Virtual Augmentation Supported Contrastive Learning of Sentence Representations

Advisor : Jia-Ling, Koh

Speaker : Ting-I, Weng

Source : ACL'22

Date : 2023/06/06

Outline

- Introduction
- Method
- Experiment
- Conclusion

Data Augmentation

- generate additional training data

	原始句子	增強後的句子1	增強後的句子2
隨機字刪除	已經決定要去看醫生了，但又怕醫藥費很貴	已經決定要去看醫生了，但又怕醫藥費__	已經決定要去看醫生了_又怕醫藥費__
隨機同義詞替換	雖然知道自己有病要看醫生，但一直提不起勇氣去看	雖然知道自己卧病要看醫生，但一直提不起勇氣去看	雖然知道和谐有病要看醫生，但一直提不起勇氣去看
隨機實體替換	星期六要去找醫生	中国通信服务要去找醫生	瑞丰光电要去找醫生
隨機近義詞替換	我覺得我該去看醫生了	腥覺得我絯去看醫生了	媷觼得我該去看缢生了

Problem

- SimCSE
  - unsupervised
    - dropout
  - supervised
    - dropout + hard negative samples

Problem

- VaSCL : Based on SimCSE
  - unsupervised
    - dropout + hard negative samples

Solution

- VaSCL
  - dropout based on unsupervised SimCSE
  - k-nearest neighbors algorithm
  - perturbations

Outline

- Introduction
- Method
- Experiment
- Conclusion

https://huggingface.co/datasets/mteb/sts12-sts

Input

Dataset : STS-12

Dataset Preview

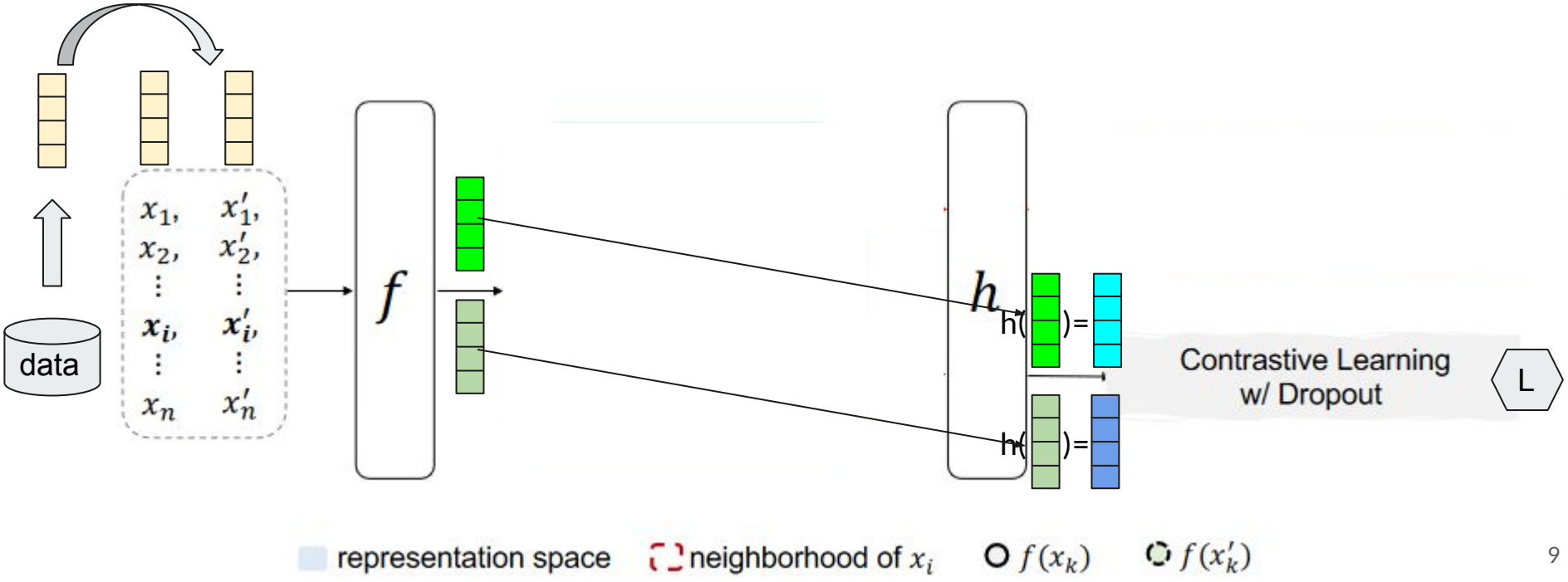
Size: 250 kB | API | Go to dataset viewer

Split

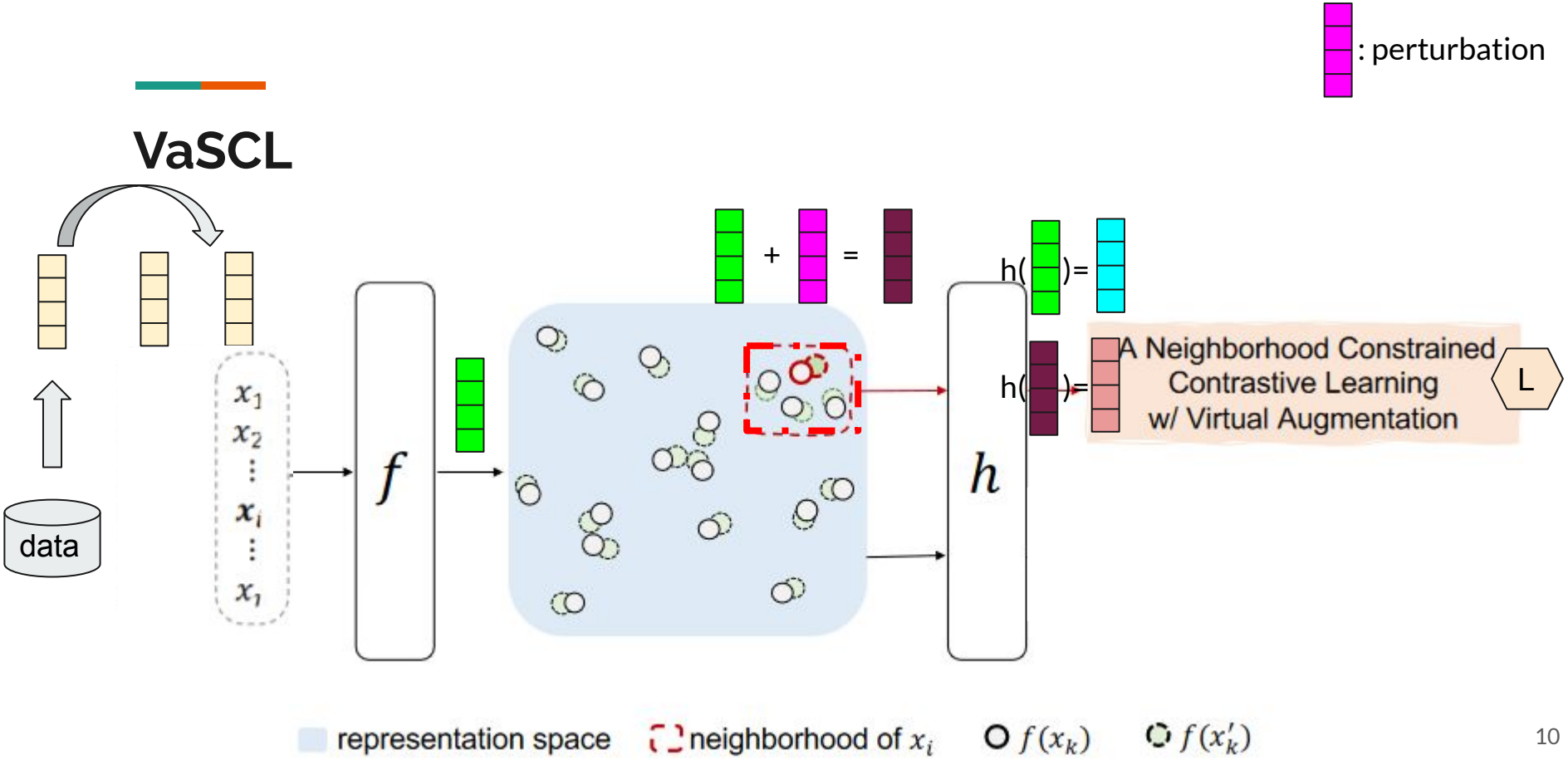
train (2.23k rows)

split (string)	sentence1 (string)	sentence2 (string)	score (float64)
"train"	"But other sources close to the sale said Vivendi was keeping the door open to further bids and…	"But other sources close to the sale said Vivendi was keeping the door open for further bids in th…	4
"train"	"Micron has declared its first quarterly profit for three years."	"Micron's numbers also marked the first quarterly profit in three years for the DRAM manufacturer."	3.75
"train"	"The fines are part of failed Republican efforts to force or entice the Democrats to return."	"Perry said he backs the Senate's efforts, including the fines, to force the Democrats to…	2.8

VaSCL : Based on SimCSE



VaSCL


: perturbation
data
x_1
x_2
x_i
x_1
f
+
=
h( )=
h( )=
h
A Neighborhood Constrained Contrastive Learning w/ Virtual Augmentation
L
representation space
neighborhood of x_i
$f(x_k)$
$f(x'_k)$

Loss : Based on SimCSE

$$\ell_{\mathcal{B}}(\mathbf{z}_i, \mathbf{z}_{i'}) = \quad (1)$$

$$-\log \frac{e^{\mathrm{sim}(\mathbf{z}_i,\mathbf{z}_{i'})/\tau}}{e^{\mathrm{sim}(\mathbf{z}_i,\mathbf{z}_{i'})/\tau} + \sum_{j\in\mathcal{B}\setminus(i,i')} e^{\mathrm{sim}(\mathbf{z}_i,\mathbf{z}_j)/\tau}}$$

postive sample

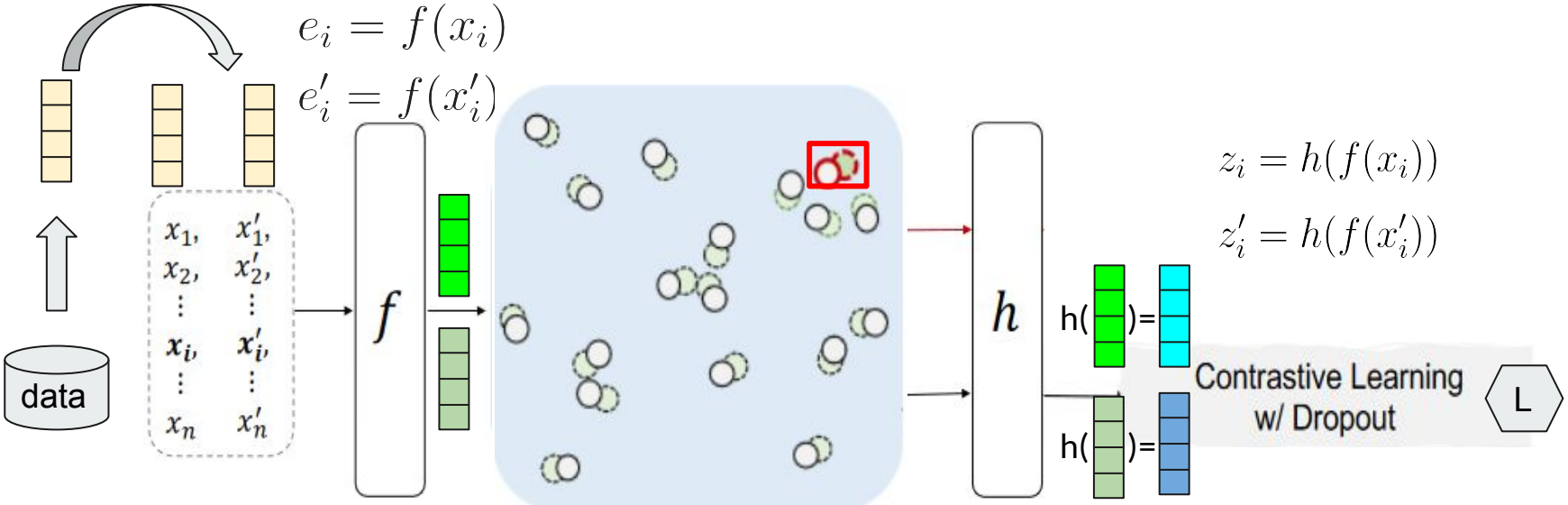
postive sample

negative sample

- Contrastive Learning w/ Dropout

$$e_i = f(x_i)$$

$$e'_i = f(x'_i)$$



$$z_i = h(f(x_i))$$

$$z'_i = h(f(x'_i))$$

Loss : VaSCL

$$\ell_{\mathcal{N}(i)}(\mathbf{z}_i^{\delta}, \mathbf{z}_i) = \tag{2}$$

$$-\log \frac{e^{\mathbf{sim}(\mathbf{z}_i^{\delta}, \mathbf{z}_i)/\tau}}{e^{\mathbf{sim}(\mathbf{z}_i^{\delta}, \mathbf{z}_i)/\tau} + \sum_{k \in \mathcal{N}(i)} e^{\mathbf{sim}(\mathbf{z}_i^{\delta}, \mathbf{z}_k)/\tau}}.$$

postive sample

postive sample

negative sample

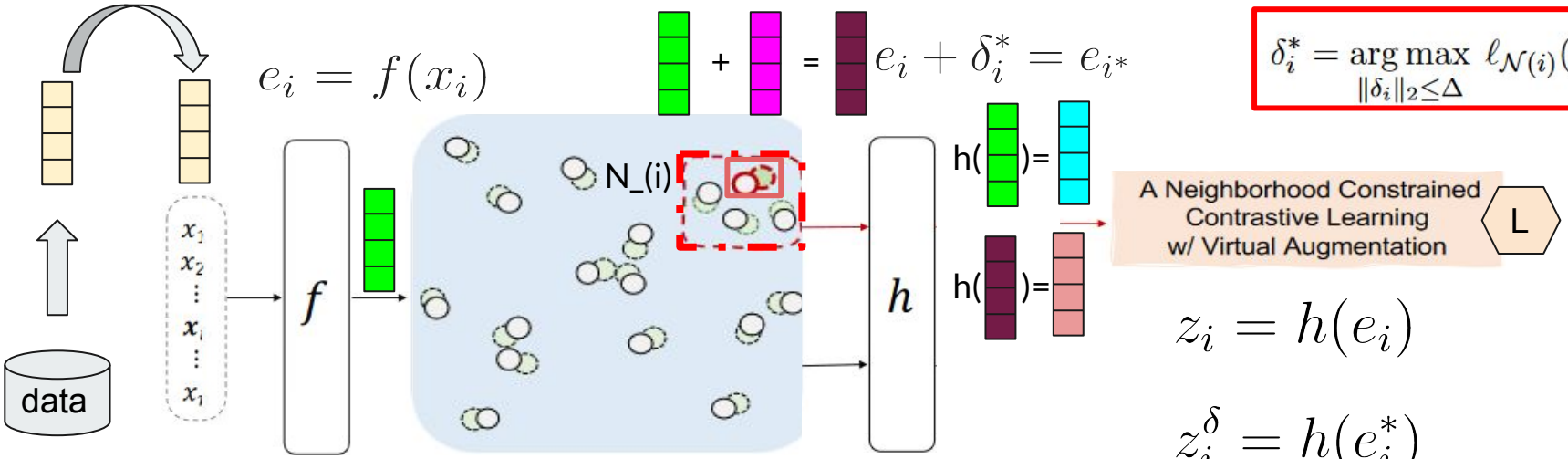
- Neighborhood Constrained Contrastive Learning
  - Find a good perturbation

$$sim(\cdot) = \frac{Z_i^T \cdot Z_i'}{\|Z_i\|_2 \, \|Z_i'\|_2}$$

$$\boxed{\delta_i^* = \underset{\|\delta_i\|_2 \leq \Delta}{\arg\max} \; \ell_{\mathcal{N}(i)}(\mathbf{z}_i^{\delta}, \mathbf{z}_i)}$$

$$e_i = f(x_i)$$

$$e_i + \delta_i^* = e_{i*}$$



$$z_i = h(e_i)$$

$$z_i^{\delta} = h(e_i^*)$$

Loss

A Neighborhood Constrained Contrastive Learning w/ Virtual Augmentation

L

+

Contrastive Learning w/ Dropout

L

=

L

L

$$\mathcal{L}_{\text{VaSCL}} = \frac{1}{2M} \sum_{i=1}^{M} \{ \ell_{\bar{\mathcal{B}}}(\mathbf{z}_i, \mathbf{z}_{i'}) + \ell_{\bar{\mathcal{B}}}(\mathbf{z}_{i'}, \mathbf{z}_i) + \ell_{\mathcal{N}_{\text{A}}(i)}(\mathbf{z}_i, \mathbf{z}_i^*) + \ell_{\mathcal{N}_{\text{A}}(i)}(\mathbf{z}_i^*, \mathbf{z}_i) \}$$

L

Outline

- Introduction
- Method
- Experiment
- Conclusion

- https://huggingface.co/mteb (STS12-16 & STS-B)
- https://github.com/facebookresearch/SentEval
- https://zenodo.org/record/2787612 (SICK)

Datasets

- semantic textual similarity (STS) related tasks

corpus	STS(2012-2016)	SICK	STS benchmark
Type	evaluate text similarity	relatedness and entailment	Include text from image captions, news headlines, and user forums

- short text clustering

corpus	AGNews	SearchSnippets	StackOverflow	Biomedical	Tweet	Google News
Source	Web	web	web	web	web	web
Task Type	News Topic Classification	Topic Classification	Short Text Classification	Short Text Classification	Short Text Classification	Short Text Classification

- https://github.com/sonos/nlu-benchmark

Datasets

- Intent Classification

corpus	SNIPS	BANKING77	HWU64	CLINC50
distinct intents	7	77	64	150

- distinct intents
  - Search Creative Work (e.g. Find me the I, Robot television show),
  - Get Weather (e.g. Is it windy in Boston, MA right now?),
  - Book Restaurant (e.g. I want to book a highly rated restaurant for me and my boyfriend tomorrow night),
  - Play Music (e.g. Play the last track from Beyoncé off Spotify),
  - Add To Playlist (e.g. Add Diamonds to my roadtrip playlist)
  - Rate Book (e.g. Give 6 stars to Of Mice and Men)
  - Search Screening Event (e.g. Check the showtimes for Wonder Woman in Paris)

- SimCSE:
  - pre-trained RoBERTa
- VaSCL
  - pre-trained RoBERTa

Semantic textual similarity (STS) related tasks

	STS12	STS13	STS14	STS15	STS16	SICK-R	STS-B	Avg.
RoBERTa$_{distil}$	54.41	46.85	56.96	65.79	64.22	61.10	59.01	58.33
SimCSE$_{distil}$	65.58	77.42	70.17	79.31	78.45	67.66	**77.98**	73.79
VaSCL$_{distil}$	**67.68**	**80.61**	**72.19**	**80.92**	**78.59**	**68.81**	77.32	**75.16**
RoBERTa$_{base}$	53.95	47.42	55.87	64.73	63.55	62.94	58.40	58.12
SimCSE$_{base}$	68.88	80.46	73.54	80.98	80.68	**69.54**	80.29	76.34
VaSCL$_{base}$	**69.02**	**82.38**	**73.93**	**82.54**	**80.96**	69.40	**80.52**	**76.96**
RoBERTa$_{large}$	55.00	50.14	54.87	62.14	62.99	58.93	54.56	56.95
SimCSE$_{large}$	69.83	81.29	74.42	83.77	79.79	68.89	80.66	76.95
VaSCL$_{large}$	**73.36**	**83.55**	**77.16**	**83.25**	**80.66**	**72.96**	**82.36**	**79.04**

Spearman rank correlation between the cosine similarity of sentence representation pairs and the ground truth similarity scores

- https://huggingface.co/datasets/ag_news
- https://www.kaggle.com/datasets/nishanthsalian/genia-biomedical-event-dataset?select=dev_data.csv

Short text clustering

	Ag News	Search Snippets	Stack Overflow	Bio-medical	Tweet	Google News	Avg
RoBERTa$_{distil}$	59.32	33.18	14.16	24.69	37.10	58.05	37.75
SimCSE$_{distil}$	**73.33**	60.74	66.97	35.69	50.68	67.55	59.16
VaSCL$_{distil}$	71.71	**62.76**	**73.98**	**38.82**	**51.35**	**67.66**	**61.05**
RoBERTa$_{base}$	66.50	30.83	15.63	26.98	37.80	58.51	39.38
SimCSE$_{base}$	65.53	**55.97**	64.18	38.12	49.16	65.69	56.44
VaSCL$_{base}$	**68.33**	47.26	**76.15**	**39.53**	**51.50**	**67.10**	**58.31**
RoBERTa$_{large}$	**69.35**	53.00	27.89	33.25	46.08	64.04	48.93
SimCSE$_{large}$	62.93	51.55	54.11	35.39	50.92	**67.86**	53.79
VaSCL$_{large}$	66.09	**61.57**	**69.04**	**42.91**	**56.74**	67.75	**60.68**

Clustering accuracy

WDel : random word deletion
WNet : synonym substitute
CTxt : contextual synonyms substitute

Explicit Data Augmentation

Outline

- Introduction
- Method
- Experiment
- Conclusion

Conclusion

1. This paper propose a virtual augmentationoriented contrastive learning framework
2. Constructing the neighborhoods of each training instance, which can, in turn, be leveraged to generate effective data augmentations