Identical and Fraternal Twins: Fine-Grained Semantic Contrastive Learning of Sentence Representations

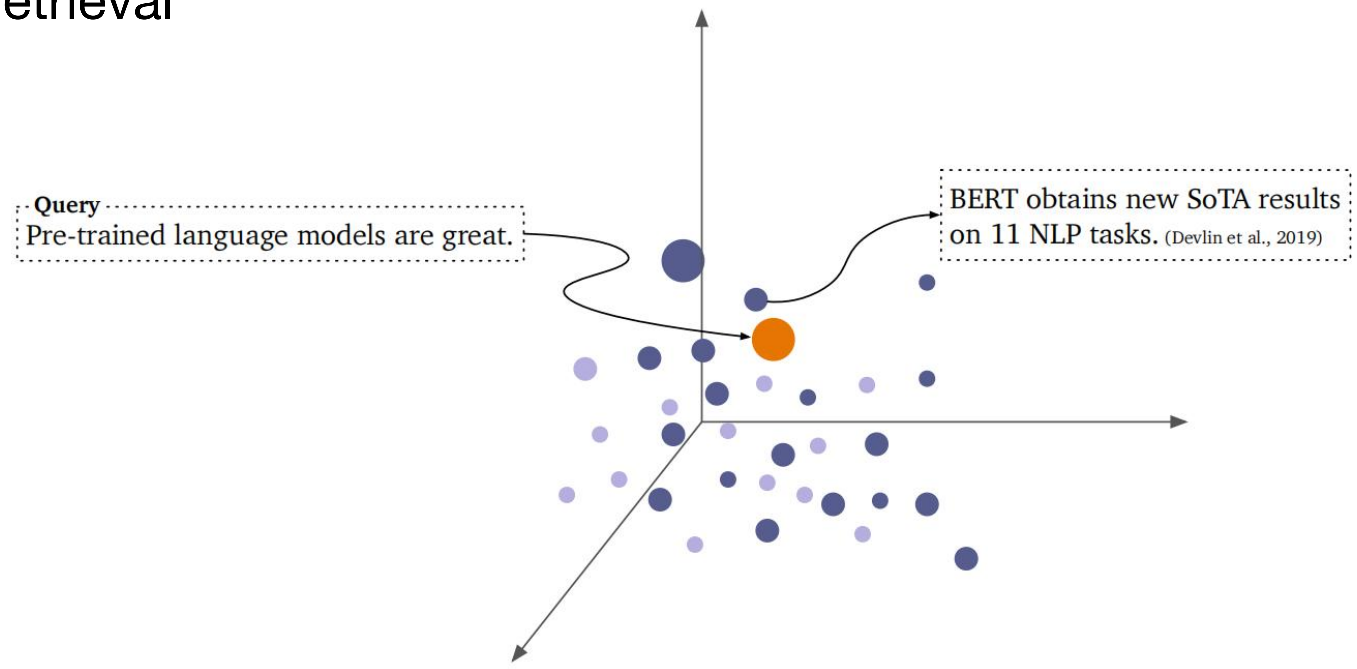
Qingfa Xiao[1,3], Shuangyin Li[1,(✉)] and Lei Chen[2,3]

[1]South China Normal University
[2]The Hong Kong University of Science and Technology
[3]The Hong Kong University of Science and Technology (Guangzhou)
qingfaxiao@m.scnu.edu.cn, shuangyinli@scnu.edu.cn, leichen@cse.ust.hk

Background: Sentence Embeddings

Learning universal representations of sentences has wide applications in NLP

- Semantic matching
- Sentence clustering
- Information retrieval
- ...

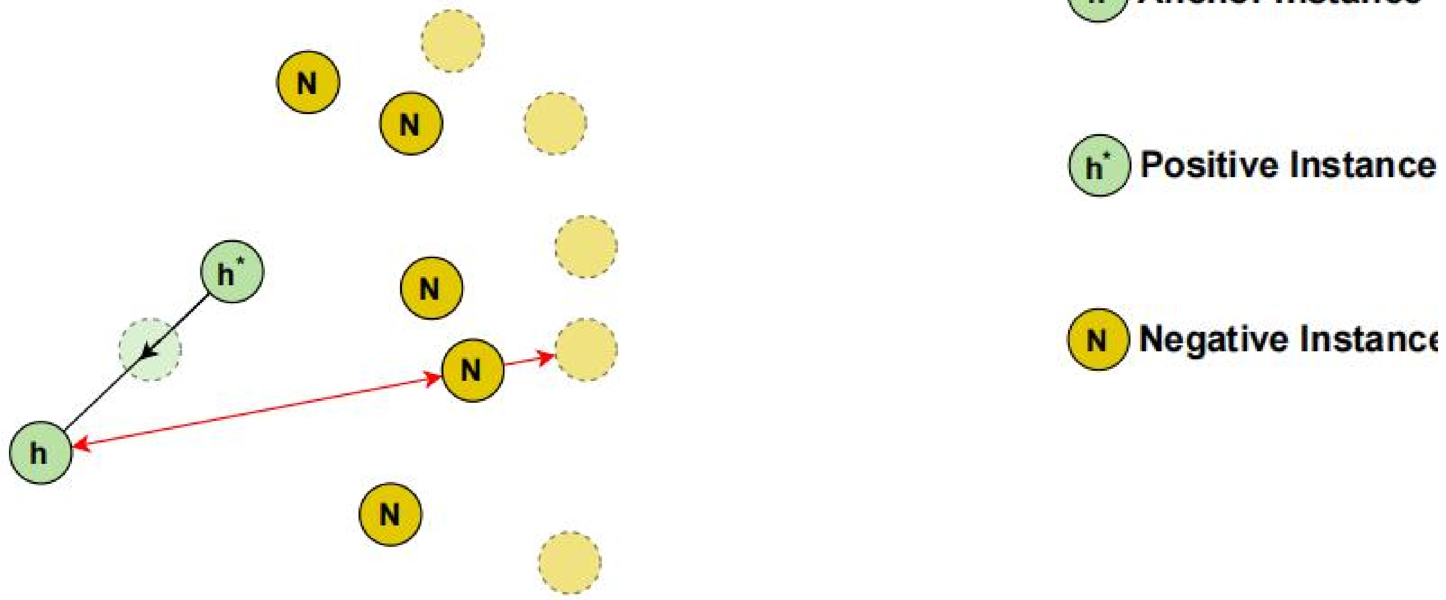


Background: Contrastive Learning

Contrastive learning is an intuitive and effective training objective that aims to create desired semantic representations by

- **bringing** semantically positive instances **closer** together
- **pushing away** those that are not semantically negative.

(a) Contrastive Learning with InfoNCE



Motivation:

The lack of fine-grained semantic discrimination ability via contrastive learning.

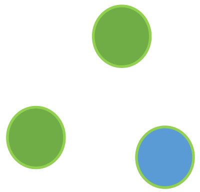
The girl with a red hat. → Data Augmentations →

The girl with ~~a~~ red hat.

The girl with a ~~red~~ hat.

The girl with a red <u>bike</u>.

→ Contastive Learning →

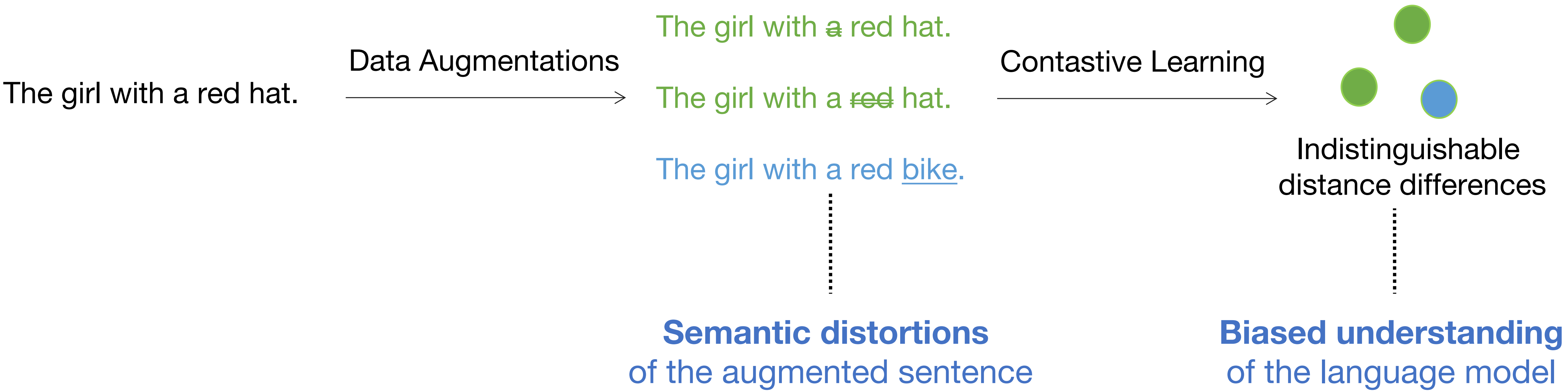


Indistinguishable distance differences

Semantic distortions of the augmented sentence

Biased understanding of the language model

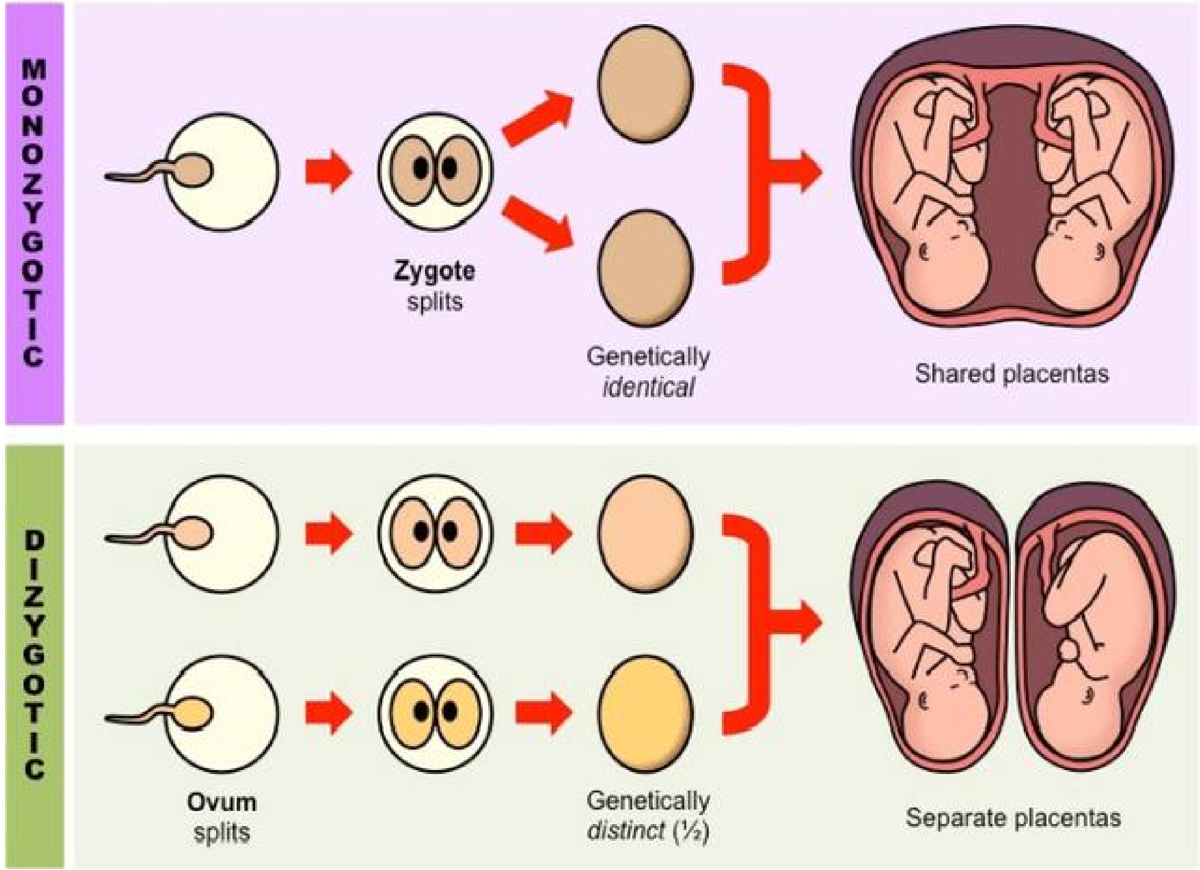
Motivation:
The lack of fine-grained semantic discrimination ability via contrastive learning.



When the different types of positive pairs come to contrastive learning,

they should be treated under the different standards.

Motivation:
Which twins are more similar?



Identical Twins

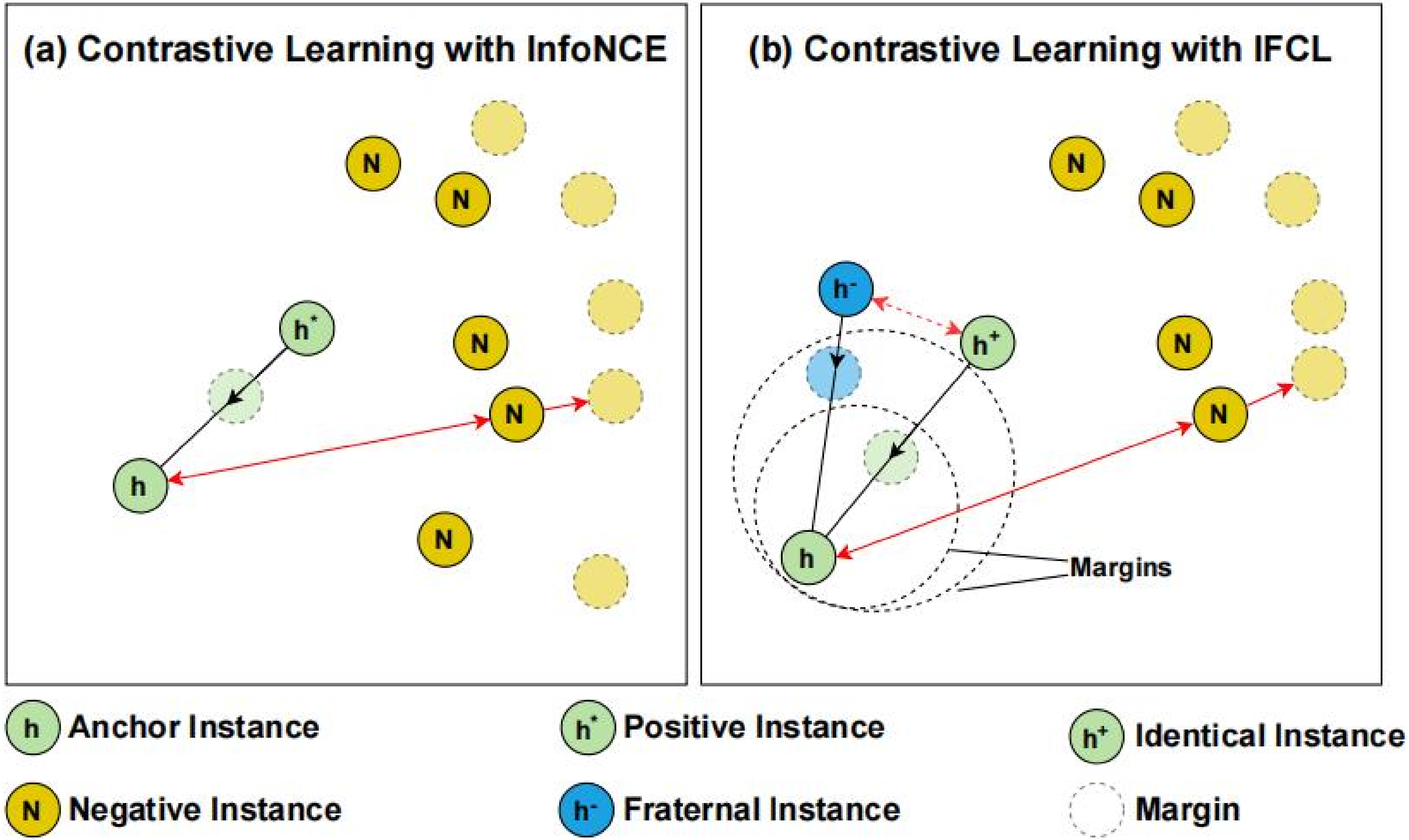
Fraternal Twins

Everyone can easily distinguish between these two pairs of twins.

But, can a language model do the same?

Objective:

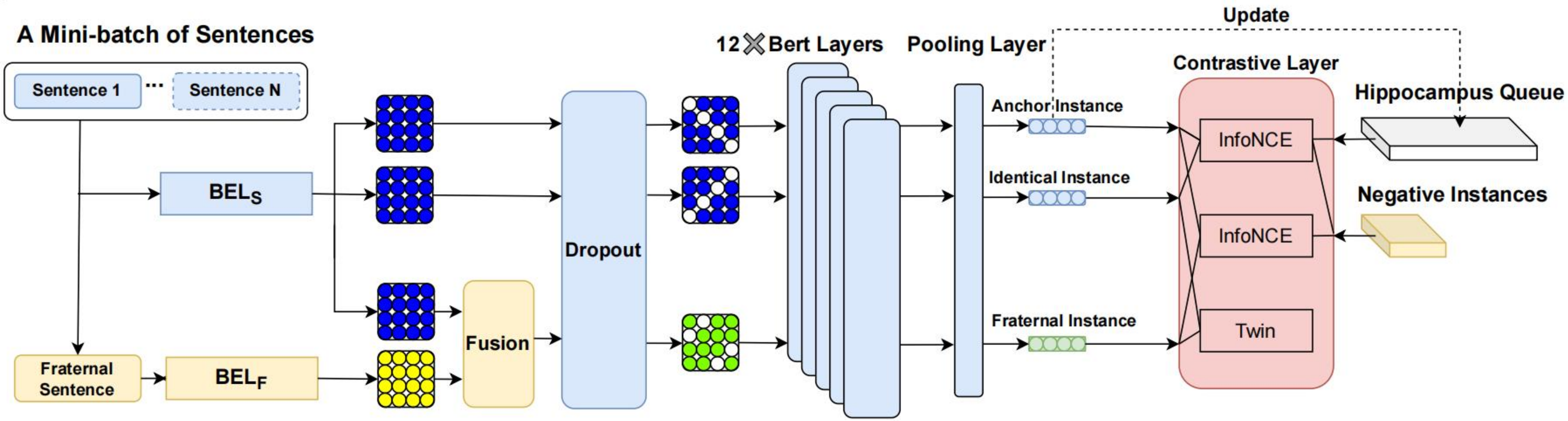
Keep the margins between the two pairs of twins to help model distinguish the subtle differences.



Challenges:
Contrastive learning with data augementation

- Data augementation for gaining positve pairs with less **semantic distortions**
- Adaptive contrastive learning for different types of positive samples to **address sub-optimal issues**

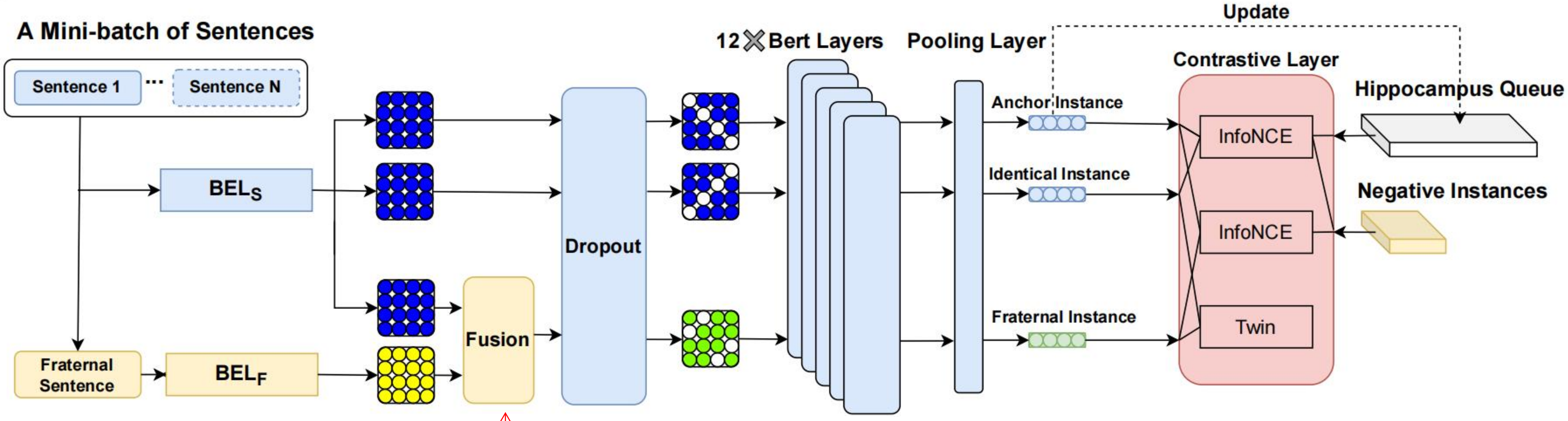
IFCL Framework:



Three main components in this IFCL framework:

- A fusion data augmentation technique
- A training loss function named Twin Loss
- A hippocampus queue mechanism

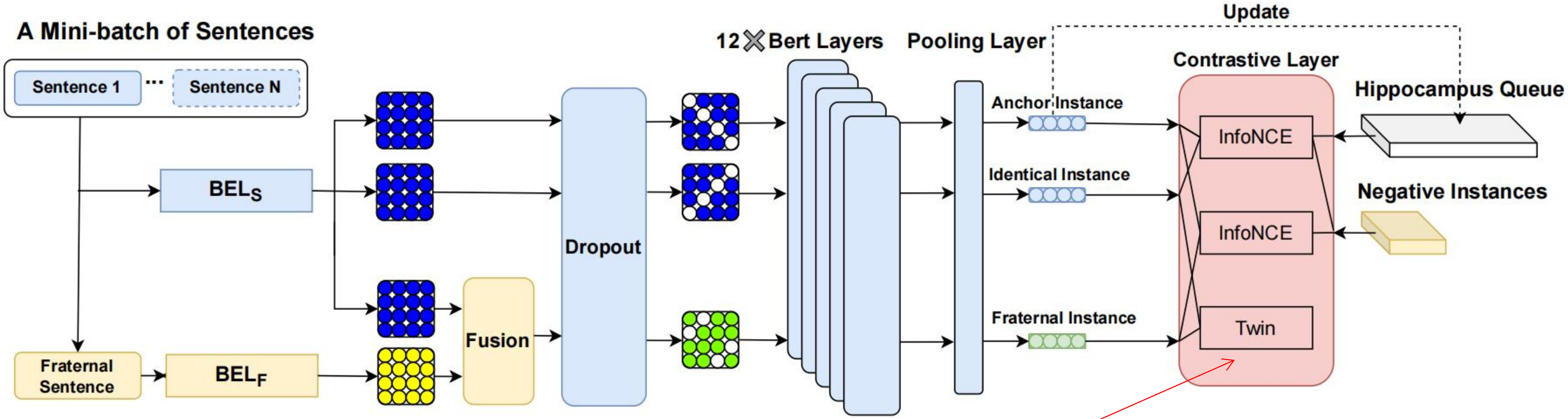
IFCL Framework:



- A fusion data augmentation technique

minimizing semantic distortions and **increasing diversity** of expressions.

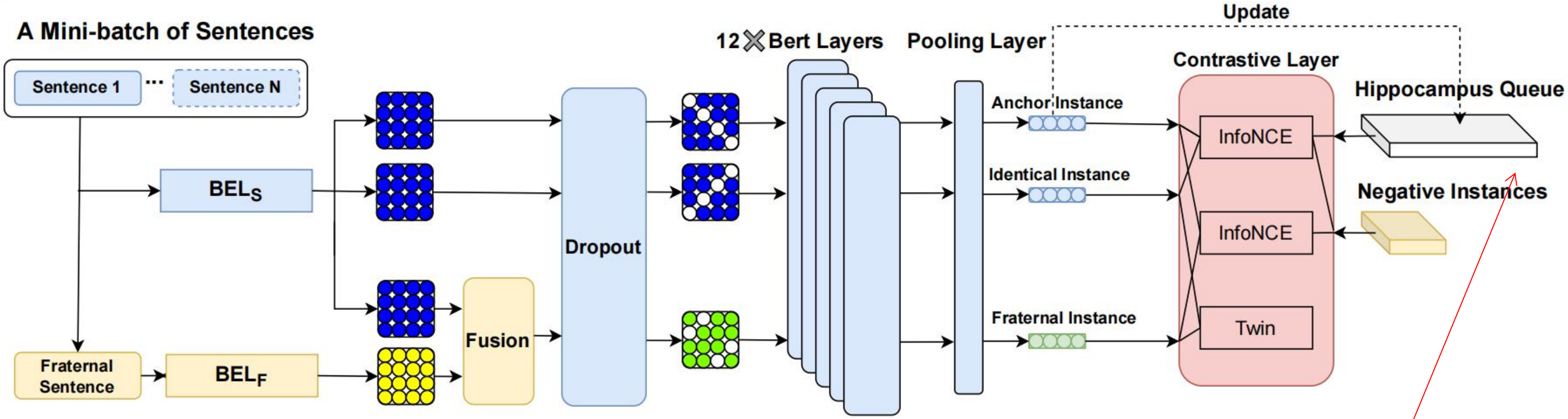
IFCL Framework:



- A training loss function named Twin Loss

capturing **fine-grained semantics** and alleviating the sub-optimal issues according to their margins

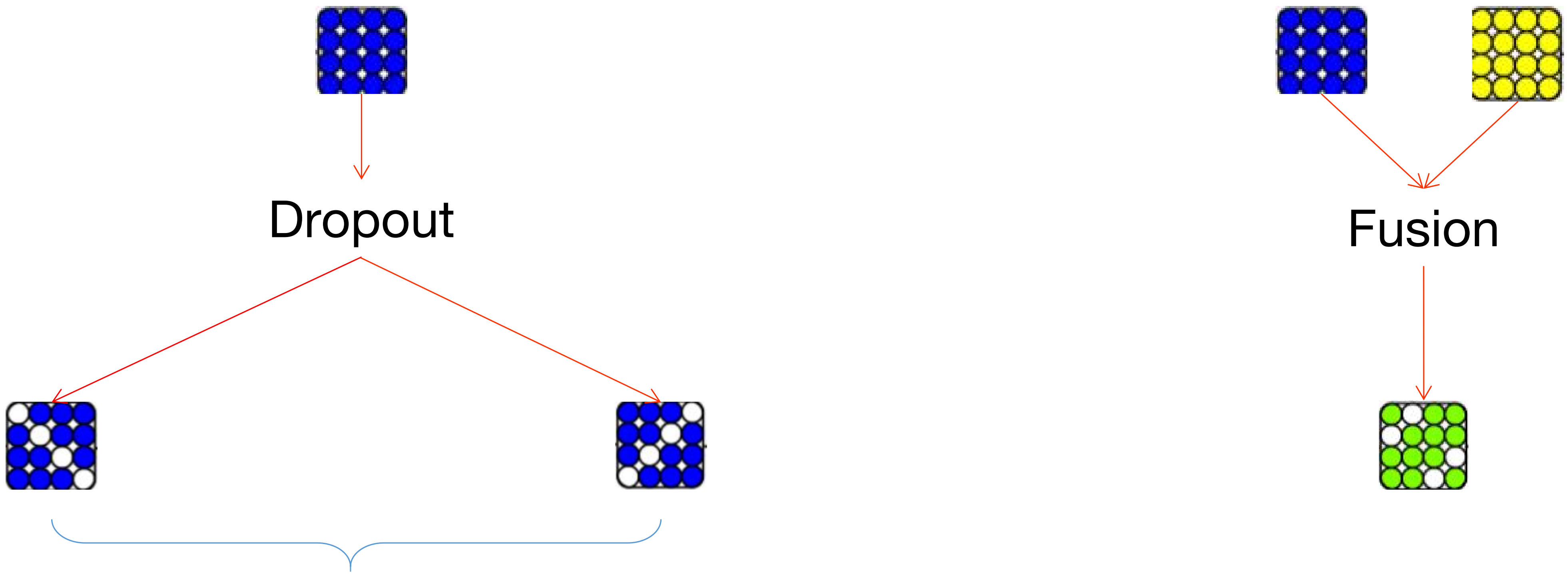
IFCL Framework:



- A hippocampus queue mechanism

storing the previous mini-batches into a short-term memory and **reusing** the negative effectively

Method:
How to generate the Identical and Fraternal Twins?



Identical twins: the most similar positive pair

Fraternal twins: the diverse pair with less semantic distortions

Method:

InfoNCE Loss with positive and negative instances

For the set of identical twins $\{h_i, h_i^+\}_{i=1}^N$ or fraternal twins $\{h_i, h_i^-\}_{i=1}^N$, we define the function by using negative instances $\{\mathbf{H}_m\}_{m=1}^{k*N}$ stored in the hippocampus queue.

$$\ell_i^I = -\log \frac{e^{\mathrm{sim}\left(\mathbf{h}_i, \mathbf{h}_i^+\right)/\tau}}{\sum_{j=1}^N e^{\mathrm{sim}\left(\mathbf{h}_i, \mathbf{h}_j^+\right)/\tau} + \varphi},$$

$$\varphi = \sum_{m=1}^{k*N} p_m * e^{\mathrm{sim}(\mathbf{h}_i, \mathbf{H}_m)/\tau},$$

Method:
Twins Loss for fine-grained semantic understanding

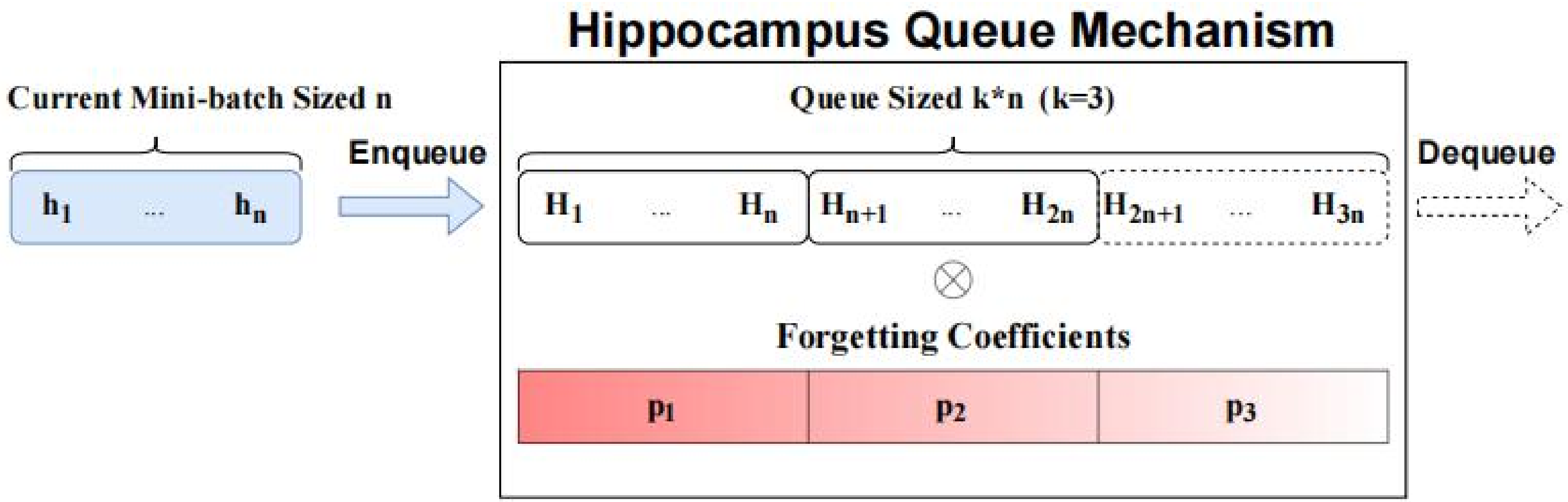
- This loss function aim to **keeping the margins** between two types of positive pairs
- M represents the **innate margins** between identical and fraternal twins
- Each M depends on the previous step to **prevent sub-optimal optimization problems**

$$\ell_i^T = \left| e^{\mathrm{sim}(\mathbf{h}_i, \mathbf{h}_i^{+})} - e^{\mathrm{sim}(\mathbf{h}_i, \mathbf{h}_i^{T})} - \mathbf{M}_i \right|,$$

$$\mathbf{M}_i = e^{\mathrm{sim}(\mathbf{emb}_i, \mathbf{emb}_i^{+})} - e^{\mathrm{sim}(\mathbf{emb}_i, \mathbf{emb}_i^{-})}$$

Method:
Hippocampus Queue Mechanism for reusing instances

- The queue storing the negative is **continuously updated**
- The sample is gradient-free to **save GPU memory**
- The forgetting coefficient focus more on the **latest instance**



Results:

Experiments on semantic textual similarity tasks

- Evaluate using Spearman's correlation metric
- Performe well in both Chinese and English tasks.

Results of Chinese tasks		
Method	**Chinese STS-B**	**SimCLUE**
BERT	55.52	29.89
BERT-whitening$^\bullet$	68.27	-
SimCSE-BERT$^\bullet$	68.91	40.74
SimCSE-BERT$^\diamond$	60.41	40.54
IFCL-BERT$^\diamond$	**71.41**	**44.42**

Results of English tasks								
Method	**STS12**	**STS13**	**STS14**	**STS15**	**STS16**	**STS-B**	**SICK-R**	**Avg.**
BERT$_{base}$	39.70	59.38	49.67	66.03	66.19	53.87	62.06	56.70
BERT-flow$_{base}$	58.40	67.10	60.85	75.16	71.22	68.66	64.47	66.55
BERT-whitening$_{base}$	57.83	66.90	60.90	75.08	71.31	68.24	63.73	66.28
ConSERT$_{base}$	64.64	78.49	69.07	79.72	75.95	73.97	67.31	72.74
SimCSE-BERT$_{base}^\diamond$	68.40	82.41	74.38	80.91	78.56	76.85	72.23	76.25
VaSCL-BERT$_{base}^\diamond$	69.08	81.95	74.64	82.64	80.57	80.23	71.23	77.19
DCLR-BERT$_{base}^\diamond$	70.81	83.73	75.11	82.56	78.44	78.31	71.59	77.22
MoCoSE-BERT$_{base}^\diamond$	71.48	81.40	74.47	**83.45**	78.99	78.68	**72.44**	77.27
PT-BERT$_{base}^\diamond$	71.20	**83.76**	**76.34**	82.63	78.90	79.42	71.94	77.74
IFCL-BERT$_{base}^\diamond$	**71.57**	82.35	75.08	83.03	**80.17**	**80.27**	72.16	**77.80**
BERT$_{large}$	57.73	61.17	61.18	68.07	70.25	59.59	60.34	62.62
ConSERT$_{large}$	70.69	82.96	74.13	82.78	76.66	77.53	70.37	76.45
SimCSE$_{large}^\diamond$	70.88	84.16	76.43	84.50	79.76	79.26	73.88	78.41
DCLR-BERT$_{large}^\diamond$	71.87	**84.83**	**77.37**	**84.70**	**79.81**	79.55	74.19	78.90
MoCoSE-BERT$_{large}^\diamond$	**74.50**	84.54	77.32	84.11	79.67	80.53	73.26	79.13
IFCL-BERT$_{large}^\diamond$	73.88	84.31	76.64	84.01	79.56	**81.37**	**76.30**	**79.44**

Analyse:
What makes the Twins Loss effective?

Reducing the mutual information between positive pairs while **preserving task-relevant information** is optimal for the task

- More diverse semantics are preserved

$\mathbb{MI}(h, h^-)$ is higher than $\mathbb{MI}(h, h^+)$

- Mutual information of positive pairs contains more task-relevant information.

$\mathbb{MI}(h, h^+) \approx \mathbb{MI}(h, h^-) \approx \mathbb{MI}_{task}$

Table 4. Mutual information and task-relevant information. The IFCL-BERT w/o TL means training IFCL-BERT without Twins Loss. The experiments are conducted with EnData and STS-B datasets on Bert-base.

Method	$\mathbb{MI}(h, h^+)$	$\mathbb{MI}(h, h^-)$	$\mathbb{MI}_{task}$
IFCL-BERT	4.15	4.17	4.31
IFCL-BERT w/o TL	4.23	4.20	4.58
SimCSE	4.24	-	4.52

Thank YOU

Q & A