Comprehensive Analysis of Negative Sampling in Knowledge Graph Representation Learning

Hidetaka Kamigaito[1], Katsuhiko Hayashi[2]

1: Nara Institute of Science and Technology (NAIST)

2: Hokkaido University

Negative Sampling (NS) Loss in Knowledge Graph Embedding (KGE)

Two types of loss functions

In KGE, we commonly use the following loss functions:

- The original NS loss by Mikolov+ 2013

$$-\frac{1}{|D|}\sum_{(x,y)\in D}\left[\log(\sigma(s_\theta(x,y))) + \sum_{y_i\sim p_n(y|x)}^{\nu}\log(\sigma(-s_\theta(x,y_i)))\right]$$

- The one used for KGE [Sun+ 2019, Ahrabian+ 2020]

$$-\frac{1}{|D|}\sum_{(x,y)\in D}\left[\log(\sigma(s_\theta(x,y)+\gamma)) + \frac{1}{\nu}\sum_{y_i\sim p_n(y|x)}^{\nu}\log(\sigma(-s_\theta(x,y_i)-\gamma))\right]$$

Observed data following $p_d(x,y)$: $D=\{(x_1,y_1),\cdots,(x_n,y_n)\}$ *Noise distribution*: $p_n(y|x)$
Score function: $s_\theta(x,y)$ *Number of negative samples*: ν *Margin term*: γ

Negative Sampling (NS) Loss in Knowledge Graph Embedding (KGE)

Two types of loss functions

In KGE, we commonly use the following loss functions:

- The original NS loss by Mikolov+ 2013

$$-\frac{1}{|D|}\sum_{(x,y)\in D}\Bigg[\log(\sigma(s_\theta(x,y))) + \sum_{y_i\sim p_n(y|x)}^{\nu}\log(\sigma(-s_\theta(x,y_i)))\Bigg]$$

- The one used for KGE [Sun+ 2019, Ahrabian+ 2020]

$$-\frac{1}{|D|}\sum_{(x,y)\in D}\Bigg[\log(\sigma(s_\theta(x,y)+\gamma)) + \frac{1}{\nu}\sum_{y_i\sim p_n(y|x)}^{\nu}\log(\sigma(-s_\theta(x,y_i)-\gamma))\Bigg]$$

Observed data following $p_d(x,y)$: $D=\{(x_1,y_1),\cdots,(x_n,y_n)\}$ *Noise distribution*: $p_n(y|x)$
Score function: $s_\theta(x,y)$ *Number of negative samples*: ν *Margin term*: γ

The two loss functions have the different terms

Differences of the Two Loss Functions

Different terms

The two NS loss functions have the following differences:

- The original NS loss does not have the margin term γ different from the NS loss in KGE.
- The NS loss in KGE has the normalization term $1/\nu$ for the number of negative samples.

We investigated the differences to understand the characteristics of the two loss functions

Our Theoretical Findings

Theoretical analysis

Our theoretical findings:

1. **Equivalence between the two loss functions**
2. **Effects of the Margin Term γ**
3. **Effects of the Number of Negative Samples ν**
4. **Relationship between the Margin Term γ and the Number of Negative Samples ν.**
5. **Relationship between the NS loss in KGE and Self-adversarial Negative Sampling (SANS) loss.**
6. **Subsampling for KGE**

Our Theoretical Findings

Theoretical analysis

Our theoretical findings:

1. **Equivalence between the two loss functions**
2.
3.
4.

> We show that the existence of v and γ has no effect on the distribution that the model will fit when the NS loss reaches the optimal solution.
>
> See Prop. 3.1 in our paper for the details.

5. **Relationship between the NS loss in KGE and Self-adversarial Negative Sampling (SANS) loss.**
6. **Subsampling for KGE**

Our Theoretical Findings

Theoretical analysis

Our theoretical findings:

1. **Equivalence between the two loss functions**
2. **Effects of the Margin Term *γ***
3. **Effects of the Number of Negative Samples *v***
4. **Relationship between the Margin Term *γ* and the Number of Negative Samples *v*.**

5.

6.

We show that to make a **distance-based scoring method** capable to reach the optimal solution, we should tune *γ* in the NS loss used for KGE and *v* in the original NS loss.

Distance-based scoring: $-||f_\theta(x, y)||_p$

- Used in TransE and RotatE

However, scoring methods with **unlimited value ranges**, such as RESCAL, ComplEx, and DistMult are **not related** to the discussion.

See Props. 3.2, 3.3, 3.4, and 3.5 in our paper for the details.

Our Theoretical Findings

Theoretical analysis

Our theoretical findings:

1. **Equivalence between the two loss functions**
2. **Effects of the Margin Term γ**
3. **Effects of the Number of Negative Samples v**
4. **Relationship between the Margin Term γ and the Number of Negative Samples v.**
5. **Relationship between the NS loss for KGE and Self-adversarial Negative Sampling (SANS) loss.**
6. **Subsamplin**

We show that we can consider the SANS loss as the NS loss for KGE when *v* is enough large and $p_n(y|x) = p_\theta(y|x)$.

$$p_\theta(y|x) = \frac{\exp\left(s_\theta(x,y)\right)}{\sum_{y' \in Y} \exp\left(s_\theta(x,y')\right)}$$

See Prop. 3.6 in our paper for the details.

Our Theoretical Findings

Theoretical analysis

Our theoretical findings:

1. **Equivalence between**
2. **Effects of the Margin**
3. **Effects of the Numbe**
4. **Relationship betweer**
5. **Relationship betweer loss.**
6. **Subsampling for KGE**

To fill in the gap between the distribution of the observed data and a true distribution behind the data, we reformulate the NS loss by introducing functions *A(x,y)* and *B(x)* as follows:

The NS loss with subsumpling

$$-\frac{1}{|D|}\sum_{(x,y)\in D}\Big[A(x,y)\log(\sigma(s_\theta(x,y)+\gamma))+\frac{1}{\nu}\sum_{y_i\sim p_n(y_i|x)}^{\nu}B(x)\log(\sigma(-s_\theta(x,y_i)-\gamma))\Big]$$

Frequency-based subsampling (Freq)

$$A(x,y) = \frac{\frac{1}{\sqrt{\#(x,y)}}}{\sum_{(x',y')\in D}\frac{1}{\sqrt{\#(x',y')}}},\ B(x) = \frac{\frac{1}{\sqrt{\#x}}}{\sum_{x'\in D}\frac{1}{\sqrt{\#x'}}}$$

$$\#(x,y)\approx\#(e_i,r_k)+\#(r_k,e_j)$$

Unique-based subsampling (Uniq)

$$A(x,y)=B(x)=\frac{\frac{1}{\sqrt{\#x}}}{\sum_{x'\in D}\frac{1}{\sqrt{\#x'}}}$$

See subsection 3.5 in our paper for the details.

Empirical Analysis

Experiments

- We examined whether our theoretical is valid for actual datasets and models shown below.

Datasets

Dataset	Entities	Relations	Tuples: Train	Tuples: Valid	Tuples: Test
FB15k-237	14,541	237	272,115	17,535	20,466
WN18RR	40,943	11	86,835	3,034	3,134
YAGO3-10	123,182	37	1,079,040	4,978	4,982

Models

Model	Score Function	Parameters
RESCAL	$\mathbf{h}^\intercal \mathbf{M}_r \mathbf{t}$	$\mathbf{h}, \mathbf{t} \in \mathbb{R}^d$, $\mathbf{M} \in \mathbb{R}^{d \times d}$
DistMult	$\mathbf{h}^\intercal \mathrm{diag}(\mathbf{r}) \mathbf{t}$	$\mathbf{h}, \mathbf{r}, \mathbf{t} \in \mathbb{R}^d$
ComplEx	$\mathrm{Re}(\mathbf{h}^\intercal \mathrm{diag}(\mathbf{r}) \bar{\mathbf{t}})$	$\mathbf{h}, \mathbf{r}, \mathbf{t} \in \mathbb{C}^d$
TransE	$-\|\|\mathbf{h} + \mathbf{r} - \mathbf{t}\|\|_p$	$\mathbf{h}, \mathbf{r}, \mathbf{t} \in \mathbb{R}^d$
RotatE	$-\|\|\mathbf{h} \circ \mathbf{r} - \mathbf{t}\|\|_p$	$\mathbf{h}, \mathbf{r}, \mathbf{t} \in \mathbb{C}^d, \|r_i\| = 1,$
HAKE	$-\|\|\mathbf{h} \circ \mathbf{r} - \mathbf{t}\|\|_p$ $-\lambda\|\| \sin((\mathbf{h}' + \mathbf{r}' - \mathbf{t}')/2)\|\|_1$	$\mathbf{h}, \mathbf{t} \in \mathbb{R}^d, \mathbf{r} \in \mathbb{R}^d_+$ $\mathbf{h}', \mathbf{r}', \mathbf{t}' \in [0, 2\pi)^d,$ $\lambda \in \mathbb{R}$

- We confirmed that the observed Mean Reciprocal Ranks (MRRs) are along with our theoretical analysis.
  - See section 4 in our paper for the details.

Conclusion

Our analysis

- We conducted a theoretical analysis for the NS loss used in KGE learning and derived theoretical facts.
- The experimental results indicate that the theoretical facts we derived are also observed in the real-world datasets.