Efficient Parallel Translating Embedding For Knowledge Graphs

Denghui Zhang, Manling Li, Yantao Jia, Yuanzhuo Wang, Xueqi Cheng
Institute of Computing Technology, Chinese Academy of Sciences
No.6 Kexueyuan South Road, Zhongguancun, Haidian, Beijing, China 100190
zhangdenghui@ict.ac.cn, limanling@software.ict.ac.cn, jiayantao@ict.ac.cn, wangyuanzhuo@ict.ac.cn, cxq@ict.ac.cn

ABSTRACT
Knowledge graph embedding aims to embed entities and relations of knowledge graphs into low-dimensional vector spaces. Translating embedding methods regard relations as the translation from head entities to tail entities, which achieve the state-of-the-art results among knowledge graph embedding methods. However, a major limitation of these methods is the time consuming training process, which may take several days or even weeks for large knowledge graphs, and result in great difficulty in practical applications. In this paper, we propose an efficient parallel framework for translating embedding methods, called ParTrans-X, which enables the methods to be paralleled without locks by utilizing the distinguished structures of knowledge graphs. Experiments on two datasets with three typical translating embedding methods, i.e., TransE [3], TransH [19], and a more efficient variant TransE-AdaGrad [11] validate that ParTrans-X can speed up the training process by more than an order of magnitude.

CCS CONCEPTS
• Computing methodologies → Reasoning about belief and knowledge;

KEYWORDS
Knowledge Graph Embedding, Translation-based, Parallel

ACM Reference format:

1 INTRODUCTION
Knowledge graphs are structured graphs with various entities as nodes and relations as edges. They are usually in form of RDF-style triples \((h, r, t)\), where \(h\) represents a head entity, \(t\) a tail entity, and \(r\) the relation between them. In the past decades, a quantity of large scale knowledge graphs have sprung up, e.g., Freebase [2], WordNet [14], YAGO [12], OpenKN [7], and have played a pivotal role in supporting many applications, such as link prediction, question answering, etc. Although these knowledge graphs are very large, i.e., usually containing thousands of relation types, millions of entities and billions of triples, they are still far from complete. As a result, knowledge graph completion (KGC) has been payed much attention to, which mainly aims to predict missing relations between entities under the supervision of existing triples.

Recent years have witnessed great advances of translating embedding methods to tackle KGC problem. The methods represent entities and relations as the embedding vectors by regarding relations as translations from head entities to tail entities, such as TransE [3], TransH [19], TransR[11], etc. However, the training procedure is time consuming, since they all employ stochastic gradient descent (SGD) to optimize a translation-based loss function, which may require days to converge for large knowledge graphs.

For instance, Table 1 shows the complexity of typical translating embedding methods, where \(T_{total}\) stands for the total training time with \(T_{epoch}\) for the time of each epoch, and one epoch is a single pass over all triples. \(n_h, n_r,\) and \(n_t\) are the number of entities, relations and triples in the knowledge graph respectively. \(d\) is the embedding dimension which is the same for entities and relations in this case, and \(e_p\) is the minimum epochs which used to be set to 1000. It can be seen that the time complexity of TransE is proportional to \(n_h, d\) and \(e_p\). When \(d\) is 100 and \(e_p\) is 1000, it will take 78 minutes for TransE to learn the embeddings of Freebase with 1.9 billion triples, which results in approximately 3932 times the training time, namely, 212 days. Furthermore, the whole Freebase contains over 3 billion triples4, and it will take about 357 days to learn the embeddings of it. Despite its large size, Freebase still suffers from data incomplete problem, e.g., 75% persons do not have nationalities in Freebase [5]. On top of that, most improved variants of TransE employ more complex loss function to better train the embedding vectors, thus they possess higher time complexity or model complexity, and the training time of them will be even unbearable. For example, it will take more than 59 years for Freebase-rdf-latest when employing TransR, which is one of the typical improved variants and achieves far better performance than TransE.

There have been attempts to resolve the efficiency issue of translating embedding methods for knowledge graphs. Pasquale[15] proposed TransE-AdaGrad to speed up the training process by leveraging adaptive learning rates. However, TransE-AdaGrad essentially reduces the number of epochs to converge, and still cannot

\footnotetext[1]{The experiments are conducted on a dual Intel Xeon E5-2640 CPUs (10 cores each \(\times\) 2 hyperthreading, running at 2.4 GHz) machine with 126GB of RAM. The kernel is Red Hat 4.4.7


\[ h, r, t \] denotes a head entity, a relation and a tail entity.}
do well with large scale knowledge graphs. In fact, with more and more computation resources available, it is natural and more effective to parallel these embedding methods, which will lead to significant improvement in training efficiency and can scale to quite large knowledge graphs if given sufficient hardware resources.

However, it is challenging to parallel the translating embedding methods, since the training processes mainly employ stochastic gradient descent algorithm (SGD) or the variants of it. SGD is inherently sequential, as a dependence exists between each iteration. Parallelizing translating embedding methods straightforwardly will result in collisions between different processors. For instance, an entity embedding vector is updated by two processors at the same time, and the gradients calculated by these processors are different. In this case, the diverse gradients are called collisions. To avoid collisions, some methods [9] lock embedding vectors, which will slow the training process greatly as there are so many vectors. On the contrary, updating vectors without locks leads to high efficiency, but should be based on specific assumptions [4, 16]. Since the lock-free training process may result in poor convergence if adopting suboptimal strategy to resolve collisions.

Our key observation of translating embedding methods is that the update performed in one iteration of SGD is based on only one triple and its corrupted sample, which is not necessarily bound up with other embedding vectors. This gives us chance to learn the embedding vectors in parallel without being locked. In this article, we analyze the distinguished data structure of knowledge graphs, and propose an efficient parallel framework for translating embedding methods, called ParTrans-X. It enables translating methods to update the embedding vectors efficiently in shared memory without locks. Thus the training process is greatly speeded up with multi-processors, which can be more than an order of magnitude faster without lowering learning quality.

The contribution of this article is:
1. We explore the law of collisions emerging along with the number of processors, by modelling the training data of knowledge graph into hypergraphs.
2. We propose ParTrans-X framework to train translating methods efficiently in parallel. It utilizes the training data sparsity of large scale knowledge graphs, and can be easily applied to many translating embedding methods.
3. We apply ParTrans-X to typical translating embedding methods, i.e., TransE [3], TransH [19], and a more efficient variant TransE-AdaGrad, and experiments validate the effectiveness of ParTrans-X on two widely used datasets.

The paper is organized as follows. With related work in Sec.2, the collision formulation is introduced in Sec.3 and ParTrans-X is proposed based on it in Sec.4. Then, experiments demonstrate training efficiency of ParTrans-X in Sec.5, with conclusions in Sec.6.

## 2 RELATED WORK
In recent years, translating embedding methods have played a pivotal role in Knowledge Graph Completion, which usually employ stochastic gradient descent algorithm to optimize a translation-based loss function, i.e.,

$$ L = \sum_{(h,r,t)} \max \{0, f_r(h,t) + M - f_r(h,t)\}, \quad (1) $$

where \((h, r, t)\) represents the positive triple that exists in the knowledge graph, while \((h', r', t')\) stands for the negative triple that is not in the knowledge graph. \(\max \{0, \cdot\}\) is the hinge loss and \(M\) is the margin between positive and negative triples. \(f_r(h,t)\) is the score function to determine whether the triple \((h, r, t)\) should exist in the knowledge graph, which varies from different translating embedding methods.

A significant work is TransE [3], which heralds the start of translating embedding methods. It looks upon a triple \((h, r, t)\) as a translation from the head entity \(h\) to the tail entity \(t\), i.e., \(h + r \approx t\), and the score function is \(f_r(h, t) = ||h+r-t||\), where \(||\cdot||\) represents L1-similarity or L2-similarity. The boldface suggests the vectors in the embedding space, namely, \(h, t \in \mathbb{R}^d\), \(r \in \mathbb{R}^d\), where \(d = d_h = d_t\) is the dimension of embedding space, \(d_h\) the dimension for entities and \(d_t\) for relations. Moreover, TransH [19] assumes that it is the projections of entities to a relation-specific hyperplane that satisfy the translation constraint, i.e., \(f_r(h, t) = ||h + r - t||\), where \(h = h + w_r h_w\) and \(t = t - w_r t_w\), with \(w_r \in \mathbb{R}^{d_e}\) as the normal vector of the hyperplane related to \(r\). Furthermore, TransR [11] employs rotation transformation to project the entities to a relation-specific space, i.e., \(f_r(h, t) = ||h_r + r - t||\), where \(h_r = M_r h\) and \(t_r = M_r t\), and \(M_r \in \mathbb{R}^{d_r \times d_e}\) is the projection matrix relation to \(r\). Some works also involves more information to better embedding, e.g., paths [10], margins [8].

Although this category of methods achieve the state-of-the-art results, the main limitation is the computationally expensive training process when facing large scale knowledge graphs. Recently, a method TransE-AdaGrad [15] was proposed to reduce the training time of TransE by employing AdaGrad [6], an variant of SGD, to adaptively modify the learning rate. Although the training time has been reduced greatly, there is still some way to go when facing large scale knowledge graphs. With the computation resources greatly enriched, training in parallel seems to be a more reliable way to relieve this issue. Actually, there are some works, e.g., [18], to parallel some graph computation paradigms, such as online query processing, offline analytics, etc. Nevertheless, it is not easy to train translating embedding methods in parallel, since the main optimization algorithm SGD is born to run in sequence. The major obstacle to parallel SGD is the collisions between updates of different processors for the same parameter [17], to overcome which there are two main branches of methods.

### Table 1: Complexity Analysis of Typical translating embedding Methods, with \(d = 100\), \(ep = 1000\)

<table>
<thead>
<tr>
<th>Model Complexity</th>
<th>Time Complexity</th>
<th>on FB15k</th>
<th>on Freebase-rdf-latest</th>
<th>on the whole Freebase</th>
</tr>
</thead>
<tbody>
<tr>
<td>TransE (O(n_t \times d \times k))</td>
<td>(O(dn_e + dn_r))</td>
<td>4.5s</td>
<td>4658s</td>
<td>17,696s</td>
</tr>
<tr>
<td>TransH (O(n_t \times d \times k))</td>
<td>(O(dn_e + 2dn_r))</td>
<td>6s</td>
<td>100 minutes</td>
<td>6.5 hours</td>
</tr>
<tr>
<td>TransR (O(n_t \times d^2 \times k))</td>
<td>(O(dn_e + dn_r + n_rd^2))</td>
<td>473s</td>
<td>5 days</td>
<td>21.5 days</td>
</tr>
</tbody>
</table>

We analyze the distinguished data structure of knowledge graphs, with other embedding vectors. This gives us chance to learn the AdaGrad, and experiments validate the effectiveness of ParTrans-X on two widely used datasets.
The first branch is to design a strategy to resolve collisions according to specific data structure. For example, Hogwild! [16] is a lock-free scheme works well for sparse data, which means that there is only a small part of parameters to update by each iteration of SGD. It has been proved that processors are unlikely to overwrite each other’s progress, and the method can achieve a nearly optimal rate of convergence. While the second branch is to split the training data to reduce collisions. Downpour SGD [4] mainly employ Distbelief [4] framework, which divides the training data into a number of subsets, then the model replicas run independently on each of these subsets, and do not communicate with each other. Inspired by this, TensorFlow [1] splits a computation graph into a subgraph for every worker and communication takes place using Send/Receive node pairs. Motivated by training large-scale convolutional neural networks for image classification, Elastic Averaging SGD (EASGD) [20] reduces the amount of communication between local workers and the master to allow the parameters of local workers to fluctuate further from the center ones. There are also works to improve the performance in parallel settings, e.g., Delay-tolerant Algorithms for SGD [13] adapts not only to the sequence of gradients, but also to the precise update delays that occur, inspired by AdaGrad.

However, these parallel framework are based on specific assumptions, and can not directly apply to translating embedding models without exploring distinguished data structures of knowledge graphs. Therefore, we shall propose a parallel framework for translating embedding models, called ParTrans-X, as knowledge graphs are mainly in form of triples, and trained triple by triple, it will lead to particular parallel framework.

3 LAW OF COLLISIONS EMERGING IN KG

As mentioned previously, there may exist collisions between processors when they update the same embedding vector, which ends up being one of the most challenging aspects of parallelizing translating embedding methods. Hence, we explore the law of collisions emerging in this section. At first we formulate the training data of knowledge graphs into hypergraphs. Then the collisions in training process are further discussed based on this formulation.

3.1 Hypergraph Formulation

Firstly, we model the knowledge graph formally as \( G = (E, R, T) \), where \( E \) is the set of entities with \( R \) the set of relations, and \( T \) is the set of triples \((h, r, t)\), in which \( h, t \in E \) and \( r \in R \). The cardinalities of \( E, R \) and \( T \) are \( n_e, n_r \) and \( n \) respectively. In this graph, nodes are entities, and edges are triples that connecting nodes with a distinguished relation. For example, the knowledge graph shown in Figure 1(a), where black nodes stand for the entities in knowledge graphs and lines for relations, can be represented as \( G = (E, R, T) \), where \( E = \{e_1, e_2, e_3, e_4, e_5\} \), \( R = \{r_1, r_2, r_3\} \) and \( T = \{(e_1, r_2, e_4), (e_1, r_1, e_2), (e_2, r_3, e_5), (e_4, r_3, e_2)\} \). In this case, \( n_e = 5, n_r = 3 \) and \( n = 4 \).

Secondly, the training data of knowledge graphs can be looked upon as hypergraphs. Recall the loss function of translating embedding methods in Eq.(1), which means in one iteration of SGD, only one positive triple \((h, r, t)\) and one negative triple \((h', r, t')\) are concerned. To be more clear, the data used in one iteration, i.e., \([\{h, r, t\}, \{h', r, t'\}]\), is called a sample. Note that \((h', r, t')\) is constructed by substituting one entity \(h' \in E \) or \(t' \in E \) for \(h \) or \(t \) respectively, contributing to a corrupted triple \((h', r, t) \) or \((h, r, t')\), which is just simply denoted by \((h', r, t')\) following [3]. Consequently, a sample corresponding to three entities, i.e., \(h, t, h' \) or \(t'\), and one relation \(r\). As a result, the training data can be formulated in to a 4-uniform hypergraph, in which all the hyperedges have the same cardinality 4. In this hypergraph, nodes are entities and relations, and edges are training samples containing 4 nodes, i.e., three entities and one relation. More formally,

\[
\text{Definition 3.1.} \quad \text{The training data to embed the knowledge graph } \ G = (E, R, T) \text{ by translating embedding methods is organized as a 4-uniform hypergraph } H = (V, S), \text{ where } V = [E \cup R] \text{ is the set of entities or relations, and } S \text{ is the set of training samples } s, \text{ where } s = \{(h, r, t, h', r, t') : h, r, t, h', t' \in E, r \in R\}.
\]

For example, the hypergraph in Figure 1(b) is one of the hypergraphs generated by Figure 1(a), where black nodes are entities and colored nodes are relations, and the colored blocks represent hyperedges. Here, different colors are related to different relations. For instance, for triple \((e_1, r_1, e_3)\), the negative triple sampled in Figure 1(b) is \((e_1, r_1, e_2)\), which contributes to a sample \(s_1 = \{e_1, r_1, e_2, e_3\}\), thus the hyperedge colored by red contains \(e_1, e_2, e_3 \) and \(r_1\). Note that many other negative triples can be constructed, e.g., \((e_1, r_1, e_3)\) for triple \((e_1, r_1, e_3)\), and the hypergraph generated in Figure 1(b) is just an example. Similarly, the other samples in Figure 1(b) are \(s_2 = \{e_1, r_2, e_4, e_5\}\), \(s_3 = \{e_2, r_3, e_3, e_5\}\) and \(s_3 = \{e_3, r_2, e_2, e_3\}\).

To better analyze the collisions between processors, we define the following statistics of the hypergraph \(H\). Given a hyperedge \(s\),

\[
\sigma(s) = \{s' : \exists \rho \in s \cap s', r \in R\}
\]

denotes the set of hyperedges containing the same relations with hyperedge \(s\).

\[
\hat{\sigma} = \max_{s \in S} \sigma(s)
\]

denotes the maximal number of hyperedges containing same relations, where \(| \cdot |\) denotes the cardinality.

\[
\rho(s) = \{|s' : \exists \rho \in s \cap s', e \in E\|
\]

denotes the set of hyperedges containing one or more same entities with hyperedge \(s\).

\[
\hat{\rho} = \max_{s \in S} \rho(s)
\]

denotes the maximal number of hyperedges containing same entities, where \(| \cdot |\) denotes the cardinality the same as before.
3.2 Collision Formulation

In this section, we will verify it is highly possible that few collisions happen when training by \( p \) processors for large and sparse knowledge graphs. Let \( X_{samp} \) represent the event that \( p \) processors select \( p \) different samples. \( X_{rel} \) represents the event that there are collisions between relations, i.e., different processors updates a same relation vector, and \( X_{ent} \) between entities similarly. The verification is decomposed into two steps, 1) to prove it is quite likely that the processors handle different samples, i.e., \( P(X_{samp} = 1) \approx 1 \), which is the prerequisite to collisions emerging; 2) to prove it is unlikely that these different samples correspond to the same relations or entities, i.e., \( P(X_{rel} = 0) \approx 1 \) and \( P(X_{ent} = 0) \approx 1 \).

Supposing that for embedding methods and the knowledge graph, the training samples \( S = \{ s_1, s_2, \ldots, s_i, \ldots, s_n \} \) of size \( n \) is drawn independently and identically distributed (i.e., i.i.d.) from some unknown distribution \( \mathcal{D} \). Therefore, the probability of \( s_i \) being selected by \( p \) processors is supposed to be

\[
P(s_i = 1) = \frac{1}{n}.
\]

Moreover, according to i.i.d., it is reasonable to assume that the sample selecting process by \( p \) processors is an observation from a Multinomial Distribution, i.e., selecting one sample from \( n \) samples and repeated \( p \) times. Let \( x_i \) denote the number of processors that select \( s_i \) during the same iteration of SGD, then the possibility of \( s_i \) being selected by \( c_1 \) processors, \( \ldots \), \( c_{n_c} \) being selected by \( c_{n_c} \) processors is as follows,

\[
P(x_1 = c_1, \ldots, x_{n_e} = c_{n_e}) = \binom{p}{c_1} \binom{p-c_1}{c_2} \cdots \binom{p-c_{n_c}}{c_{n_c}} \sum_{i=1}^{n_c} c_i = p;
\]

where \( \sum_{i=1}^{n_c} c_i = p \) indicates that there are \( p \) and only \( p \) samples being selected in the same iteration of SGD.

Theorem 3.2. For a knowledge graph with \( n \) triples and training in \( p \) processors in parallel, when \( n \) is large and \( p \) is relatively small, the possibility that \( p \) processors select \( p \) different samples is

\[
P(X_{samp} = 1) \approx 1
\]

with probability at least \( \gamma \), where

\[
\gamma = \prod_{i=1}^{p-1} \left( 1 - \frac{i}{n} \right).
\]

Proof. Given that \( p \) processors select \( p \) different samples, the possibility of relations in a collision can be deduced according to conditional probability as follows,

\[
P(X_{rel} = 0) = P(X_{rel} = 0|X_{samp} = 1) \cdot P(X_{samp} = 1),
\]

where \( P(X_{rel} = 0|X_{samp} = 1) \) is the possibility of \( p \) samples containing distinct relations being selected, which is supposed to be similar to sampling without replacement. More precisely, assuming a sample \( s \) is selected randomly, then the next sample selected \( s' \) should be from \( S - \sigma(s) \), and the third sample \( s'' \) should be selected from \( S - \sigma(s) \cup \sigma(s') \). Accordingly, \( P(X_{rel} = 0|X_{samp} = 1) \) is deduced as follows when \( p < \frac{n}{\sigma} + 1 \) is satisfied,

\[
P(X_{rel} = 0|X_{samp} = 1) = \sum_{s_1, s_2, \ldots, s_p \in S} \frac{1}{n} \frac{|S - \sigma(s_1)| \cdot |S - \sigma(s_1) \cup \sigma(s_2)| \cdots |S - \sigma(s_1) \cup \sigma(s_2) \cdots \cup \sigma(s_p-1)|}{n - (p-1)}
\]

\[
= \prod_{i=1}^{p-1} \left( 1 - \frac{i}{n} \right).
\]

By Eq.(12), the possibility of no collisions between relations in different processors is

\[
P(X_{rel} = 0) = \prod_{i=1}^{p-1} \left( 1 - \frac{i}{n} \right) \prod_{i=1}^{p-1} \left( 1 - \frac{i}{n} \right) = \prod_{i=1}^{p-1} \left( 1 - \frac{i^2}{n^2} \right).
\]

Note that \( p > \frac{n}{\sigma} + 1 \) results in \( 1 - \frac{(p-1)^2}{n^2} < 0 \), which means \( \sigma \) is so large that one or more processors will definitely select the same relation among \( p \) processors, namely, \( P(X_{rel} = 0) = 0 \). Furthermore, when \( \frac{n}{\sigma} \) is relatively small, \( P(X_{rel} = 0) \approx 1 \).
Similarly, the possibility of no entities in a collision can be derived as follows, and no more tautology here due to the limitation of length.

**Theorem 3.4.** For a knowledge graph with $n$ triples and training in $p$ processors in parallel, when $\frac{p}{n}$ is relatively small and $p < \frac{\log n}{\rho} + 1$, the possibility of no collisions between entities is

$$P(X_{ent} = 0) \approx 1$$

with probability at least $\gamma$, where

$$\gamma = \prod_{i=1}^{p-1} \left(1 - \frac{i\rho}{n}\right).$$

It is verified in Theorem 3.2, Theorem 3.3 and Theorem 3.4 that if $n$ is large and $\delta$ and $\rho$ are relatively small, i.e., the knowledge graph is large and sparse, the number of processors $p$ can be very large with supportable collisions, which enables the training process to run in parallel. Motivated by this, we define sparsity of training data in a knowledge graph by $\min\left(\frac{D}{n}, \frac{\rho}{n}\right)$. The smaller its value is, the more processors can be used to parallel the training process. Actually, it is the large and sparse knowledge graphs that are in dire need of parallel translating embedding methods. Since they are far from completion, but are too large to train in serial. Besides, since $\delta$ and $\rho$ is deduced by the worst case, it is reasonable to assume that the average $\delta$ and $\rho$ can better reflect the general structures in knowledge graphs, and the collisions will be less in practice. As a result, we suppose that it would still work well if the average $\delta$ and $\rho$ are relatively small, as a few collisions will not affect the consistency.

### 3.3 Special Insights on Parallelizing TransE

There is an interesting finding that TransE can be further parallelized than other translating embedding methods, since there are less collisions due to the distinguished score function $f_r(h,t) = \|h + r - t\|$. More precisely, the gradient calculation of TransE using $L2$-similarity is as follows,

$$h_k := h_k - \eta \cdot 2((h_k + r_k - t_k) \cdot h'_k := h'_k + \eta \cdot 2(h'_k + r'_k - t'_k)$$

$$r_k := r_k - \eta \cdot 2((h_k + r_k - t_k), r'_k := r'_k + \eta \cdot 2(h'_k + r'_k - t'_k)$$

$$t_k := t_k - \eta \cdot 2((h_k + r_k - t_k), t'_k := t'_k - \eta \cdot 2(h'_k + r'_k - t'_k)$$

where $h_k$ represents the $k$-th dimension of embedding vector $h$, $k \in \{0, 1, \ldots, d\}$, and $d$ is the dimension of embedding space. It can be seen that in TransE, the gradient of each dimension is independent of other dimensions, which means that the collisions between different dimensions of the same embedding vector will not disturb each other. That is to say, only the collisions between the same dimension of the same embedding vector will matter in the training process of TransE.

For example, Figure 2 shows the updating of $h$ by two processors (Processor 1 and Processor 2) at the same time, where $\nabla h$ is the gradient of $h$ calculated by Processor 1, and $\nabla h'$ by Processor 2. Normally, when Processor 2 calculates the gradient $\nabla h'$, the whole embedding vector $h$ will be involved, which is half updated by Processor 1. Obviously, this will result in training errors. On the contrary, if it is the training process of TransE in Figure 2, the calculation of $\nabla h'$ by Processor 2 only concerns the $k$-th dimension $h_k$. As a result, there will be no disturbance between Processor 1 and Processor 2, as long as the two processors are not performing update to the same dimension of the same embedding vector.

Consequently, the possibility of collisions emerging is greatly decreased for TransE. Since not only the entities or relations are the same one, but also the dimensions being updated are the same. Namely, the maximal degree of parallelism is far larger than other translating embedding methods. This indicates that parallelizing without locks is ideally suited for TransE, and may scale well to extremely large knowledge graphs by given sufficient computation resources.

### 4 THE PARTRANS-X FRAMEWORK

Inspired by the findings that collisions between processors are negligible when a knowledge graph is large and sparse, they are stored in shared memory and every processors can perform updates to them at will.

The training process of ParTrans-X starts with initializing the embedding vectors according to Uniform or Bernoulli Distribution, where no parallel section is needed since it takes constant time. However we can parallel the learning process of each epoch, which is the most time consuming part. Running by $p$ processors in parallel can decrease the training epochs by $p$ times, i.e., the parallel training epoch is $ep^p = \frac{ep}{p}$. To do this, we first determine the random sampling seed $seed[i]$ by calling `SEED_RAND` for the $i$-th processor. The random sampling seeds differ from each other to avoid same pseudo-random sequence for different processors. Then, each processor performs embedding learning procedure epoch by epoch asynchronously (lines 5-12). One epoch is a loop over all triples. Each loop is done by firstly normalizing the entity embedding vectors following [3]. Then a positive triple $P[i][j] = (h, r, t)$ is sampled from shared memory, where $i$ means that the
current processor is $i$-th processor, and superscript $j$ stands for $j$-th epoch. According to $P[i][j]$, a negative triple $N[i][j] = (h', r, t')$ is generated by sampling a corrupted entity $h'$ (or $t'$) from shared memory, where $i$ and $j$ are the same as before. That is to say, a sample $S[i][j]$ is constructed by $P[i][j]$ and $N[i][j]$, which then be used to calculate the gradient $\nabla [f_r(h, t) + M - f_r(h', t')]$ according to Eq.1, and update the embeddings of entities and relations $(h, r, t, h', t')(j+1) \leftarrow (h, r, t, h', t')(j) - \eta \nabla [i][j]$.

Algorithm 1 ParTrans-X

Require:
- Training triples $T = \{(h, r, t)\}$, entities and relations set $E$ and $R$, embedding dimension $d$, margin $M$, training epochs $ep$, the number of processors $p$;

Ensure:
- Embeddings of entities and relations;
1. Initialize $r \in R$ and $e \in E$ by uniform distribution and persist them in the shared memory
2. for $i \leftarrow 0$ to $p$ do
3. $e' \leftarrow \frac{e}{\|e\|}$ for each entity $e \in E$
4. $P[i][j] \leftarrow \text{SAMPLE}(T, \text{seed}[i])$
5. $N[i][j] \leftarrow \text{SAMPLE}_\text{NEG}(P[i][j], T, \text{seed}[i])$
6. $S[i][j] \leftarrow [P[i][j], N[i][j]]$
7. $\nabla [i][j] \leftarrow \nabla [f_r(h, t) + M - f_r(h', t')]$,
where $k, r, t, h', t' \in S[i][j]$
8. $(h, r, t, h', t')(j+1) \leftarrow (h, r, t, h', t')(j) - \eta \nabla [i][j]$
9. end loop
10. end for
11. Generate embeddings of $E$ and $R$ after all processors finish

4.2 Application to Typical translating embedding Methods

The framework can be applied to many translating embedding methods, which employ SGD or its variants to optimize the hinge loss with similar algorithm framework, and are only different in the score function $f_r(h, t)$ as mentioned in Sec.2, e.g., TransE, TransH and so on. Hence, the parallel algorithm of them can be obtained by applying the corresponding score function in Lines 11-12 of the pseudocode in Algorithm 1.

For example, for TransE, the gradient updating procedure in Lines 11 is performed according to Eq.(16). For TransH, which employs the score function $f_r(h, t) = \|h + r - t\|_2$, the gradient updating procedure of $h$ in Lines 11 is as follows,

$$h_k := h_k - \eta \cdot 2 \left(\langle h_k - \text{wr}^\top \text{w}_{rk} \rangle + \alpha_k - \langle t_k - \text{wr}^\top \text{w}_{tk} \rangle\right).$$

Namely, ParTrans-X has the flexibility to parallel many translating embedding methods, since they possess similar training process. Moreover, ParTrans-X can be directly applied to the improved variant TransE-AdaGrad, since the training data sparsity of knowledge graph still holds. In one iteration of AdaGrad, it updates the embedding vectors according to the gradient from the previous iteration. Highly similar to SGD, AdaGrad can be easily parallelled using our framework by only performing a learning rate calculation procedure during the gradient update procedure, i.e., Line 12 of the pseudocode in Algorithm 1. For example, to parallel TransE-AdaGrad, the learning rate is determined adaptively by adding

$$\eta(j) := \frac{\sqrt{\sum_{k=1}^{j} (\nabla(k))^2}}{\eta^*}$$

before Line 12 in Algorithm 1, where $j$ is the current epoch, with $\eta(j)$ the learning rate of $j$-th epoch. $\nabla(k), k < j$ represents all the previous gradient before $j$-th epoch. $\eta^*$ is the initial learning rate.

5 EXPERIMENT

Firstly, we apply ParTrans-X to TransE, TransH and TransE-Adagrad in Sec.5.1. In Sec.5.2, experiment results demonstrate excessive decline in training time by ParTrans-X, with scaling performance along with increasing number of processors shown in Sec.5.3.

5.1 Experimental Settings

The datasets employed are two representative datasets WN18 and FB15k, which are subsets of well-known knowledge graphs WordNet and Freebase respectively, and have been widely used by translating embedding methods [3, 8, 11, 19]. Table 2 shows the statistics of them. Without loss of generality, $\frac{p}{n}$ and $\frac{\sigma}{n}$ are also shown, and they are both small on WN18 and FB15k. Furthermore, it can be seen that the two datasets possess different characteristics. Namely, WN18 possesses only 18 relations, which results in large possibility of collisions between relations. On the contrary, FB15k is less unbalanced in the number of entities and relations.

<table>
<thead>
<tr>
<th>Dataset</th>
<th># Rel</th>
<th># Ent</th>
<th># Train</th>
<th># Valid</th>
<th># Test</th>
<th>$\sigma/n$</th>
<th>$p/n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>WN18</td>
<td>18</td>
<td>40,943</td>
<td>141,442</td>
<td>5,000</td>
<td>5,000</td>
<td>1.7e-1</td>
<td>5.6e-4</td>
</tr>
<tr>
<td>FB15k</td>
<td>1,345</td>
<td>14,951</td>
<td>483,142</td>
<td>50,000</td>
<td>59,071</td>
<td>9.3e-3</td>
<td>2.3e-3</td>
</tr>
</tbody>
</table>

To tackle the KGC problem, experiments are conducted on the link prediction task which aims to predict the missing entities $h$ or $t$ for a triple $(h, r, t)$. Namely, it predicts $t$ given $(h, r)$ or predict $h$ given $(r, t)$. Similar to the setting in [3], the task returns a list of candidate entities from the knowledge graph.

To evaluate the performance of link prediction, we adopt Mean Rank and Hits@10 under “Raw” and “Filter” settings as evaluation measure following [3]. Mean Rank is the average rank of the correct entities, and Hits@10 is proportion of correct entities ranked in top-10. It is clear that a good predictor has low mean rank and high Hits@10. This is called “Raw” setting, and “Filter” setting filters out the corrupted triples which are correct.

To evaluate the speed up performance, we adopt Training Time and Speed-up Ratio as evaluation measures, where Training Time is measured using wall-clock seconds. Speed-up Ratio is

$$\text{Speedup Ratio} = \frac{t_{\text{serial}}}{t_{\text{parallel}}}.$$
where $t_{\text{serial}}$ is the training time in serial, and $t_{\text{parallel}}$ is the training time under parallel methods.

Baselines include typical translating embedding methods, TransE, TransH and TransE-Adagrad, which can all be trained in parallel using the ParTrans-X framework, denoted by ParTransE, ParTransH and ParTransE-Adagrad respectively in Table 3. Note that TransE and TransH adopt the programs publicly available\(^5\), which are the most efficient serial versions to our knowledge, and TransE-Adagrad is implemented based on TransE.

Each experiment is conducted 10 times and the average is taken as results, with all time measured in wall-clock seconds. Our experiments are carried out on dual Intel Xeon E5-2640 CPUs, and each of them possesses 10 physical cores 20 logical cores and runs at 2.4 GHz. The machine has 128 GB RAM and runs Red Hat. We use OpenMP for multithreading, each thread binds a processor.

### 5.2 Link Prediction Performance of ParTrans-X

Experiments on each baseline and its parallel implementation in ParTrans-X employ the same hyper-parameters, which are decided during the validation set. The learning rate $\eta$ during the stochastic gradient descent process is selected among $[0, 0.01, 0.001, 0.0001]$, the embedding dimension $d_e$ and $d_r$ are selected in $[20, 50, 100]$, the margin $M$ between positive and negative triples is selected among $[1, 2, 3, 4]$. For TransE and ParTransE, the parameters are $\eta = 0.01, d_e = d_r = 20, M = 3$ on WN18, and $\eta = 0.001, d_e = d_r = 100, M = 4$ on Fb15k. For TransH and ParTransH, the parameters are $\eta = 0.01, d_e = d_r = 20, M = 3$ on WN18, and $\eta = 0.001, d_e = d_r = 100, M = 3$ on Fb15k. For TransEAdaGrad and ParTransE-Adagrad, the parameters are $\eta^* = 0.3, d_e = d_r = 50, M = 4$ on WN18, and $\eta^* = 0.1, d_e = d_r = 100, M = 3$ on Fb15k. All the experiments employ L1-similarity. ParTransE, ParTransH and ParTransE-Adagrad all run in 20 processors for both datasets.

It can be observed from Table 3 that:

1. Link prediction performance in parallel is as good as the serial counterparts on both WN18 and FB15k, which demonstrates that ParTrans-X will not affect embedding performance.

2. The training time is greatly reduced by ParTrans-X. On WN18, TransE-AdaGrad only speeds up TransE by 4.7 times, compared to our 28 times. On FB15k, the training time of TransE is reduced from more than 1 hour to less than 1 minute by ParTransE-AdaGrad.

3. ParTrans-X achieves higher speedup ratio on Fb15k than on WN18. Since Fb15k has far more training triples than WN18, the time of each epoch on Fb15k is much longer than WN18. As a result, the overhead of multi-threading is less important compared to the whole training time on Fb15k, which leads to a higher speedup ratio. It further validates the superiority of ParTrans-X to handle the data with large size.

4. ParTrans-X achieves enormous improvement on training time when applying to TransEAdaGrad, especially on Fb15k, where the speedup ratio has been improve to 111 from 9. Since AdaGrad decreases the total epochs needed by making the convergence come earlier, and ParTrans-X reduce training time by running in parallel, the two different strategies can achieve higher speedup ratio when combined.

![Figure 3: The descent process of loss.](image)

Moreover, the descent process of loss for the three algorithms on WN18 and FB15k is shown in Figure 3. It can be seen that, for both datasets, the loss optimizing by ParTrans-X has already fallen sharply in the preceding epochs, and it yields sensibly lower values of the loss than TransEAdaGrad and TransE even after a few iterations(< 5 epoches). Still, ParTrans-X performs better on Fb15k than WN18, shows that it is more effective on large data size.

### 5.3 Scaling Results for Multi-Processors

Furthermore, we carry out a number of experiments to test if the implementations scale with increasing number of processors. We mainly analyze two aspects of experiment results, i.e., the training time and the link prediction performance.

\(^5\)https://github.com/thunlp/KB2E
It employs the intrinsic sparsity of training data in large knowledge graphs, which enables the embedding vectors to be learnt without locks and not inducing errors. Experiments validate that ParTrans-X can speed up the training process by more than an order of magnitude, without lowering embedding performance.

7 ACKNOWLEDGE

We thank Jun Xu and the anonymous reviewers for valuable suggestions. The work was funded by National Natural Science Foundation of China (No. 61572469, 61402442, 91646120, 61572473, 61402022), the National Key R&D Program of China (No. 2016QY02D0405, 2016YFB1000902), and National Grand Fundamental Research 973 Program of China (No. 2013CB329602, 2014CB340401).

REFERENCES


