InteractE: Improving Convolution-based Knowledge Graph Embeddings by Increasing Feature Interactions

Shikhar Vashishth[1*] **Soumya Sanyal**[1*] **Vikram Nitin**[2†]
Nilesh Agrawal[1] **Partha Talukdar**[1]
[1]Indian Institute of Science, [2]Columbia University
{shikhar, soumyasanyal, anilesh, ppt}@iisc.ac.in
vikram.nitin@columbia.edu

Proof of Propositions

Proposition 6.1. *For any kernel $\boldsymbol{w}$ of size k, for all $n \geq \left(\frac{5k}{3} - 1\right)$ if k is odd and $n \geq \frac{(5k+2)(k-1)}{3k}$ if k is even, the following statement holds:*

$$\mathcal{N}_{het}(\phi_{alt}, k) \geq \mathcal{N}_{het}(\phi_{stk}, k)$$

Proof. Any $M_k \in \mathbb{R}^{k \times k}$, $M_k \subseteq \phi_{alt}$ contains $\left\lfloor \frac{k}{2} \right\rfloor$ rows of elements of $\boldsymbol{e}_s$, and $\left\lfloor \frac{k+1}{2} \right\rfloor$ rows of elements of $\boldsymbol{e}_r$, or vice-versa.

For a single fixed M_k, the total number of triples (a_i, b_j, M_k) and (b_j, a_i, M_k) is

$$2 \times k \left\lfloor \frac{k}{2} \right\rfloor \times k \left\lfloor \frac{k+1}{2} \right\rfloor$$

The number of possible M_k matrices is $(n - k + 1)^2$. Hence the total number of heterogeneous interactions is

$$\mathcal{N}_{het}(\phi_{alt}, k) = (n - k + 1)^2 k^2 \times 2 \left\lfloor \frac{k}{2} \right\rfloor \left\lfloor \frac{k+1}{2} \right\rfloor$$

Any $M_k \in \mathbb{R}^{k \times k}$, $M_k \subseteq \phi_{stk}$ contains l rows of elements of $\boldsymbol{e}_s$, and $k - l$ rows of elements of $\boldsymbol{e}_r$, where $0 \leq l \leq k$.

For a fixed l, the number of different possible M_k matrices is $(n - k + 1)$. Hence, the number of heterogeneous interactions is

$$\begin{aligned}
\mathcal{N}_{het}(\phi_{stk}, k) &= (n - k + 1) \left(\sum_{l=0}^{k} 2 \times kl \times k(k - l) \right) \\
&= (n - k + 1) \cdot k^2 \cdot \left(k^2(k+1) - \frac{k(k+1)(2k+1)}{3} \right) \\
&= (n - k + 1) \cdot k^2 \cdot \left(\frac{k(k+1)(k-1)}{3} \right) \qquad (1)
\end{aligned}$$

We need to check whether,

$$\mathcal{N}_{het}(\phi_{alt}, k) \geq \mathcal{N}_{het}(\phi_{stk}, k)$$

$$\implies (n - k + 1) \left\lfloor \frac{k}{2} \right\rfloor \left\lfloor \frac{k+1}{2} \right\rfloor \geq \frac{k(k+1)(k-1)}{6}$$

For odd k, this becomes

$$\begin{aligned}
(n - k + 1) \left(\frac{k-1}{2} \right) \left(\frac{k+1}{2} \right) &\geq \frac{k(k+1)(k-1)}{6} \\
n - k + 1 &\geq \frac{2k}{3} \\
n &\geq \frac{5k}{3} - 1
\end{aligned}$$

For even k,

$$\begin{aligned}
(n - k + 1) \left(\frac{k}{2} \right)^2 &\geq \frac{k(k+1)(k-1)}{6} \\
nk - k(k-1) &\geq \frac{2(k+1)(k-1)}{3} \\
n &\geq \frac{(5k+2)(k-1)}{3k}
\end{aligned}$$

□

Proposition 6.2. *For any kernel $\boldsymbol{w}$ of size k and for all $\tau < \tau'$ ($\tau, \tau' \in \mathbb{N}$), the following statement holds:*

$$\mathcal{N}_{het}(\phi^{\tau}_{alt}, k) \geq \mathcal{N}_{het}(\phi^{\tau'}_{alt}, k)$$

Proof. For simplicity, let us assume that $\alpha = n/(2\tau) \in \mathbb{N}$, i.e., ϕ^{τ}_{alt} is composed of exactly α blocks of τ rows of $\boldsymbol{e}_s$ and $\boldsymbol{e}_r$ stacked alternately. Also, when $\tau < k$, we assume that $k/\tau \in \mathbb{N}$. Now, for any $M_k \in \mathbb{R}^{k \times k}$, $M_k \subseteq \phi^{\tau}_{alt}$, we consider the following two cases:

Case 1. $\boldsymbol{\tau \geq k - 1}$: It is easy to see that this case can be split into n/τ subproblems, each of which is similar to ϕ_{stk}. Hence,

$$\mathcal{N}_{het}(\phi^{\tau}_{alt}, k) = \left(\frac{n}{\tau} \right) \mathcal{N}_{het}(\phi_{stk}, k)$$

Clearly, $\mathcal{N}_{het}(\phi^{\tau}_{alt}, k)$ is monotonically decreasing with increasing τ.

[*] contributed equally to this paper.
[†] Research done at Indian Institute of Science.

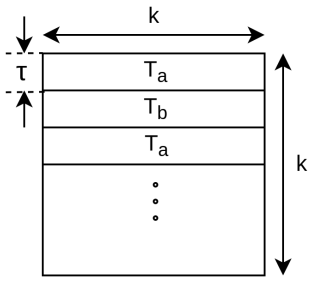

Figure 1: The figure depicts a $k \times k$ matrix M_k. T_a, T_b are reshaped matrices each containing τk components of $\boldsymbol{e}_s, \boldsymbol{e}_r$ respectively.

Case 2. $\boldsymbol{\tau < k - 1}$: As shown in Fig. 1, let T_a, $T_b \in \mathbb{R}^{\tau \times k}$ denote a submatrix formed by components of $\boldsymbol{e}_s$, $\boldsymbol{e}_r$ respectively. Note that if k is even, then for any $M_k \subseteq \phi^{\tau}_{alt}$, the number of components of $\boldsymbol{e}_s$ and $\boldsymbol{e}_r$ are always equal to $k^2/2$ each. For odd k, the number of T_a's and T_b's are $\frac{(k/\tau)+1}{2}$ and $\frac{(k/\tau)-1}{2}$ in some order. Now, if we move M_k down by i rows ($i \leq \tau$), the total number of heterogeneous interactions across all such positions is:

$$\begin{aligned}
&\frac{n}{\tau}\sum_{i=0}^{\tau-1} k^2 \left(\left(\frac{k/\tau+1}{2}\right)\tau - i\right)\left(\left(\frac{k/\tau-1}{2}\right)\tau + i\right) \\
&= \frac{nk^2}{\tau}\sum_{i=0}^{\tau-1}\frac{(k+\tau-2i)(k-\tau+2i)}{4} \\
&= \frac{nk^2}{4\tau}\sum_{i=0}^{\tau-1} k^2 - (\tau^2 + 4i^2 - 4i\tau) \\
&= \frac{nk^2}{4}\left((k^2-\tau^2) - \frac{4(\tau-1)(2\tau-1)}{6} + \frac{4\tau(\tau-1)}{2}\right) \\
&= C\left(k^2 - \frac{\tau^2}{3} - \frac{2}{3}\right)
\end{aligned}$$

We can see that this is also monotonically decreasing with increasing τ. It is also evident that the above expression is maximum at $\tau = 1$ (since $\tau \in \mathbb{N}$).

□

Proposition 6.3. *For any kernel $\boldsymbol{w}$ of size k and for all reshaping functions $\phi : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^{n \times n}$, the following statement holds:*

$$\mathcal{N}_{het}(\phi_{chk}, k) \geq \mathcal{N}_{het}(\phi, k)$$

Proof. For any ϕ, and for any $M_k \in \mathbb{R}^{k \times k}$ such that $M_k \subseteq \phi$, let x, y be the number of components of $\boldsymbol{e}_s$ and $\boldsymbol{e}_r$ in M_k respectively. Then $\mathcal{N}_{het}(M_k, k) = 2xy$. Also, since total number of elements in M_k is fixed, we have $x + y = k^2$.

Using the AM-GM inequality on x, y we have,

$$xy \leq \left(\frac{x+y}{2}\right)^2 = \frac{k^4}{4}$$

If k is odd, since $x, y \in \mathbb{N}$,

$$xy \leq \frac{k^4 - 1}{4}$$

Therefore, the maximum interaction occurs when $x = y = \frac{k^2}{2}$ (for even k), or $x = \frac{k^2+1}{2}, y = \frac{k^2-1}{2}$ (for odd k). It can be easily verified that this property holds for all $M_k \subseteq \phi_{chk}$. Hence,

$$\mathcal{N}_{het}(\phi, k) = \sum_{M_k} 2xy \leq \sum_{M_k} \frac{2k^4}{4} = \mathcal{N}_{het}(\phi_{chk}, k)$$

□

Proof. If M_k contains x components of $\boldsymbol{e}_s$ and y components of $\boldsymbol{e}_r$, then $\mathcal{N}_{het}(M_k, k) = 2xy$, and $\mathcal{N}_{het}(M'_k, k) = 2(x-l)(y-(p-l))$ for some $l \leq p$ and $l \leq x$. We observe that

$$\begin{aligned}
\mathcal{N}_{het}(M'_k, k) &= \mathcal{N}_{het}(M_k, k) - 2(x-l)(p-l) - 2ly \\
&\leq \mathcal{N}_{het}(M_k, k)
\end{aligned}$$

□

Proposition 6.4. *Let Ω_0, $\Omega_c : \mathbb{R}^{n \times n} \to \mathbb{R}^{(n+p)\times(n+p)}$ denote zero padding and circular padding functions respectively, for some $p > 0$. Then for any reshaping function ϕ,*

$$\mathcal{N}_{het}(\Omega_c(\phi), k) \geq \mathcal{N}_{het}(\Omega_0(\phi), k)$$

Proof. Given $\Omega_c(\phi))$, we know that we can obtain $\Omega_0(\phi)$ by replacing certain components of $\Omega_c(\phi))$ with 0. So for every $M_k \subseteq \Omega_c(\phi)$, there is a corresponding $M'_k \subseteq \Omega_0(\phi)$ which is obtained by replacing some p components ($p \geq 0$) of M_k with 0.

Using the above Lemma, we can see that

$$\begin{aligned}
\mathcal{N}_{het}(\Omega_c(\phi), k)) &= \sum_{M_k \subseteq \Omega_c(\phi)} \mathcal{N}_{het}(M_k, k)) \\
&\geq \sum_{M'_k \subseteq \Omega_0(\phi)} \mathcal{N}_{het}(M'_k, k)) \\
&= \mathcal{N}_{het}(\Omega_0(\phi), k))
\end{aligned}$$

□

Hyperparameters

We use the standard training, validation and test splits provided with the datasets. A detailed description of the datasets is included in the main paper. We select the best model using the validation data on the hyperparameters listed in Table 1. Most of the hyperparamters are adopted from ConvE (Dettmers et al. 2018) model. In this paper, we explore both 1-1 (Bordes et al. 2013) and 1-N (Dettmers et al. 2018) scoring techniques. In 1-N scoring, each (s, r) pair is scored against all the entities $o \in \mathcal{E}$ simultaneously. For training, we use Adam optimizer (Kingma and Ba 2014) and Xavier initialization (Glorot and Bengio 2010) for initializing parameters.

Hyperparameter	Values
Learning rate	{0.001, 0.0001}
Label smoothening	{0.0, 0.1}
Batch size	{16, 64, 256}
Negative Samples	{100, 500, 1000, 4000}
l_2 regularization	$\{0, 10^{-5}\}$
Hidden dropout	{0, 0.3, 0.5}
Feature dropout	{0, 0.2, 0.5}
Input dropout	{0, 0.2}
k_w	{10}
k_h	{20}
Number of convolutional filters	{32, 48, 64, 96}
Convolutional kernal size k	{3, 5, 7, 9, 11}
Number of feature permutations t	{1, 2, 3, 4, 5}

Table 1: Details of hyperparameters used. Please refer to Section for more details.

References

[Bordes et al. 2013] Bordes, A.; Usunier, N.; Garcia-Duran, A.; Weston, J.; and Yakhnenko, O. 2013. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems*. 2787–2795.

[Dettmers et al. 2018] Dettmers, T.; Pasquale, M.; Pontus, S.; and Riedel, S. 2018. Convolutional 2d knowledge graph embeddings. In *Proceedings of the 32th AAAI Conference on Artificial Intelligence*, 1811–1818.

[Glorot and Bengio 2010] Glorot, X., and Bengio, Y. 2010. Understanding the difficulty of training deep feedforward neural networks. In Teh, Y. W., and Titterington, M., eds., *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, 249–256. Chia Laguna Resort, Sardinia, Italy: PMLR.

[Kingma and Ba 2014] Kingma, D., and Ba, J. 2014. Adam: A method for stochastic optimization.