Supporting Clustering with Contrastive Learning

Advisor: Jia-Ling, Koh

Speaker: Zi-Xin Chen

Source: NAACL'22

Date: 2023/04/25

Outline

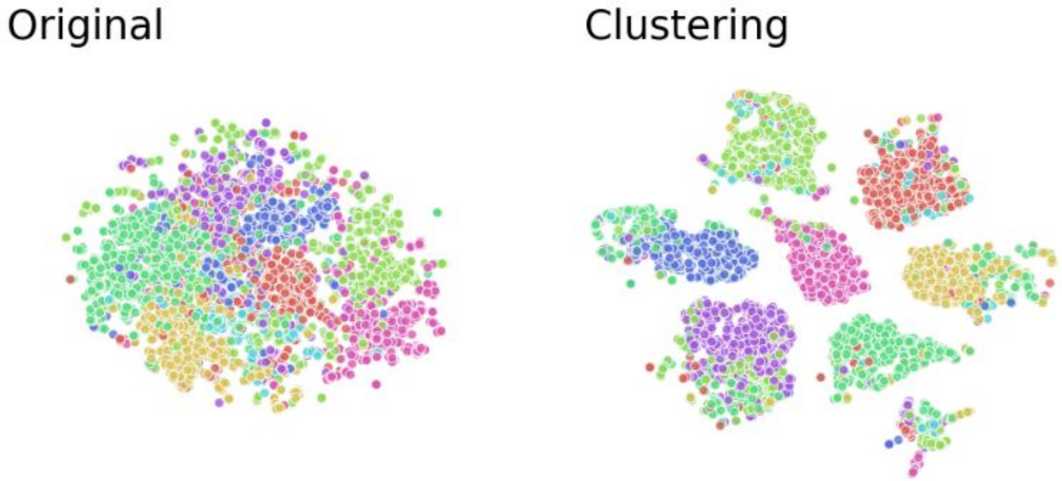
- **Introduction**
- Method
- Experiment
- Conclusion

Deep clustering

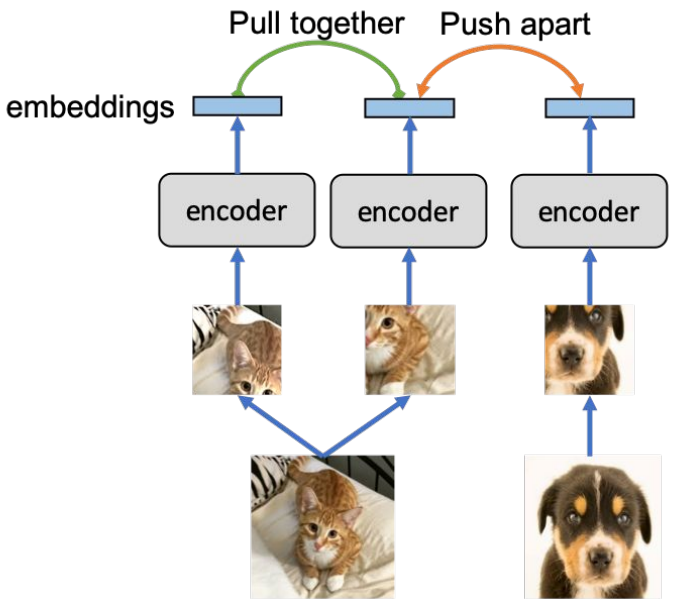
By optimizing a clustering objective function to learn better data representation for clustering.

Problem

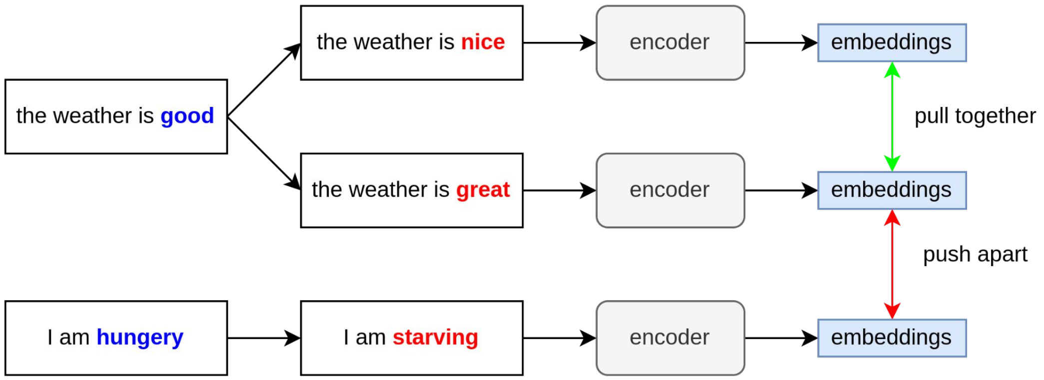
Even with deep neural networks, data still has significant **overlap** across categories before clustering starts.



Contrastive learning



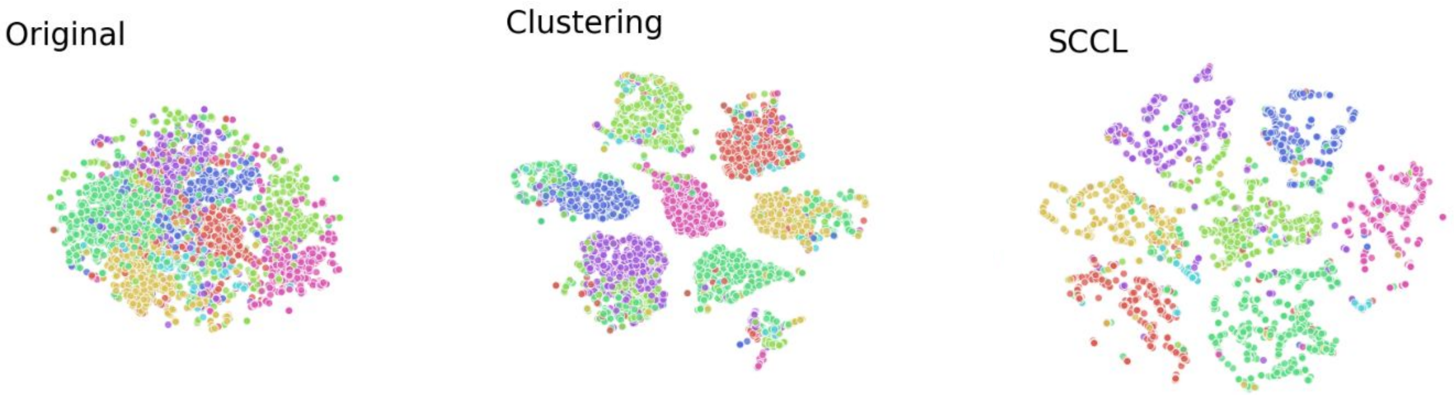
▲Contrastive learning for image



▲Contrastive learning for text

Supporting Clustering with Contrastive Learning (SCCL)

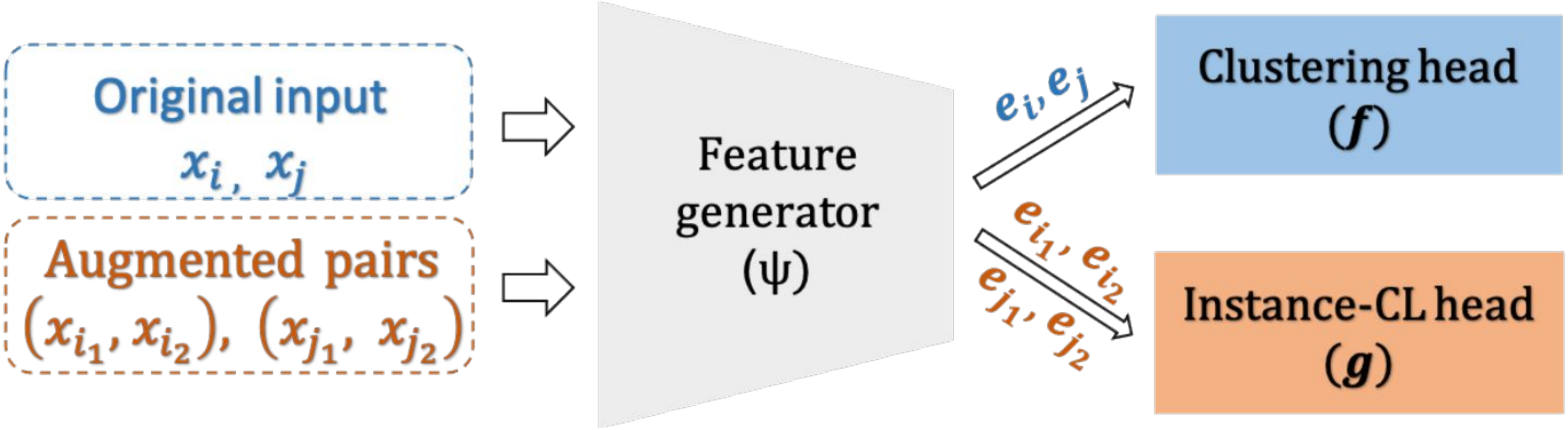
Use contrastive learning to promote better separation in clustering.



Outline

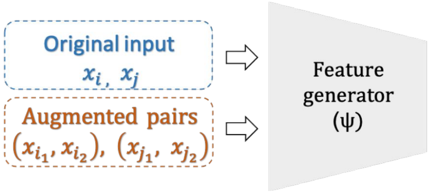
- Introduction
- **Method**
- Experiment
- Conclusion

SCCL

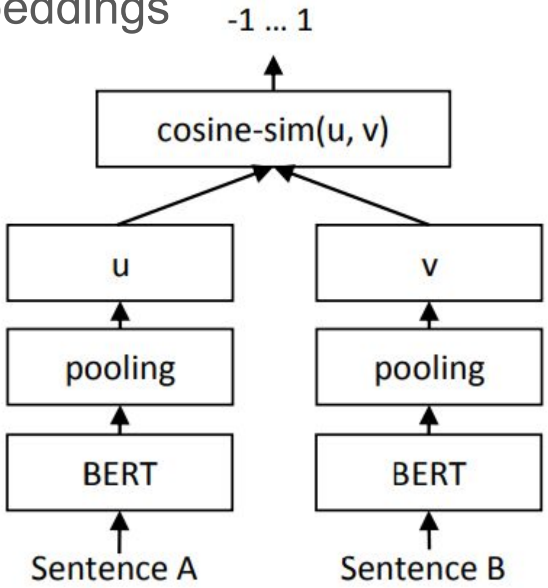


Original input
x_i, x_j
Augmented pairs
(x_{i_1}, x_{i_2}), (x_{j_1}, x_{j_2})
Feature generator
(ψ)
e_i, e_j
e_{i_1}, e_{i_2}
e_{j_1}, e_{j_2}
Clustering head
(f)
Instance-CL head
(g)

Feature Generator - Sentence-BERT



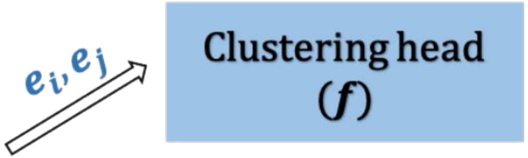
- Use siamese networks
- Use pre-trained BERT networks and only fine-tune it to yield useful sentence embeddings



Example fine-tune dataset - SNLI:

a collection of 570,000 sentence pairs annotated with the labels ***contradiction***, ***eintailment***, and ***neutral***.

Deep Clustering



Target distribuion

$$\ell_j^C = \mathrm{KL}\left[p_j \| q_j\right] = \sum_{k=1}^{K} p_{jk} \log \frac{\boxed{p_{jk}}}{\boxed{q_{jk}}}$$

Soft clustering

Deep Clustering

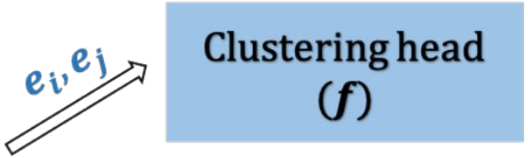
e_i, e_j → **Clustering head (f)**

$$e_j = \psi(x_j)$$

$$q_{jk} = \frac{\left(1 + \|\overset{\text{data}}{\boxed{e_j}} - \overset{\text{centroid}}{\boxed{\mu_k}}\|_2^2/\alpha\right)^{-\frac{\alpha+1}{2}}}{\sum_{k'=1}^{K}\left(1 + \|e_j - \mu_{k'}\|_2^2/\alpha\right)^{-\frac{\alpha+1}{2}}}$$

➡ The probability of data ***j*** assign to cluster ***k*** (soft clustering)

Deep Clustering



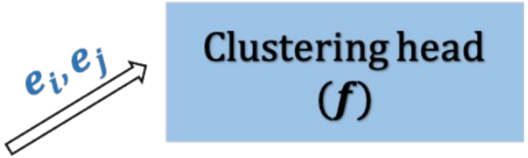
$$p_{jk} = \frac{q_{jk}^2 / f_k}{\sum_{k'} q_{jk'}^2 / f_{k'}}$$

Target distribution

$$f_k = \sum_{j=1}^{M} q_{jk}, k = 1, \ldots, K$$

Cluster frequency

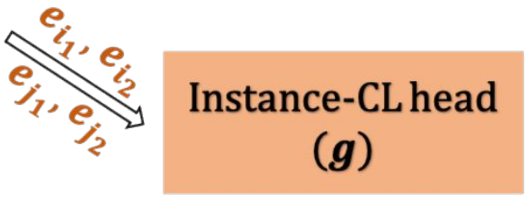
Deep Clustering



$$\ell_j^C = \mathrm{KL}\left[p_j \| q_j\right] = \sum_{k=1}^{K} p_{jk} \log \frac{p_{jk}}{q_{jk}}$$

$$\mathcal{L}_{\text{Cluster}} = \sum_{j=1}^{M} \ell_j^C / M$$

Instance-wise Contrastive Learning



Randomly sampled minibatch: $\mathcal{B} = \{x_i\}_{i=1}^{M}$

Pairs of augmentations for each data instance in ***B***: $\mathcal{B}^a = \{\tilde{x}_i\}_{i=1}^{2M}$

Positive pairs: $\tilde{x}_{i^1}, \tilde{x}_{i^2} \in \mathcal{B}^a$

Negative pairs: other 2M-2 examples in ***B^a***

Instance-wise Contrastive Learning

$$e_{i_1}, e_{i_2},\ e_{j_1}, e_{j_2} \longrightarrow \boxed{\text{Instance-CL head } (g)}$$

$$\tilde{z}_j = g(\psi(\tilde{x}_j)), j = i^1, i^2$$

$$\ell_{i^1}^I = -\log \frac{\boxed{\exp(\mathrm{sim}(\tilde{z}_{i^1}, \tilde{z}_{i^2})/\tau)}}{\boxed{\sum_{j=1}^{2M} \mathbb{1}_{j \neq i^1} \cdot \exp(\mathrm{sim}(\tilde{z}_{i^1}, \tilde{z}_j)/\tau)}}$$

$$\mathcal{L}_{\text{Instance-CL}} = \sum_{i=1}^{2M} \ell_i^I / 2M$$

Objective Function

$$\mathcal{L} = \mathcal{L}_{\text{Instance-CL}} + \eta \mathcal{L}_{\text{Cluster}}$$

$$= \sum_{j=1}^{M} \ell_j^C / M + \eta \sum_{i=1}^{2M} \ell_i^I / 2M$$

Outline

- Introduction
- Method
- **Experiment**
- Conclusion

Dataset

Dataset	$\|V\|$	Documents		Clusters	
		N^D	Len	N^C	L/S
AgNews	21K	8000	23	4	1
StackOverflow	15K	20000	8	20	1
Biomedical	19K	20000	13	20	1
SearchSnippets	31K	12340	18	8	7
GooglenewsTS	20K	11109	28	152	143
GooglenewsS	18K	11109	22	152	143
GooglenewsT	8K	11109	6	152	143
Tweet	5K	2472	8	89	249

Evaluation Metric

- Accuracy for clustering

Permutes clustering labels to match the ground truth labels

$$\mathrm{ACC} = \max_{m} \frac{\sum_{i=1}^{n} \mathbf{1}\{l_i = \boxed{m}(c_i)\}}{n}$$

- NMI

Entropy before clustering

Entropy after clustering

$$I(C,T) = \boxed{H(T)} - \boxed{H(T|C)}$$

$$\mathrm{NMI} = \frac{I(C,T)}{\sqrt{H(T) \cdot H(C)}}$$

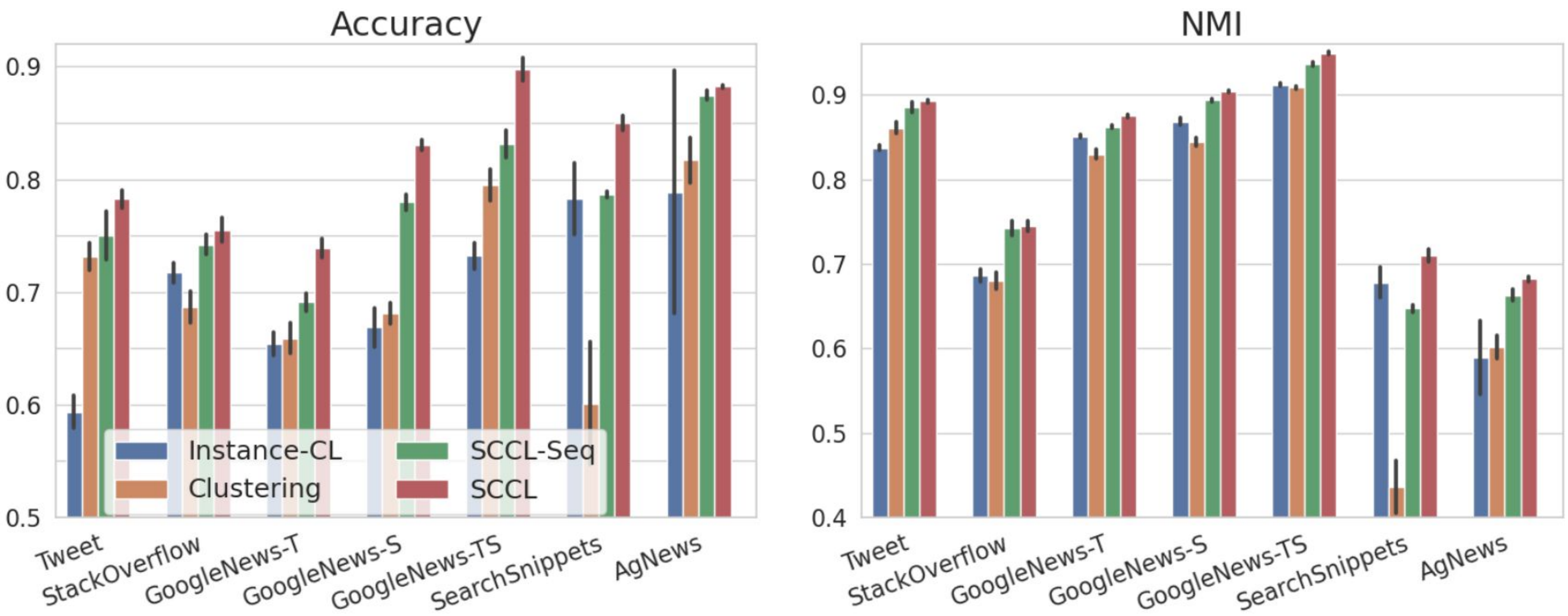
Comparison

	AgNews		SearchSnippets		StackOverflow		Biomedical	
	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI
BoW	27.6	2.6	24.3	9.3	18.5	14.0	14.3	9.2
TF-IDF	34.5	11.9	31.5	19.2	58.4	58.7	28.3	23.2
STCC	-	-	77.0	63.2	51.1	49.0	43.6	38.1
Self-Train	-	-	77.1	56.7	59.8	54.8	**54.8**	**47.1**
HAC-SD	81.8	54.6	82.7	63.8	64.8	59.5	40.1	33.5
SCCL	**88.2**	**68.2**	**85.2**	**71.1**	**75.5**	**74.5**	46.2	41.5

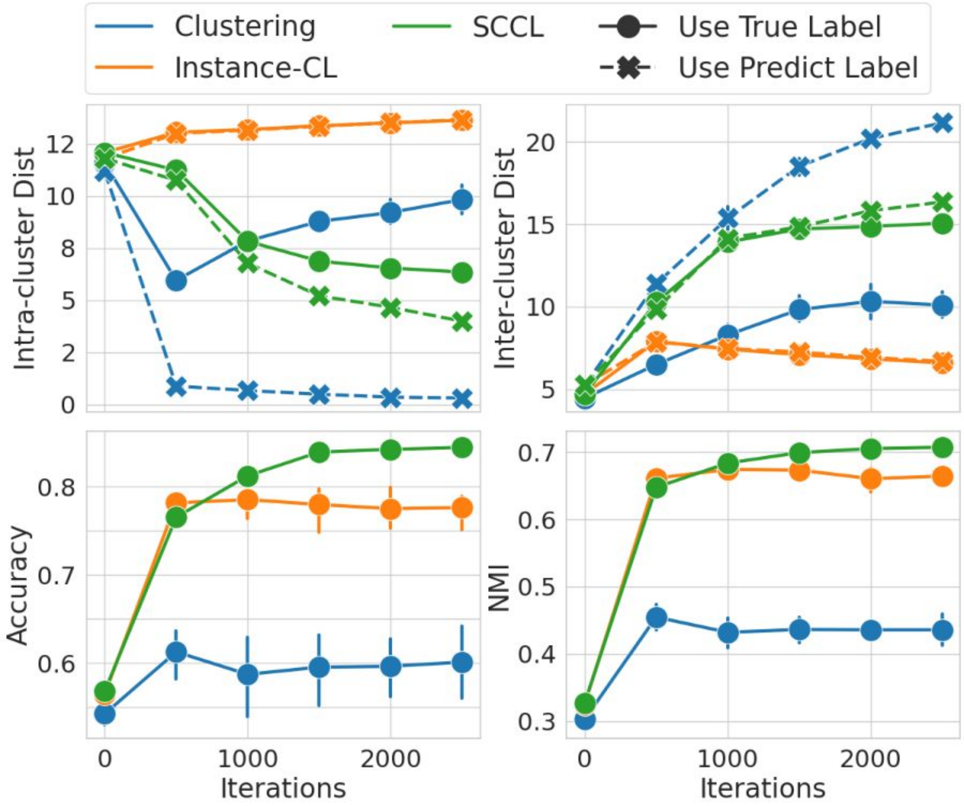
Comparison

	GoogleNews-TS		GoogleNews-T		GoogleNews-S		Tweet	
	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI
BoW	57.5	81.9	49.8	73.2	49.0	73.5	49.7	73.6
TF-IDF	68.0	88.9	58.9	79.3	61.9	83.0	57.0	80.7
STCC	-	-	-	-	-	-	-	-
Self-Train	-	-	-	-	-	-	-	-
HAC-SD	85.8	88.0	**81.8**	84.2	80.6	83.5	**89.6**	85.2
SCCL	**89.8**	**94.9**	75.8	**88.3**	**83.1**	**90.4**	78.2	**89.2**

Ablation Study



Intra-cluster and Inter-cluster Distance



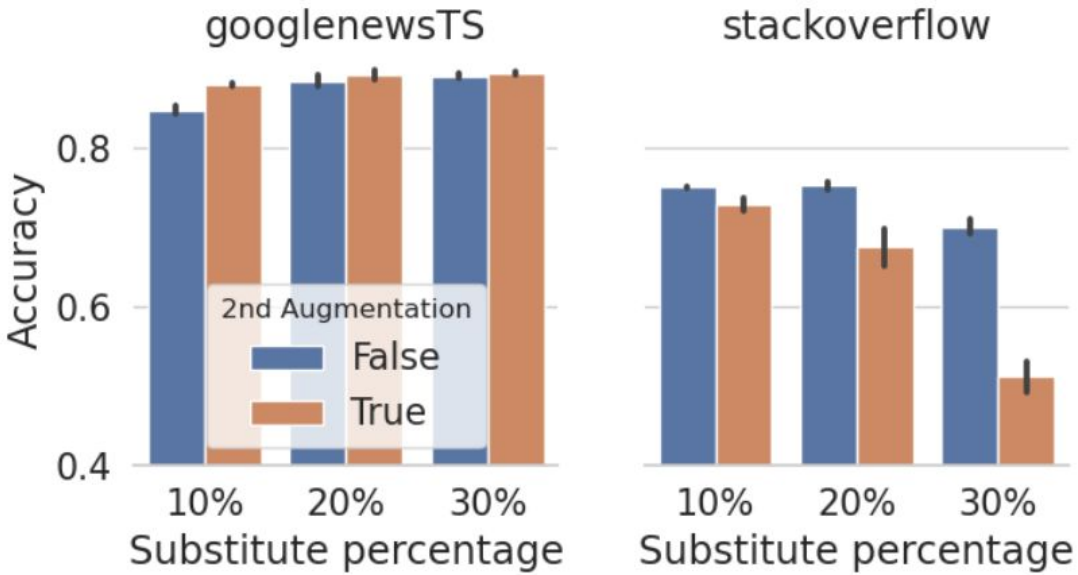
Text Data Augmentations

WNet	Replace text with WordNet synonyms
Ctxt	Use transformer to find top-n words for substitution
Para	Translate text to another language then translate back to English

Dataset	**Accuracy**			**NMI**		
	WNet	**Para**	**Ctxt**	**WNet**	**Para**	**Ctxt**
AgNews	86.6	86.5	**88.2**	66.0	65.2	**68.2**
SearchSnippets	78.1	83.7	**85.0**	61.9	68.1	**71.0**
StackOverflow	69.1	73.3	**75.5**	69.9	72.7	**74.5**
Biomedical	42.8	43.0	**46.2**	38.0	39.5	**41.5**
GooglenewsTS	82.1	83.5	**89.8**	92.1	92.9	**94.9**
GooglenewsS	73.0	75.3	**83.1**	86.4	87.4	**90.4**
GooglenewsT	66.3	67.5	**73.9**	83.4	83.6	**87.5**
Tweet	70.6	73.7	**78.2**	86.2	86.4	**89.2**

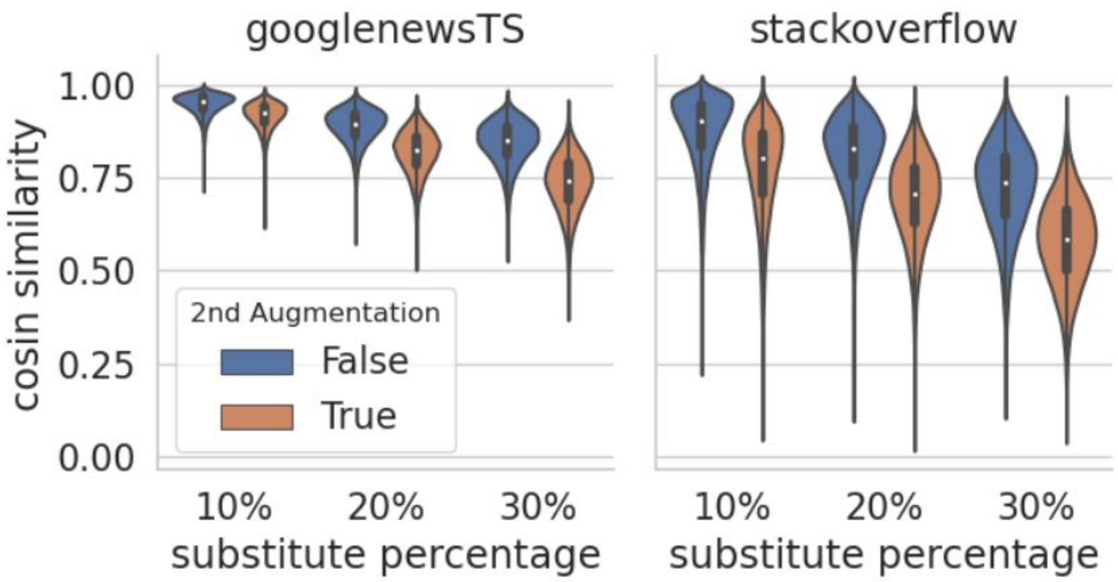
Composition of Data Augmentations

2nd Augmentation: Contextual + CharSwap



Composition of Data Augmentations

Similarity between original text and augmented text (2nd)



Outline

- Introduction
- Method
- Experiment
- **Conclusion**

Conclusion

- SCCL outperforms or performs highly comparably to the state-of-the-art methods.
- SCLL is capable of generating high-quality clusters by integrating the deep clustering and contrastive learning.