiv:2403.11311*, 2024.
- [161] Y. Chai, Q. Yin, and J. Zhang, "Improved training of mixture-of-experts language gans," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023.
- [162] Y. Zhang, S. Sun, X. Gao, Y. Fang, C. Brockett, M. Galley, J. Gao, and B. Dolan, "Retgen: A joint framework for retrieval and grounded text generation modeling," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, pp. 11739–11747, 2022.
- [163] J. Wu and M. Hou, "Enhancing diversity for logical <span style='font-variant:small-caps;'>table-to-text</span> generation with mixture of experts," vol. 41, no. 4, p. e13533.
- [164] E. Frantar and D. Alistarh, "Qmoe: Practical sub-1-bit compression of trillion-parameter models," *arXiv preprint arXiv:2310.16795*, 2023.
- [165] M. R. Costa-Jussà, J. Cross, O. Çelebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi, J. Lam, D. Licht, J. Maillard, et al., "No language left behind: Scaling human-centered machine translation," *arXiv preprint arXiv:2207.04672*, 2022.
- [166] NLLB Team, M. R. Costa-jussà, J. Cross, O. Çelebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi, J. Lam, D. Licht, J. Maillard, A. Sun, S. Wang, G. Wenzek, A. Youngblood, B. Akula, L. Barrault, G. M. Gonzalez, P. Hansanti, J. Hoffman, S. Jarrett, K. R. Sadagopan, D. Rowe, S. Spruit, C. Tran, P. Andrews, N. F. Ayan, S. Bhosale, S. Edunov, A. Fan, C. Gao, V. Goswami, F. Guzmán, P. Koehn, A. Mourachko, C. Ropers, S. Saleem, H. Schwenk, and J. Wang, "Scaling neural machine translation to 200 languages," vol. 630, no. 8018, pp. 841–846.
- [167] H. Huang, N. Ardalani, A. Sun, L. Ke, H.-H. S. Lee, A. Sridhar, S. Bhosale, C.-J. Wu, and B. Lee, "Towards moe deployment: Mitigating inefficiencies in mixture-of-expert (moe) inference," *arXiv preprint arXiv:2303.06182*, 2023.
- [168] S. Chen, Z. Jie, and L. Ma, "Llava-mole: Sparse mixture of lora experts for mitigating data conflicts in instruction finetuning mllms," *arXiv preprint arXiv:2401.16160*, 2024.
- [169] B. Mustafa, C. Riquelme, J. Puigcerver, R. Jenatton, and N. Houlsby, "Multimodal contrastive learning with limoe: the language-image mixture of experts," *Advances in Neural Information Processing Systems*, vol. 35, pp. 9564–9576, 2022.
- [170] X. Sun, Y. Chen, Y. Huang, R. Xie, J. Zhu, K. Zhang, S. Li, Z. Yang, J. Han, X. Shu, et al., "Hunyuan-large: An open-source moe model with 52 billion activated parameters by tencent," *arXiv preprint arXiv:2411.02265*, 2024.
- [171] A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei, et al., "Qwen2.5 technical report," *arXiv preprint arXiv:2412.15115*, 2024.
- [172] T. Wei, B. Zhu, L. Zhao, C. Cheng, B. Li, W. Lü, P. Cheng, J. Zhang, X. Zhang, L. Zeng, et al., "Skywork-moe: A deep dive into training techniques for mixture-of-experts language models, 2024," URL <https://arxiv.org/abs/2406.06563>, 2024.
- [173] N. Gupta and J. Yip, "Dbrx: Creating an llm from scratch using databricks," in *Databricks Data Intelligence Platform: Unlocking the GenAI Revolution*, pp. 311–330, Springer, 2024.
- [174] B. Zoph, I. Bello, S. Kumar, N. Du, Y. Huang, J. Dean, N. Shazeer, and W. Fedus, "St-moe: Designing stable and transferable sparse expert models," *arXiv preprint arXiv:2202.08906*, 2022.
- [175] T. Zhu, X. Qu, D. Dong, J. Ruan, J. Tong, C. He, and Y. Cheng, "Llama-moe: Building mixture-of-experts from llama with continual pre-training," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 15913–15923, 2024.
- [176] X. Qu, D. Dong, X. Hu, T. Zhu, W. Sun, and Y. Cheng, "Llama-moe v2: Exploring sparsity of llama from perspective of mixture-of-experts with post-training," *arXiv preprint arXiv:2411.15708*, 2024.
- [177] A. Vavre, E. He, D. Liu, Z. Yan, J. Yang, N. Tajbakhsh, and A. Aithal, "Llama 3 meets moe: Efficient upcycling," *arXiv preprint arXiv:2412.09952*, 2024.
- [178] F. Xue, Z. Zheng, Y. Fu, J. Ni, Z. Zheng, W. Zhou, and Y. You, "Open-moe: An early effort on open mixture-of-experts language models," *arXiv preprint arXiv:2402.01739*, 2024.
- [179] B. Lenz, O. Lieber, A. Araz, A. Bergman, A. Manevich, B. Peleg, B. Aviram, C. Almagor, C. Fridman, D. Padnos, et al., "Jamba: Hybrid transformer-mamba language models," in *The Thirteenth International Conference on Learning Representations*, 2025.
- [180] J. Team, B. Lenz, A. Araz, A. Bergman, A. Manevich, B. Peleg, B. Aviram, C. Almagor, C. Fridman, D. Padnos, et al., "Jamba-1.5: Hybrid transformer-mamba models at scale," *arXiv preprint arXiv:2408.12570*, 2024.
- [181] O. Lieber, B. Lenz, H. Bata, G. Cohen, J. Osin, I. Dalmedigos, E. Safahi, S. Meirum, Y. Belinkov, S. Shalev-Shwartz, et al., "Jamba: A hybrid transformer-mamba language model," *arXiv preprint arXiv:2403.19887*, 2024.
- [182] S. Wu, J. Luo, X. Chen, L. Li, X. Zhao, T. Yu, C. Wang, Y. Wang, F. Wang, W. Qiao, et al., "Yuan 2.0-m32: Mixture of experts with attention router," *arXiv preprint arXiv:2405.17976*, 2024.
- [183] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of machine learning research*, vol. 21, no. 140, pp. 1–67, 2020.
- [184] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, "SuperGlue: Learning feature matching with graph neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4938–4947, 2020.
- [185] D. Dai, C. Deng, C. Zhao, R. Xu, H. Gao, D. Chen, J. Li, W. Zeng, X. Yu, Y. Wu, et al., "Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models," *arXiv preprint arXiv:2401.06066*, 2024.
- [186] DeepSeek-AI, A. Liu, B. Feng, B. Wang, B. Wang, B. Liu, C. Zhao, C. Deng, C. Ruan, D. Dai, D. Guo, D. Yang, D. Chen, D. Ji, E. Li, F. Lin, F. Luo, G. Hao, G. Chen, G. Li, H. Zhang, H. Xu, H. Yang, H. Zhang, H. Ding, H. Xin, H. Gao, H. Li, H. Qu, J. L. Cai, J. Liang, J. Guo, J. Ni, J. Li, J. Chen, J. Yuan, J. Qiu, J. Song, K. Dong, K. Gao, K. Guan, L. Wang, L. Zhang, L. Xu, L. Xia, L. Zhao, L. Zhang, M. Li, M. Wang, M. Zhang, M. Zhang, M. Tang, M. Li, N. Tian, P. Huang, P. Wang, P. Zhang, Q. Zhu, Q. Chen, Q. Du, R. J. Chen, R. L. Jin, R. Ge, R. Pan, R. Xu, R. Chen, S. S. Li, S. Lu, S. Zhou, S. Chen, S. Wu, S. Ye, S. Ma, S. Wang, S. Zhou, S. Yu, S. Zhou, S. Zheng, T. Wang, T. Pei, T. Yuan, T. Sun, W. L. Xiao, W. Zeng, W. An, W. Liu, W. Liang, W. Gao, W. Zhang, X. Q. Li, X. Jin, X. Wang, X. Bi, X. Liu, X. Wang, X. Shen, X. Chen, X. Chen, X. Nie, X. Sun, X. Wang, X. Liu, X. Xie, X. Yu, X. Song, X. Zhou, X. Yang, X. Lu, X. Su, Y. Wu, Y. K. Li, Y. X. Wei, Y. X. Zhu, Y. Xu, Y. Huang, Y. Li, Y. Zhao, Y. Sun, Y. Li, Y. Wang, Y. Zheng, Y. Zhang, Y. Xiong, Y. Zhao, Y. He, Y. Tang, Y. Piao, Y. Dong, Y. Tan, Y. Liu, Y. Wang, Y. Guo, Y. Zhu, Y. Wang, Y. Zou, Y. Zha, Y. Ma, Y. Yan, Y. You, Y. Liu, Z. Z. Ren, Z. Ren, Z. Sha, Z. Fu, Z. Huang, Z. Zhang, Z. Xie, Z. Hao, Z. Shao, Z. Wen, Z. Xu, Z. Zhang, Z. Li, Z. Wang, Z. Gu, Z. Li, and Z. Xie, "DeepSeek-v2: A strong, economical, and efficient mixture-of-experts language model," *arXiv preprint arXiv:2405.04434*, no. arXiv:2405.04434.
- [187] DeepSeek-AI, D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi, X. Zhang, X. Yu, Y. Wu, Z. F. Wu, Z. Gou, Z. Shao, Z. Li, Z. Gao, A. Liu, B. Xue, B. Wang, B. Wu, B. Feng, C. Lu, C. Zhao, C. Deng, C. Zhang, C. Ruan, D. Dai, D. Chen, D. Ji, E. Li, F. Lin, F. Dai, F. Luo, G. Hao, G. Chen, G. Li, H. Zhang, H. Bao, H. Xu, H. Wang, H. Ding, H. Xin, H. Gao, H. Qu, H. Li, J. Guo, J. Li, J. Wang, J. Chen, J. Yuan, J. Qiu, J. Li, J. L. Cai, J. Ni, J. Liang, J. Chen, K. Dong, K. Hu, K. Gao, K. Guan, K. Huang, K. Yu, L. Wang, L. Zhang, L. Zhao, L. Wang, L. Zhang, L. Xu, L. Xia,

- M. Zhang, M. Zhang, M. Tang, M. Li, M. Wang, M. Li, N. Tian, P. Huang, P. Zhang, Q. Wang, Q. Chen, Q. Du, R. Ge, R. Zhang, R. Pan, R. Wang, R. J. Chen, R. L. Jin, R. Chen, S. Lu, S. Zhou, S. Chen, S. Ye, S. Wang, S. Yu, S. Zhou, S. Pan, S. S. Li, S. Zhou, S. Wu, S. Ye, T. Yun, T. Pei, T. Sun, T. Wang, W. Zeng, W. Zhao, W. Liu, W. Liang, W. Gao, W. Yu, W. Zhang, W. L. Xiao, W. An, X. Liu, X. Wang, X. Chen, X. Nie, X. Cheng, X. Liu, X. Xie, X. Liu, X. Yang, X. Li, X. Su, X. Lin, X. Q. Li, X. Jin, X. Shen, X. Chen, X. Sun, X. Wang, X. Song, X. Zhou, X. Wang, X. Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. Zhang, Y. Xu, Y. Li, Y. Zhao, Y. Sun, Y. Wang, Y. Yu, Y. Zhang, Y. Shi, Y. Xiong, Y. He, Y. Piao, Y. Wang, Y. Tan, Y. Ma, Y. Liu, Y. Guo, Y. Ou, Y. Wang, Y. Gong, Y. Zou, Y. He, Y. Xiong, Y. Luo, Y. You, Y. Liu, Y. Zhou, Y. X. Zhu, Y. Xu, Y. Huang, Y. Li, Y. Zheng, Y. Zhu, Y. Ma, Y. Tang, Y. Zha, Y. Yan, Z. Z. Ren, Z. Ren, Z. Sha, Z. Fu, Z. Xu, Z. Xie, Z. Zhang, Z. Hao, Z. Ma, Z. Yan, Z. Wu, Z. Gu, Z. Zhu, Z. Liu, Z. Li, Z. Xie, Z. Song, Z. Pan, Z. Huang, Z. Xu, Z. Zhang, and Z. Zhang, “DeepSeek-r1: Incentivizing reasoning capability in LLMs via reinforcement learning,” *arXiv preprint arXiv:2501.12948*, no. arXiv:2501.12948, 2025.
- [188] R. Akrou, D. Tateo, and J. Peters, “Continuous action reinforcement learning from a mixture of interpretable experts,” vol. 44, no. 10, pp. 6795–6806.
- [189] T. C. Fung, A. L. Badescu, and X. S. Lin, “A class of mixture of experts models for general insurance: Theoretical developments,” vol. 89, pp. 111–127.
- [190] S. E. Yuksel, J. N. Wilson, and P. D. Gader, “Twenty years of mixture of experts,” *IEEE transactions on neural networks and learning systems*, vol. 23, no. 8, pp. 1177–1193, 2012.
- [191] S. Masoudnia and R. Ebrahimpour, “Mixture of experts: a literature survey,” *Artificial Intelligence Review*, vol. 42, pp. 275–293, 2014.
- [192] W. Cai, J. Jiang, F. Wang, J. Tang, S. Kim, and J. Huang, “A survey on mixture of experts,” *arXiv preprint arXiv:2407.06204*, 2024.
- [193] W. Gan, Z. Ning, Z. Qi, and P. S. Yu, “Mixture of experts (moe): A big data perspective,” *arXiv preprint arXiv:2501.16352*, 2025.
- [194] J. Liu, P. Tang, W. Wang, Y. Ren, X. Hou, P.-A. Heng, M. Guo, and C. Li, “A survey on inference optimization techniques for mixture of experts models,” *arXiv preprint arXiv:2412.14219*, 2024.
- [195] A. Vats, R. Raja, V. Jain, and A. Chadha, “The evolution of mixture of experts: A survey from basics to breakthroughs,” *Preprints (August 2024)*, 2024.
- [196] V. Dimitri, B. Regina, and M. Alfonz, “A survey on mixture of experts: Advancements, challenges, and future directions,” *Authorea Preprints*, 2025.
- [197] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.
- [198] S. Ingrassia, S. C. Minotti, and G. Vittadini, “Local statistical modeling via a cluster-weighted approach with elliptical distributions,” *Journal of classification*, vol. 29, pp. 363–401, 2012.
- [199] J. Geweke, M. Keane, and D. Runkle, “Alternative computational approaches to inference in the multinomial probit model,” *The review of economics and statistics*, pp. 609–632, 1994.
- [200] M. P. Keane, “A note on identification in the multinomial probit model,” *Journal of Business & Economic Statistics*, vol. 10, no. 2, pp. 193–200, 1992.
- [201] W. Peebles and S. Xie, “Scalable diffusion models with transformers,” in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.
- [202] B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, et al., “Language models are few-shot learners,” *arXiv preprint arXiv:2005.14165*, vol. 1, p. 3, 2020.
- [203] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, et al., “Lora: Low-rank adaptation of large language models,” *ICLR*, vol. 1, no. 2, p. 3, 2022.
- [204] S. Lin, L. Yang, D. Fan, and J. Zhang, “Trgp: Trust region gradient projection for continual learning,” *arXiv preprint arXiv:2202.02931*, 2022.
- [205] S. Lin, L. Yang, D. Fan, and J. Zhang, “Beyond not-forgetting: Continual learning with backward knowledge transfer,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 16165–16177, 2022.
- [206] R. Aljundi, K. Kelchtermans, and T. Tuytelaars, “Task-free continual learning,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11254–11263, 2019.
- [207] D.-W. Zhou, H.-L. Sun, J. Ning, H.-J. Ye, and D.-C. Zhan, “Continual learning with pre-trained models: A survey,” *arXiv preprint arXiv:2401.16386*, 2024.
- [208] Q. Sun, Y. Liu, T.-S. Chua, and B. Schiele, “Meta-transfer learning for few-shot learning,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 403–412, 2019.
- [209] Y. Chen, Z. Liu, H. Xu, T. Darrell, and X. Wang, “Meta-baseline: Exploring simple meta-learning for few-shot learning,” in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9062–9071, 2021.
- [210] M. Ren, E. Triantafillou, S. Ravi, J. Snell, K. Swersky, J. B. Tenenbaum, H. Larochelle, and R. S. Zemel, “Meta-learning for semi-supervised few-shot classification,” *arXiv preprint arXiv:1803.00676*, 2018.
- [211] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap, “Meta-learning with memory-augmented neural networks,” in *International conference on machine learning*, pp. 1842–1850, PMLR, 2016.
- [212] J. Snell, K. Swersky, and R. Zemel, “Prototypical networks for few-shot learning,” *Advances in neural information processing systems*, vol. 30, 2017.
- [213] C. Finn, P. Abbeel, and S. Levine, “Model-agnostic meta-learning for fast adaptation of deep networks,” in *International conference on machine learning*, pp. 1126–1135, PMLR, 2017.
- [214] A. Nichol and J. Schulman, “Reptile: a scalable metalearning algorithm,” *arXiv preprint arXiv:1803.02999*, vol. 2, no. 3, p. 4, 2018.
- [215] M. Garnelo, J. Schwarz, D. Rosenbaum, F. Viola, D. J. Rezende, S. Eslami, and Y. W. Teh, “Neural processes,” *arXiv preprint arXiv:1807.01622*, 2018.
- [216] N. Göring, F. Hess, M. Brenner, Z. Monfared, and D. Durstewitz, “Out-of-domain generalization in dynamical systems reconstruction,” *arXiv preprint arXiv:2402.18377*, 2024.
- [217] R. D. Nzoym, D. A. Barton, and T. Deakin, “Extending contextual self-modulation: Meta-learning across modalities, task dimensionalities, and data regimes,” *arXiv preprint arXiv:2410.01655*, 2024.
- [218] A. Farahani, S. Voghoci, K. Rasheed, and H. R. Arabnia, “A brief review of domain adaptation,” *Advances in data science and information engineering: proceedings from ICDATA 2020 and IKE 2020*, pp. 877–894, 2021.
- [219] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra, et al., “Matching networks for one shot learning,” *Advances in neural information processing systems*, vol. 29, 2016.
- [220] K. Rakelly, A. Zhou, C. Finn, S. Levine, and D. Quillen, “Efficient off-policy meta-reinforcement learning via probabilistic context variables,” in *International conference on machine learning*, pp. 5331–5340, PMLR, 2019.
- [221] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, “Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor,” in *International conference on machine learning*, pp. 1861–1870, Pmlr, 2018.
- [222] R. Caruana, “Multitask learning,” *Machine learning*, vol. 28, pp. 41–75, 1997.
- [223] K.-H. Thung and C.-Y. Wee, “A brief review on multi-task learning,” *Multimedia Tools and Applications*, vol. 77, no. 22, pp. 29705–29725, 2018.
- [224] Y. Zhang and Q. Yang, “An overview of multi-task learning,” *National Science Review*, vol. 5, no. 1, pp. 30–43, 2018.
- [225] N. Vithayathil Varghese and Q. H. Mahmoud, “A survey of multi-task deep reinforcement learning,” *Electronics*, vol. 9, no. 9, p. 1363, 2020.
- [226] C. D’Eramo, D. Tateo, A. Bonarini, M. Restelli, and J. Peters, “Sharing knowledge in multi-task deep reinforcement learning,” *arXiv preprint arXiv:2401.09561*, 2024.
- [227] L. Sun, H. Zhang, W. Xu, and M. Tomizuka, “Paco: Parameter-compositional multi-task reinforcement learning,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 21495–21507, 2022.
- [228] C. Devin, A. Gupta, T. Darrell, P. Abbeel, and S. Levine, “Learning modular neural network policies for multi-task and multi-robot transfer,” in *2017 IEEE international conference on robotics and automation (ICRA)*, pp. 2169–2176, IEEE, 2017.
- [229] R. Yang, H. Xu, Y. Wu, and X. Wang, “Multi-task reinforcement learning with soft modularization,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 4767–4777, 2020.
- [230] Q. Zhang, X. Liao, Q. Liu, J. Xu, and B. Zheng, “Leaving no one behind: A multi-scenario multi-task meta learning approach for advertiser modeling,” in *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pp. 1368–1376, 2022.
- [231] A. Graves and A. Graves, “Long short-term memory,” *Supervised sequence labelling with recurrent neural networks*, pp. 37–45, 2012.



- [232] R. Cui, D. Cao, Q. Liu, Z. Zhu, and Y. Jia, "Vp and vs prediction from digital rock images using a combination of u-net and convolutional neural networks," *Geophysics*, vol. 86, no. 1, pp. MR27–MR37, 2021.
- [233] S. Karimpouli, S. Khoshlesan, E. H. Saenger, and H. H. Koochi, "Application of alternative digital rock physics methods in a real case study: a challenge between clean and cemented samples," *Geophysical Prospecting*, vol. 66, no. 4, pp. 767–783, 2018.
- [234] J. Kober, J. A. Bagnell, and J. Peters, "Reinforcement learning in robotics: A survey," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1238–1274, 2013.
- [235] I. Szita, "Reinforcement learning in games," in *Reinforcement Learning: State-of-the-art*, pp. 539–577, Springer, 2012.
- [236] B. R. Kiran, I. Sobh, V. Talpaert, P. Mannion, A. A. Al Sallab, S. Yogamani, and P. Pérez, "Deep reinforcement learning for autonomous driving: A survey," *IEEE transactions on intelligent transportation systems*, vol. 23, no. 6, pp. 4909–4926, 2021.
- [237] C. Yu, J. Liu, S. Nemati, and G. Yin, "Reinforcement learning in healthcare: A survey," *ACM Computing Surveys (CSUR)*, vol. 55, no. 1, pp. 1–36, 2021.
- [238] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.
- [239] V. Feinberg, A. Wan, I. Stoica, M. I. Jordan, J. E. Gonzalez, and S. Levine, "Model-based value estimation for efficient model-free reinforcement learning," *arXiv preprint arXiv:1803.00101*, 2018.
- [240] S. Levine, A. Kumar, G. Tucker, and J. Fu, "Offline reinforcement learning: Tutorial, review, and perspectives on open problems," *arXiv preprint arXiv:2005.01643*, 2020.
- [241] O. Celik, A. Taranovic, and G. Neumann, "Acquiring diverse skills using curriculum reinforcement learning with mixture of experts," *arXiv preprint arXiv:2403.06966*, 2024.
- [242] T. Kurutach, I. Clavera, Y. Duan, A. Tamar, and P. Abbeel, "Model-ensemble trust-region policy optimization," *arXiv preprint arXiv:1802.10592*, 2018.
- [243] K. Lee, M. Laskin, A. Srinivas, and P. Abbeel, "Sunrise: A simple unified framework for ensemble learning in deep reinforcement learning," in *International Conference on Machine Learning*, pp. 6131–6141, PMLR, 2021.
- [244] Z. Yang, K. Ren, X. Luo, M. Liu, W. Liu, J. Bian, W. Zhang, and D. Li, "Towards applicable reinforcement learning: Improving the generalization and sample efficiency with policy ensemble," *arXiv preprint arXiv:2205.09284*, 2022.
- [245] X. Lou, J. Guo, J. Zhang, J. Wang, K. Huang, and Y. Du, "Pecan: Leveraging policy ensemble for context-aware zero-shot human-ai coordination," *arXiv preprint arXiv:2301.06387*, 2023.
- [246] H. Mohammadi, E. Nazerfard, and T. Firoozi, "Reinforcement learning-based mixture of vision transformers for video violence recognition," *arXiv preprint arXiv:2310.03108*, 2023.
- [247] Z. Wei, D. Chen, Y. Zhang, D. Wen, X. Nie, and L. Xie, "Deep reinforcement learning portfolio model based on mixture of experts," vol. 55, no. 5, p. 347.
- [248] D. Eigen, M. Ranzato, and I. Sutskever, "Learning factored representations in a deep mixture of experts," *arXiv preprint arXiv:1312.4314*, 2013.
- [249] M. Aitkin and N. Longford, "Statistical modelling issues in school effectiveness studies," *Journal of the Royal Statistical Society: Series A (General)*, vol. 149, no. 1, pp. 1–26, 1986.
- [250] H. Goldstein, "Multilevel mixed linear model analysis using iterative generalized least squares," *Biometrika*, vol. 73, no. 1, pp. 43–56, 1986.
- [251] H. Goldstein, *Multilevel statistical models*. John Wiley & Sons, 2011.
- [252] M. McCloskey and N. J. Cohen, "Catastrophic interference in connectionist networks: The sequential learning problem," in *Psychology of learning and motivation*, vol. 24, pp. 109–165, Elsevier, 1989.
- [253] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, et al., "Overcoming catastrophic forgetting in neural networks," *Proceedings of the national academy of sciences*, vol. 114, no. 13, pp. 3521–3526, 2017.
- [254] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," *Advances in neural information processing systems*, vol. 30, 2017.
- [255] X. Zhou, V. Koltun, and P. Krähenbühl, "Simple multi-dataset detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7571–7580, 2022.
- [256] X. Wang, Z. Cai, D. Gao, and N. Vasconcelos, "Towards universal object detection by domain attention," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7289–7298, 2019.
- [257] M. Ding, B. Xiao, N. Codella, P. Luo, J. Wang, and L. Yuan, "Davvit: Dual attention vision transformers," in *European conference on computer vision*, pp. 74–92, Springer, 2022.
- [258] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?," *Advances in neural information processing systems*, vol. 30, 2017.
- [259] M. Abbasi and C. Gagné, "Robustness to adversarial examples through an ensemble of specialists," *arXiv preprint arXiv:1702.06856*, 2017.
- [260] S. Kariyappa and M. Qureshi, "Improving adversarial robustness of ensembles with diversity training. 2019," *Available: arXiv*, 1901.
- [261] T. Pang, K. Xu, C. Du, N. Chen, and J. Zhu, "Improving adversarial robustness via promoting ensemble diversity," in *International Conference on Machine Learning*, pp. 4970–4979, PMLR, 2019.
- [262] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.
- [263] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al., "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4681–4690, 2017.
- [264] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer gan to bridge domain gap for person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 79–88, 2018.
- [265] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.
- [266] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.
- [267] S. Gu, D. Chen, J. Bao, F. Wen, B. Zhang, D. Chen, L. Yuan, and B. Guo, "Vector quantized diffusion model for text-to-image synthesis," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10696–10706, 2022.
- [268] P. Esser, R. Rombach, A. Blattmann, and B. Ommer, "Imagebart: Bidirectional context with multinomial diffusion for autoregressive image synthesis," *Advances in neural information processing systems*, vol. 34, pp. 3518–3532, 2021.
- [269] S. Bond-Taylor, P. Hessey, H. Sasaki, T. P. Breckon, and C. G. Willcocks, "Unleashing transformers: Parallel token prediction with discrete absorbing diffusion for fast high-resolution image generation from vector-quantized codes," in *European Conference on Computer Vision*, pp. 170–188, Springer, 2022.
- [270] C. Manning and H. Schütze, *Foundations of statistical natural language processing*. MIT press, 1999.
- [271] D. Jurafsky, *Speech & language processing*. Pearson Education India, 2000.
- [272] J. Hirschberg and C. D. Manning, "Advances in natural language processing," *Science*, vol. 349, no. 6245, pp. 261–266, 2015.
- [273] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent trends in deep learning based natural language processing," *IEEE Computational Intelligence Magazine*, vol. 13, no. 3, pp. 55–75, 2018.
- [274] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.
- [275] D. Adiwardana, M.-T. Luong, D. R. So, J. Hall, N. Fiedel, R. Thoppilan, Z. Yang, A. Kulshreshtha, G. Nemade, Y. Lu, et al., "Towards a human-like open-domain chatbot," *arXiv preprint arXiv:2001.09977*, 2020.
- [276] L. Zhang, S. Wang, and B. Liu, "Deep learning for sentiment analysis: A survey," *Wiley interdisciplinary reviews: data mining and knowledge discovery*, vol. 8, no. 4, p. e1253, 2018.
- [277] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural architectures for named entity recognition," *arXiv preprint arXiv:1603.01360*, 2016.
- [278] D. Yu and L. Deng, *Automatic speech recognition*, vol. 1. Springer, 2016.
- [279] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

- [280] T.-H. Wen, M. Gasic, N. Mrksic, P.-H. Su, D. Vandyke, and S. Young, "Semantically conditioned lstm-based natural language generation for spoken dialogue systems," *arXiv preprint arXiv:1508.01745*, 2015.
- [281] E. Reiter and R. Dale, "Building applied natural language generation systems," *Natural Language Engineering*, vol. 3, no. 1, pp. 57–87, 1997.
- [282] A. M. Rush, S. Chopra, and J. Weston, "A neural attention model for abstractive sentence summarization," *arXiv preprint arXiv:1509.00685*, 2015.
- [283] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.
- [284] P. F. Brown, J. Cocke, S. A. Della Pietra, V. J. Della Pietra, F. Jelinek, J. Lafferty, R. L. Mercer, and P. S. Roossin, "A statistical approach to machine translation," *Computational linguistics*, vol. 16, no. 2, pp. 79–85, 1990.
- [285] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, *et al.*, "Google's neural machine translation system: Bridging the gap between human and machine translation," *arXiv preprint arXiv:1609.08144*, 2016.