LowFER: Low-rank Bilinear Pooling for Link Prediction

Saadullah Amin 1  Stalin Varanasi 1 2  Katherine Ann Dunfield 1 2  Günter Neumann 1 2

Abstract
Knowledge graphs are incomplete by nature, with only a limited number of observed facts from the world knowledge being represented as structured relations between entities. To partly address this issue, an important task in statistical relational learning is that of link prediction or knowledge graph completion. Both linear and non-linear models have been proposed to solve the problem. Bilinear models, while expressive, are prone to overfitting and lead to quadratic growth of parameters in number of relations. Simpler models have become more standard, with certain constraints on bilinear map as relation parameters. In this work, we propose a factorized bilinear pooling model, commonly used in multi-modal learning, for better fusion of entities and relations, leading to an efficient and constraint-free model. We prove that our model is fully expressive, providing bounds on the embedding dimensionality and factorization rank. Our model naturally generalizes Tucker decomposition based TuckER model, which has been shown to generalize other models, as efficient low-rank approximation without substantially compromising the performance. Due to low-rank approximation, the model complexity can be controlled by the factorization rank, avoiding the possible cubic growth of TuckER. Empirically, we evaluate on real-world datasets, reaching on par or state-of-the-art performance. At extreme low-ranks, model preserves the performance while staying parameter efficient.

1. Introduction
Knowledge graphs (KGs) are large collections of structured knowledge, organized as subject and object entities and relations, in the form of fact triples $\langle \text{sub}, \text{rel}, \text{obj} \rangle$. The usefulness of knowledge graphs, however, is affected primarily by their incompleteness. The task of link prediction or knowledge graph completion (KGC) aims to infer missing facts from existing ones, by essentially scoring a relation and entities triple for use in predicting its validity, and thereby avoiding the cost and time of extending knowledge graphs manually. To accomplish this, several models have been proposed, including linear and non-linear models. Bilinear models have additionally been used in multi-modal learning due to their expressive nature, where the fusion of features from different modalities plays a key role towards the performance of a model, with concatenation or element-wise summation being commonly used fusion techniques. The underlying assumption is that the distributions of features across modalities may vary significantly, and the representation capacity of the fused features may be insufficient, therefore limiting the final prediction performance (Yu et al., 2017). In this work, we apply this assumption to knowledge graphs by considering that the entities and relations come from different multi-modal distributions and good fusion between them can potentially construct a KG.

A major drawback of using bilinear modeling methods is the quadratic growth of parameters, which results in high computational and memory costs and risks overfitting. In multi-modal learning, factorization techniques have therefore been researched to address these challenges (Kim et al., 2016; Fukui et al., 2016; Yu et al., 2017; Ben-Younes et al., 2017; Li et al., 2017; Liu et al., 2018), and constraints based bilinear maps have become a more prevalent standard in link prediction (Yang et al., 2015; Trouillon et al., 2016; Kazemi & Poole, 2018). Applying constraints can be seen as hard regularization since it allows for incorporating background knowledge (Kazemi & Poole, 2018), but restricts the learning potential of the model due to limited parameter sharing (Balažević et al., 2019a). We focus on a constraint-free approach, using the low-rank factorization of bilinear models, as it offers flexibility and generalizes well, naturally leading to other models under certain conditions. Our work extends the multi-modal factorized bilinear pooling (MFB) model, introduced by Yu et al. (2017), and applies it to the link prediction task.

Our contributions are outlined as follows:
• We propose a simple and parameter efficient linear model by extending multi-modal factorized bilinear (MFB) pooling (Yu et al., 2017) for link prediction.

• We prove that our model is fully expressive and provide bounds on entity and relation embedding dimensions, along with the factorization rank.

• We provide relations to the family of bilinear link prediction models and Tucker decomposition (Tucker, 1966) based TuckER model (Balazevic et al., 2019a), generalizing them as special cases. We also show the relation to 1D convolution based HypER model (Balazevic et al., 2019b), bridging the gap from bilinear to convolutional link prediction models.

• On real-world datasets the model achieves on par or state-of-the-art performance, where at extreme low-ranks, with limited number of parameters, it outperforms most of the prior arts, including deep learning based models.

2. Related Work

Given a set of entities $E$ and relations $R$ in a knowledge graph $KG$, the task of link prediction is to assign a score $s$ to a triple $(e_s, r, e_o)$:

$$s = f(e_s, r, e_o)$$

where $e_s \in E$ is the subject entity, $e_o \in E$ is the object entity and $r \in R$ is the relation between them. The scoring function $f$ estimates the general binary tensor $T \in |E| \times |R| \times |E|$, by assigning a score of 1 to $T_{ijk}$ if relation $r_j$ exists between entities $e_i$ and $e_k$, 0 otherwise. The scoring function can be a linear or non-linear model, trained to predict true triples in a $KG$.

Deep learning based scoring functions such as ConvE (Dettmers et al., 2018) and HypER (Balazevic et al., 2019b) use 2D and 1D convolution on subject entity and relation representations respectively. Both perform well in practice and are efficient, but the former lacks direct interpretation, whereas the latter has shown to be related to tensor factorization. Transitional methods (Bordes et al., 2013; Wang et al., 2014; Ji et al., 2015; Lin et al., 2015; Nguyen et al., 2016; Feng et al., 2016) use additive dissimilarity scoring functions, whereby they differ in terms of the constraints applied to the projection matrices. While interpretable, they are theoretically limited as they have shown to be not fully expressive (Wang et al., 2018; Kazemi & Poole, 2018). There are several other related works (Nickel et al., 2016; Das et al., 2017; Yang et al., 2017; Shen et al., 2018; Schlichtkrull et al., 2018; Ebisu & Ichise, 2018; Sun et al., 2019), but we will mainly focus on different types of linear models here, as they are more relevant to our work.

All discussed linear models can be seen as a decomposition of the tensor $T$, using different factorization methods. One way to factorize this tensor is to factorize its slices in the relation dimension with DEDICOMP (Harshman, 1978). RESCAL (Nickel et al., 2011), a relaxed version of DEDICOMP, decomposes using a scoring function that consists of a bilinear product between subject and object entity vectors with a relation specific matrix. RESCAL, however, tends to overfit due to the quadratic growth of parameters in number of relations. Others use Canonical Polyadic decomposition (CPD or simply CP) (Hitchcock, 1927; Harshman & Lundy, 1994) to factorize the binary tensor. In CP, each value in the tensor is obtained as a sum of multiple Hadamard products of three vectors, representing subject, object and relation. DistMult (Yang et al., 2015), equivalent to INDSCAL (Carroll & Chang, 1970), is as such and uses a diagonal relation matrix, unlike RESCAL, to account for overfitting. ComplEx (Trouillon et al., 2016; Trouillon & Nickel, 2017) uses complex valued vectors for entities and relations to explicitly model asymmetric relations. SimplE (Kazemi & Poole, 2018) extends CP by introducing two vectors (head and tail) for each entity and two for relations (including the inverse). Tucker decomposition (Tucker, 1966) based TuckER (Balazevic et al., 2019a) learns a 3D core tensor, which is multiplied with a matrix along each mode to approximate the binary tensor. A key difference between CP based methods and TuckER is that it learns representations not only via embeddings, but also through a shared core tensor.

3. Model

Downstream performance for tasks such as visual question answering strongly depends on the multi-modal fusion of features to leverage the heterogeneous data (Liu et al., 2018). Bilinear models are expressive as they allow for pairwise interactions between the feature dimensions but also introduce huge number of parameters that lead to high computational and memory costs and the risk of overfitting (Fukui et al., 2016). Substantial research has therefore focused on efficiently computing the bilinear product. In multi-modal compact bilinear (MCB) pooling (Gao et al., 2016; Fukui et al., 2016), authors employ a sampling based approximation that uses the property that the tensor sketch projection (Charikar et al., 2004; Pham & Pagh, 2013) of the outer product of two vectors can be represented as its sketches convolution. Multi-modal low-rank bilinear (MLB) pooling (Kim et al., 2016) uses two low-rank projection matrices to transform the features from the original space to a common space, followed by the Hadamard product, which was later generalized by the multi-modal factorized bilinear (MFB) pooling (Yu et al., 2017). Our work is based on the MFB model but can also be seen as related to Liu et al. (2018). In contrast to KGC bilinear models, these bilinear models allow for parameter sharing and generally, are constraint-free.
3.1. Multi-modal Factorized Bilinear Pooling (MFB)

Given two feature vectors \( x \in \mathbb{R}^m, y \in \mathbb{R}^n \) and a bilinear map \( W \in \mathbb{R}^{m \times n} \), the bilinear transformation is defined as
\[
z = x^T W y = 1^T (U^T x \odot V^T y)
\]
where \( U \in \mathbb{R}^{m \times k} \), \( V \in \mathbb{R}^{n \times k} \), \( k \) is the factorization rank, \( \odot \) is the element-wise product of two vectors and \( 1 \in \mathbb{R}^k \) is vector of all ones. Therefore, to obtain an output feature vector \( z \in \mathbb{R}^e \), two 3D tensors are required, \( W_x = [U_1, U_2, \ldots, U_o] \) \( x \) reshape, \( W_y \) and \( W_y = [V_1, V_2, \ldots, V_o] \) \( y \) reshape, where \( W_x \in \mathbb{R}^{m \times k \times o} \), \( W_y \in \mathbb{R}^{n \times k \times o} \) are 3D tensors and \( W_x' \in \mathbb{R}^{m \times k} \), \( W_y' \in \mathbb{R}^{n \times k} \) are their reshaped 2D matrices respectively. The final (fused) vector \( z \) is then obtained by summing non-overlapping windows of size \( k \) over the Hadamard product of projected vectors using \( W_x' \) and \( W_y' \):
\[
z = \text{SumPool}(W_x' \odot W_y' y, k)
\]
At \( k = 1 \), MFB reduces to MLB, which converges slowly, and MCB requires very high-dimensional vectors to perform well (Yu et al., 2017). Further, MFB significantly lowers the number of parameters with low-rank factorized matrices and leads to better performance.

3.2. Low-rank Bilinear Pooling for Link Prediction

Consider that entities and relations are not intrinsically bound and come from two different modalities, such that good fusion between them can potentially result in a knowledge graph of fact triples. Entities and relations can be shown to possess certain properties that allow them to function similarly to others within the same modality. For example, the relation place-of-birth shares inherent properties with the relation residence. As such, similar entity pairs can yield similar relations, given appropriate shared properties. Like in multi-modal auditory-visual fusion, where the sound of a roar may better predict a resulting image within the distribution of animals that roar, a relation such as place-of-birth, can better predict an entity pair within a distribution of person, place entity pairs. In link prediction, we assume that the latent decomposition with MFB can help the model capture different aspects of interactions between an entity and a relation, which can lead to better scoring with the missing entity. We therefore, apply the Low-rank Factorization trick of bilinear maps with \( k \)-sized non-overlapping summation pooling (section 3.1) to Entities and Relations (LowFER).

More formally, for an entity \( e \in \mathcal{E} \), we represent its embedding vector \( e \in d_e \) dimension as a look-up from entity embedding matrix \( E \in \mathbb{R}^{n_e \times d_e} \) and relation vector \( r \in \mathbb{R}^{d_r} \) from relation embedding matrix \( R \in \mathbb{R}^{n_r \times d_r} \), where \( n_e \) and \( n_r \) are number of entities and relations in \( KG \). LowFER projects \( e \) and \( r \) into a common space \( \mathbb{R}^{kd_e} \) followed by Hadamard product and \( k \)-summation pooling, where \( k \) is the factorization rank. The output vector \( z \) is then matched against target entity \( e_o \) to give final score.

Using this compact notation in Eq. 2, the final scoring function of LowFER is obtained as:
\[
f(e_s, r, e_o) = (S^k \text{diag}(U^T e_s) V^T r)^T e_o
\]
\[\footnotesize1\]Note that at \( k = 1 \), \( S^1 = I_{d_e} \)
Table 1. Bounds for fully expressive linear models, where \( n \) is the number of true facts and the trivial bound is given by \( n_e^2 n_r \).

<table>
<thead>
<tr>
<th>Model</th>
<th>Full Expressibility Bounds</th>
</tr>
</thead>
<tbody>
<tr>
<td>RESCAL (Nickel et al., 2011)</td>
<td>((d_e, d_r) = (n_e, n_r^2))</td>
</tr>
<tr>
<td>HoIE (Nickel et al., 2016)</td>
<td>(d_e = d_r = 2n_e n_r + 1)</td>
</tr>
<tr>
<td>ComplEx (Trouillon et al., 2016)</td>
<td>(d_e = d_r = n_e n_r)</td>
</tr>
<tr>
<td>SimplE (Kazemi &amp; Poole, 2018)</td>
<td>(d_e = d_r = \min(n_e n_r, n + 1))</td>
</tr>
<tr>
<td>TuckER (Balazevic et al., 2019a)</td>
<td>((d_e, d_r) = (n_e, n_r))</td>
</tr>
<tr>
<td>LowFER</td>
<td>((d_e, d_r) = (n_e, n_r^k)) for (k = \min(n_e, n_r))</td>
</tr>
</tbody>
</table>

3.3. Training LowFER

To train the LowFER model, we follow the setup of Balazevic et al. (2019a). First, we apply sigmoid non-linearity after Eq. 5 to get the probability \( p(y_{(e,r,e_o)}) = \sigma(f(e, r, e_o)) \) of a triple belonging to \( K \mathcal{G} \). Then, for every triple \((e, r, e_o)\) in the dataset, a reciprocal relation is added by generating a synthetic example \((e_o, r^{-1}, e)\) (Dettmers et al., 2018; Lacroix et al., 2018) to create the training set \( \mathcal{D} \). For faster training, Dettmers et al. (2018) introduced 1-N scoring, where each tuple \((e, r)\) and \((e_o, r^{-1})\) is simultaneously scored against all entities \(e \in \mathcal{E}\) to predict 1 if \(e = e_o\) or \(e_o\) respectively and 0 elsewhere (see Trouillon et al. (2017) and Sun et al. (2019) for other methods to collect negative samples). The model is trained with binary cross-entropy instead of margin based ranking loss (Bordes et al., 2013), which is prone to overfitting for link prediction (Trouillon & Nickel, 2017; Kazemi & Poole, 2018). For a mini-batch \( \mathcal{B}\) of size \(m\) drawn from \( \mathcal{D}\), we minimize:

\[
\min_{\Theta} \frac{1}{m} \sum_{(e,r) \in \mathcal{E}} \frac{1}{n_e} \sum_{i=1}^{n_e} (y_i \log(p(y_{(e,r,e_i)})))
+ (1 - y_i) \log(1 - p(y_{(e,r,e_i)})))
\]

where \(y_i\) is a target label for a given entity-relation pair \((e, r)\) for entity \(e_i\), \(y_{(e,r,e_i)}\) is the model prediction and \(\Theta\) represents model parameters. Following Yu et al. (2017), we also apply the power normalization \(x \leftarrow \text{sign}(x)/|x|^{0.5}\) and \(l_2\)-normalization \(x \leftarrow x/||x||\) before summation pooling to stabilize the training from large output values as a result of Hadamard product in Eq. 3.

4. Theoretical Analysis

4.1. Full Expressibility

A key theoretical property of link prediction models is their ability to be fully expressive, which we define formally as:

**Definition 1.** Given a set of entities \(\mathcal{E}\), relations \(\mathcal{R}\), correct triples \(\mathcal{T} \subseteq \mathcal{E} \times \mathcal{R} \times \mathcal{E}\) and incorrect triples \(\mathcal{T}' = \mathcal{E} \times \mathcal{R} \times \mathcal{E} \setminus \mathcal{T}\), then a model \(M\) with scoring function \(f(e, r, e_o)\) is said to be fully expressive if it can accurately separate \(\mathcal{T}\) from \(\mathcal{T}'\) for all \(e, e_o \in \mathcal{E}\) and \(r \in \mathcal{R}\).
4.2. Relation with TuckER

Initially, it was shown by Kazemi & Poole (2018) that RESCAL, DistMult, ComplEx and SimplE belong to a family of bilinear models with different set of constraints. Later, Balažević et al. (2019a) established that TuckER generalizes all of these models as special cases. In this section, we will formulate relation between our model and TuckER (Balažević et al., 2019a), followed by relations with the family of bilinear models in the next section. This provides a unifying view and shows LowFER’s ability to generalize.

TuckER’s scoring function is defined as follows (Balažević et al., 2019a):

$$\phi_t(e_s, r, e_o) = W \times_1 e_s \times_2 r \times_3 e_o$$

(6)

where $W \in \mathbb{R}^{d_e \times d_r \times d_e}$ is the core tensor, $e_s, e_o \in \mathbb{R}^{d_e}$ and $r \in \mathbb{R}^{d_r}$ are subject entity, object entity and the relation vectors respectively. $\times_n$ denotes the tensor product along the $n$-th mode. First, note that Eq. 4 can be expanded as:

$$S^k(U^T e_s \circ V^T r) = \begin{bmatrix} e_s^T (\sum_{i=1}^k u_i \otimes v_i) r \\ \vdots \\ e_s^T (\sum_{i=(j-1)k+1}^{jk} u_i \otimes v_i) r \\ \vdots \\ e_s^T (\sum_{i=(k-1)d_e}^{kd_e} u_i \otimes v_i) r \end{bmatrix}$$

where $u_i \in \mathbb{R}^{d_u}$ and $v_i \in \mathbb{R}^{d_v}$ are column vectors of $U$ and $V$ respectively and $\otimes$ represents the outer product of two vectors. To take the vectors $e_s$ and $r$ out, we realize the above matrix operations in a different way. We first create $k$ distance apart in $U$ (and $V$), i.e., for the l-th slice, we have $W_U^{(l)} = [u_{1l}, u_{k+l}, \ldots, u_{(k-1)d_u+l}] \in \mathbb{R}^{d_e \times d_u}$ and $W_V^{(l)} = [v_{1l}, v_{k+l}, \ldots, v_{(k-1)d_v+l}] \in \mathbb{R}^{d_v \times d_v}$. Taking the column-wise outer product of these sliced matrices forms a 3D tensor in $\mathbb{R}^{d_e \times d_r \times d_e}$. With slight abuse of notation, we also use $\otimes$ to represent this tensor operation. It can be viewed as transforming the matrix obtained by mode-2 Khatri-Rao product into a 3D tensor (Cichocki et al., 2016). Now consider a 3D tensor $W \in \mathbb{R}^{d_e \times d_r \times d_e}$ as the sum of these $k$ products:

$$W = \sum_{i=1}^k W_U^{(i)} \otimes W_V^{(i)}$$

(7)

Figure 3 shows these operations. With this tensor, the scoring function $f$ in Eq. 5 can be re-written as TuckER’s scoring function as follows:

$$\hat{\phi}_t(e_s, r, e_o) = W \times_1 e_s \times_2 r \times_3 e_o$$

(8)

It should be noted that $W$ in Eq. 8 is obtained as a summation of $k$ low-rank 3D tensors, each of which is obtained by stacking rank-1 matrices in contrast to TuckER’s core tensor $W$ in Eq. 6, which can be a full rank 3D tensor. Our model can therefore approximate TuckER and can be viewed as a generalization of TuckER (Balažević et al., 2019a). We further show that we can accurately obtain $W$ with appropriate $W_U^{(i)}$’s and $W_V^{(i)}$’s in Eq. 7 (proof in Appendix A.2).

Proposition 2. Given a TuckER model with entity embedding dimension $d_e$, relation embedding dimension $d_r$, and core tensor $W$, there exists a LowFER model with $k \leq \min(d_e, d_r)$, entity embedding dimension $d_e$ and relation embedding dimension $d_r$ that accurately represents the former.

LowFER and TuckER parameters grow linearly in the number of entities and relations as $O(n_e d_e + n_r d_r)$. However, LowFER’s shared parameters complexity can be controlled by decoupled low-rank matrices through the factorization rank, making it more flexible, e.g., consider $d = d_e = d_r$, the core tensor $W$ of TuckER grows as $O(d^3)$, whereas LowFER grows only as $O(kd^2)$. As an example, in Lacroix
et al. (2018) authors used \(d_e = d_r = 2000\) which would require more than 8 billion parameters to model with TuckER compared to only 4k million for LowFER, with \(k\) controlling the growth. More generally, at \(k = d_e/2\), LowFER has equal number of parameters as TuckER therefore, we expect similar performance at such rank values. In practice, \(k = \{1, 10, 30\}\) performs extremely well (section 5.1).

4.3. Relations with the Family of Bilinear Models

In this section, we will establish relations between LowFER and other bilinear models. For simplicity, we consider the relation embedding to be a constant matrix \(R = I_n\) in all the cases and use \(V\) to model relation parameters. However, the conditions presented here can be extended otherwise, with the remark that they are not unique.

**RESCAL** (Nickel et al., 2011): scoring function is defined as:

\[
\phi_r(e_s, r_1, e_o) = e_s^T W_r e_o
\]

where \(W_r \in \mathbb{R}^{d_r \times d_r}\) is \(l\)-th relation matrix. For LowFER to encode RESCAL with Eq. 5, we set \(k = d_r, d_e = n_r\) and \(U = \{I_{d_e} | I_{d_e} | \ldots | I_{d_e}\} \in \mathbb{R}^{d_r \times d_e^2}\) (block matrix partitioned as \(d_e\) identity matrices of size \(d_e \times d_e\)). This is effectively taking a row \(l\) from \(V \in \mathbb{R}^{n_r \times d_e^2}\), reshaping it to \(d_e \times d_e\) matrix and then taking the transpose to get the equivalent \(W_r\) in RESCAL’s scoring function.

**DISTMULT** (Yang et al., 2015): scoring function is defined as:

\[
\phi_d(e_s, r_1, e_o) = e_s^T \text{diag}(w_l)e_o
\]

where \(w_l \in \mathbb{R}^{d_e}\) is the vector for \(l\)-th relation. For LowFER to encode DistMult with Eq. 5, we set \(k = 1, d_r = n_r\) and \(U = I_{d_e}\). This is effectively taking a row \(l\) from \(V \in \mathbb{R}^{n_r \times d_e}\) and creating a diagonal matrix of it to get the equivalent \(\text{diag}(w_l)\) in DistMult’s scoring function.

**SIMPLE** (Kazemi & Poole, 2018): scoring function is defined as:

\[
\phi_s(e_s, r_1, e_o) = \frac{1}{2} (h_s^T \text{diag}(r_l) t_{e_o} + h_o^T \text{diag}(r_l^{-1}) t_{e_s})
\]

where \(h_s, h_o \in \mathbb{R}^d\) are subject, object entities head vectors, \(t_s, t_o \in \mathbb{R}^d\) are subject, object entities tail vectors and \(r_1, r_l^{-1} \in \mathbb{R}^d\) are relation and inverse relation vectors. Let \(e_s = \{h_s, t_s\} \in \mathbb{R}^{2d}, e_o = \{h_o, t_o\} \in \mathbb{R}^{2d}\) and \(r_l = \{r_1; r_l\} \in \mathbb{R}^{2d}\) then Simple scoring is equivalent to \(\frac{1}{2} e_s^T \text{diag}(r_l)e_o\), where \(e_s\) and \(r_l\) are obtained by swapping the head, tail vectors in \(e_s = \{h_s, t_s\}\) and relation, inverse relation vectors in \(r_l = \{r_1; r_l\}\) respectively. For LowFER to encode Simple, \(U\) becomes a permutation matrix (ignoring the \(\frac{1}{2}\) scaling factor), swapping the first \(d\)-half with the second \(d\)-half of a given vector in \(\mathbb{R}^{2d}\) and \(l\)-th row in \(V\) is \(r_l\), more specifically, with Eq. 5, we set \(k = 1, d_e = 2d, d_r = n_r\) and \(U \in \mathbb{R}^{2d \times 2d}\) is a block matrix with four partitions such that, \(U_{12} = U_{21} = \frac{1}{2}I_d\) and 0s elsewhere.

**COMPLEX** (Trouillon et al., 2016) scoring function is defined as:

\[
\phi_c(e_s, r_1, e_o) = \text{Re}(e_s)^T \text{diag(Re}(r_l))\text{Re}(e_o) + \text{Im}(e_s)^T \text{diag(Im}(r_l))\text{Im}(e_o) + \text{Re}(e_s)^T \text{diag(Im}(r_l))\text{Im}(e_o) - \text{Im}(e_s)^T \text{diag(Re}(r_l))\text{Re}(e_o)
\]

where \(\text{Re}(\cdot)\) and \(\text{Im}(\cdot)\) represents the real and imaginary parts of a complex vector. Consider \(e_s = [\text{Re}(e_s); \text{Im}(e_s)] \in \mathbb{R}^{2d}\) and \(e_o = [\text{Re}(e_o); \text{Im}(e_o)] \in \mathbb{R}^{2d}\) then the ComplEx scoring function can be obtained as \(e_s^T W_l e_o\), where \(W_l \in \mathbb{R}^{2d \times 2d}\) represents the \(l\)-th relation matrix such that its diagonal is \([\text{Re}(r_1); \text{Re}(r_l)]\), the \(d\) offset diagonal is \(\text{Im}(r_1)\) and \(-d\) offset diagonal is \(-\text{Im}(r_1)\). For LowFER to encode ComplEx, similar to Simple, we will use two permutation matrices to obtain the above four terms. That is, in Eq. 8, we have \(k = 2, d_e = 2d, d_r = n_r\) and \(U \in \mathbb{R}^{2d \times 4d}\) is such that \(W_l^{(1)}\) is a block matrix with \(W_{U_{11}}^{(1)} = W_{U_{12}}^{(1)} = I_d\) and 0 elsewhere. Further, \(W_{U_{22}}^{(2)}\) is also a block matrix with \(W_{U_{22}}^{(2)} = -I_d, W_{U_{22}}^{(2)} = I_d\) and 0 elsewhere. Lastly, \(V \in \mathbb{R}^{n_r \times 4d}\) is such that \(W_l^{(1)}\) row \(l\) has \([\text{Re}(r_1); \text{Im}(r_1)]\) and \(W_l^{(2)}\) row \(l\) has \([\text{Im}(r_1); \text{Re}(r_1)]\), i.e.,
\[ W^{(2)} = W^{(1)} P, \] where \( P \in \mathbb{R}^{2d \times d} \) is the \( d \)-half swapping permutation matrix. Figure 4 demonstrates how \( W \) parameters for the family of bilinear models under the conditions discussed in this section.

4.4. Relation to HypER

HypER (Balažević et al., 2019b) is a convolutional model based on hypernetworks (Ha et al., 2017), where the relation specific 1D filters are generated by the hypernetwork and convolved with the subject entity vector. Balažević et al. (2019b) showed that it can be understood in terms of tensor factorization up to a non-linearity. With a similar argument, we show that LowFER encodes HypER, bringing it closer to the convolutional approaches as well.

HypER scoring function is defined as (Balažević et al., 2019b):

\[
\phi_h(e_s, r, e_o) = h(\text{vec}(e_s \ast F_r) W) e_o \tag{9}
\]

where \( F_r = \text{vec}^{-1}(H_r) \in \mathbb{R}^{n_f \times 1}, H \in \mathbb{R}^{n_f \times d_r} \) (hypernetwork), \( W \in \mathbb{R}^{n \times d \times d} \), \( \ast \) transforms \( n \times m \) matrix to \( nm \)-sized vector, \( \text{vec}^{-1}(\cdot) \) does the reverse operation, \( h(\cdot) \) is the convolution operator, \( h(\cdot) \) is ReLU non-linearity and \( n_f, l_f \) and \( l_m = d_e - l_f + 1 \) are number of filters, filter length and output length of convolution. The convolution between a filter and the subject entity embedding can be seen as a matrix multiplication, where the filter is converted to a Toeplitz matrix of size \( l_m \times d_e \). With \( n_f \) filters, we can realize a 3D tensor of size \( n_f \times l_m \times d_e \). Since the filters are generated by the hypernetwork, we have \( d_e \) such 3D tensors, resulting in a 4D tensor of size \( n_f \times l_m \times d_e \times d_r \) (Balažević et al., 2019b).

5. Experiments and Results

We conducted the experiments on four benchmark datasets: WN18 (Bordes et al., 2013), WN18RR (Dettmers et al., 2018), FB15k (Bordes et al., 2013) and FB15k-237 (Toutanova et al., 2015) (see Appendix B for the details, including best hyperparameters and additional experiments).

5.1. Link Prediction

Table 2 shows our main results, where LowFER-1, LowFER-10 and LowFER-\( k^* \) represent our model for \( k = 1, k = 10 \) and \( k = \text{best} \). We choose LowFER-1 and LowFER-10 as baselines. Overall, LowFER reaches competitive performance on all the datasets with state-of-the-art results on FB15k and FB15k-237. On WN18 and WN18RR, TuckER is marginally better than LowFER.

LowFER performs well at low-ranks with significantly less number of parameters compared to other linear models (Table 3). At \( k = 1 \), it performs better than or on par with both non-linear and linear models (including ComplEx and Simple) except HypER and TuckER. For FB15k-237, LowFER-1 (\( \sim 3 \)M parameters) outperforms R-GCN, RotatE, DistMult and ComplEx by an average of 5.9% on MRR, and it additionally outperforms convolutional models (ConvE, HypER) at \( k = 10 \) with only +0.8M parameters. On FB15k, the best reported TuckER model is improved upon, with absolute +1.9% increase on toughest Hits@1 metric. This already achieves state-of-the-art with almost half the parameters, \( \sim 5.5 \)M in contrast to TuckER’s \( \sim 11.3 \)M. On WN18RR and WN18, LowFER-1 outperforms all the models excluding TuckER and HypER. With LowFER-\( k^* \), we marginally reach state-of-the-art performance on WN18RR and FB15k-237. On FB15k, we reach new state-of-the-art for \( \sim 9.51 \)M parameters with +2.9% and +4.1% improvement on MRR and Hits@1.

The empirical gains can be attributed to LowFER’s ability to perform good fusion between entities and relations while avoiding overfitting through low-rank matrices remaining parameter efficient, with strong performance even at ex-
Table 3. Comparison between the number of parameters in millions (M) of strong linear models. For LowFER-\(k^*\), the \(k\) values are 10, 100, 30 and 50 for WN18, FB15k-237, WN18RR and FB15k respectively.

<table>
<thead>
<tr>
<th>Model</th>
<th>WN18</th>
<th>FB15k-237</th>
<th>WN18RR</th>
<th>FB15k</th>
</tr>
</thead>
<tbody>
<tr>
<td>ComplEx</td>
<td>16.4</td>
<td>6.0</td>
<td>16.4</td>
<td>6.5</td>
</tr>
<tr>
<td>SimplE</td>
<td>16.4</td>
<td>-</td>
<td>16.4</td>
<td>6.5</td>
</tr>
<tr>
<td>TuckER</td>
<td>9.4</td>
<td>11.0</td>
<td>9.4</td>
<td>11.3</td>
</tr>
<tr>
<td>LowFER-1</td>
<td>8.2</td>
<td>3.0</td>
<td>8.2</td>
<td>4.6</td>
</tr>
<tr>
<td>LowFER-10</td>
<td>8.6</td>
<td>3.8</td>
<td>8.6</td>
<td>5.5</td>
</tr>
<tr>
<td>LowFER-(k^*)</td>
<td>8.6</td>
<td>11.3</td>
<td>9.6</td>
<td>9.5</td>
</tr>
</tbody>
</table>

Table 4. Link prediction results on FB15k with \(d_e = d_r = 200\).

<table>
<thead>
<tr>
<th>(k)</th>
<th>Params (M)</th>
<th>MRR</th>
<th>Hits@1</th>
<th>Hits@3</th>
<th>Hits@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3.60</td>
<td>0.634</td>
<td>0.538</td>
<td>0.695</td>
<td>0.803</td>
</tr>
<tr>
<td>5</td>
<td>3.92</td>
<td>0.720</td>
<td>0.641</td>
<td>0.776</td>
<td>0.860</td>
</tr>
<tr>
<td>10</td>
<td>4.33</td>
<td>0.742</td>
<td>0.667</td>
<td>0.790</td>
<td>0.871</td>
</tr>
<tr>
<td>30</td>
<td>5.93</td>
<td>0.774</td>
<td>0.709</td>
<td>0.817</td>
<td>0.885</td>
</tr>
<tr>
<td>50</td>
<td>7.53</td>
<td>0.776</td>
<td>0.713</td>
<td>0.818</td>
<td>0.886</td>
</tr>
<tr>
<td>100</td>
<td>11.53</td>
<td>0.779</td>
<td>0.717</td>
<td>0.821</td>
<td>0.887</td>
</tr>
</tbody>
</table>

Figure 5. Influence of increasing the factorization rank on MRR and Hits@1 scores for FB15k.

Figure 6. Influence of changing the entity embedding dimension \(d_e\) on Hits@1 metric and growth of parameters in million (M).

Empirically, we found when \(d_e = d_r\), taking \(k = d_e/2\) performs nearly the same as TuckER (Balažević et al., 2019a). This can be observed in LowFER-\(k^*\) for FB15k-237 (\(d_e = d_r = 200, k = 100\)), where our results are almost indistinguishable from TuckER’s. This can be expected as the number of parameters in both models are almost the same (~11M). It should be noted that in practice when we train LowFER, we initialize with two i.i.d matrices, which are not shared, compared to TuckER’s core tensor (Eq. 6), allowing us to reach almost the same performance despite less parameter sharing.

5.3. Effect of Embedding Dimension

The size of entity embedding dimension \(d_e\) accounts for the significant number of parameters in LowFER, growing linearly with number of entities \(n_e\). To study the effect, we trained our models on FB15k, with \(d_r = 30, k = 50\) constant, and varying \(d_e\) in \{30, 50, 100, 150, 200, 250, 300, 350, 400\}. As can be seen in Figure 6, increasing the entity embedding dimension significantly increases the Hits@1 metric, for almost linear growth in number of parameters. However, it only improves till 300 and starts overfitting afterwards.

In Balažević et al. (2019a), authors reported \(d_e = d_r = 200\) as best choice of dimensions for TuckER on FB15k, however, we found using \(d_e = 300\) and \(d_r = 30\) better with lesser number of parameters for LowFER. For fair compari-
We expected LowFER to match TuckER’s performance as noted above. We saw an improvement over all the metrics. At $k=100$, where we expected LowFER to match TuckER’s performance (MRR=0.795, Hits@1=0.741, ~11 million parameters), it was lower (−1.6% on MRR and −2.4% on Hits@1). In comparison, our model with $d_e=300$, $d_r=30$ and $k=10$ with ~5.6 million parameters only, gives better results than this setting and TuckER. Therefore, at $d_e=d_r=200$, our model is most likely overfitting.

As noted above that it could be that LowFER is overfitting therefore, we did coarse grid search over relation embedding dimension in {30, 50, 100, 150, 200} and $k$ in {1, 5, 10, 30, 50, 100, 150, 200} while keeping $d_e=200$ fixed. We found $d_e=50$ at $k=150$ reaches almost the same performance as TuckER with ~10.6M parameters compared to TuckER’s ~11.3M parameters. We also experimented with $l_2$-regularization (Reg) and noted minor improvements, with regularization strength 0.0005. Table 5 summarizes these results. According to the experiments reported in main results (Table 2) were without any regularization. In general, we only noticed slight improvements in FB15k with $l_2$-regularization.

### 5.4. Analysis of Relation Results

Link prediction models that can discover relation types automatically without prior knowledge indicate better generalization. As shown, and discussed in section 4, LowFER, among other models (Table 1), can learn to capture all relation types without additional constraints. However, in practice, these bounds are loose and require very large dimensions, raising an inspection into their performance on different relation types. In Kazemi & Poole (2018), it was identified that WN18 contains redundant relations, i.e., $\forall e_i, e_j \in E : (e_i, r_1, e_j) \in T \iff (e_j, r_2, e_i) \in T$, such as <hypernym, hyponym>, <meronym, holonym> etc. To alleviate this, Dettmers et al. (2018) proposed WN18RR with such relations removed, since knowledge about one can help infer the knowledge about the other. Table 6 shows the per relation results of LowFER and TuckER on WN18 and WN18RR. We see that performance drops for 7 relations, with an average performance decrease of −70.6% and −69.3% for LowFER and TuckER respectively (with highest decrease on member_of_domain_usage for both). For symmetric relations (such as derivationally_related_form), the performance is approximately the same where we observe severe limitation to model asymmetry. We believe this is because LowFER (also TuckER) is constraint-free and adding certain constraints based on background knowledge is necessary to improve the model’s accuracy. Simple is the only fully expressive model that has formally shown to address these limitations (cf. Proposition 3, 4 and 5 in Kazemi & Poole (2018)). Since LowFER subsumes Simple therefore, such rules can be studied for extending LowFER to incorporate the background knowledge.

### 6. Conclusion

This work proposes a simple and parameter efficient fully expressive linear model that is theoretically well sound and performs on par or state-of-the-art in practice. We showed that LowFER generalizes to other linear models in KGC, providing a unified theoretical view. It offers a strong baseline to the deep learning based models and raises further interest into the study of linear models. We also highlighted some limitations with respect to gains on harder relations, which still pose a challenge. We conclude that the constraint-free and parameter efficient linear models, which allow for parameter sharing, are better from a modeling perspective, but are still similarly limited in learning difficult relations. Therefore, studying the trade-off between parameters sharing and constraints becomes an important future work.

### Acknowledgements

The authors would like to thank the anonymous reviewers for helpful feedback and gratefully acknowledge the use of code released by Balažević et al. (2019a). The work was partially funded by the European Union’s Horizon 2020 research and innovation programme under grant agreement No. 777107 through the project Precise4Q and by the German Federal Ministry of Education and Research (BMBF) through the project DEELEE (01IW17001).
References


A. Proofs

A.1. Proposition 1

Proof. First, we will prove the case for \( k = d_e \), with the proof for the case \( k = d_r \) following a similar argument. For both cases, we represent entity embedding vector as \( e_i \in \{0, 1\}^{\vert E\vert} \), such that only \( i \)-th element is 1, and similarly, relation embedding vector as \( r_j \in \{0, 1\}^{\vert R\vert} \), such that only \( j \)-th element is 1. We represent with \( U \in \mathbb{R}^{d_e \times d_e} \) and \( V \in \mathbb{R}^{d_r \times d_r} \) the model parameters, then, given any triple \((e_i, r_j, e_l)\) \( \in T \) with indices \((i, j, l)\), such that \( 1 \leq i, l \leq \vert E\vert \) and \( 1 \leq j \leq \vert R\vert \):

For \( k = d_e \): We let \( U_{m n} = 1 \) for \( n = m + (o - 1)d_e \), for all \( m \) in \{1, ..., \( d_e \)\} and for all \( o \) in \{1, ..., \( k \)\} and 0 otherwise. Further, let \( V_{p q} = 1 \) for \( p = j \) and \( q = (l - 1)d_e + i \) and 0 otherwise. Applying \( g(e_i, r_j) \) and taking dot product of the resultant vector with \( e_l \) (Eq. 5) perfectly represents the ground truth as 1. Also, for any triple in \( T' \), a score of 0 is assigned.

For \( k = d_r \): We let \( U_{m n} = 1 \) for \( n = i \) and \( n = (l - 1)d_e + j \) and 0 otherwise. Further, let \( V_{p q} = 1 \) for \( q = p + (o - 1)d_e \), for all \( p \) in \{1, ..., \( d_e \)\} and for all \( o \) in \{1, ..., \( k \)\} and 0 otherwise. Rest of the argument follows the same as for \( k = d_e \). ☐

A.2. Proposition 2

Proof. From Eq. 7 and 8, observe that the \( m \)-th slice of the core tensor \( W \) on object dimension is approximated by adding \( k \) rank-1 matrices, each of which is a cross product between \( m \)-th column in \( W^{(l)}_{U} \) and \( m \)-th column in \( W^{(l)}_{V} \), for all \( l \) in \{1, ..., \( d_e \)\}. Each slice of the core tensor \( W \) on object dimension has a maximum rank \( \min(d_e, d_r) \) and from Singular Value Decomposition (SVD), there exists \( n \) \( (\leq \min(d_e, d_r)) \) scaled left singular and scaled right singular vectors whose sum of the cross products is equal to the slice. By choosing these scaled left singular vectors, scaled right singular vectors and zero vectors (in case the rank of the corresponding slice is less than the maximum rank of any such slice) as columns for matrices \( W^{(l)}_{U}, W^{(l)}_{V} \) for all \( l \) in \{1, ..., \( d_e \)\}, the core tensor \( W \) is obtained from Eq. 7 with \( k \leq \min(d_e, d_r) \). ☐

Please note that the bounds presented in Table 1 are weak and in general, not very useful. They are derived only for checking the full expressibility of a model, which is also referred to as model being universal in Wang et al. (2018), to handle all-types of relations with zero error, i.e., perfect reconstruction of the binary tensor \( T \) for a given \( KG \). Since factorization based methods can be seen as approximating the true binary tensor, more useful bounds can be derived by studying the quality of the approximations for a given accuracy level. The bounds for RESCAL, ComplEx and HoIE are reported from Wang et al. (2018) while for SimpLE (Kazemi & Poole, 2018) and TuckER (Balažević et al., 2019a), from their respective papers.

As discussed in section 4.1, it was first shown in Wang et al. (2018) that TransE is not universal, which was later generalized to other translational models by Kazemi & Poole (2018). RotatE (Sun et al., 2019), a state-of-the-art dissimilarity based model, alleviates the issues of TransE by learning counterclockwise rotations in the complex space. For a triple \((h, r, t)\), RotatE models the tail entity as \( t = h \circ r \), where \( h, r, t \in \mathbb{C}^{d} \) are head and tail embeddings and \( r \in \mathbb{C}^{d} \) is the relation embedding with a restriction on the element-wise modulus, \( |r| = 1 \). Therefore, it only affects the phases of the entity embeddings in the complex vector space. Sun et al. (2019) showed that it can learn symmetric, asymmetric, inverse and composition relations (cf. Lemma 1, 2, 3) and degenerates to TransE (cf. Theorem 4). However, we note that RotatE is also not fully expressive due to its inability to model the transitive relations in the general case, i.e., irrespective of the size of embedding dimension.

Proposition 3. RotatE is not fully expressive due to a limitation on the transitive relations.

Proof. Consider \( \{e_1, e_2, e_3\} = \Delta \subset \mathcal{E} \) and \( r \in \mathcal{R} \) be a transitive relation on \( \Delta \) such that \( r(e_1, e_2), r(e_2, e_3) \) and \( r(e_1, e_3) \) belong to the ground truth. Let \( e_1, e_2, e_3 \in \mathbb{C}^{d} \) be the embedding vectors for RotatE. Let us assume that \( r(e_1, e_2) \) and \( r(e_2, e_3) \) hold with RotatE, then we get \( e_2 = r \circ e_1 \) and \( e_3 = r \circ e_2 \). From definition of transitive relation we know that \( r(e_1, e_2) \land r(e_2, e_3) \implies r(e_1, e_3) \), here we obtain \( e_3 = r \circ r \circ e_1 \). Therefore for \( r(e_1, e_3) \) to hold with RotatE, we must have \( r \circ r = r \implies r = 1 \), which in turn suggest \( e_1 = e_2 = e_3 \) but \( e_1, e_2, e_3 \) are distinct entities. More concretely, this condition requires that for all elements of relation embedding \( r, \cos(\theta_{r,i}) + i\sin(\theta_{r,i}) \) should match \( \cos(2\theta_{r,i}) + i\sin(2\theta_{r,i}) \), which is only possible when \( \theta_{r,i} \in \{0, 2\pi\} \), effectively no rotation. ☐
on the core tensor of the Tucker decomposition (Tucker, 1966), it can subsume the family of bilinear models. In this work, we showed that LowFER subsumes TuckER and can be seen as providing low-rank approximation of the core tensor with accurate representation under certain conditions (Proposition 2). We also showed that LowFER can subsume the family of bilinear models and HypER up to a non-linearity. Figure A.1 provides a network style map for the models discussed here.

B. Experiments

In this section, we will present the details of the datasets, evaluation metrics, model implementation, the choice of hyperparameters and report additional experiments.

B.1. Data

We conducted the experiments on four benchmark datasets: WN18 (Bordes et al., 2013) - a subset of Wordnet, WN18RR (Dettmers et al., 2018) - a subset of WN18 created through the removal of inverse relations from validation and test sets, FB15k (Bordes et al., 2013) - a subset of Freebase, and FB15k-237 (Toutanova et al., 2015) - a subset of FB15k created through the removal of inverse relations from validation and test sets. Table A.1 shows the statistics of all the datasets.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>n_e</th>
<th>n_r</th>
<th>n_e/n_r</th>
<th>Training</th>
<th>Validation</th>
<th>Testing</th>
</tr>
</thead>
<tbody>
<tr>
<td>WN18</td>
<td>40,943</td>
<td>18</td>
<td>2275</td>
<td>144</td>
<td>5,000</td>
<td>5,000</td>
</tr>
<tr>
<td>WN18RR</td>
<td>40,943</td>
<td>11</td>
<td>3722</td>
<td>86</td>
<td>835</td>
<td>3,034</td>
</tr>
<tr>
<td>FB15k</td>
<td>14,951</td>
<td>345</td>
<td>11</td>
<td>483</td>
<td>142</td>
<td>50,000</td>
</tr>
<tr>
<td>FB15k-237</td>
<td>14,541</td>
<td>237</td>
<td>61</td>
<td>272</td>
<td>115</td>
<td>17,535</td>
</tr>
</tbody>
</table>

Table A.1. Datasets used for link prediction experiments, where n_e = number of entities, n_r = number of relations and the entities-to-relations ratio n_e/n_r is approximated to the nearest integer.

B.2. Evaluation Metrics

We report the standard metrics of Mean Reciprocal Rank (MRR) and Hits@k for k ∈ {1, 3, 10}. For each test triple (e_o, r, e_a), we score all the triples (e_a, r, e) for all e ∈ E. We then compute the inverse rank of true triple and average them over all examples. However, Bordes et al. (2013) identified an issue with this evaluation and introduced filtered MRR.

---

1For x, y ∈ ℝ^d, it states that x^T y = 1/2 |ℱ(x)|^2 |ℱ(y)|, where ℱ : ℝ^d → ℂ^d is the discrete Fourier transform (DFT).

2The rank of a tensor is the minimal number of rank-1 tensors that yield it in a linear combination. It is known that the tensor rank is NP-hard to compute, and for a 3rd-order tensor n × m × k, it can be more than min(n, m, k) but no more than min(nm, mk, nk) (Miettinen, 2011). Whereas, the r-rank of a tensor W is the dimension of the vector space spanned by the r-mode vectors, which are the columns of the matrix unfolding W_r (De Lathauwer et al., 2000).

3https://bit.ly/3k641Ba
Table A.2. Best performing hyper-parameter values for LowFER, where lr=learning rate, dr=decay rate, de=entity embedding dimension, dl=relation embedding dimension, k=LowFER factorization rank, dE=entity embedding dropout, dMB=MBF dropout, dOut=output dropout and ls=label smoothing. Please note that dE, dMBF and dOut are the same as #1, #2 and #3 as in TuckER (see Appendix A in Balajić et al. (2019a)) respectively.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>lr</th>
<th>dr</th>
<th>de</th>
<th>dl</th>
<th>k</th>
<th>dE</th>
<th>dMBF</th>
<th>dOut</th>
<th>ls</th>
</tr>
</thead>
<tbody>
<tr>
<td>WN18</td>
<td>0.005</td>
<td>0.995</td>
<td>200</td>
<td>30</td>
<td>10</td>
<td>0.2</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>WN18RR</td>
<td>0.01</td>
<td>1.0</td>
<td>200</td>
<td>30</td>
<td>30</td>
<td>0.2</td>
<td>0.2</td>
<td>0.3</td>
<td>0.1</td>
</tr>
<tr>
<td>FB15k</td>
<td>0.003</td>
<td>0.99</td>
<td>300</td>
<td>30</td>
<td>50</td>
<td>0.2</td>
<td>0.2</td>
<td>0.3</td>
<td>0.0</td>
</tr>
<tr>
<td>FB15k-237</td>
<td>0.0005</td>
<td>1.0</td>
<td>200</td>
<td>200</td>
<td>100</td>
<td>0.4</td>
<td>0.4</td>
<td>0.5</td>
<td>0.1</td>
</tr>
</tbody>
</table>

where we only consider triples of the form \( \{ (e_s, r, e) \mid \forall e \in \mathcal{E} \text{ s.t.} \ (e_s, r, e) \notin \text{train} \cup \text{valid} \cup \text{test} \} \) during evaluation. We therefore reported filtered MRR for all the experiments. The Hits@k metric computes the percentage of test triples whose ranking is less than or equal to k.

B.3. Implementation and Hyperparameters

We implemented LowFER\(^4\) using the open-source code released by TuckER (Balajić et al., 2019a)\(^5\). We did random search over the embedding dimensions in \{30, 50, 100, 200, 300\} for \( d_e \) and \( d_r \). Further, we varied the factorization rank \( k \) in \{1, 5, 10, 30, 50, 100, 150, 200\}, with \( k = 1 \) (LowFER-1) and \( k = 10 \) (LowFER-10) as baselines. For WN18RR and WN18, we found best \( d_e = 200 \) and \( d_r = 30 \) with \( k = 30 \) and \( 10 \) respectively. For FB15k-237, we found best \( d_e = 200 \) at \( k = 100 \). All of these embedding dimensions match the best reported in TuckER (Balajić et al., 2019a). However, for FB15k, we found using the configuration of \( d_e = 300 \) and \( d_r = 30 \) to be consistently better than \( d_e = d_r = 200 \). For fair comparison, we also reported the results for \( d_e = d_r = 200 \) and the best configuration when \( d_e = 200 \) and \( d_r, k \leq 200 \) (Table 5).

Similar to Balajić et al. (2019a), we used Batch Normalization (Ioffe & Szegedy, 2015) but additionally power normalization and \( l_2\)-normalization to stabilize training from large outputs following the Hadamard product in main scoring function (Yu et al., 2017)\(^6\). We tested the best reported hyperparameters of Balajić et al. (2019a) with random search and observed good performance in initial testing. With \( d_e, d_r \) and \( k \) selected, we used fixed set of values for rest of the hyperparameters reported in Balajić et al. (2019a), including learning rate, decay rate, entity embedding dropout, MFB dropout, output dropout and label smoothing (Szegedy et al., 2016; Pereyra et al., 2017) (see Table A.2 for the best hyperparameters). We used Adam (Kingma & Ba, 2015) for optimization. In all the experiments, we trained the models for 500 epochs with batch size 128 and reported the final results on test set.

B.4. Results on YAGO3-10

We report additional results on YAGO3-10, which is a subset of YAGO3 (Mahdisoltani et al., 2013), consisting of 123,182 entities and 37 relations such that have each entity has at least 10 relations. We used the same best hyperparameters as for WN18RR. Table A.3 shows that our model outperforms state-of-the-art models including RotatE and HypER. It is worth noting that LowFER-\( k^* \) on YAGO3-10 has only \~62M parameters compared to \~61M parameters of RotatE (Sun et al., 2019), which also includes their self-adversarial negative sampling.

B.5. LowFER with Non-linearity

Similar to Kim et al. (2016), we perform a simple ablation study by adding non-linearity to the LowFER scoring function as follows:

\[
\tilde{f}(e_s, r, e_o) = (\sigma(S^k \text{diag}(U^T e_s)V^T r))^T e_o
\]

where we use hyperbolic tangent \( \sigma = \tanh \) non-linearity. Applying non-linear activation function can be seen as increasing the representation capacity of the model but Table

Table A.3. Link prediction results on YAGO3-10. Results for DistMult, ComplEx and ConvE are taken from Detmers et al. (2018) and for RotatE (Sun et al., 2019) (with self-adversarial negative sampling) and HypER (Balajić et al., 2019b) are taken from respective papers.

<table>
<thead>
<tr>
<th>Model</th>
<th>MRR</th>
<th>Hits@1</th>
<th>Hits@3</th>
<th>Hits@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>DistMult</td>
<td>0.340</td>
<td>0.240</td>
<td>0.380</td>
<td>0.540</td>
</tr>
<tr>
<td>ComplEx</td>
<td>0.360</td>
<td>0.260</td>
<td>0.400</td>
<td>0.550</td>
</tr>
<tr>
<td>ConvE</td>
<td>0.440</td>
<td>0.350</td>
<td>0.400</td>
<td>0.620</td>
</tr>
<tr>
<td>RotatE</td>
<td>0.495</td>
<td>0.402</td>
<td>0.550</td>
<td>0.670</td>
</tr>
<tr>
<td>HypER</td>
<td>0.533</td>
<td>0.455</td>
<td>0.580</td>
<td>0.678</td>
</tr>
<tr>
<td>LowFER-( k^* )</td>
<td>0.537</td>
<td>0.457</td>
<td>0.583</td>
<td>0.688</td>
</tr>
</tbody>
</table>

Table A.4. Link prediction results with LowFER-\( k^* \) and additional \( \tanh \) non-linearity. The \( \downarrow \) shows that the performance went down compared to the linear counterparts reported in Table 2.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>MRR</th>
<th>Hits@1</th>
<th>Hits@3</th>
<th>Hits@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>FB15k-237</td>
<td>0.345</td>
<td>0.256</td>
<td>0.378</td>
<td>0.526</td>
</tr>
<tr>
<td>FB15k</td>
<td>0.818</td>
<td>0.771</td>
<td>0.850</td>
<td>0.898</td>
</tr>
<tr>
<td>WN18RR</td>
<td>0.457</td>
<td>0.429</td>
<td>0.469</td>
<td>0.511</td>
</tr>
<tr>
<td>WN18</td>
<td>0.950</td>
<td>0.946</td>
<td>0.952</td>
<td>0.957</td>
</tr>
</tbody>
</table>

\(^4\)https://github.com/suamin/LowFER  
\(^5\)https://github.com/ibalazevic/TuckER  
\(^6\)We observed no performance degradation by removing these additional normalization techniques but we used it in all the experiments to be consistent with prior work of Yu et al. (2017).
A.4 shows that the general performance of LowFER goes down.

B.6. Models Comparison

We compared LowFER with non-linear models including ConvE (Dettmers et al., 2018), R-GCN (Schlichtkrull et al., 2018), Neural LP (Yang et al., 2017), RotatE (Sun et al., 2019), TransE (Bordes et al., 2013), TorusE (Ebisu & Ichise, 2018) and HypER (Balažević et al., 2019b). In linear models, we compared against DistMult (Yang et al., 2015), HolE (Nickel et al., 2016), ComplEx (Trouillon et al., 2016), ANALOGY (Liu et al., 2017), SimplE (Kazemi & Poole, 2018) and state-of-the-art TuckER (Balažević et al., 2019a) model. Results for the Canonical Tensor Decomposition (Lacroix et al., 2018) were not included due to the uncommon choice of extremely large embedding dimensions of $d_c = d_r = 2000$.

Additional models that were not reported in the main results (Table 2) due to partial results but were still outperformed by LowFER include M-Walk (Shen et al., 2018) with their reported metrics of MRR=0.437, Hits@1=0.414 and Hits@3=0.445 on WN18RR and MINERVA (Das et al., 2017) with Hits@10=0.456 on FB15k-237. The results in Table 2 for all the models were taken from Balažević et al. (2019b) and Balažević et al. (2019a). Lastly, in the section 5.4, to perform per relations comparisons, we trained the TuckER models with the best reported configurations in Balažević et al. (2019a) for WN18 and WN18RR.

---

7Where we reported their results in Table 2 without the self-adversarial negative sampling. For fair comparison, see Appendix H in their paper.