
MoEMeta: Mixture-of-Experts Meta Learning for Few-Shot Relational Learning

Han Wu^{1,2}, Jie Yin^{1*}

¹The University of Sydney, Australia

²Peking University, China

{han.wu, jie.yin@sydney.edu.au}

Abstract

Few-shot knowledge graph relational learning seeks to perform reasoning over relations given only a limited number of training examples. While existing approaches largely adopt a meta-learning framework for enabling fast adaptation to new relations, they suffer from two key pitfalls. First, they learn relation meta-knowledge in isolation, failing to capture common relational patterns shared across tasks. Second, they struggle to effectively incorporate local, task-specific contexts crucial for rapid adaptation. To address these limitations, we propose MoEMeta, a novel meta-learning framework that disentangles globally shared knowledge from task-specific contexts to enable both effective model generalization and rapid adaptation. MoEMeta introduces two key innovations: (i) a mixture-of-experts (MoE) model that learns globally shared relational prototypes to enhance generalization, and (ii) a task-tailored adaptation mechanism that captures local contexts for fast task-specific adaptation. By balancing global generalization with local adaptability, MoEMeta significantly advances few-shot relational learning. Extensive experiments and analyses on three KG benchmarks show that MoEMeta consistently outperforms existing baselines, achieving state-of-the-art performance.¹

1 Introduction

Knowledge graphs (KGs) encode real-world knowledge in the form of factual triplets (h, r, t) , where each triplet represents a relationship r between a head entity h and a tail entity t (Bollacker et al., 2008; Suchanek et al., 2007; Vrandečić and Krötzsch, 2014). KGs are crucial for powering various intelligent applications, such as question answering (Wu et al., 2016), Web search (Fang et al., 2017), and advanced recommendation systems (Cao et al., 2019). Despite their widespread adoption, real-world KGs are inherently incomplete. Although numerous KG completion methods (Bordes et al., 2013; Yang et al., 2015; Trouillon et al., 2016) have been proposed to infer missing facts based on observed triplets, their effectiveness is severely hindered by data sparsity, especially for long-tail or emerging relations with only a handful of known training triplets.

In response, *few-shot relational learning* (FSRL) has emerged as a promising direction for reasoning over few-shot relations with limited training triplets, enabling generalization to novel relations in data-scarce scenarios. Most existing FSRL methods (Chen et al., 2019; Jiang et al., 2021; Niu et al., 2021; Wu et al., 2023) are built upon model-agnostic meta-learning (MAML) (Finn et al., 2017), a gradient-based meta-learning framework that learns a set of global parameters as an initialization for rapid task adaptation using only one or a few steps of gradient descent updates. In FSRL, each task is relation-centric, comprising only a few training triplets of a single relation, and the objective is to

*Corresponding author.

¹The code is available at: <https://github.com/alexhw15/MoEMeta>.

predict instances of novel relations not seen during meta-training. These methods focus on learning and encoding meta-knowledge among tasks to support rapid adaptation with minimal supervision.

Limitations of prior work. Despite promising results, current meta-learning based FSRL methods face two key limitations. First, they learn about meta-knowledge within individual tasks while neglecting shared patterns that exist across tasks. This limitation arises from the assumption that meta-training tasks are independent and identically distributed (i.i.d.), where a base model is adapted to each meta-training task using its support set, and then updated based on the performance on its query set. However, KG relations are heterogeneous and often semantically clustered, with some relations more closely related than others. For example, on `Nell-One` (Mitchell et al., 2018), `FatherOfPerson` and `BrotherOf` share a strong thematic overlap (family ties), whereas `ColorOf` pertains to a different type of attribute (physical properties). Ignoring such underlying commonalities among related tasks hinders model generalization in few-shot scenarios. While accounting for task variance has been explored in other few-shot domains like image classification (Li et al., 2019; Zhou et al., 2021), it remains largely unaddressed in the unique, non-i.i.d. context of relational learning.

Second, current methods are limited by their reliance on a single set of global parameters to encode meta-knowledge. With a shared initialization, a uniform adaptation process lacks the flexibility to capture the unique characteristics of individual tasks from only a few support triplets. This limitation is critical because relations and entities in KGs exhibit diverse interaction patterns (Wang et al., 2014; Ji et al., 2015), such as one-to-one, one-to-many, many-to-one, and many-to-many. An entity often reveals entirely different aspects depending on its relational context. For example, the entity `Elon Musk` participates in the `CeoOf` relation (a professional role) and the `FatherOfPerson` relation (a family role). Methods constrained by a shared initialization struggle to adapt to such locally distant contexts, and failing to modulate meta-knowledge for a specific task results in suboptimal adaptation.

Present work and contributions. To overcome these limitations, we propose **MoEMeta**, a novel meta-learning framework to tackle a critical, non-trivial challenge not addressed by prior FSRL work: *how to disentangle globally shared knowledge from task-specific contexts to achieve effective generalization and rapid adaptation under few-shot settings*. MoEMeta provides a principled solution through two novel components that work in synergy.

- **Global knowledge generalization:** To capture **what to generalize**, we introduce a **mixture-of-experts (MoE) model** to dynamically learn common relational patterns shared across tasks, referred to as **relational prototypes**, from a global pool of experts. The MoE is globally optimized, enabling the dynamic selection of experts that specialize in modeling the compositional nature of relations in KGs—an explicit and novel design for enhancing generalization.
- **Local context adaptation:** To address **what to adapt**, we propose a novel **task-tailored adaptation** mechanism that leverages task-specific structural contexts to fine-tune relation-meta and entity embeddings for local adaptation. Unlike general task-aware modulation methods (Vuorio et al., 2019; Abdollahzadeh et al., 2021), which rely on task mode prediction to modulate the meta-learned prior parameters, our adaptation operates at the embedding level, offering a lightweight yet effective way to capture diverse entity-relation interactions unique to individual tasks.

Together, these two innovations strike an effective balance between global knowledge generalization and fine-grained task-specific adaptation, which is crucial for FSRL with limited supervision. Experimental results demonstrate that MoEMeta achieves state-of-the-art performance on three KG benchmarks, significantly outperforming existing methods. In-depth ablation studies and visualizations further validate the efficacy of each component in MoEMeta.

2 Related Work

Few-Shot Relation Learning. Existing approaches to FSRL can be categorized into *metric-learning methods* and *meta-learning methods*. Metric-learning methods learn a similarity metric between the support and query sets for adaptation. Approaches in this category include GMatching (Xiong et al., 2018), designed for one-shot relation learning; FSKGC (Zhang et al., 2020), which extends to few-shot settings using a recurrent autoencoder to aggregate support instances; FAAN (Sheng et al., 2020), which employs a dynamic attention mechanism for improved one-hop neighbor aggregation; CSR (Huang et al., 2022), which utilizes subgraph matching for adaptation; and NP-FKGC (Luo et al., 2023), which leverages neural processes to address distribution shifts in unseen relations.

Most mainstream FSRL methods are based on meta-learning (Finn et al., 2017), which learns an initial meta-prior from existing tasks for rapid adaptation to new few-shot relations. MetaR (Chen et al., 2019) learns relation-specific meta-information by simply averaging support triplet representations. MetaP (Jiang et al., 2021) utilizes meta-pattern learning under one-shot settings. GANA (Niu et al., 2021) integrates meta-learning with TransH (Wang et al., 2014) and devises a gated, attentive neighbor aggregator to handle sparse neighborhood information. HiRe (Wu et al., 2023) exploits multi-granularity relational information to learn more generalizable meta-knowledge. RelAdapter (Ran et al., 2024) adopts a shallow feedforward network with a residual connection to adapt relation meta-knowledge within each target task. Despite their progress, these methods face two key limitations: (1) they often neglect common relational patterns when learning meta-knowledge during meta-training, and (2) they fail to effectively incorporate local, task-specific contexts essential for rapid adaptation. Our work addresses these critical gaps by proposing a principled meta-learning framework that disentangles globally shared knowledge from task-specific adaptation, enabling both effective generalization and fast adaptation in FSRL. For a more extensive review of related work, including KG embedding learning and general few-shot learning techniques, please refer to Appendix ??.

Mixture of Experts. The mixture-of-experts (MoE) architecture, originally introduced by Jacobs et al. (1991), is a modular machine learning framework designed to improve performance by dynamically combining multiple specialized models, known as “experts”. Sparse MoEs and their “soft” variants (Jordan and Jacobs, 1994; Puigcerver et al., 2024) have been widely adopted in deep learning, especially in large-scale models for natural language processing and computer vision, to improve model training efficiency and scalability through adaptive subnetwork selection or dynamic pruning (Shazeer et al., 2017; Du et al., 2022). In a related KG context, MoE has been applied to fuse different semantics of relations (structural ontology and textual descriptions) at the triplet level (Song et al., 2022). In contrast, our work innovatively proposes MoE as a meta-learner for FSRL that operates at the task level to learn specialized experts capturing diverse relational patterns, thereby enabling superior cross-task knowledge transfer and task adaptation.

3 Preliminaries

3.1 FSRL Problem Formulation

A knowledge graph (KG) can be represented as a collection of factual triplets $\mathcal{G} = \{(h, r, t) \in \mathcal{E} \times \mathcal{R} \times \mathcal{E}\}$, where $\mathcal{E}$ denotes the set of entities, and $\mathcal{R}$ denotes the set of relations. Each triplet (h, r, t) indicates a relationship r between the head entity h and the tail entity t .

Given a novel relation $r \notin \mathcal{R}$ and its corresponding support set $\mathcal{S}_r = \{(h_i, r, t_i)\}_{i=1}^K$, where $h_i, t_i \in \mathcal{E}$, the objective of *few-shot relational learning* (FSRL) is to predict the missing tail entities in a query set $\mathcal{Q}_r = \{(h_j, r, ?)\}$, given a head entity $h_j \in \mathcal{E}$ and a relation r . For each query triplet $(h_j, r, ?) \in \mathcal{Q}_r$, a candidate set $\mathcal{C}_{h_j, r}$ is provided, and the goal is to rank the true tail entity highest among all candidates. The support set $\mathcal{S}_r$ and the query set $\mathcal{Q}_r$ jointly define a task $\mathcal{T}_r = (\mathcal{S}_r, \mathcal{Q}_r)$ for a relation r . In K -shot relation learning, the support set $\mathcal{S}_r$ contains only a few training triplets (typically $K = 1, 3, 5$).

3.2 Meta-Learning based FSRL

Given the strength of meta-learning in enabling rapid adaptation with minimal data, we build our approach upon a meta-learning based framework, involving two stages: meta-training and meta-testing. The model is expected to learn a meta-prior from seen tasks in the meta-training stage, which can then be adapted to novel tasks during the meta-testing stage.

Meta-Training Stage. During meta-training, an FSRL model is trained on a set of relations $\mathcal{R}_{tr}$ with the corresponding task data $\mathcal{D}_{tr} = \{\mathcal{T}_r \mid r \in \mathcal{R}_{tr}\}$. The model aims to learn a prior, which initializes a relation-meta learner for a new task. For each task $\mathcal{T}_r = (\mathcal{S}_r, \mathcal{Q}_r) \in \mathcal{D}_{tr}$, the learner extracts a task-specific relation-meta $\mathbf{R}_{\mathcal{T}_r} \in \mathbb{R}^d$ based on the support set $\mathcal{S}_r$, which is rapidly updated using a gradient meta $\mathbf{G}_{\mathcal{T}_r}$ computed from the loss on the support set:

$$\mathbf{R}_{\mathcal{T}_r} \leftarrow \mathbf{R}_{\mathcal{T}_r} - \alpha \mathbf{G}_{\mathcal{T}_r}, \quad (1)$$

where α is the learning rate. The updated relation-meta $\mathbf{R}_{\mathcal{T}_r}$ is then applied to the query set with the ground-truth tail entities, where the query loss is backpropagated to update both the prior and the entity embedding matrix.

Meta-Testing Stage. During meta-testing, the model is evaluated on novel relations $\mathcal{R}_{\text{te}}$ and their associated task data $\mathcal{D}_{\text{te}} = \{\mathcal{T}_r \mid r \in \mathcal{R}_{\text{te}}\}$, where $\mathcal{R}_{\text{tr}} \cap \mathcal{R}_{\text{te}} = \emptyset$. For each novel task $\mathcal{T}_r = (\mathcal{S}_r, \mathcal{Q}_r) \in \mathcal{D}_{\text{te}}$, the model uses the learned prior to initialize the relation-meta learner and compute an initial relation-meta $\mathbf{R}_{\mathcal{T}_r}$, which is then adapted using its support set $\mathcal{S}_r$ to obtain $\mathbf{R}'_{\mathcal{T}_r}$. Finally, for each query triplet $(h, r, ?)$, the model ranks all candidate tail entities t in the candidate set $\mathcal{C}_{h,r}$ by computing a score $s(h, \mathbf{R}'_{\mathcal{T}_r}, t)$ for each candidate $t \in \mathcal{C}_{h,r}$ and sorting them in ascending order.

4 The Proposed MoEMeta Framework

This section presents our proposed **MoEMeta** framework. Its design is underpinned by two key insights: First, relations in real-world KGs exhibit heterogeneous characteristics, with some subsets of relations sharing greater similarity and stronger interconnections; Second, relations in KGs manifest diverse interaction patterns, such as one-to-many, many-to-one, and many-to-many. Building upon these insights, we introduce MoEMeta, a novel meta-learning framework that (1) utilizes a MoE model to dynamically learn relational prototypes and generate relation-meta for each task; and (2) incorporates a task-tailored adaptation mechanism to facilitate local adaptation to each task.

MoEMeta consists of three core components: (1) attentive neighbor aggregation, (2) MoE-based meta-knowledge learning, and (3) task-tailored local adaptation. Figure 1 illustrates the detailed process of calculating the loss on the support set, through these three components. In essence, MoEMeta(Φ, η) contains two sets of learnable parameters: the parameters in attentive neighbor aggregation and MoE-based meta-knowledge learning are globally optimized during meta-training, denoted as Φ ; while the parameters in task-tailored local adaptation are locally optimized within each task, denoted as η .

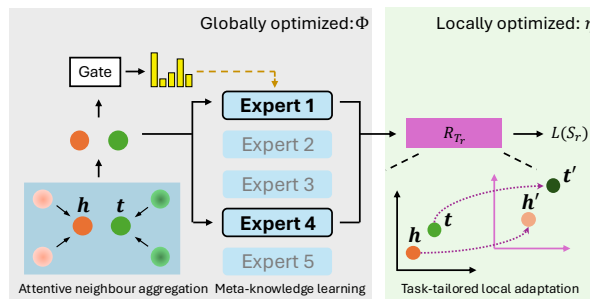


Figure 1: The computation of the support loss. The support loss $\mathcal{L}(\mathcal{S}_r)$ is used to update relation-meta $\mathbf{R}_{\mathcal{T}_r}$, which is applied to the query set for model optimization.

4.1 Attentive Neighbor Aggregation

Given a K -shot task $\mathcal{T}_r$ with its support set $\mathcal{S}_r = \{(h_i, r, t_i)\}_{i=1}^K$, we propose an attentive neighbor encoder that aggregates the neighborhood information for both head entity h_i and tail entity t_i . Specifically, for each entity, the encoder processes its neighboring tuples (neighboring entities and relations), combined with its own embeddings, to generate an enriched representation.

For a given entity e , its one-hop neighbors are represented as a set of neighboring tuples $\{(r_i, e'_i)\}_{i=1}^n$, where n is the total number of neighbors, e'_i is the neighboring entity and r_i denotes the relation between e'_i and e . We first concatenate the embeddings of r_i and e'_i to form the embedding of the i -th tuple, $\mathbf{c}_i = \mathbf{r}_i \oplus \mathbf{e}'_i$, where $\oplus$ denotes concatenation of two vectors. To capture interactions among neighbors, we then apply a linear transformation followed by a non-linear activation function:

$$\mathbf{c}'_i = \text{ReLU}(\mathbf{W}\mathbf{c}_i), \quad (2)$$

where $\mathbf{W}$ is a learnable weight matrix. To enable the encoder to attend to informative neighbors, we compute an attention-like gate value g_i for each neighboring tuple using a sigmoid function:

$$g_i = \sigma(\beta^T \mathbf{c}'_i), \quad (3)$$

where β is a learnable parameter vector. The gate values control the influence of different neighbors in the aggregation process. Finally, the weighted neighbor embeddings are aggregated and combined with the entity's own embedding $\mathbf{e}$ to obtain the final embedding:

$$\mathbf{e} = \text{Aggregate}(\{g_i \cdot \mathbf{c}'_i\}_{i=1}^n) + \mathbf{e}. \quad (4)$$

The neighbor encoder is applied to head entity h and tail entity t independently to ensure that the final embedding reflects both the entity's intrinsic information and its local neighborhood structure.

4.2 MoE-based Meta-Knowledge Learning

To overcome the limitation of existing methods that neglect shared meta-knowledge across related tasks, we introduce a globally learned MoE model that captures common relational patterns, termed *relational prototypes*. For each task, a subset of relevant experts is dynamically selected from a shared pool based on its support triplets to learn the relation-meta. This mechanism enables adaptive relation decomposition and expert specialization, facilitating the learning of diverse, compositional relational patterns and promoting effective generalization across tasks.

Specifically, we design a sparsely-gated MoE model consisting of M experts. Each expert within the MoE is architecturally identical to a standard MLP (multilayer perceptron). For a given support triplet, a gating mechanism dynamically routes the triplet to a selected subset of experts, assigning weights to each expert in the subset based on its relevance. The outputs of the top- N experts are then aggregated according to their weights, resulting in a representation of the relation associated with the triplet. To compute the relation-meta for a support set, we average the relation representations derived from all triplets within the support set. This approach enables the MoE to retrieve the most relevant meta-knowledge from the shared expert pool to model diverse relational patterns.

Given a K -shot task $\mathcal{T}_r$ and its support set $\mathcal{S}_r = \{(h_i, r, t_i)\}_{i=1}^K$, the head and tail entities of each triplet, h_i and t_i , are augmented via our attentive neighbor aggregator to obtain the embeddings $\mathbf{h}_i$ and $\mathbf{t}_i$. The relation representation $\mathbf{r}_i$ derived from (h_i, t_i) can be computed by our MoE as follows:

$$\mathbf{r}_i = \frac{1}{N} \sum_{j=1}^M g_{i,j} f_j(\mathbf{h}_i, \mathbf{t}_i; \boldsymbol{\theta}_j), \quad (5)$$

$$g_{i,j} = \begin{cases} s_{i,j}, & \text{if } s_{i,j} \in \text{Top}_i(\{s_{i,j} \mid 1 \leq j \leq M\}, N), \\ 0, & \text{otherwise,} \end{cases} \quad (6)$$

$$s_{i,j} = \text{softmax}(\text{Gate}(\mathbf{h}_i, \mathbf{t}_i; \boldsymbol{\theta}_g)), \quad (7)$$

where each $f_j(\cdot; \boldsymbol{\theta}_j)$ is an expert network and M denotes the total number of experts. For each support triplet i , a gating network $\text{Gate}(\cdot; \boldsymbol{\theta}_g)$ computes a relevance score $s_{i,j}$ for each expert j , conditioned on the corresponding inputs $\mathbf{h}_i$ and $\mathbf{t}_i$. These scores are then normalized via a softmax function over all M experts. The function $\text{Top}_i(\cdot, N)$ selects only the top N highest scores among M experts for each triplet i , setting the remaining scores to zero. This sparsity mechanism enables the model to selectively activate only a subset of experts for each support triplet, facilitating the learning of more specialized relational prototypes globally shared across tasks.

The final relation-meta is computed as the average of relation representations from all support triplets:

$$\mathbf{R}_{\mathcal{T}_r} = \frac{1}{K} \sum_{i=1}^K \mathbf{r}_i \quad (8)$$

4.3 Task-Tailored Local Adaptation

As discussed earlier, the global parameters alone— $\mathbf{W}$ and β in attentive neighbor aggregation, and $\{\boldsymbol{\theta}_i\}_{i=1}^M$ and $\boldsymbol{\theta}_g$ in meta-knowledge learning—may be insufficient for effective task-specific adaptation from only a few support triplets. The unique characteristics of individual tasks are difficult to capture with a shared initialization. To address this issue, we propose a dynamic learnable projection mechanism that facilitates local adaptation, allowing task-specific relational learning tailored to each task.

Specifically, for each task, MoEMeta maintains a task-specific set of local parameters, denoted as $\boldsymbol{\eta}$, namely three randomly initialized projection vectors, $\mathbf{p}_h$, $\mathbf{p}_r$, and $\mathbf{p}_t$, which project the embeddings $\mathbf{h}_i$, $\mathbf{R}_{\mathcal{T}_r}$, and $\mathbf{t}_i$ into a task-specific subspace. Inspired by TransD (Ji et al., 2015), which addresses the limitations of fixed mapping strategies using per-entity and per-relation projections, our approach leverages task-specific projections to avoid overparametrization under few-shot settings. As all triplets within a task share the same relation, we adopt a unified projection vector $\mathbf{p}_h$ for all head entities, and $\mathbf{p}_t$ for all tail entities within the task. This provides a lightweight solution for enabling efficient task-specific adaptation with minimal parameter overhead.

Unlike previous FSRL methods including RelAdapter (Ran et al., 2024), which do not account for diverse interactions between entities and relations tailored to different tasks, our approach allows for a relation-specific adaptation, with projection operations defined below:

$$\mathbf{h}'_i = \mathbf{h}_i + (\mathbf{p}_h^\top \mathbf{R}_{\mathcal{T}_r}) \cdot \mathbf{R}_{\mathcal{T}_r}, \quad (9)$$

$$\mathbf{R}'_{\mathcal{T}_r} = \mathbf{R}_{\mathcal{T}_r} + (\mathbf{p}_r^\top \mathbf{R}_{\mathcal{T}_r}) \cdot \mathbf{R}_{\mathcal{T}_r}, \quad (10)$$

$$\mathbf{t}'_i = \mathbf{t}_i + (\mathbf{p}_t^\top \mathbf{R}_{\mathcal{T}_r}) \cdot \mathbf{R}_{\mathcal{T}_r}. \quad (11)$$

Above, the dot product $\mathbf{p}_x^\top \mathbf{R}_{\mathcal{T}_r}$ yields a scalar alignment score between each projection vector $\mathbf{p}_x, x \in \{h, r, t\}$, and the relation-meta $\mathbf{R}_{\mathcal{T}_r}$. This scalar score is then used to scale $\mathbf{R}_{\mathcal{T}_r}$, yielding an offset vector that modulates corresponding features in the entity embeddings and relation-meta. The updated embeddings $(\mathbf{h}'_i, \mathbf{R}'_{\mathcal{T}_r}, \mathbf{t}'_i)$ thus capture task-specific nuances while maintaining consistency with the generalizable meta-knowledge. Although the three projection vectors could functionally be merged into a single perturbation term, maintaining separate projections for the head entity, relation-meta, and tail entity enhances modularity and flexibility. This design allows for more expressive or heterogeneous projection functions that can incorporate richer inductive biases or task-specific modeling needs in future extensions. Overall, this local adaptation scheme ensures that each task can adaptively focus on its unique aspects while preserving shared commonalities across related tasks.

To further optimize the adapted embeddings $\mathbf{h}'_i$, $\mathbf{R}'_{\mathcal{T}_r}$, and $\mathbf{t}'_i$, we define a scoring function in the projected space to measure the plausibility of each triplet (h_i, r, t_i) , which is given by:

$$\text{score}(h_i, t_i) = \|\mathbf{h}'_i + \mathbf{R}'_{\mathcal{T}_r} - \mathbf{t}'_i\|_2, \quad (12)$$

where $\|\cdot\|_2$ denotes the ℓ_2 norm. $\mathbf{h}'_i$, $\mathbf{R}'_{\mathcal{T}_r}$, and $\mathbf{t}'_i$ are the projected embeddings of the head entity, relation-meta, and tail entity, respectively. This formulation ensures that, in the task-specific projection space, the distance between the adapted head entity embedding $\mathbf{h}'_i$ and the adapted tail entity embedding $\mathbf{t}'_i$ is minimized after being translated by the relation-meta embedding $\mathbf{R}'_{\mathcal{T}_r}$.

Consider the entire support set, we define a margin-based loss function as follows:

$$\mathcal{L}(\mathcal{S}_r) = \sum_{(h_i, r, t_i) \in \mathcal{S}_r} \max\{0, \text{score}(h_i, t_i) + \gamma - \text{score}(h_i, t'_i)\}, \quad (13)$$

where γ is a margin hyperparameter, and (h_i, r, t'_i) represents a negative triplet generated by replacing the tail entity t_i with a randomly sampled entity $t'_i \in \mathcal{E}$ such that $(h_i, r, t'_i) \notin \mathcal{G}$.

Once the gradients of the support loss $\mathcal{L}(\mathcal{S}_r)$ are computed, we update the projection vectors $\mathbf{p}_h$, $\mathbf{p}_r$, $\mathbf{p}_t$, together with the relation-meta $\mathbf{R}_{\mathcal{T}_r}$, to ensure generalization to the support set. Specifically, the projection vectors can be updated in a unified manner as follows:

$$\mathbf{p}_x \leftarrow \mathbf{p}_x - \alpha \frac{\partial \mathcal{L}(\mathcal{S}_r)}{\partial \mathbf{p}_x}, \quad x \in \{h, r, t\}. \quad (14)$$

Similarly, the relation-meta $\mathbf{R}_{\mathcal{T}_r}$ is updated as:

$$\mathbf{R}_{\mathcal{T}_r} \leftarrow \mathbf{R}_{\mathcal{T}_r} - \alpha \frac{\partial \mathcal{L}(\mathcal{S}_r)}{\partial \mathbf{R}_{\mathcal{T}_r}}, \quad (15)$$

where α is the learning rate. These updates adapt the model using the support set, which is then evaluated on the query set $\mathcal{Q}_r$. For each query triplet $(h_q, r, t_q) \in \mathcal{Q}_r$, we evaluate its plausibility for each t_q from the candidate set $\mathcal{C}_{h_q, r}$, using the score function Eq. (12). Then, the query loss is similarly defined as:

$$\mathcal{L}(\mathcal{Q}_r) = \sum_{(h_q, r, t_q) \in \mathcal{Q}_r} \max\{0, \text{score}(h_q, t_q) + \gamma - \text{score}(h_q, t'_q)\}, \quad (16)$$

where (h_q, r, t'_q) is a negative triplet generated by corrupting the tail entity t_q with a randomly sampled entity $t'_q \in \mathcal{E}$. The optimization objective for training the entire model is to minimize $\mathcal{L}(\mathcal{Q}_r)$.

The task-specific projection vectors $\mathbf{p}_h$, $\mathbf{p}_r$, and $\mathbf{p}_t$ are randomly initialized and trained for each task on the support set $\mathcal{S}_r$, enabling effective local adaptation. This mechanism provides two key advantages. (1) Task-specific adaptation: By projecting embeddings into a task-specific latent space, the model captures fine-grained task-specific local contexts. (2) Complementarity to MoE: While the MoE captures globally shared relational prototypes, dynamic learnable projections provide finer-grained local adaptation, thereby improving model generalization under few-shot settings.

4.4 MoEMeta Meta-Training and Meta-Testing

During meta-training, MoEMeta optimizes two sets of parameters: global parameters $\Phi = \{\mathbf{W}, \beta, \{\theta\}_{i=1}^M, \theta_g\}$, for attentive neighbor aggregation and MoE-based meta-knowledge learning, which are globally optimized; and task-specific adaptation parameters $\eta = \{\mathbf{p}_h, \mathbf{p}_r, \mathbf{p}_t\}$, which are optimized locally for each task. The meta-training algorithm of MoEMeta is detailed in Algorithm 1.

During meta-testing, the learned global parameters Φ are used to initialize the model and remain frozen. For each novel task, only the local parameters η are fine-tuned on the support set before evaluation on the query set by ranking candidate tail entities, following the same inference procedure as in meta-training. The complexity analysis of MoEMeta’s components is provided in Appendix A.

Algorithm 1: Meta-training Algorithm of MoEMeta

Input: Meta-training task set $\mathcal{D}_t$, knowledge graph $\mathcal{G}$
Output: Optimized global parameters Φ

- 1 Initialize global parameters $\Phi = \{\mathbf{W}, \beta, \{\theta\}_{i=1}^M, \theta_g\}$;
- 2 **while** not converged **do**
- 3 Sample a task $\mathcal{T}_r = \{\mathcal{S}_r, \mathcal{Q}_r\} \in \mathcal{D}_t$;
- 4 **foreach** $(h_i, r, t_i) \in \mathcal{S}_r$ **do**
- 5 // Attentive Neighbor Aggregation
 Attentively aggregate neighbors using Eq. 2–Eq. 4;
- 6 // MoE-based Meta-Knowledge Learning
 Compute expert scores via Eq. 7;
- 7 Select top- N experts and compute relation representation $\mathbf{r}_i$;
- 8 Compute the initial relation-meta $\mathbf{R}_{\mathcal{T}_r}$ via Eq. 8;
- 9 // Inner Loop: Task-Tailored Local Adaptation
 Initialize task-specific adaptation parameters $\eta = \{\mathbf{p}_h, \mathbf{p}_r, \mathbf{p}_t\}$;
- 10 **foreach** $(h_i, r, t_i) \in \mathcal{S}_r$ **do**
- 11 Compute projected embeddings using Eq. 9–Eq. 11;
- 12 Compute the support loss $\mathcal{L}(\mathcal{S}_r)$ via Eq. 13;
- 13 Update task-specific projection vectors $\mathbf{p}_h, \mathbf{p}_r, \mathbf{p}_t$ via Eq. 14;
- 14 Update relation-meta $\mathbf{R}_{\mathcal{T}_r}$ via Eq. 15;
- 15 // Outer Loop: Meta-Update
 Compute the query loss $\mathcal{L}(\mathcal{Q}_r)$ over all $(h_q, r, t_q) \in \mathcal{Q}_r$ via Eq. 16;
- 16 Update global parameters Φ via gradient descent on $\mathcal{L}(\mathcal{Q}_r)$.

5 Experiments

Datasets. We evaluate our method on three widely used KG benchmarks specifically for few-shot relational learning: Nell-One and Wiki-One (Xiong et al., 2018), as well as FB15K-One (Ran et al., 2024). The training/validation/test splits include 51/5/11, 133/16/34, and 75/11/33 tasks on Nell-One, Wiki-One, and FB15K-One, respectively. Detailed dataset statistics are summarized in Table 1.

Evaluation Metrics. We report both MRR (mean reciprocal rank) and Hits@ k ($k = 1, 5, 10$). MRR captures the average reciprocal rank of correct entities, while Hits@ k measures the proportion of correct entities appearing in the top k predictions. Following prior studies, our method is evaluated against baselines under the 1-shot and 5-shot settings

on Nell-One and Wiki-One, which are commonly used for evaluating FSRL, and under the 3-shot setting on FB15K-One, following the setup of RelAdapter (Ran et al., 2024).

Baselines. We compare our proposed method with two categories of state-of-the-art baselines:

- **Conventional KG embedding methods:** This group focuses on modeling relational structures in KGs to learn entity and relation embeddings. We compare with four widely-used baselines including TransE (Bordes et al., 2013), TransD (Ji et al., 2015), DistMult (Yang et al., 2015), and

Table 1: Statistics of datasets.

Dataset	#Relations	#Entities	#Triplets	#Tasks
Nell-One	358	68,545	181,109	67
Wiki-One	822	4,838,244	5,859,240	183
FB15K-One	237	14,541	281,624	119

Table 2: Comparison against state-of-the-art methods on Nell-One and Wiki-One. MetaR-I and MetaR-P indicate the In-train and Pre-train of MetaR (Chen et al., 2019).

Methods	Nell-One								Wiki-One							
	MRR		Hits@1		Hits@5		Hits@10		MRR		Hits@1		Hits@5		Hits@10	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
TransE	0.105	0.168	0.041	0.082	0.111	0.186	0.226	0.345	0.036	0.052	0.011	0.042	0.024	0.057	0.059	0.090
TransD	0.159	0.276	0.125	0.159	0.162	0.322	0.234	0.431	0.063	0.101	0.034	0.061	0.062	0.104	0.135	0.181
DistMult	0.165	0.214	0.106	0.140	0.174	0.246	0.285	0.319	0.046	0.077	0.014	0.035	0.034	0.078	0.087	0.134
ComplEx	0.179	0.239	0.112	0.176	0.212	0.253	0.299	0.364	0.055	0.070	0.021	0.030	0.044	0.063	0.100	0.124
GMatching	0.185	0.201	0.119	0.143	0.260	0.264	0.313	0.311	0.200	-	0.120	-	0.272	-	0.336	-
FSKGC	-	0.184	-	0.136	-	0.234	-	0.272	-	0.158	-	0.097	-	0.206	-	0.287
FAAN	-	0.279	-	0.200	-	0.364	-	0.428	-	0.341	-	0.281	-	0.395	-	0.436
MetaR-I	0.250	0.261	0.170	0.168	0.336	0.350	0.401	0.437	0.193	0.221	0.152	0.178	0.233	0.264	0.280	0.302
MetaR-P	0.164	0.209	0.093	0.141	0.238	0.280	0.331	0.355	0.314	0.323	0.266	0.270	0.375	0.385	0.404	0.418
MetaP†	0.232	-	0.179	-	0.281	-	0.330	-	-	-	-	-	-	-	-	-
GANa	0.236	0.245	0.173	0.166	0.293	0.334	0.347	0.390	0.260	0.261	0.221	0.317	0.307	0.333	0.334	0.384
HiRe	<u>0.288</u>	<u>0.306</u>	<u>0.184</u>	<u>0.207</u>	<u>0.403</u>	<u>0.439</u>	<u>0.472</u>	<u>0.520</u>	<u>0.322</u>	<u>0.371</u>	<u>0.271</u>	<u>0.319</u>	<u>0.383</u>	<u>0.419</u>	<u>0.433</u>	<u>0.469</u>
RelAdapter†	-	-	-	-	-	-	-	-	0.247	0.305	0.209	0.245	0.281	0.365	0.307	0.415
MoEMeta	0.322*	0.339*	0.228*	0.236*	0.429*	0.440	0.501*	0.523	0.325*	0.377*	0.276	0.324	0.385	0.420	0.434	0.473

†: MetaP and RelAdapter’s results are not fully available due to unreleased pre-trained information.

ComplEx (Trouillon et al., 2016). Using OpenKE (Han et al., 2018), these models are reproduced with hyperparameters specified in the original papers.

- **Few-shot relational learning methods:** We compare our method against state-of-the-art FSRL baselines evaluated under the same framework, including (1) metric-learning methods: GMatching (Xiong et al., 2018), FAAN (Sheng et al., 2020), FSKGC (Zhang et al., 2020); and (2) strong meta-learning competitors: MetaR (Chen et al., 2019), GANA (Niu et al., 2021), HiRe (Wu et al., 2023), and RelAdapter (Ran et al., 2024). We exclude methods like CSR (Huang et al., 2022) and NP-FKGC (Luo et al., 2023) as they employ a different training and evaluation protocol.

All results are obtained after models are trained using all triplets from background relations (Xiong et al., 2018) and meta-training tasks, and then evaluated on meta-testing tasks.

Experimental Setup. Following prior work, we set the embedding dimension to 100 for Nell-One and FB15K-One, and 50 for Wiki-One, initialized using TransE-pretrained weights provided by GMatching (Xiong et al., 2018). The number of neighbors per entity is capped at 50 for aggregation. For MoE, each expert network is a two-layer MLP (hidden dimension: 64, ReLU). The gating network is also a two-layer MLP (hidden dimension: 64, output dimension: 1, and ReLU). We set the number of experts M as 32 and the number of selected experts N as 5. The margin γ in Eq. 16 is set to 1. MoEMeta is trained using the Adam optimizer (batch size: 1,024, learning rate: 0.001). All results are averaged over 5 independent runs with different random seeds. All models are implemented in PyTorch and trained on a single Tesla P100 GPU. More details for reproducing our results are provided in Appendix B.1.

5.1 Comparison with Baseline Methods

Tables 2 and 3 compare MoEMeta against state-of-the-art methods on Nell-One and Wiki-One under the 1-shot and 5-shot settings, and on FB15K-One under the 3-shot setting. The best and the second best performers are respectively marked by grey and underline. Our proposed MoEMeta consistently achieves state-of-the-art results across all datasets and settings. On Nell-One, MoEMeta surpasses HiRe, the second best performer, by 11.8% MRR, 23.9% Hits@1, 6.4% Hits@5, and 6.1% Hits@10 under 1-shot setting, and 10.8% MRR, 14.0% Hits@1, 2.3% Hits@5, and 0.6% Hits@10 under 5-shot setting. On FB15K-One, MoEMeta outperforms the second best method by 4.4% MRR, 0.3% Hits@1, and 13.2% Hits@5 under 3-shot setting. We further conduct paired t-tests between MoEMeta and the second best performers on all datasets. The improvements of MoEMeta over the second best method with statistical significance at the 0.05 level are marked with •. These results highlight

Table 3: Comparison under 3-shot setting on FB15K-One.

Methods	FB15K-One		
	MRR	Hits@1	Hits@10
TransE	0.294	0.204	0.437
TransD	0.301	0.203	0.439
DistMult	0.234	0.208	0.364
ComplEx	0.239	0.205	0.359
GMatching	0.309	0.245	0.441
MetaR	0.368	0.251	0.536
FAAN	0.363	0.279	0.542
GANa	0.388	<u>0.301</u>	0.553
HiRe	0.378	0.281	0.571
RelAdapter	<u>0.405</u>	0.297	<u>0.575</u>
MoEMeta	0.423 •	0.302	0.651 •

MoEMeta’s effectiveness in learning

Table 4: Ablation study on Nell-One and FB15K-One.

Ablation on	Nell-One								FB15K-One		
	1-shot				5-shot				3-shot		
	MRR	Hits@1	Hits@5	Hits@10	MRR	Hits@1	Hits@5	Hits@10	MRR	Hits@1	Hits@10
MoEMeta	0.322	0.228	0.429	0.501	0.339	0.236	0.440	0.523	0.422	0.298	0.652
w/o N.A	0.311	0.219	0.418	0.489	0.328	0.231	0.436	0.509	0.413	0.286	0.631
w/o MoE	0.291	0.199	0.392	0.467	0.293	0.201	0.399	0.484	0.401	0.273	0.624
w/o L.A	0.301	0.206	0.398	0.476	0.315	0.213	0.415	0.501	0.408	0.285	0.628

N.A indicates “attentive neighbor aggregation” and L.A indicates “task-tailored local adaptation”.

generalizable meta-knowledge that adapts well to novel relations and proving its competitiveness against other baselines.

Across three datasets, the most substantial improvements are on Nell-One, particularly in MRR under both 1-shot and 5-shot settings. The primary gain is attributed to the increase in the challenging Hits@1 metric, which gauges the model’s ability to rank the correct tail entity as the top candidate among all choices. This is especially notable since triplets in Nell-One are constructed in natural language, providing richer semantics that can be more effectively captured by our relational prototypes. These results validate the efficacy of our proposed MoE-based relation-meta learner in learning globally shared, fine-grained relational patterns and adapting these prototypes to new, unseen tasks. We additionally provide a runtime analysis for both meta-training and meta-testing in Appendix B.2.

5.2 Ablation Studies

To assess the contribution of MoEMeta’s three key components, attentive neighbor aggregation, MoE-based meta-knowledge learning, and task-tailored local adaptation, we report ablation results on Nell-One under the 1-shot and 5-shot settings, and FB15K-One under the 3-shot setting in Table 4.

- **Removing attentive Neighbor Aggregation (w/o N.A)**, which removes the first term in Eq. 4, results in a moderate performance decrease across all settings. This suggests the importance of incorporating neighboring information to enhance entity representations.
- **Replacing MoE (w/o MoE)** with an MLP-based relation-meta learner from (Chen et al., 2019) leads to a significant performance decline; MRR drops by 9.6% and 13.6% under 1-shot and 5-shot settings on Nell-One, and 5.0% under 3-shot setting on FB15K-One. This underscores MoE’s indispensable role in acquiring relational prototypes as generalizable meta-knowledge.
- **Ablating task-tailored Local Adaptation (w/o L.A)**, where the learned relation-meta is directly applied to the query set, also causes a significant performance drop across datasets. This suggests that global meta-knowledge alone is insufficient, highlighting the crucial role of task-tailored adaptation for capturing local distinct patterns and improving generalization across tasks.

5.3 Analysis on Complex Relations

To further evaluate the significance of task-tailored local adaptation for complex relations, we conduct a detailed analysis on Nell-One under the 5-shot setting by categorizing the meta-testing set into three relation types: 1-N, N-1, and N-N. In particular, we compare MoEMeta with a variant that removes task-tailored local adaptation (w/o L.A). As shown in Table 5, the removal of local adaptation leads to a substantially larger performance drop on 1-N and N-N relations than on N-1 relations. This is because, in N-1 relations, the tail entity is relatively predictable, with a smaller search space and more stable mapping patterns. In contrast, 1-N and N-N relations exhibit higher ambiguity and a more dispersed distribution of tail entities, making them more difficult to predict. Notably, removing task-tailored adaptation results in the largest performance drop for the N-N relation type. This analysis confirms the critical role of our task-tailored adaptation mechanism in enhancing generalization to complex and highly ambiguous relational patterns.

In addition, we visualize the distributions of gating values of different relations in Figure 2 to provide a more intuitive understanding of why learning relational prototypes works. As observed, relations with similar patterns (in the upper figure) exhibit a comparable gating value distribution and activate similar subsets of experts, whereas dissimilar relations (in the lower figure) demonstrate a markedly different pattern. This observation highlights our model’s ability to distinguish diverse relational patterns but capture globally shared relational prototypes across related tasks.

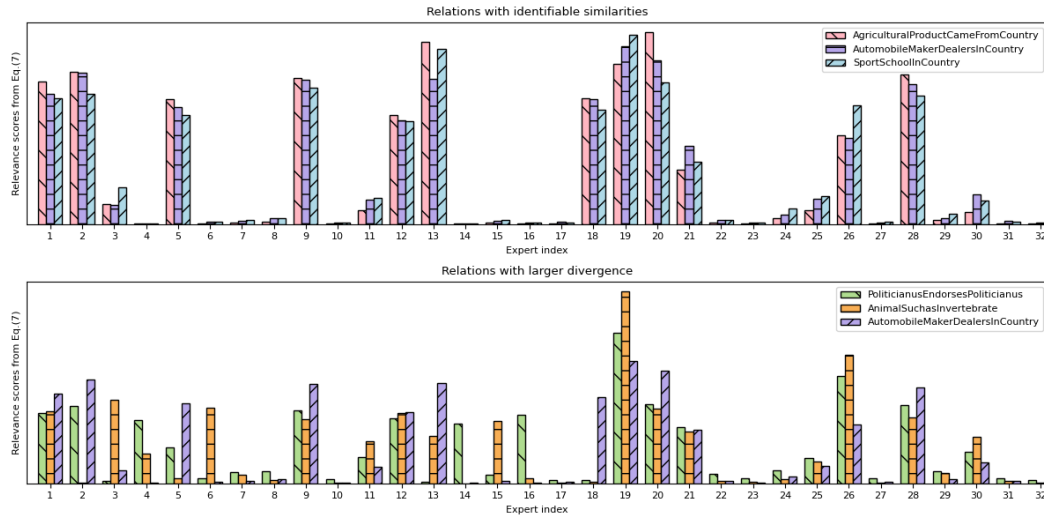


Figure 2: Two gating value distributions over 32 experts generated by our MoE-based meta-knowledge learner. The upper figure shows that three similar relations exhibit similar gating value distributions, suggesting that a similar set of experts are activated. In contrast, the lower figure demonstrates that for relations with distinct relational patterns, the activated experts differ significantly. This highlights the presence of shared relational patterns, underscoring the necessity of learning relation prototypes.

Table 5: Performance analysis under 5-shot setting on Nell-One for 1-N, N-1 and N-N relation types.

Methods	Rel. types	MRR	Hits@1	Hits@5	Hits@10
MoEMeta	1-N	0.406	0.310	0.506	0.577
	N-1	0.945	0.913	0.986	0.986
	N-N	0.404	0.279	0.536	0.618
w/o L.A	1-N	0.391	0.292	0.500	0.565
	N-1	0.874	0.871	0.986	0.986
	N-N	0.372	0.256	0.485	0.586

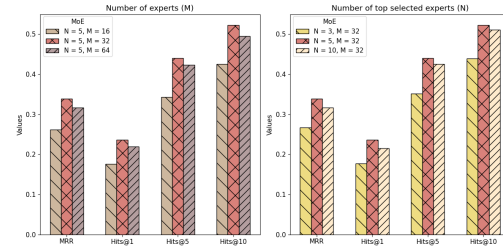


Figure 3: Hyperparameter sensitivity study of MoE under 5-shot setting on Nell-One.

5.4 Hyperparameter Analysis on MoE

To assess the impact of two MoE hyperparameters, the number of experts M and the number of top selected experts N , we conduct a sensitivity analysis on Nell-One under the 5-shot setting as a case study. When fixing $N = 5$ and varying $M \in \{16, 32, 64\}$, MoEMeta achieves the best performance with 32 experts. This aligns with the guideline of setting the number of experts M smaller than the total number of tasks to encourage learning shared relational prototypes and prevent overfitting with limited support triplets. When fixing $M = 32$ and varying $N \in \{3, 5, 10\}$, we observe that selecting slightly more experts (e.g., 5 vs. 3) leads to improved performance, yet increasing N to 10 causes only a slight drop, suggesting that MoEMeta is relatively insensitive to N within a reasonable range.

6 Conclusion

This work presents MoEMeta, a novel meta-learning framework for few-shot relational learning in KGs. By integrating an MoE model to capture globally shared relational prototypes with a task-tailored adaptation mechanism to refine embeddings locally, MoEMeta effectively balances cross-task generalization and rapid task-specific adaptation. Extensive experiments on three real-world KG datasets demonstrate that MoEMeta consistently outperforms state-of-the-art methods, showcasing its effectiveness for few-shot relation learning. Comprehensive ablation studies and visualizations further validate the importance of MoEMeta's key components, confirming its efficacy in modeling complex relational data in KGs under limited supervision.

References

- Milad Abdollahzadeh, Toubia Malekzadeh, and Ngai-Man Cheung. 2021. Revisit Multimodal Meta-Learning through the Lens of Multi-Task Learning. In *NeurIPS*, Vol. 34. 14632–14644.
- Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *SIGMOD*. 1247–1250.
- Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *NeurIPS*. 2787–2795.
- Yixin Cao, Xiang Wang, Xiangnan He, Zikun Hu, and Tat-Seng Chua. 2019. Unifying knowledge graph learning and recommendation: Towards a better understanding of user preferences. In *WWW*. 151–161.
- Mingyang Chen, Wen Zhang, Wei Zhang, Qiang Chen, and Huajun Chen. 2019. Meta Relational Learning for Few-Shot Link Prediction in Knowledge Graphs. In *EMNLP-IJCNLP*. 4217–4226.
- Nan Du, Yanping Huang, Andrew M. Dai, Quan Hoang, Youlong Cheng, Zhifeng Chen, Claire Cui, Quoc V. Le, and Yonghui Wu. 2022. Mixture-of-Experts with Expert Choice Routing. In *NeurIPS*. 7103–7114.
- Yuan Fang, Kingsley Kuan, Jie Lin, Cheston Tan, and Vijay Chandrasekhar. 2017. Object detection meets knowledge graphs. In *IJCAI*. 1661–1667.
- Li Fei-Fei, R. Fergus, and P. Perona. 2006. One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28, 4 (2006), 594–611.
- Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*. 1126–1135.
- Chelsea Finn, Kelvin Xu, and Sergey Levine. 2018. Probabilistic model-agnostic meta-learning. In *NeurIPS*. 9537–9548.
- Xu Han, Shulin Cao, Xin Lv, Yankai Lin, Zhiyuan Liu, Maosong Sun, and Juanzi Li. 2018. Openke: An open toolkit for knowledge embedding. In *EMNLP*. 139–144.
- Qian Huang, Hongyu Ren, and Jure Leskovec. 2022. Few-shot relational reasoning via connection subgraph pretraining. In *NeurIPS*. 6397–6409.
- Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. 1991. Adaptive Mixtures of Local Experts. *Neural Computation* 3, 1 (1991), 79–87.
- Kevin Swersky, Jake Snell, and Richard Zemel. 2017. Prototypical networks for few-shot learning. In *NeurIPS*, Vol. 10. 4077–4087.
- Guoliang Ji, Shizhu He, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Knowledge graph embedding via dynamic mapping matrix. In *ACL*. 687–696.
- Zhiyi Jiang, Jianliang Gao, and Xinqi Lv. 2021. MetaP: Meta Pattern Learning for One-Shot Knowledge Graph Completion. In *SIGIR*. 2232–2236.
- Michael I. Jordan and Robert A. Jacobs. 1994. Hierarchical Mixtures of Experts and the EM Algorithm. *Neural Computation* 6, 2 (1994), 181–214.
- Timothée Lacroix, Nicolas Usunier, and Guillaume Obozinski. 2018. Canonical tensor decomposition for knowledge base completion. In *ICML*. 2863–2872.
- Hongyang Li, David Eigen, Samuel Dodge, Matthew Zeiler, and Xiaogang Wang. 2019. Finding task-relevant features for few-shot learning by category traversal. In *CVPR*. 1–10.
- Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. 2015. Learning entity and relation embeddings for knowledge graph completion. In *AAAI*, Vol. 29. 2181–2187.

- Hao Liu, Jiarui Feng, Lecheng Kong, Dacheng Tao, Yixin Chen, and Muhan Zhang. 2024. Graph Contrastive Learning Meets Graph Meta Learning: A Unified Method for Few-shot Node Tasks. In *The Web Conference*. 365–376.
- Yonghao Liu, Mengyu Li, Ximing Li, Fausto Giunchiglia, Xiaoyue Feng, and Renchu Guan. 2022. Few-shot node classification on attributed networks with graph meta-learning. In *SIGIR*. 471–481.
- Linhao Luo, Yuan-Fang Li, Gholamreza Haffari, and Shirui Pan. 2023. Normalizing Flow-based Neural Process for Few-Shot Knowledge Graph Completion. In *SIGIR*. 900–910.
- T. Mitchell, W. Cohen, E. Hruschka, P. Talukdar, B. Yang, J. Betteridge, A. Carlson, B. Dalvi, M. Gardner, B. Kisiel, J. Krishnamurthy, N. Lao, K. Mazaitis, T. Mohamed, N. Nakashole, E. Platanios, A. Ritter, M. Samadi, B. Settles, R. Wang, D. Wijaya, A. Gupta, X. Chen, A. Saparov, M. Greaves, and J. Welling. 2018. Never-Ending Learning. *Commun. ACM* 61, 5 (Apr 2018), 103–115.
- Alex Nichol and John Schulman. 2018. Reptile: a scalable metalearning algorithm. *arXiv preprint arXiv:1803.02999* 2, 3 (2018), 4.
- Guanglin Niu, Yang Li, Chengguang Tang, Ruiying Geng, Jian Dai, Qiao Liu, Hao Wang, Jian Sun, Fei Huang, and Luo Si. 2021. Relational Learning with Gated and Attentive Neighbor Aggregator for Few-Shot Knowledge Graph Completion. In *SIGIR*. 213–222.
- Joan Puigcerver, Carlos Riquelme, Basil Mustafa, and Neil Houlsby. 2024. From Sparse to Soft Mixtures of Experts. In *ICLR*.
- Liu Ran, Zhongzhou Liu, Xiaoli Li, and Yuan Fang. 2024. Context-Aware Adapter Tuning for Few-Shot Relation Learning in Knowledge Graphs. In *EMNLP*. 17525–17537.
- Sachin Ravi and Hugo Larochelle. 2017. Optimization as a model for few-shot learning. In *ICLR*.
- Mengye Ren, Eleni Triantafillou, Sachin Ravi, Jake Snell, Kevin Swersky, Joshua B Tenenbaum, Hugo Larochelle, and Richard S Zemel. 2018. Meta-learning for semi-supervised few-shot classification. *arXiv preprint arXiv:1803.00676* (2018).
- Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. 2016. Meta-learning with memory-augmented neural networks. In *ICML*. 1842–1850.
- Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. In *ICLR*.
- Jiawei Sheng, Shu Guo, Zhenyu Chen, Juwei Yue, Lihong Wang, Tingwen Liu, and Hongbo Xu. 2020. Adaptive Attentional Network for Few-Shot Knowledge Graph Completion. In *EMNLP*. 1681–1691.
- Ran Song, Shizhu He, Suncong Zheng, Shengxiang Gao, Kang Liu, Zhengtao Yu, and Jun Zhao. 2022. Decoupling Mixture-of-Graphs: Unseen Relational Learning for Knowledge Graph Completion by Fusing Ontology and Textual Experts. In *COLING*. 2237–2246.
- Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: a core of semantic knowledge. In *WWW*. 697–706.
- Siyuan Sun and Hongyang Gao. 2023. Meta-AdaM: A Meta-Learned Adaptive Optimizer with Momentum for Few-Shot Learning. In *NeurIPS*, Vol. 36. 65441–65455.
- Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. 2019. RotatE: Knowledge Graph Embedding by Relational Rotation in Complex Space. In *ICLR*.
- Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. 2016. Complex embeddings for simple link prediction. In *ICML*. 2071–2080.
- Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. 2016. Matching networks for one shot learning. In *NeurIPS*. 3637–3645.

- Johannes Von Oswald, Dominic Zhao, Seijin Kobayashi, Simon Schug, Massimo Caccia, Nicolas Zucchet, and João Sacramento. 2021. Learning where to learn: gradient sparsity in meta and continual learning. In *NeurIPS*. 5250–5263.
- Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM* 57, 10 (2014), 78–85.
- Risto Vuorio, Shao-Hua Sun, Hexiang Hu, and Joseph J. Lim. 2019. Multimodal model-agnostic meta-learning via task-aware modulation. In *NeurIPS*. 1–12.
- Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014. Knowledge graph embedding by translating on hyperplanes. In *AAAI*, Vol. 28. 1112–1119.
- Han Wu, Jie Yin, Bala Rajaratnam, and Jianyuan Guo. 2023. Hierarchical Relational Learning for Few-Shot Knowledge Graph Completion. In *ICLR*.
- Qi Wu, Peng Wang, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. 2016. Ask me anything: Free-form visual question answering based on knowledge from external sources. In *CVPR*. 4622–4630.
- Wenhan Xiong, Mo Yu, Shiyu Chang, Xiaoxiao Guo, and William Yang Wang. 2018. One-Shot Relational Learning for Knowledge Graphs. In *EMNLP*. 1980–1990.
- Bishan Yang, Scott Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2015. Embedding Entities and Relations for Learning and Inference in Knowledge Bases. In *ICLR*.
- Chuxu Zhang, Huaxiu Yao, Chao Huang, Meng Jiang, Zhenhui Li, and Nitesh V Chawla. 2020. Few-shot knowledge graph completion. In *AAAI*, Vol. 34. 3041–3048.
- Pan Zhou, Yingtian Zou, Xiao-Tong Yuan, Jiashi Feng, Caiming Xiong, and Steven Hoi. 2021. Task similarity aware meta learning: Theory-inspired improvement on MAML. In *UAI*. PMLR, 23–33.
- Yingtian Zou, Fusheng Liu, and Qianxiao Li. 2021. Unraveling model-agnostic meta-learning via the adaptation learning rate. In *ICLR*.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: The main claims in the abstract and introduction accurately reflect the paper's contributions and scope. Specifically, we propose MoEMeta with a mixture-of-experts model and a task-tailored adaptation mechanism to learn relational prototypes shared across tasks and improve few-shot relational learning.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: The paper discusses the limitations of the proposed method in various sections. For instance, in Section 4.3 we discuss the three projection vectors in the task-tailored local adaptation mechanism and potential future extension. In Appendix B.2, the paper discusses that the proposed method incurs slightly higher computational overhead to achieve substantial performance gains in the FSRL tasks.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: We provide essential assumptions and validations in both the main paper and appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: For the reproducibility, we provide the detailed configuration of hyperparameters used in the experiments in the main paper and other details for reproducing all reported results in Appendix B.1.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

- (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The data used in the experiments (Nell-One, Wiki-One, and FB15K-One) are all publicly available. The code of our work is provided via an anonymous github repository, with instructions to reproduce the main results of our method.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We present the details of all the experimental settings in Section 5 of the main paper.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: The paper reports the statistical significance results demonstrating the improvements of the proposed method over the second best method are statistically significant at the 0.05 level.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: All experiments are conducted using 1 Tesla P100 GPU. We have included the information in our paper.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: All research conducted in the paper conform with the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: There is no societal impact of this work needed to be specified.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper poses no such risks. All data are publicly available. We also provide every detail of our model.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The creators or original owners of assets are properly cited and credited.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.

- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: This paper does not release new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: This paper does not involve such aspects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: This paper does not involve such aspects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core method development in this paper does not involve LLMs as any components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.

Appendix

A Complexity Analysis

In this section, we analyze the computational complexity of our proposed MoEMeta for its key components. Table 6 summarizes the computational complexity of each component, where d denotes the dimension of entity embeddings, n denotes the number of neighbors in attentive neighbor aggregation, and K denotes the number of support triplets per task. For MoE-based meta-knowledge learning, M indicates the total number of experts, and N indicates the number of experts activated in each forward pass.

As shown in Table 6, the complexity of MoEMeta scales linearly with the number of neighbors per entity during the aggregation stage ($\mathcal{O}(Knd^2)$) and linearly with the number of support triplets during MoE-based meta-knowledge learning ($\mathcal{O}(NKd^2)$). The task-tailored local adaptation introduces only a minimal overhead of ($\mathcal{O}(d)$) in both parameters and computation. Overall, the computational cost is dominated by neighbor aggregation ($\mathcal{O}(Knd^2)$), which is an essential component in current state-of-the-art FSRL methods. The additional computational and parameter overhead introduced by MoEMeta is negligible, making it efficient for FSRL in KGs.

Table 6: Complexity analysis of MoEMeta components with respect to the number of model parameters and the number of multiplication operations in each epoch. d denotes the dimension of entity embeddings. n denotes the number of neighbors per entity, and K denotes the number of support triplets per task, where $K \ll n, d$. M is the total number of experts, while N is the number of experts activated in each forward pass.

Component	# Parameters	# Multiplication Operations
Attentive neighbor aggregation	$\mathcal{O}(d^2)$	$\mathcal{O}(Knd^2)$
MoE-based meta-knowledge learning	$\mathcal{O}(Md^2)$	$\mathcal{O}(NKd^2)$
Task-tailored local adaptation	$\mathcal{O}(d)$	$\mathcal{O}(d)$
Total	$\mathcal{O}(Md^2)$	$\mathcal{O}(nKd^2)$

B Additional Experimental Details and Results

B.1 Reproducibility Details

In our experiments, our method and all baselines are compared using the same evaluation protocol. The results of baseline methods are reproduced by using their official open-sourced implementations or adopted from their respective papers.

For MetaR (Chen et al., 2019) (both In-Train and Pre-Train)², FAAN (Sheng et al., 2020)³, MetaP (Jiang et al., 2021)⁴ and HiRe (Wu et al., 2023)⁵, we directly use the results reported in their papers. The results for GMatching (Xiong et al., 2018)⁶ under 1-shot and 5-shot settings are taken from (Chen et al., 2019). Since FSKGC (Zhang et al., 2020)⁷ was originally evaluated under different conditions (with a smaller candidate set), we use the re-implemented results provided by FAAN to ensure consistency. For GANA (Niu et al., 2021)⁸, we reproduce its results by setting the number of neighbors to 50, aligning with the settings of other methods. We reproduce the results of RelAdapter (Ran et al., 2024)⁹ under the 1-shot and 5-shot settings on Wiki-One. However, as the pre-trained contextual information required for Nell-One is not released, the performance of RelAdapter on this dataset is unavailable for evaluation. Following RelAdapter, the comparison

²MetaR: <https://github.com/AnselCmy/MetaR>

³FAAN: <https://github.com/JiaweiSheng/FAAN>

⁴MetaP: <https://github.com/jzystc/MetaP>

⁵HiRe: <https://github.com/alexhw15/HiRe>

⁶GMatching: <https://github.com/xwhan/One-shot-Relational-Learning>

⁷FSKGC: https://github.com/chuxuzhang/AAAI2020_FSRL

⁸GANA: <https://github.com/ngl567/GANA-FewShotKGC>

⁹RelAdapter: <https://github.com/smufang/RelAdapter>

results on FB15K-One under 3-shot setting are reported in Table 3, with all baseline results sourced from RelAdapter (Ran et al., 2024). All models are implemented in PyTorch and trained on a single Tesla P100 GPU.

B.2 Runtime Analysis

We also perform a runtime analysis comparing MoEMeta with other baselines. For a fair comparison, we evaluate MoEMeta on an RTX3090 GPU to keep consistency with baseline runtimes reported in RelAdapter (Ran et al., 2024). Table 7 reports the total runtime of meta-training and the inference time per triplet during meta-testing in seconds.

During meta-training, MoEMeta is only slightly slower than MetaR and RelAdapter, while being faster than other baselines such as GANA and FSKGC. MoEMeta requires 26,277 seconds to train on the NELL-One dataset, compared to 22,691 seconds for MetaR and 24,085 seconds for RelAdapter. The moderate increase in training time mainly comes from the two additional MLPs used in the MoE. Despite this additional overhead, MoEMeta achieves significantly better performance. For example, under the 1-shot setting on NELL-One, MoEMeta outperforms MetaR, by 28.8% MRR, 34.1% Hits@1, 27.7% Hits@5, and 24.9% Hits@10, respectively (see Table 2). For the inference during meta-testing, all methods including MoEMeta complete in milliseconds per triplet.

Table 7: Runtime comparison of meta-training and meta-testing time on Wiki-One and FB15K-One.

Methods	Meta-training Time (seconds)		Meta-testing Time per Triplet (seconds)	
	Wiki-One	FB15K-One	Wiki-One	FB15K-One
GMatching	28,941	19,678	0.009	0.004
FSKGC	28,742	20,765	0.013	0.004
GANA	35,374	28,167	0.017	0.006
FAAN	32,036	23,675	0.016	0.005
HiRe	34,257	27,736	0.036	0.007
MetaR	22,691	16,504	0.012	0.005
RelAdapter	24,085	17,529	0.045	0.008
MoEMeta	26,277	18,321	0.019	0.006

C Detailed and Additional Related Work

For a more comprehensive discussion of related literature beyond the main paper, we provide additional related work in this section.

C.1 Supervised KG Embedding Learning

KG embedding methods aim to learn a low-dimensional embedding space for entities and relations in KGs using input triplets as supervision. They broadly fall into two categories for KG relational learning: (1) Translation-based methods, such as TransE (Bordes et al., 2013), TransH (Wang et al., 2014), TransD (Ji et al., 2015), and TransR (Lin et al., 2015), which assume a translational relationship between entities and relations for learning a unified embedding space; and (2) Semantic matching methods, such as ComplEx (Trouillon et al., 2016), RotatE (Sun et al., 2019), and ComplEx-N3 (Lacroix et al., 2018), which improve the modeling of relational patterns in a vector or complex space. However, these methods rely on large amounts of triplets for training, and their performance degrades significantly under few-shot settings, where only a handful of training triplets are available for each relation.

C.2 Few-Shot Learning and Meta Learning

Few-shot learning aims to train a supervised model that can generalize to new tasks with only limited amounts of training data (Fei-Fei et al., 2006). Meta-learning (Finn et al., 2017; Santoro et al., 2016; Ravi and Larochelle, 2017; Nichol and Schulman, 2018; Ren et al., 2018) has emerged as a powerful paradigm for enabling rapid learning on few-shot tasks by leveraging meta-knowledge

learned from meta-training tasks. Relative to metric-based approaches (Jake Snell and Zemel, 2017; Vinyals et al., 2016), optimization-based meta-learning methods have gained significant attention due to their strong generalization capabilities. These methods cast meta-learning into a bi-level optimization problem; the inner loop fine-tunes a base model on individual tasks, and the outer loop optimizes the base model for quick adaptation across tasks. MAML (Finn et al., 2017) is the most representative algorithm in this category, which learns global parameters as an initialization to be shared among task-specific models for fast adaptation using one or a few steps of gradient descent updates. Due to its flexible formulation, MAML has been applied to various domains, including computer vision, robotics, lifelong learning, and beyond. Subsequent variants of MAML (Finn et al., 2018; Von Oswald et al., 2021) improve upon its adaptability by optimizing model initialization or learning adaptive learning rates for faster convergence (Sun and Gao, 2023; Zou et al., 2021; Finn et al., 2018). Our work focuses on learning well-generalized meta-knowledge across tasks while integrating task-specific local adaptation to improve few-shot relational learning in KGs.

C.3 Few-Shot Relation Learning

Few-shot relation learning (FSRL) addresses the challenges of learning with limited training examples and relational information in KGs. Current FSRL approaches can be broadly categorized into two groups: metric-learning methods and meta-learning methods. Metric-learning methods focus on learning a similarity metric between the support and query sets for adaptation. GMatching (Xiong et al., 2018) introduces one-shot relation learning, where a one-hop neighbor encoder with equal weights is used to learn entity embeddings and a matching network is designed to compare similarity between support and query entity pairs. FSKGC (Zhang et al., 2020) generalizes the setting to more shots and employs a recurrent autoencoder to aggregate few-shot instances in the support set, yet imposing an unrealistic sequential dependencies among support triplets. FAAN (Sheng et al., 2020) proposes a dynamic attention mechanism to improve the aggregation of one-hop neighbors. CSR (Huang et al., 2022) employs subgraph matching as a hypothesis testing to adapt from the support triplets to the query set. NP-FKGC (Luo et al., 2023) employs neural process to better handle distribution shifts for unseen relations.

Most mainstream FSRL methods are based on meta-learning (Finn et al., 2017), which focuses on learning an initial meta-prior from existing tasks that can generalize to new few-shot relations. MetaR (Chen et al., 2019) learns relation-specific meta-information for adaptation, by simply averaging the representations of all support triplets. GANA (Niu et al., 2021) integrates meta-learning with TransH (Wang et al., 2014) and devises a gated and attentive neighbor aggregator to address sparse neighbors for learning informative entity embeddings. HiRe (Wu et al., 2023) exploits multi-granularity relational information to learn meta-knowledge that better generalizes to unseen relations. RelAdapter (Ran et al., 2024) adopts a shallow feedforward network with a residual connection to adapt relation-meta within each target task. Additionally, pre-trained context information of related entities is used to augment entity embeddings.

Despite their great progress, the aforementioned meta-learning based methods suffer from two key limitations. First, these methods neglect common relational patterns shared across tasks when learning meta-knowledge during meta-training. Second, they fail to effectively incorporate local, task-specific information essential for fast adaptation, as global meta-knowledge may not be optimally tailored to the unique characteristics of individual tasks. Our work is thus proposed to address these critical gaps within the meta-learning framework, enabling more effective model generalization and adaptation under few-shot settings.

Our work is also related to recent graph-based and contrastive meta-learning methods such as (Liu et al., 2022) and (Liu et al., 2024). While these approaches are relevant to few-shot learning and meta-learning on node-level tasks for homogeneous graphs, they are not directly applicable to our setting, which focuses on few-shot relational learning in KGs—that centers on relation-level tasks. Our method, MoEMeta, specifically targets the key challenges of few-shot relational learning, such as the complex and diverse relational patterns in KGs. This distinction sets our work apart from general graph few-shot learning or meta-learning methods that target node-level tasks.