Incorporating Global Information in Local Attention for Knowledge Representation Learning

Yu Zhao\textsuperscript{1,}\textsuperscript{*}, Han Zhou\textsuperscript{2,}\textsuperscript{*}, Ruobing Xie\textsuperscript{3}, Fuzhen Zhuang\textsuperscript{4,5}, Qing Li\textsuperscript{1}, Ji Liu\textsuperscript{6}

\textsuperscript{1}Fintech Innovation Center, Financial Intelligence and Financial Engineering Key Laboratory, Southwestern University of Finance and Economics, Chengdu, China
\textsuperscript{2}Department of Computer Science, The University of Hong Kong, China
\textsuperscript{3}WeChat Search Application Department, Tencent, Beijing, China
\textsuperscript{4}Institute of Artificial Intelligence, Beihang University, Beijing 100191, China
\textsuperscript{5}Xiamen Data Intelligence Academy of ICT, CAS, China
\textsuperscript{6}AI platform and Seattle AI Lab, Kwai Inc.

Abstract

Graph Attention Networks (GATs) have proven a promising model that takes advantage of localized attention mechanism to perform knowledge representation learning (KRL) on graph-structure data, e.g., Knowledge Graphs (KGs). While such approaches model entities’ local pairwise importance, they lack the capability to model global importance relative to other entities of KGs. This causes such models to miss critical information in tasks where global information is also a significant component for the task, such as in knowledge representation learning. To address the issue, we allow the proper incorporation of global information into the GAT family of models through the use of scaled entity importance, which is calculated by an attention-based global random walk algorithm. In the context of KRL, incorporating global information boosts performance significantly. Experimental results on KG entity prediction against the state-of-the-arts sufficiently demonstrate the effectiveness of our proposed model.

1 Introduction

Graph Attention Networks (GATs) have been successfully applied to various tasks over graphs (Velickovic et al., 2018; Lee et al., 2018b), such as graph classification (Wu et al., 2019b; Lee et al., 2018a), link prediction (Abu-El-Haija et al., 2018), and node classification (Lee et al., 2019; Zhang et al., 2020a). GATs learn from the underlying graph structure by making use of localized attention mechanism (Wu et al., 2019a; Xu et al., 2019; Vashishth et al., 2020b), where the hidden representation of each node is computed by recursively aggregating and attending over its corresponding local neighbors’ features, and the weighting coefficients are calculated inductively with self-attention strategy (Thekumparampil et al., 2018; Qian et al., 2018; Zhang et al., 2018). The original GATs perform only on single-relational homogeneous graphs (Velickovic et al., 2018; Wang et al., 2019b). Recent advancements were proposed to operate on more general and prevalent multi-relational graphs (Wang et al., 2019b; Hong et al., 2020; Nathani et al., 2019; Zhang et al., 2020c), such as the representative Knowledge Graphs (KGs) which contain multiple types of entities (nodes) and relationships (edges) (Zhou et al., 2018; Han et al., 2018; Wang et al., 2019a; Zhao et al., 2020). However, these approaches can only exploit localized features within the neighborhood of individual entities (Nathani et al., 2019; Busbridge et al., 2019; Zhang et al., 2020c). For some tasks, such simplified localized feature aggregation may be sufficient, but insufficient for knowledge representation learning (KRL) tasks that also need exploring global information (Xie et al., 2020).

In this paper, we concentrate on how to incorporate global information in local attention for knowledge representation learning. Specifically, we allow the proper incorporation of global information into the GAT family of models through the use of scaled entity importance, which is estimated by a global random walk algorithm upon the whole graph structural information. In KGs, entity importance\textsuperscript{1} indicates the global significance or authority of an entity. Intuitively, it can be quite beneficial if an entity attends more to its “authoritative” neighbors that have high scores of global entity importance. For instance, a movie “Titanic” links to different actors, among which a superstar

---

\textsuperscript{*} Equal Contribution. Corresponding author: Y. Zhao (zhaoyu@swufe.edu.cn).

\textsuperscript{1}The notions of its counterparts, e.g., global node importance or object authority, have been widely studied in graphs (Li et al., 2012; Liu et al., 2017; Park et al., 2019), which enable a number of applications such as Web search (Brin and Page, 1998; Kleinberg, 1999), social network analysis (Weng et al., 2010), RecSys (Jing et al., 2014), query disambiguation (Makris et al., 2012; Saxena et al., 2020).

1341

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 1341–1351
August 1–6, 2021. ©2021 Association for Computational Linguistics
In this paper, we propose a novel Entity Importance-aware Graph Attention Networks, EIGAT, which incorporates global entity importance in local attention mechanism for learning effective knowledge representations. As shown in Figure 1, we give a brief illustration of our proposed EIGAT, which is compared to early proposed GCN (Kipf and Welling, 2017) and GAT (Velickovic et al., 2018). In EIGAT, the importance scores of all entities are expected to be estimated upon global information and to be incorporated in local entity aggregation (Equation 5) for building better entity embeddings. In particular, we provide an attention-based random walk approach to estimate entity importance upon global structural information for serving EIGAT. We conduct extensive experiments on several different types of KGs by entity prediction against state-of-the-art methods, which sufficiently demonstrate our proposed EIGAT can successfully incorporating global information in local attention to improve knowledge representation learning.

The contributions of this paper are threefold:

- We propose to incorporate global information in local attention for knowledge representation learning.
- We propose EIGAT, a novel entity importance-aware graph attention networks which incorporate global entity importance into local entity aggregation.
- The extensive experimental results demonstrate the efficacy of our proposed model in link prediction.

2 Related Work

To make this paper self-contained, we introduce some related topics here on Knowledge Representation Learning and Graph Neural Networks (GNNs).

2.1 Knowledge Representation Learning (KRL)

In recent years, knowledge representation learning on KGs has been a hot research topic (Xiao et al., 2017; Shi and Weninger, 2017; Ebisu and Ichise, 2019; Balazevic et al., 2019; Zhang et al., 2020b). These methods roughly fall into four categories: (i) Translational-based models, which view relations as translations from a head entity to a tail entity, such as Trans(E, H, R, D and G) (Bordes et al., 2013; Wang et al., 2014; Lin et al., 2015; Ji et al., 2015; Xiao et al., 2016), ComplEx (Trouillon et al., 2016), JoBi ComplEx (Balazevic et al., 2019). (ii) Tensor factorization based models, which assume the score of a triple can be factorized into several tensors, such as RESCAL (Nickel et al., 2011), NTN (Socher et al., 2013), DistMult (Yang et al., 2015), HOPE (Nickel et al., 2016). (iii) CNN-based models, which use convolution over embed-
Table 1: Different variants of Graph Neural Networks: GNN (Scarselli et al., 2008), GCN (Kipf and Welling, 2017), GAT (Velickovic et al., 2018), and our proposed model EIGAT that incorporates global information in local attention.

| GNN-based Models | Node Aggregation | Operation | Key Concepts |
|------------------|------------------|-----------|--------------|
| GNN (Scarselli et al., 2008) | $x_j = f \left( l_j, l_{co}, x_{ne}, l_{ne} \right)$ | $f(\cdot)$ | Transduction function |
| GCN (Kipf and Welling, 2017) | $E^\ell = \sigma \left( D^{\frac{1}{2}} A D^{\frac{1}{2}} E^{\ell-1} W^{\ell-1} \right)$ | $D^{\frac{1}{2}} A D^{\frac{1}{2}}$ | Convolutional operation |
| GAT (Velickovic et al., 2018) | $a^\ell_{ij} = \sigma \left( \sum_{e_{kj} \in \mathcal{E}_{ij}} \alpha^\ell_{ikj} \cdot W^{\ell-1} e^{\ell-1} \right)$ | $\alpha_{ij}$ | Local attention |
| EIGAT (our proposed) | $a^\ell_{ij} = \sigma \left( \sum_{e_{kj} \in \mathcal{E}_{ij}} REI_{\ell} (e_i) \sum_{k \in \mathcal{N}_{ij}} \alpha^\ell_{ikj} \cdot v_{ikj}^\ell \right)$ | $REI_{\ell} (e_i), \alpha_{ikj}$ | Global information + Local attention |

...ings to predict links, such as ConvE (Dettmers et al., 2018), ConvKB (Nguyen et al., 2018), InteractE (Vashishth et al., 2020a), ReInceptionE (Xie et al., 2020), and ParamE (Che et al., 2020). (iv) Graph neural network-based models, such as R-GCN (Schlichtkrull et al., 2018), A2N (Bansal et al., 2019), Nathani’s (Nathani et al., 2019), RGHAT (Zhang et al., 2020c), ATT (Chami et al., 2020), which yielded state-of-the-art performance for KRL. Along this line, we focus on a GNN-based approach to deal with knowledge representation learning task.

2.2 Graph Attention Networks (GATs)

Graph Neural Networks (GNNs) develop a deep neural network to deal with arbitrary graphs for representation learning (Scarselli et al., 2008; Zhou et al., 2019; Hou et al., 2020). Graph Convolutional Networks (GCNs) are one of their most prominent progress (Schlichtkrull et al., 2018; Wu et al., 2019a; Xu et al., 2019; Vashishth et al., 2020b), which generalize local convolutional operation on the graph-structured data, i.e. gather information from one-hop neighbors and all neighbors contribute equally in the message passing. Inspired by the successful development of the attention mechanism in NLP and CV, Velickovic et al. (2018) proposed Graph Attention Networks (GATs) by incorporating local attention mechanism (Vaswani et al., 2017; Qian et al., 2018; Lu and Li, 2020) into GCNs, which calculate the hidden states of each node by attending over its neighbors (Thekumparampil et al., 2018; Lee et al., 2018b; Yang et al., 2019).

Recently, several advanced extensions of GATs were proposed for operating on knowledge graphs. Han et al. (2018) proposed to jointly apply attention to KGs and external text data. Busbridge et al. (2019) proposed RGAT by extending non-relational GATs to incorporate relational information, but with poor performance. Nathani et al. (2019) proposed a triple-level attention model that captures the integrated features of both entity and relation in a given entity’s neighborhood, and Zhang et al. (2020c) proposed a two-level hierarchical attention mechanism. These studies are related to our work in the sense that we all use GNNs to capture more structural information in KGs. However, all of them ignore global information in local attention computation.

Most recently, (Xu et al., 2020) proposed a Transformer-based model to enhance the copy mechanism for abstractive summarization by considering the global importance of each source word based on the degree centrality in the Transformer, which inspires our idea of incorporating global information in local attention for KRL. Table 1 summarizes the key concepts and other different settings of GNNs.

3 Methodology

In this section, we introduce the details of the proposed EIGAT model that incorporates global information in local attention for knowledge representation learning on KGs. We start by describing a single entity importance-aware graph attention layer, which is the building block of our model’s overall architecture. Before that, we briefly introduce the notations of this paper.

**Notations.** In a graph attention networks with $L$ layers, the input to $\ell$-th layer ($\ell = 1, \ldots, L$) are two embedding sets: (I) the output entity embeddings from $(\ell-1)$-th layer which is represented by a matrix $E^{\ell-1} \in \mathbb{R}^{N_e \times N_\ell}$, $E^{\ell-1} = \{e^1_1, e^2_1, \ldots, e^N_{N_{\ell-1}}\}$, where $N_e$ is the num-
ber of entities, and \( \eta^\ell-1 \) is the dimension of output entity embedding in \((\ell-1)\)-th layer. (2) the output relationship embeddings from \((\ell-1)\)-th layer, denoted by a matrix \( \mathbf{R}^\ell-1 \in \mathbb{R}^{\ell-1 \times N_r} \), \( \mathbf{R}^\ell-1 = \{ \mathbf{r}_1^\ell-1, \mathbf{r}_2^\ell-1, \ldots, \mathbf{r}_{N_r}^\ell-1 \} \), where \( N_r \) and \( \zeta^\ell-1 \) represent the number of relationships and the output relationship’s feature dimension in \((\ell-1)\)-th layer, respectively. The \( \ell \)-th layer then produces the corresponding new output embedding matrices (of potentially different cardinality), \( \mathbf{E}^\ell \in \mathbb{R}^{N_e \times N_r} \) and \( \mathbf{R}^\ell \in \mathbb{R}^{g \times N_r} \). Specifically, we describe the \( \ell \)-th graph attention layer.

3.1 Local Attention Evaluation

A triple relation \( t_{ij}^k = (e_i, r_k, e_j) \) indicates a relationship \( r_k \) between head entity \( e_i \) and tail entity \( e_j \). Following (Nathani et al., 2019), the representation \( \mathbf{v}_{ikj}^\ell \) of the triple \( t_{ij}^k \) is built as follows:

\[
\mathbf{v}_{ikj}^\ell = \mathbf{W}_1^\ell \cdot [\mathbf{e}_i^{\ell-1} \parallel \mathbf{r}_k^{\ell-1} \parallel \mathbf{e}_j^{\ell-1}],
\]

where \( \mathbf{W}_1^\ell \) denotes a linear transformation matrix in \( \ell \)-th layer, \( \mathbf{e}_i^{\ell-1} \), \( \mathbf{r}_k^{\ell-1} \) and \( \mathbf{e}_j^{\ell-1} \) denote the output embeddings of \( e_i, r_k \) and \( e_j \) in \((\ell-1)\)-th layer, respectively. \( \parallel \) represents concatenation. We then calculate the absolute relation attention value \( b_{ikj}^\ell \) of the triple \( t_{ij}^k \).

\[
b_{ikj}^\ell = \text{LeakyReLU} \left( \mathbf{W}_2^\ell \cdot \mathbf{v}_{ikj}^\ell \right),
\]

where \( \mathbf{W}_2^\ell \) and LeakyReLU are a linear weight vector in \( \ell \)-th layer and a non-linearity active function respectively that act upon the embedding \( \mathbf{v}_{ikj}^\ell \) in turn. We then utilize softmax to evaluate the relative relation attention value \( \alpha_{ikj}^\ell \) of the triple \( t_{ij}^k \) in \( \ell \)-th layer.

\[
\alpha_{ikj}^\ell = \text{softmax}_{ik} \left(b_{ikj}^\ell \right) = \frac{\exp \{b_{ikj}^\ell \}}{\sum_{e_m \in \text{In}(e_i), r_r \in \mathcal{R}_{nj}} \exp \{b_{irm}^\ell \}}.
\]

\( \text{In}(e_j) \) denotes the neighbors pointing to targeted tail entity \( e_j \). \( \mathcal{R}_{nj} \) denotes the set of relationships between \( e_n \) and \( e_j \).

3.2 Global Entity Importance Estimation

To obtain global entity importance \( EI(e_i) \) of an entity \( e_i \), we formally introduce a relation attention-based global random walk method, as follows:

\[
EI(e_i)^\ell = (1 - d) + d \times \sum_{e_m \in \text{In}(e_i), r_r \in \mathcal{R}_{mi}} \sum_{e_n \in \text{Out}(e_m), r_r \in \mathcal{R}_{mn}} b_{mrm}EI(e_m)^{\ell-1},
\]

where \( d \) is a hyperparameter denoting the probability that an imaginary surfer randomly moves to a neighboring entity. \( (1 - d) \) denotes the probability of teleporting to any other entities randomly, which is able to alleviate the information island problem caused by the isolated entities that lack of any in-degree or out-degree neighbors (e.g. \# median in-degree=0 in NELL-995 in Table 2). \( Out(e_m) \) denotes the neighborhoods that an entity \( e_m \) points to. \( EI(e_m)^{\ell-1} \) denotes the EI score of the entity \( e_m \) in \((\ell-1)\)-th iteration. The random walk distance \( t \) depends on both the number of attention layers \( L \) and training epochs \( C, t \in (1, L \times C) \). The relation weights (e.g. \( b_{mri} \)) are calculated by Equation (2). Unlike conventional fixed weights-based random walk methods (Mihalcea and Tarau, 2004; Florescu and Caragea, 2017), a novelty is that the dynamic relation weights (e.g. \( b_{mri} \)) are iteratively and automatically optimized during training by the graph attention mechanism. In line with the theoretical desiderata for modeling node importance in MRGs, this method develops the following essential characteristics: (i) Neighborhood-awareness, i.e. neighboring EI scores can be taken into account when a given entity’s importance score is modeled. (ii) Relationship-awareness, i.e. different relationships could play a different role in propagating EI score. (iii) Centrality-awareness, i.e. more central nodes inherently and reasonably would be more important than less central nodes. (iv) Universal and flexible, i.e. it utilizes only graph global structural information.

3.3 Incorporate Global Information in Local Attention

Though attention mechanism can assign different importance to nodes via learned weights, it is still a local computation. The attention value, e.g., \( \alpha_{ikj}^\ell \) in Equation (3), is the function of pairwise feature interaction within local neighborhood and do not take account of entity importance from global graph structure. To this end, we incorporate global information in local attention computation, as shown in Figure 1 (EIGAT).

Specifically, to generate the output embedding \( \mathbf{e}_j^\ell \) of tail entity \( e_j \) in \( \ell \)-th layer, we incorporate global relative head entity importance \( REI_{e_j}(e_i)^\ell \) in local attention to conduct entity aggregation with its associated triple representations \( \mathbf{v}_{ikj}^\ell \) weighted
by their relative attention values $\alpha_{ikj}^L$, as follows:

$$e_j^L = \sigma \left( \sum_{e_i \in In(e_j)} REI_{e_j}(e_i)^\ell \cdot \sum_{k \in I_{ikj}} \alpha_{ikj}^L \cdot v_{ikj}^L \right), \quad (5)$$

and

$$REI_{e_j}(e_i)^\ell = \text{softmax}_{I_{e_j}(e_i)}((EI(e_i))^\ell) = \frac{\exp{(EI(e_i))^\ell}}{\sum_{e_i' \in I_{e_j}(e_i)} \exp{(EI(e_i')^\ell)}}, \forall e_i \in In(e_j). \quad (6)$$

In Eq. (5), we bring in global relative entity importance $REI_{e_j}(e_i)^\ell$ of different head entities in $In(e_j)$ for learning more about those significant neighboring entities, and thus could get better knowledge representations for the targeted tail entity $e_j$.

To stabilize the learning process of self-attention, as suggested by (Velickovic et al., 2018), we employ multi-head attention. Specifically, $M$ independent attention mechanisms execute the transformation of Eq. (5), and then their features are concatenated as:

$$\tilde{e}_j^\ell = \sigma \left( \sum_{m=1}^{M} \sum_{e_i \in In(e_j)} REI_{e_j}(e_i)^\ell m \cdot \sum_{k \in I_{ikj}} \alpha_{ikj}^m \cdot v_{ikj}^m \right) \quad (7)$$

We conduct a linear transformation on input relationship embedding $r_k^{L-1} \in \mathbb{R}^{\zeta^{L-1}}$ in the $L$-th layer as:

$$r_k^L = W^{\tau,\Omega} \cdot r_k^{L-1}, \quad (8)$$

where $W^{\tau,\Omega} \in \mathbb{R}^{\zeta^L \times \zeta^{L-1}}$ is a weight matrix, $\zeta^{L-1}$ and $\zeta^L$ are dimensions of input and output relationship embeddings, respectively. $r_k^L \in \mathbb{R}^{\zeta^L}$ represents the output relationship embedding in the $L$-th layer.

4 Model Architecture

Our model follows an encoder-decoder framework: (i) the encoder model includes $L$ attention layers, (ii) the decoder model provides a scoring function (Eq. 11) to calculate the likelihood of given triples being valid. Based on it, the KG incompleteness issue is expected to be alleviated by link prediction (Section 5), i.e., inferring possible missing relations, e.g. $(e_i, r_k, ?)$ or $(?, r_k, e_j)$.

4.1 Encoder

Based on a single attention layer introduced above, we build the overall architecture of our encoder model with $L$ layers. In practice, we set $L=2$ for our encoder model. In the final $L$-th layer, instead of concatenation (Equation 7), we employ averaging and delay applying the final non-linearity activation:

$$e_j^L = \sigma \left( \frac{1}{M} \sum_{m=1}^{M} \sum_{e_i \in In(e_j)} REI_{e_j}(e_i)^L m \cdot \sum_{k \in I_{ikj}} \alpha_{ikj}^m v_{ikj}^L \right) \quad (9)$$

To keep initial entity information in the final embedding, we obtain the final entity embedding $\tilde{e}_j^* \in \mathbb{R}^{\eta^L}$ by combining the transformed initial embeddings $\tilde{e}_0^j \in \mathbb{R}^{\eta^0}$ and the output entity embedding $\tilde{e}_j^L \in \mathbb{R}^{\eta^L}$ of the $L$-th layer, as follows:

$$\tilde{e}_j^* = W^* \cdot \tilde{e}_0^j + \tilde{e}_j^L, \quad \forall e \in \mathcal{E}. \quad (10)$$

$W^* \in \mathbb{R}^{\eta^L \times \eta^0}$ is a projecting matrix. The initial entity embeddings (i.e. $\tilde{e}_0^j, \forall e \in \mathcal{E}$) and relationship embeddings (i.e. $\tilde{r}_0^\tau, \forall \tau \in \mathcal{R}$) are pre-trained by Bordes et al. (2013).

4.2 Decoder

Among the existing KG completion (KGC) models, we utilize the most recent model ConvKB (Nguyen et al., 2018) as decoder model3. Given a triple $t_{ij}^k$, the scoring function is formally defined as:

$$f(t_{ij}^k) = \left( \prod_{m=1}^{[\Omega]} g((e_i^*, r_k^L, e_j^*) \ast \omega^m) \right) \cdot W, \quad (11)$$

where $\Omega$ denotes the set of filters, $\tau=[\Omega]$ and $\omega \in \Omega$. $\Omega$ and $W$ are shared parameters and independent of $e_i, r_k$ and $e_j$. $g(\cdot)$ is an activation function such as ReLU. $\ast$ denotes a convolution operator. These $\tau$ feature maps are concatenated into a single vector $\in \mathbb{R}^{\tau \xi}$ which is then computed with a weight vector $W \in \mathbb{R}^{\tau \omega}$ via a dot product to give a likelihood score for the triple $t_{ij}^k$. $\phi$ denotes the dimension of entity and relation embeddings. In practice, we set $\phi=\eta^L \ast \zeta^L$ for ConvKB.

4.3 Optimization

We utilize a two-step training procedure for the encoder-decoder framework, which is a routine optimization way for it (Zhou et al., 2019). (i) We first train the encoder model to learn the embeddings of entities and relationships, by minimizing

---

3We choose ConvKB here, and it is not difficult for other KGC methods, such as CapsE (Nguyen et al., 2019), ConvE (Dettmers et al., 2018), etc. Note that we also tried different models as decoder, but found that ConvKB performs best.
a hinge-loss function, as follows:

\[ L_1 = \sum_{t_{ij} \in G} \sum_{t'_{ij} \in G'} \max \left\{ h_{t_{ij}}, -h_{t'_{ij}} + \gamma, 0 \right\}. \]  

(12)

Here, \( h_{t_{ij}} = \| e_i^* + r_k^* - e_j^* \|_{\ell_1} \) indicates the translational scoring function of the triple \( t_{ij} \) (Bordes et al., 2013). \( \gamma > 0 \) is a margin hyper-parameter. 

(ii) We then train and learn the parameters of the decoder model ConvKB for link prediction, by minimizing a soft-margin loss function, as follows:

\[ L_2 = \sum_{t'_{ij} \in (G \cup G')} \log \left( 1 + \exp \left( l_{t'_{ij}} \cdot f(t'_{ij}) \right) \right) + \lambda \| W \|_2^2, \]  

(13)

in which, \( l_{t'_{ij}} = \begin{cases} 1, & t'_{ij} \in G \setminus G' \setminus \mathcal{G} \\ -1, & t'_{ij} \in G' \setminus \mathcal{G} \end{cases} \) and \( G' \) are the sets of positive triples and negative triples, respectively.

\[ G' = \{ t'_{ij} | e_i \in \mathcal{E} \setminus e_i \} \cup \{ t'_{ij} | e_j \in \mathcal{E} \setminus e_i \}. \]  

(14)

5 Experiments

We evaluate the effectiveness of our proposed model EIGAT by link prediction (determined by Equation 11), which aims to infer possible missing relations, i.e., predict \( e_j \) given \( (e_i, r_k, ?) \) or predict \( e_i \) given \( (? , r_k, e_j) \).

5.1 Datasets

We use three public benchmark datasets for link prediction experiments, including: Kinship (Lin et al., 2018), NELL-995 (Xiong et al., 2017), FB15K-237 (Toutanova et al., 2015), where we discard another popular dataset WN18RR due to its too sparse to learn global information. The basic statistics of all datasets are included in Table 2. To explore the performance of our proposed model on different datasets with different global topology characteristics, we compute their density value (Coleman and Moré, 1983) and report them in Table 2. Since the densities in NELL-995 is sparser than Kinship and FB15K-237, and its median in-degree even is 0, it is relative hard for global entity importance estimation in NELL-995.

### Table 2: Statistics of datasets.

| Datasets       | #Entities | #Rel | Train | Validation | Test | Total | Mean in-degree | Median in-degree | Density |
|----------------|-----------|------|-------|------------|------|-------|----------------|------------------|---------|
| Kinship        | 104       | 25   | 8544  | 1068       | 1074 | 10,686| 82.15          | 82.5             | 0.998   |
| FB15K-237      | 14,541    | 237  | 272,115 | 17,535  | 20,466 | 310,116 | 18.71           | 8                | 0.001   |
| NELL-995       | 75,492    | 200  | 149,678 | 543      | 3992  | 154,213| 1.98            | 0                | 2.71E-5 |

**Definition 1. (Graph Density).** Graph density aims to measure how sparse a graph is. Similar to (Coleman and Moré, 1983), given a graph \( G \), it’s formally defined as follows:

\[ D(G) = \frac{E}{N(N-1)}, \]  

(15)

where \( N \) denotes the number of nodes in \( G \), and \( E \) denotes the number of edges in \( G \). The lower the \( D(G) \), the sparser the graph is.

**Table 3: Hyperparameters for the encoder model EIGAT on all datasets.**

| Datasets       | Kinship | NELL-995 | FB15K-237 |
|----------------|---------|----------|-----------|
| Learning rate  | 1e-2    | 1e-3     | 1e-3      |
| Weight decay   | 1e-6    | 1e-6     | 1e-6      |
| Epochs         | 4000    | 3000     | 3200      |
| Dropouts       | 0.3     | 0.5      | 0.5       |
| Leaky Relu     | 0.2     | 0.2      | 0.2       |
| nheads         | 2       | 2        | 2         |
| Final dimensions | 200 | 200     | 200       |
| Negative ratio | 1       | 1        | 1         |
| Margin         | 1       | 1        | 1         |
| RW parameter   | 0.85    | 0.85     | 0.85      |

**Table 4: Hyperparameters for the decoder model EIGAT on all datasets.**

| Datasets       | Kinship | NELL-995 | FB15K-237 |
|----------------|---------|----------|-----------|
| Learning rate  | 1e-2    | 1e-3     | 1e-3      |
| Weight decay   | 1e-5    | 5e-6     | 5e-7      |
| Epochs         | 400     | 200      | 200       |
| Dropouts       | 0.0     | 0.3      | 0.2       |
| Filters        | 50      | 400      | 50        |

5.2 Baselines

To demonstrate the effectiveness of our proposed model EIGAT for link prediction, we compare it with the following state-of-the-art (SOTA) baselines:

- TransE (Bordes et al., 2013): a most widely used and early KGC models.
- DistMult (Yang et al., 2015): a popular tensor factorization-based KGC model which uses a bi-linear scoring function to calculate knowledge triples’ scores.
Table 5: **Link prediction results** on *Kinship* and *NELL-995*. The results of baselines are directly taken from the original papers. The best scores are in **bold**.

|          | Kinship | NELL-995 |
|----------|---------|----------|
|          | MR  | MRR  | Hits@1 (%) | Hits@3 (%) | Hits@10 (%) | MR  | MRR  | Hits@1 (%) | Hits@3 (%) | Hits@10 (%) |
| TransE (Bordes et al., 2013) | 6.8 | 0.309 | 0.9 | 64.3 | 84.1 | 2100 | 0.401 | 34.4 | 47.2 | 50.1 |
| DistMult (Yang et al., 2015) | 5.26 | 0.516 | 36.7 | 58.1 | 86.7 | 4213 | 0.485 | 39.9 | 52.8 | 61.6 |
| ComplEx (Trouillon et al., 2016) | 2.48 | 0.823 | 73.3 | 89.9 | 97.11 | 4600 | 0.482 | 40.3 | 53.1 | 61.3 |
| ConvE (Dettmers et al., 2018) | 2.03 | 0.833 | 73.8 | 91.7 | 98.14 | 3560 | 0.491 | 40.3 | 53.1 | 61.3 |
| ConvKB (Nguyen et al., 2018) | 3.3 | 0.614 | 43.62 | 75.5 | 95.3 | **600** | 0.43 | 37.0 | 47 | 54.5 |
| R-GCN (Schlichtkrull et al., 2018) | 25.92 | 0.109 | 3 | 8.8 | 23.9 | 7600 | 0.12 | 8.2 | 12.6 | 18.8 |
| Nathani’s (Nathani et al., 2019) | 1.94 | 0.904 | 85.9 | 94.1 | 98 | 965 | 0.530 | 44.7 | 56.4 | 69.5 |
| EIGA T (Ours) | **1.66** | **0.963** | **94.8** | **96.6** | **98.4** | | | | | |

- ComplEx (Trouillon et al., 2016): an advanced extension of DistMult which encodes entities and relationships into complex vector space instead of real-valued vector space.
- ConvE (Dettmers et al., 2018): a popular convolutional network-based KGC model.
- ConvKB (Nguyen et al., 2018): another SOTA convolutional network-based KGC model.
- R-GCN (Schlichtkrull et al., 2018): an advanced extension of GCN that can effectively model multi-relational data.
- Nathani’s (Nathani et al., 2019): a recent KGC model that models the local neighborhood via graph relational attention network.
- A2N (Bansal et al., 2019): a recent model that learns query-dependent representations of entities based on a GNN structure.
- HAKE (Zhang et al., 2020b): a SOTA KGC model that models semantic hierarchies
- InteractE (Vashishth et al., 2020a): a recent extension of ConvE that increase the interaction between relation and entity embeddings.
- ReInceptionE (Xie et al., 2020): a recent extension of ConvE that uses local-global structural information.
- ParamE (Che et al., 2020): another extension of ConvE that use relation embeddings.
- ATTH (Chami et al., 2020): a SOTA model that use hyperbolic space and attention-based geometric transformation.
- RGHAT (Zhang et al., 2020c): a SOTA KGC model that models the local neighborhood via hierarchical attention mechanism.

### 5.3 Evaluation Protocol

We utilize ranking criteria for evaluation. For each testing triple, we remove the head entity or tail entity and replace it by each of the entities in $E$ in turn. The model scores of the corrupted triples would be computed by the decoder model (Eq. 11) and then sorted by descending order. We can obtain the exact rank of the correct triple in the candidates. Similar to most baselines, we report the experimental results in “Filter” setting, i.e. removing corrupted triples that are already present in datasets during ranking. The evaluation metrics include: the mean reciprocal rank (MRR), mean rank (MR), and the proportion of correct entities ranked in the top N (HITS@N, N=1, 3, 10).

### 5.4 Training Protocol

Table 3 and Table 4 report the detailed hyperparameter settings of encoder and decoder models for EIGA T, respectively. In the training, we set $M=2$ heads attention mechanism. The final dimensions of entity and relation embeddings are set to 200. The slop parameter $\alpha$ of LeakyReLU in Eq. (2) is set as 0.2 on all datasets. We use auxiliary relations from 2-hop neighborhood to aggregate more information about the neighborhoods. EI scores are initialized randomly in (0,1). We utilize a typical value for $d = 0.85$ (Mihalcea and Tarau, 2004; Florescu and Caragea, 2017).

### 5.5 Results and Analysis

Table 5 and Table 6 demonstrate the results of link prediction (significance level of 0.05). We can observe that: (i) The results clearly indicate that
Table 6: Link prediction results on FB15K-237. The results of baselines are directly taken from the original papers. The best scores are in **bold**.

|           | FB15K-237 | Hits@N (%) |
|-----------|------------|------------|
|           | MR  | MRR  | @1 | @3 | @10 |
| TransE (Bordes et al., 2013) | 323 | 0.279 | 19.8 | 37.6 | 44.1 |
| DistMult (Yang et al., 2015) | 512 | 0.281 | 19.9 | 30.1 | 44.6 |
| ComplEx (Trouillon et al., 2016) | 546 | 0.278 | 19.4 | 29.7 | 45 |
| ConvE (Dettmers et al., 2018) | 245 | 0.312 | 22.5 | 34.1 | 49.7 |
| ConvKB (Nguyen et al., 2018) | 216 | 0.289 | 19.8 | 32.4 | 47.1 |
| R-GCN (Schlichtkrull et al., 2018) | 600 | 0.164 | 10 | 18.1 | 30 |
| ATT (Chami et al., 2020) | - | 0.348 | 25.2 | 38.4 | 54.0 |
| HAKE (Zhang et al., 2020b) | - | 0.346 | 25.0 | 38.1 | 54.2 |
| R-GCN (Schlichtkrull et al., 2018) | 172 | 0.354 | 26.3 | - | 53.5 |
| InteractE (Xue et al., 2020) | 173 | 0.349 | - | - | 52.8 |
| ParamE (Che et al., 2020) | - | 0.399 | 31.0 | 43.8 | 57.6 |
| RGHA T (Zhang et al., 2020c) | 196 | 0.522 | 46.2 | 54.6 | 63.1 |
| EIGAT (Ours) | 154 | 0.541 | 47.6 | 57.1 | 66.1 |

EIGAT significantly and consistently outperforms all state-of-the-art baselines on most metrics in all benchmark datasets, which demonstrate the effectiveness of our proposed model. (ii) The advantages EIGAT compared to baselines on NELL-995 seem to be smaller than others. It is because that rich global structural information in relative dense graphs, i.e., Kinship, and FB15K-237, leads to more effective entity importance estimation by global random walk methods, comparing with less global structural knowledge in relative sparse graphs, i.e., NELL-995. The results demonstrate NELL-995 is more difficult than others for EIGAT to learn, but the comparable results also verify the effectiveness and robustness of our model on both scenarios.

5.6 Ablation Study

To analyze the behavior of global information in EIGAT, we compare EIGAT with EIGAT-Remove-global (i.e., removing global entity importance from EIGAT). The comparison results in Table 7 indicate that EIGAT achieves improvements against EIGAT-Remove-global on all metrics. In particular, on MR, EIGAT surpasses EIGAT-Remove-global by a large margin 56. The results demonstrate our model can successfully take account of global information in local attention to aggregate more effective entity representations.

|           | FB15K-237 | Hits@N (%) |
|-----------|------------|------------|
|           | MR  | MRR  | @1 | @3 | @10 |
| EIGAT-Remove-global | 210 | 0.518 | 46 | 54 | 62.6 |
| EIGAT | 154 | 0.541 | 47.6 | 57.1 | 66.1 |

5.7 Case Study

Table 8 gives examples of entity prediction results of EIGAT on the FB15k-237 testing set (predicting tail entities). This illustrates the efficacy of our proposed EIGAT. Given a head entity and a relation, the top predicted tail entities (and the true one) are depicted. Even if the true fact is not always at the best front, the predicted results can still reflect common-sense.

Table 8: Example predictions on the FB15K-237 test set using EIGAT. **Bold** indicates the test triplet’s true tail and italics other true tails present in the training set.

| Head Entity | Relation | Tail Entities |
|-------------|----------|---------------|
| X-Men       | production companies | Marvel Entertainment, DC Comics, 20th Century Studios, American Zoetrope |
| United States of America | form of government | presidential regime, Democracy, republic, parliamentary monarchy, parliamentary system |
| Belgium     | time zones | Central European Time, Atlantic Time Zone, Belgium |

6 Conclusion and Future Work

In this paper, we propose to incorporate global information in local attention for knowledge representation learning and introduce a novel GAT-based model that incorporates global entity importance. In particular, we provide an attention-based global random walk approach to estimate entity importance. The experimental results of entity prediction demonstrate that our model can successfully take into account global information in local attention to improve knowledge representation learning. Interesting future work directions include generalizing EIGAT to other relational graphs (e.g. heterogeneous information network (HIN), user-item graph in recommendation system), and exploring an advanced variant of EIGAT in a semi-supervised learning scenario.

Acknowledgements

The authors would like to thank all anonymous reviewers for their insightful comments. This work was supported by the National Natural Science Foundation of China under Grant No. 61906159, U1836206, 61773361, 62072379.
References

Sami Abu-El-Haija, Bryan Perozzi, Rami Al-Rfou, and Alex Alemi. 2018. Watch your step: Learning node embeddings via graph attention. In Proceedings of NeurIPS, pages 9180–9190.

Ivana Balazevic, Carl Allen, and Timothy M. Hospedales. 2019. Tucker: Tensor factorization for knowledge graph completion. In Proceedings of EMNLP, pages 5185–5194.

Esma Balkir, Masha Naslidnyk, Dave Palfrey, and Arpit Mittal. 2019. Using pairwise occurrence information to improve knowledge graph completion on large-scale datasets. In Proceedings of EMNLP, pages 3591–3596.

Trapit Bansal, Da-Cheng Juan, Sujith Ravi, and Andrew McCallum. 2019. A2N: Attending to neighbors for knowledge graph inference. In Proceedings of ACL, pages 4387–4392.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In Proceedings of NeurIPS, pages 2787–2795.

Sergey Brin and Lawrence Page. 1998. The anatomy of a large-scale hypertextual web search engine. Computer Networks and ISDN Systems, 30:1–7.

Dan Busbridge, Dane Sherburn, Pietro Cavallo, and Sergey Brin. 1998. The anatomy of a large-scale hypertextual web search engine. Journal of the ACM (JACM), 46(5):604–632.

Jon M. Kleinberg. 1999. Authoritative sources in a hyperlinked environment. Journal of the ACM (JACM), 46(5):604–632.

John Boaz Lee, Ryan Rossi, and Xiangnan Kong. 2018a. Graph classification using structural attention. In Proceedings of SIGKDD, pages 1–9.

John Boaz Lee, Ryan A. Rossi, Sungchul Kim, and Anup Rao. 2019. Graph convolutional networks with motif-based attention. In Proceedings of ICLR.

Xu Han, Zhiyuan Liu, and Maosong Sun. 2018. Neural knowledge acquisition via mutual attention between knowledge graph and text. In Proceedings of AAAI.

Huiting Hong, Hantao Guo, Yucheng Lin, Xiaoqing Yang, Zang Li, and Jieping Ye. 2020. An attention-based graph neural network for heterogeneous structural learning. In Proceedings of AAAI.

Yifan Hou, Jian Zhang, James Cheng, Kaili Ma, Richard T. B. Ma, Hongzhi Chen, and Ming-Chang Yang. 2020. Measuring and improving the use of graph information in graph neural networks. In Proceedings of ICLR.

Guoliang Ji, Shizhu He, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Knowledge graph embedding via dynamic mapping matrix. In Proceedings of ACL, pages 687–696.

Yuchen Jing, Xiuzehe Zhang, Lifang Wu, Jinqiao Wang, Zemeng Feng, and Dan Wang. 2014. Recommendation on flickr by combining community user ratings and item importance. In Proceedings of ICME, pages 1–6.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In Proceedings of ICLR.

Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. 2015. Learning entity and relation embeddings for knowledge graph completion. In Proceedings of AAAI, pages 2181–2187.

Qi Liu, Biao Xiang, Nicholas Jing Yuan, Enhong Chen, Hui Xiong, Yi Zheng, and Yu Yang. 2017. An influence propagation view of pagerank. ACM TKDD, 11(3):1–30.
Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into text. In Proceedings of EMNLP, pages 404–411.

Deepak Nathani, Jatin Chauhan, Charu Sharma, and Manohar Kaul. 2019. Learning attention-based embeddings for relation prediction in knowledge graphs. In Proceedings of NAACL.

Dai Quoc Nguyen, Tu Dinh Nguyen, Dat Quoc Nguyen, and Dinh Phung. 2018. A novel embedding model for knowledge base completion based on convolutional neural network. In Proceedings of NAACL, pages 327–333.

Dai Quoc Nguyen, Thanh Vu, Tu Dinh Nguyen, Dat Quoc Nguyen, and Dinh Phung. 2019. A capsule network-based embedding model for knowledge graph completion and search personalization. In Proceedings of NAACL, pages 2180–2189.

Maximilian Nickel, Lorenzo Rosasco, and Tomaso Poggio. 2016. Holographic embeddings of knowledge graphs. In Proceedings of AAAI, pages 1955–1961.

Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. 2011. A three-way model for collective learning on multi-relational data. In Proceedings of ICML, pages 809–816.

Namyoung Park, Andrey Kan, Xin Luna Dong, Tong Zhao, and Christos Faloutsos. 2019. Estimating node importance in knowledge graphs using graph neural networks. In Proceedings of SIGKDD, pages 596–606.

Wei Qian, Cong Fu, Yu Zhu, Deng Cai, and Xiaofei He. 2018. Translating embeddings for knowledge graph completion with relation attention mechanism. In Proceedings of IJCAI, pages 4286–4292.

Apoorv Saxena, Aditay Tripathi, and Partha Talukdar. 2020. Improving multi-hop question answering over knowledge graphs using knowledge base embeddings. In Proceedings of ACL.

Franco Scarselli, Marco Gori, Ah Chung Tsoi, and Gabriele Monfardini. 2008. The graph neural network model. IEEE TNN, 20(1):61–80.

Michael Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In Proceedings of ESWC, pages 593–607.

Baoxu Shi and Tim Weninger. 2017. Proje: Embedding projection for knowledge graph completion. In Proceedings of AAAI, pages 1236–1242.

Richard Socher, Danqi Chen, Christopher D. Manning, and Andrew Y. Ng. 2013. Reasoning with neural tensor networks for knowledge base completion. In Proceedings of NeurIPS, pages 926–934.

Kiran K. Thekumparampil, Chong Wang, Sewoong Oh, and Li-Jia Li. 2018. Attention-based graph neural network for semi-supervised learning. In arXiv:1803.03735v1.

Kristina Toutanova, Danqi Chen, Patrick Pantel, Hoi-fung Poon, Pallavi Choudhury, and Michael Gamon. 2015. Representing text for joint embedding of text and knowledge bases. In Proceedings of EMNLP, pages 1499–1509.

Theo Trouillon, Johannes Welbl, Sebastian Riedel, Eric Gaussier, and Guillaume Bouchard. 2016. Complex embeddings for simple link prediction. In Proceedings of ICML, pages 2071–2080.

Shikhar Vasishth, Soumya Sanyal, Vikram Nitin, Nilesh Agrawal, and Partha P. Talukdar. 2020a. Interacte: Improving convolution-based knowledge graph embeddings by increasing feature interactions. In Proceedings of AAAI, pages 3009–3016.

Shikhar Vasishth, Soumya Sanyal, Vikram Nitin, and Partha Talukdar. 2020b. Composition-based multi-relational graph convolutional networks. In Proceedings of ICLR.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Proceedings of NeurIPS, pages 5998–6008.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2018. Graph attention networks. In Proceedings of ICLR.

Xiang Wang, Xiangnan He, Yixin Cao, Meng Liu, and Tat-Seng Chua. 2019a. KGAT: Knowledge graph attention network for recommendation. In Proceedings of SIGKDD, pages 950–958.

Xiao Wang, Houye Ji, Chuan Shi, Bai Wang, Peng Cui, P. Yu, and Yanfang Ye. 2019b. Heterogeneous graph attention network. In Proceedings of WWW, pages 2022–2032.

Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014. Knowledge graph embedding by translating on hyperplanes. In Proceedings of AAAI, pages 1112–1119.

Jianshu Weng, Ee-Peng Lim, Jing Jiang, and Qi He. 2010. TwitterRank: Finding topic-sensitive influential twitterers. In Proceedings of WSDM, pages 261–270.
Felix Wu, Tianyi Zhang, Amauri Holanda de Souza Jr., Christopher Fifty, Tao Yu, and Kilian Q. Weinberger. 2019a. Simplifying graph convolutional networks. In Proceedings of ICML.

Jun Wu, Jingrui He, and Jiejun Xu. 2019b. Demo-net: Degree-specific graph neural networks for node and graph classification. In Proceedings of SIGKDD, pages 406–415.

Han Xiao, Minlie Huang, and Xiaoyan Zhu. 2016. TransG: A generative model for knowledge graph embedding. In Proceedings of ACL, pages 2316–2325.

Han Xiao, Minlie Huang, and Xiaoyan Zhu. 2017. SSP: Semantic space projection for knowledge graph embedding with text descriptions. In Proceedings of AAAI, pages 3104–3110.

Zhiwen Xie, Guangyou Zhou, Jin Liu, and Jimmy Xiangji Huang. 2020. ReInceptionE: Relation-aware inception network with joint local-global structural information for knowledge graph embedding. In Proceedings of ACL, pages 5929–5939.

Wenhan Xiong, Thien Hoang, , and William Yang Wang. 2017. DeepPath: A reinforcement learning method for knowledge graph reasoning. In Proceedings of EMNLP.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. 2019. How powerful are graph neural networks? In Proceedings of ICLR. arXiv preprint arXiv:1810.00826.

Song Xu, Haoran Li, Peng Yuan, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2020. Self-attention guided copy mechanism for abstractive summarization. In Proceedings of ACL, pages 1355–1362.

Bishan Yang, Wen tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2015. Embedding entities and relations for learning and inference in knowledge bases. In Proceedings of ICLR.

Yiding Yang, Xinchao Wang, Mingli Song, Junsong Yuan, and Dacheng Tao. 2019. SPAGAN: Shortest path graph attention network. In Proceedings of IJCAI, pages 4099–4105.

Jiani Zhang, Xingjian Shi, Junyuan Xie, Hao Ma, Irwin King, and Dit-Yan Yeung. 2018. GAAN: Gated attention networks for learning on large and spatiotemporal graphs. In Proceedings of UAI.

Kai Zhang, Yaokang Zhu, Jun Wang, and Jie Zhang. 2020a. Adaptive structural fingerprints for graph attention networks. In Proceedings of ICLR.

Zhanqiu Zhang, Jianyu Cai, Yongdong Zhang, and Jie Wang. 2020b. Learning hierarchy-aware knowledge graph embeddings for link prediction. In Proceedings of AAAI, pages 3065–3072.