# Exploring the Unknown :
# Negative Sampling for Knowledge Graph Completion

2025-04-10

presenter : Sooho Moon

DMAIS

# INDEX

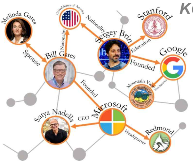
# INTRODUCTION

## ▪ What is a knowledge graph(KG)?

□ A hetrogeneous graph that consists of **entities**(nodes) and **relations**(edges)

□ Stores information in **triples** → (subject entity, relation, object entity)

□ Applied in various domains
*(recommendation, drug prediction, GraphRAG, information retrieval, question-answering, etc.)*




KG

triples

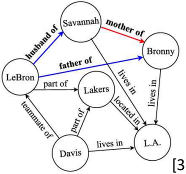
$$(Bill\ Gates, Founded, Microsoft)$$
$$(Sergey\ Brin, Founded, Google)$$
$$(Microsoft, CEO, Satya\ Nadella)$$
$$\vdots$$

# INTRODUCTION

## ▪ Knowledge Graph Completion(KGC)

- Despite its power, KG suffers from sparsity [1, 2] issue
- For example, in Freebase and DBpedia more than **66%** of the person entries are missing a birthplace [1]
- Thus building KGC models to automatically fill in(connect) missing links has been extensively studied
- Normally two task exists, link prediction(entity prediction) and relation prediction

  *link prediction : (h, r, ?) or (?, r, t) → predict "?"*
  *relation prediction : (h, ?, t) → predict "?"*
  *we only deal with link prediction for this work*

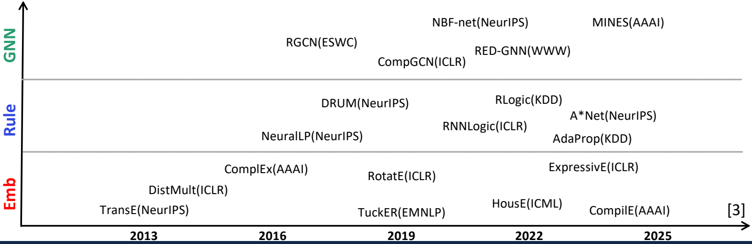



[3]

[1] Dettmers, Tim, et al. "Convolutional 2d knowledge graph embeddings." Proceedings of the AAAI conference on artificial intelligence. Vol. 32. No. 1. 2018.
[2] Sun, Zhiqing, et al. "Rotate: Knowledge graph embedding by relational rotation in complex space." ICLR arXiv:1902.10197 (2019).

# INTRODUCTION

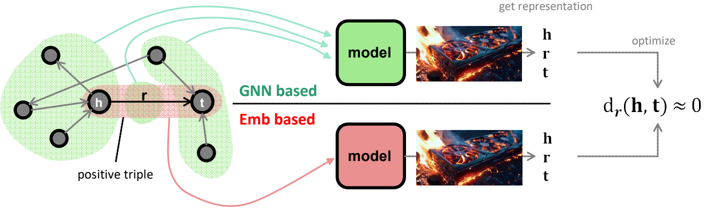
- **“Big 3” paradigm of KGC models(2013 ~ 2025)**
  - □ **Embedding base**, **Rule base**, **GNN base**
  - □ Each paradigm has its own learning method to predict links




[3] Liang, Ke, et al. "A survey of knowledge graph reasoning on graph types: Static, dynamic, and multi-modal." IEEE Transactions on Pattern Analysis and Machine Intelligence (2024).

# INTRODUCTION

- **Generlized training phase of Emb base and GNN base**
  - □ Similarity : create **embeddings** for each entity and relation **to minimize distance function**
  - □ Difference : **how** the embeddings are produced




**learning only positive links... is that it?**

**We no longer talk about 'Rule base' methods in this presentation since it has a totally different training phase**

# INTRODUCTION

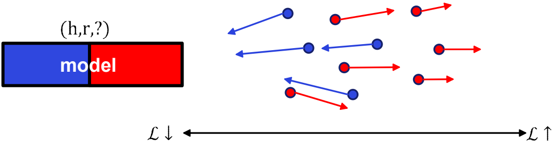
- **Contrastive learning approach in both Emb base and GNN base**
  - □ Training a model to be able to contrast between **true** and **false**
  - □ In other words, minimizing the positive samples(PS) loss and maximizing the negative samples(NS) loss
  - □ Thus the model learns like a magnet, pulling and pushing at the same time

$$PS = \{(h, r, t) \mid (h, r, t) \in G_{trn}\}$$
$$NS = \{(h', r', t') \mid (h', r', t') \notin G_{trn}\}$$

$$\mathrm{argmin}_{\theta}(\mathbb{E}[\mathcal{L}(s_{\theta}(PS), s_{\theta}(NS))])$$

$(h, r, \bullet) \in G_{trn}$

$(h, r, \bullet) \notin G_{trn}$



# INTRODUCTION

- **Using negative samples : crucial part for model performance**
  - □ KGC, CV, RS, NLP rely heavily on NS for training a robust model
  - □ Below figures show the power of NS

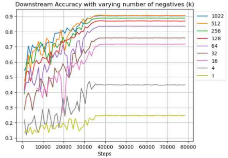
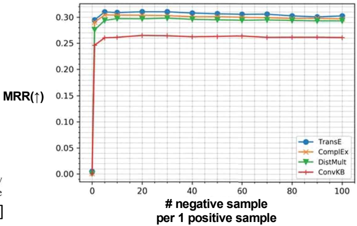



*Figure 2.* Downstream classification accuracy of contrastively learnt representations on CIFAR-10 improves with increasing the number of negative examples $k$.

[4]




[5]


[4] Awasthi, Pranjal, Nishanth Dikkala, and Pritish Kamath. "Do more negative samples necessarily hurt in contrastive learning?." International conference on machine learning. PMLR, 2022.
[5] Bayrak, Betül. "Effects of negative sampling on knowledge graph completion." 2020 5th International Conference on Computer Science and Engineering (UBMK). IEEE, 2020..

# INTRODUCTION

## Recent concerns in the KGC feild

- □ Despite the high value, negative sampling methods are not well explored
- □ Random or heuristic NS methods are still widely adopted and they are questionable in quality
- □ Thus providing better(harder) negatives to the model has huge potential

PS : (New York, location adjoining, New Jersey)

NS1 : (New York, location adjoining, Avatar Movie)

**easy negative**(model can't learn meaningful representation)

NS2 : (New York, location adjoining, New Caledonia)

**hard negative**(helps model learn better semantic)

[6]

[6] Zhang, Honggen, June Zhang, and Igor Molybog. "HaSa: Hardness and Structure-Aware Contrastive Knowledge Graph Embedding." Proceedings of the ACM Web Conference 2024. 2024.

# INTRODUCTION

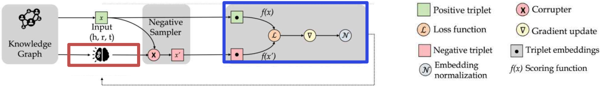
## ▪ Our objective and goal

□ Build a **sub-model**(NS Generator) that learns how to generate hard NS

□ Feed hard NS to the **main-model**(Emb base or GNN base, etc.) to enhance performance

□ Sub-model should be applicable to any main-model that uses NS

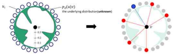
□ Make the sub-model and main-model to help each other(similar machanism to GAN)



# INTRODUCTION

- **Composition of todays seminar**

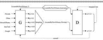
Noise Contrasive Estimation(NCE) :
**the birth of negative sampling(NS)**



why NS is important(theoretical proof)
and its **dilemma**

$$\mathbb{E}[\|(\theta_T - \theta^*)_u\|^2] = \frac{1}{T}(\frac{1}{p_d(u|v)} - 1 + \frac{1}{kp_n(u|v)} - \frac{1}{k})$$

recent works on **NS methods**



recent works on **loss function** design
from the perspective of NS

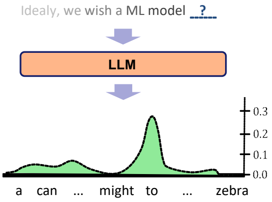
$$\frac{p_d(y|x)/p_n(y|x)}{\sum_{y' \in Y} (p_d(y'|x)/p_n(y'|x))} \qquad \mathcal{L}_N = -\mathbb{E}_X\left[\log \frac{f_k(x_{t+k}, c_t)}{\sum_{x_j \in X} f_k(x_j, c_t)}\right]$$

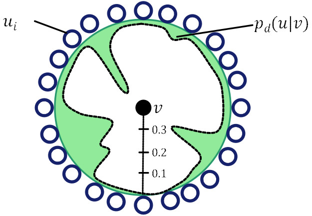
**our approach** to find **better** negative samples

**future questions and plans**

# NOISE CONTRASIVE ESTIMATION

- **Lets discuss 'training' from a lower level(feat. distribution)**
  - □ Data distribution($p_d(.)$) : underlying distribution of the true patterns
  - □ Idealy, we wish a ML model to model a certain data distribution
    *(left fig) data distribution in NLP* **/** *(right fig) data distribution in link prediction*



# NOISE CONTRASIVE ESTIMATION

## ▪ The antagonist "partition"

- □ If we want to model the data distribution directly(itself an impossible thing), below distribution should be calculated

$$p_d(u|v) = \frac{s_\theta(u|v)}{\underbrace{\sum_i s_\theta(u_i|v)}_{\text{partition function (normalization constant)}}}$$

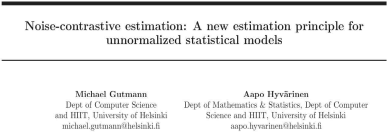
**partition function**(normalization constant)

- □ The partition function is impossible to calculate for real life data
- □ Even worse, we only have samples(=training data, $\hat{p}_d$ ) from the data distribution
- □ How can we make a model learn without needing to calculate the partition function?

# NOISE CONTRASIVE ESTIMATION

- **NCE : A new perspective to mitigate from partition**
  - □ *"The basic idea is to estimate the parameters by* ***learning to discriminate*** *between the data* ***x*** *and some artificially generated noise* ***y****"* $x \sim \hat{p}_d \quad y \sim p_n$
  - □ Distribution modeling → binary classification
  - □ No need for huge partitions

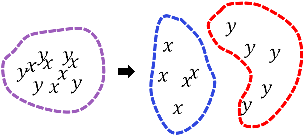
$$\frac{s_\theta(u|v)}{s_\theta(x|v) + \sum_{y \sim p_n} s_\theta(y|v)}$$



Noise-contrastive estimation: A new estimation principle for unnormalized statistical models

Michael Gutmann
Dept of Computer Science and HIIT, University of Helsinki
michael.gutmann@helsinki.fi

Aapo Hyvärinen
Dept of Mathematics & Statistics, Dept of Computer Science and HIIT, University of Helsinki
aapo.hyvarinen@helsinki.fi




[8] Gutmann, Michael, and Aapo Hyvärinen. "Noise-contrastive estimation: A new estimation principle for unnormalized statistical models." Proceedings of the thirteenth international conference on artificial intelligence and statistics. JMLR Workshop and Conference Proceedings, 2010.

# NOISE CONTRASIVE ESTIMATION

- **However, noise distribution is unknown**

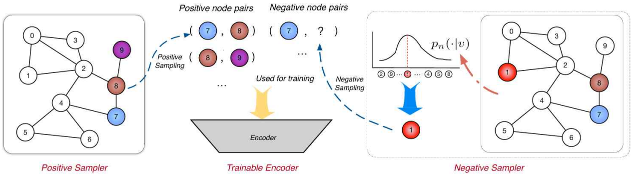



Figure 1: The SampledNCE framework. Positive pairs are sampled implicitly or explicitly according to the graph representation methods, while negative pairs are from a pre-defined distribution, both composing the training data of contrastive learning.

define an unknown distribution

[9] Yang, Zhen, et al. "Understanding negative sampling in graph representation learning." Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining. 2020.

# NOISE CONTRASIVE ESTIMATION

## ▪ Dilemma of designing noise distribution

□ [9] proved theoretical background of graph representation learning under SampledNCE

THEOREM 2. *The random variable $\sqrt{T}(\boldsymbol{\theta}_T - \boldsymbol{\theta}^*)$ asymptotically converges to a distribution with zero mean vector and covariance matrix*

$$Cov\big(\sqrt{T}(\boldsymbol{\theta}_T - \boldsymbol{\theta}^*)\big) = \boldsymbol{diag}(\boldsymbol{m})^{-1} - (1 + 1/k)\mathbf{1}^\top \mathbf{1}, \quad (5)$$

*where* $\boldsymbol{m} = \left[\frac{kp_d(u_0|v)p_n(u_0|v)}{p_d(u_0|v)+kp_n(u_0|v)}, \ldots, \frac{kp_d(u_{N-1}|v)p_n(u_{N-1}|v)}{p_d(u_{N-1}|v)+kp_n(u_{N-1}|v)}\right]^\top$ *and* $\mathbf{1} = [1, \ldots, 1]^\top$.

$$\mathbb{E}\left[\|(\boldsymbol{\theta}_T - \boldsymbol{\theta}^*)_u\|^2\right] = \frac{1}{T}\left(\underline{\frac{1}{p_d(u|v)}} - 1 + \underline{\frac{1}{kp_n(u|v)}} - \frac{1}{k}\right)$$

$\theta^*$ : *optimum parameters when trained by NCE*
$\theta_T$ : *optimum parameters when trained by SampledNCE*
$k$ : *# negative samples per 1 positive sample*

dilemma occurs between $p_d$, $p_n$

we can't maximize **both** at once

[9] Yang, Zhen, et al. "Understanding negative sampling in graph representation learning." Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining. 2020.

# NOISE CONTRASIVE ESTIMATION

- **Compensation is required when modeling noise distribution**
  - □ We would like the noise distribution to be similar to the data distribution
  - □ At the same time, don't want them to be that similar(false negative might be generated)
  - □ Building negative generator needs a different perspective & approach
  - □ The definition and nature of negative samples are scarcely studied compared to positive samples

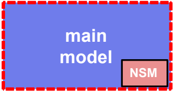
| ***positive samples*** | ***negative samples*** |
|---|---|
| • *just static data that are fed to the model*<br>• *relative quality is not defined(no one cares)* | • *imaginary data*<br>• *relative quality is not defined(definition is required)*<br>• *the effect of one self can only be seen indirectely* |

***how did recent works treat NS and combine them into training?***

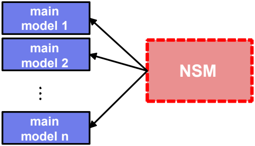
# RELATED WORKS : SAMPLING MEHODS

- **On generating harder negatives(generally two types of work)**
  - □ Negative sample method as a **side kick** → explored quite a bit since 2017
  - □ Negative sample method as a **module** → **hardely explored**

*<side kick>*




*<module>*



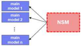
# RELATED WORKS : SAMPLING MEHODS

- **On generating harder negatives(generally two types of work)**
  - □ Both have different ways of showing their contributions

***<side kick>***


| Method | FB15k-237 | | WN18 | | WN18RR | |
|---|---|---|---|---|---|---|
| | MRR | H@10 | MRR | H@10 | MRR | H@10 |
| TransE | - | 42.8$^{\dagger}$ | - | 89.2 | - | 43.2$^{\dagger}$ |
| TransD | - | 45.3$^{\dagger}$ | - | 92.2 | - | 42.8$^{\dagger}$ |
| DistMult | 24.1$^{\ddagger}$ | 41.9$^{\ddagger}$ | 82.2 | 93.6 | 42.5$^{\ddagger}$ | 49.1$^{\ddagger}$ |
| ComplEx | 24.0$^{\ddagger}$ | 41.9$^{\ddagger}$ | **94.1** | 94.7 | **44.4**$^{\ddagger}$ | **50.7**$^{\ddagger}$ |
| TransE (pre-trained) | 24.2 | 42.2 | 43.3 | 91.5 | 18.6 | 45.9 |
| KBGAN (TransE + DistMult) | 27.4 | 45.0 | 71.0 | **94.9** | 21.3 | 48.1 |
| KBGAN (TransE + ComplEx) | **27.8** | 45.3 | 70.5 | **94.9** | 21.0 | 47.9 |
| TransD (pre-trained) | 24.5 | 42.7 | 49.4 | 92.8 | 19.2 | 46.5 |
| KBGAN (TransD + DistMult) | **27.8** | **45.8** | 77.2 | 94.8 | 21.4 | 47.2 |
| KBGAN (TransD + ComplEx) | 27.7 | **45.8** | 77.9 | 94.8 | 21.5 | 46.9 |

***<module>***




| TransE | FB13 | | | FB15K237 | | |
|---|---|---|---|---|---|---|
| | MRR↑ | MR↓ | Hit@10↑ | MRR↑ | MR↓ | Hit@10↑ |
| Uniform | 0.0820 | 4472 | 15.69 | 0.2188 | 282 | 38.48 |
| Bernoulli | 0.2460 | 5638 | 36.53 | 0.2257 | 268 | 39.56 |
| NSCaching | 0.3087 | **3804** | 40.12 | 0.3067 | 188 | 48.05 |
| GN+DN | 0.3137 | 4752 | 40.62 | 0.2895 | 190 | 47.45 |
| IF-NS | **0.3219** | 4870 | **42.22** | **0.3095** | **174** | **48.85** |
| TransH | FB13 | | | FB15K237 | | |
| | MRR↑ | MR↓ | Hit@10↑ | MRR↑ | MR↓ | Hit@10↑ |
| Uniform | 0.1041 | 12 315 | 16.49 | 0.2212 | 283 | 39.18 |
| Bernoulli | 0.2375 | 4802 | 35.35 | 0.2363 | 200 | 40.08 |
| NSCaching | 0.2891 | **3163** | 39.58 | 0.2931 | 199 | 47.81 |
| GN+DN | 0.3022 | 4585 | 40.15 | 0.2925 | 196 | 47.89 |
| IF-NS | **0.3058** | 4010 | **40.79** | **0.3089** | **180** | **48.93** |
| TransD | FB13 | | | FB15K237 | | |
| | MRR↑ | MR↓ | Hit@10↑ | MRR↑ | MR↓ | Hit@10↑ |
| Uniform | 0.1495 | 13 033 | 22.29 | 0.2175 | 295 | 38.10 |
| Bernoulli | 0.2468 | 4341 | 36.06 | 0.2354 | 240 | 40.73 |
| NSCaching | 0.3124 | **3817** | 41.30 | 0.3071 | 187 | 48.17 |
| GN+DN | 0.3145 | 4920 | 40.99 | 0.2907 | 198 | 47.54 |
| IF-NS | **0.3282** | 5064 | **42.25** | **0.3109** | **167** | **49.05** |


[10] Cai, Liwei, and William Yang Wang. "Kbgan: Adversarial learning for knowledge graph embeddings." arXiv preprint arXiv:1711.04071 (2017).
[12] Cai, M., Deng, Z., & Xiong, C. (2025). IF-NS: A New negative sampling framework for knowledge graph embedding using influence function. Knowledge-Based Systems, 315, 113258.

# RELATED WORKS : SAMPLING MEHODS

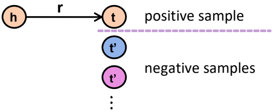
- **On generating harder negatives(SIDEKICK/heuristic)**

  □ **Random :** assume noise distribution is uniform

  □ **Batch NS :** only pick negative entities from the same mini-batch(memory efficient)

  □ **K-hop :** assume hard negative is arround k-hop of the target entity

  □ **NMiss :** selects negative candidates that are ranked higher than positive entity

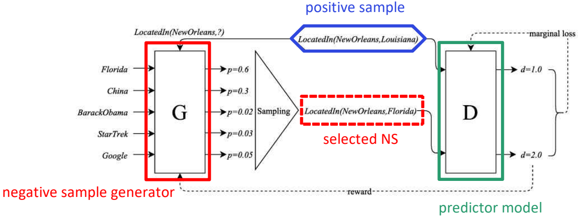



*#PS : #NS = 1 :* ***A***

***A*** *is a hyperparameter, usually selected from* ***256 ~ 1024***

# RELATED WORKS : SAMPLING MEHODS

- **On generating harder negatives(SIDEKICK/model)**
  - □ **KBGAN(arxiv'17)** : although proposed as a predictor model, generator can be used as a module

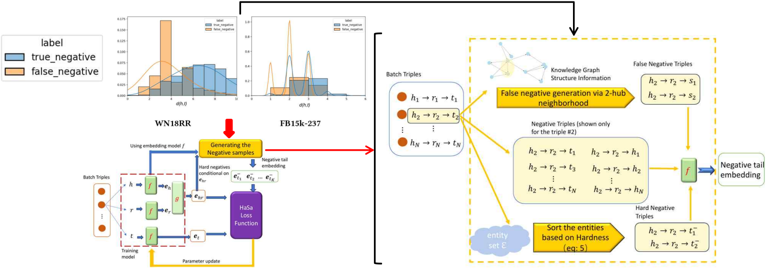




[10] Cai, Liwei, and William Yang Wang. "Kbgan: Adversarial learning for knowledge graph embeddings." arXiv preprint arXiv:1711.04071 (2017).

# RELATED WORKS : SAMPLING MEHODS

- **On generating harder negatives(module/model)**
  - □ **HaSa(WWW'24)** : introduces a false negative aware loss function and detection method




[6] Zhang, Honggen, June Zhang, and Igor Molybog. "HaSa: Hardness and Structure-Aware Contrastive Knowledge Graph Embedding." Proceedings of the ACM Web Conference 2024. 2024.

# RELATED WORKS : SAMPLING MEHODS

- **On generating harder negatives(deviating from false negatives)**
  - □ Use the nature of **relation cardinality** to mitigate from generating false negatives
  - □ A relation is either 1-1, 1-N, N-1, N-N
  - □ Only change the '1' side when generating negatives
  - □ Effective, but **1) 1-1 relations are minority**, **2) can't change 1 side if target entity is on the other side**

***Relation cardinality of FB15k237***

*tail side*

| *head side* | **1** | **N** |
|---|---|---|
| **1** | 7.17 | 10.97 |
| **N** | 36.28 | 45.56 |

*%*

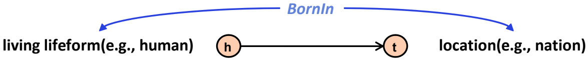
$h \xrightarrow{r(1:N)} t$

**if** head-batch → '1'part should be negative sampled

**if** tail-batch → 'N'part should be negative sampled

# RELATED WORKS : SAMPLING MEHODS

- **On generating harder negatives(deviating from false negatives)**
  - □ Regardless of relation cardinality, **entity type** can be used to deviate from generating false negatives
  - □ If a relation is given, head and tail's type is restricted
  - □ By not sampling the same nature entity for negative, false negative is prevented
  - □ Flexible than cardinality aware approach, but is the generated negative even hard?

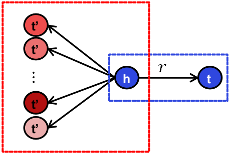


# RELATED WORKS : SAMPLING MEHODS

- **On generating harder negatives(deviating from false negatives)**
  - □ A more novel and reasonable approach would be to train the NS generator to handle this issue
  - □ We haven't dove into this topic that much, lets discuss it next time

# RELATED WORKS : LOSS FUNCTION

- **General form of NS incorporated loss function**
  - □ **1** positive and **k** negatives
  - □ Negative part is usually meaned to prevent underfitting

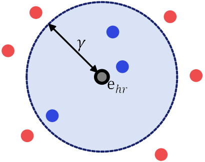



$$argmax_{\theta}(\underline{s_{\theta}(h,r,t)} - \underline{\mathbb{E}_{(h,r,t')\sim NS}[s_{\theta}(h,r,t')]})$$

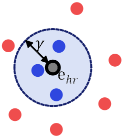
# RELATED WORKS : LOSS FUNCTION

- **Distance based loss function with margin terms**
  - □ **Margin(γ)** defines the boundary between 'right' and 'wrong'
  - □ One of the most commonly used loss function in KGC feild

$$L = \underbrace{-\log \sigma(\gamma - d_r(\mathbf{h}, \mathbf{t}))}_{\textit{PS loss}} - \underbrace{\sum_{i=1}^{n} \frac{1}{k} \log \sigma(d_r(\mathbf{h}_i', \mathbf{t}_i') - \gamma)}_{\textit{NS loss(usually k=n)}}$$




*big* **γ**



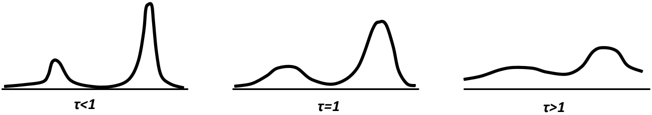

*small* **γ**

$\in$ PS

$\in$ NS

# RELATED WORKS : LOSS FUNCTION

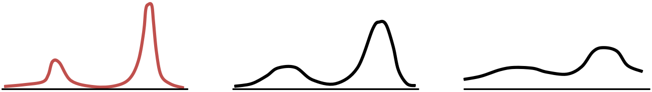
- **InfoNCE(Information Noise Contrasive Estimation)**
  - □ A softmax form loss function
  - □ A temperature hyper-parameter **τ** to control the distribution sharpness
    *some view* ***τ*** *as a learnable parameter*




$$\mathcal{L}_{InfoNCE} = -\log \frac{exp(s_\theta(h,r,t)/\tau)}{exp(s_\theta(h,r,t)/\tau) + \sum_{i=1}^{n} exp(s_\theta(h,r,t_i^{'})/\tau)}$$

# RELATED WORKS : LOSS FUNCTION

- **InfoNCE(Information Noise Contrasive Estimation)**
  - □ Why doesn't InfoNCE calculate the mean over the negative part?
  - □ ***τ** is usually set to a small value(**0.07~0.1**)*



***τ<1*** ***τ=1*** ***τ>1***

$$\mathcal{L}_{InfoNCE} = -\log \frac{exp(s_\theta(h,r,t)/\tau)}{exp(s_\theta(h,r,t)/\tau) + \sum_{i=1}^{n} exp(s_\theta(h,r,t_i^{'})/\tau)}$$

# RELATED WORKS : LOSS FUNCTION

- **SANS(Self-Adversarial Negative Sampling)**
  - □ Weighted mean over NS
  - □ Weight is determined by the current model's state
  - □ **α** is the temperature hyper-parameter(similar role with InfoNCE's **τ**)

$$p(h'_j, r, t'_j | \{(h_i, r_i, t_i)\}) = \frac{\exp \alpha f_r(\mathbf{h}'_j, \mathbf{t}'_j)}{\sum_i \exp \alpha f_r(\mathbf{h}'_i, \mathbf{t}'_i)}$$

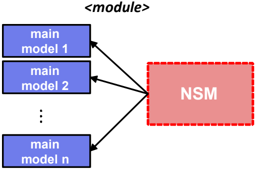
$$L = -\log \sigma(\gamma - d_r(\mathbf{h}, \mathbf{t})) - \sum_{i=1}^{n} p(h'_i, r, t'_i) \log \sigma(d_r(\mathbf{h}'_i, \mathbf{t}'_i) - \gamma)$$

[2] Sun, Zhiqing, et al. "Rotate: Knowledge graph embedding by relational rotation in complex space." ICLR arXiv:1902.10197 (2019).

# OUR APPROACH

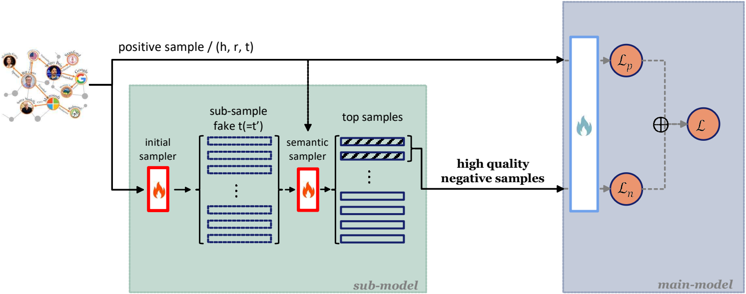
## NS generator

- □ Several criteria for “good hard negative generator”
- □ **1)** improve the main model performance
- □ **2)** mitigate from false negatives
- □ **3)** efficiency
- □ **4)** generalizability(model & dataset)



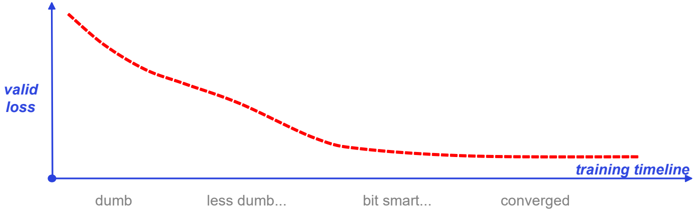
# OUR APPROACH

- **Skeleton of child-model(NS generator)** : sub-model training : main-model training



# DISCUSSION

- **(1) Should we always give hard negatives to the model?**
  - □ Will hard negative help even in the early stage of training(when the model is still dumb)?
  - □ If we think from a human perspective...



# DISCUSSION

## (2) 'Diversity vs Hard negative'(tradeoff)

□ Random sampling provides diverse negatives for the model

□ Hard negative sampling sacrifices diversity over harder samples

□ However, can we catch both diversity and hard negative?

# DISCUSSION

## (3) New loss function

- □ Conventional loss functions restrict the influence of negative loss part
- □ This was a reasonable technique when negatives were too easy to discriminate
- □ However, as importance of negative sampling grows, the loss function must change as well
- □ From the perspective of NCE, what is the idle loss function?

$$L = -\log \sigma(\gamma - d_r(\mathbf{h}, \mathbf{t})) - \sum_{i=1}^{n} \frac{1}{k} \log \sigma(d_r(\mathbf{h}'_i, \mathbf{t}'_i) - \gamma)$$

$$\mathcal{L}_{InfoNCE} = -\log \frac{exp(s_\theta(h,r,t)/\tau)}{exp(s_\theta(h,r,t)/\tau) + \sum_{i=1}^{n} exp(s_\theta(h,r,t_i^{'})/\tau)}$$

# DISCUSSION

## (4) When will 'good negative sample generator' prevail?

- Good NS generator's contribution might prevail when the dataset is
  - **Difficult** : low quality negatives won't help the model that much
  - **Huge in volume** : where neg batch size is restricted, quality of each NS will be more important

**ICLR'24 reviewer comment on negative sample generator paper [11]**
***"Moreover, the study should use datasets that are so large that negative sampling is actually needed."***

[11] https://openreview.net/forum?id=rcsV1C2eyk

# Thank You!


**Contact**: Sooho Moon (Email: moonwalk725@cau.ac.kr)