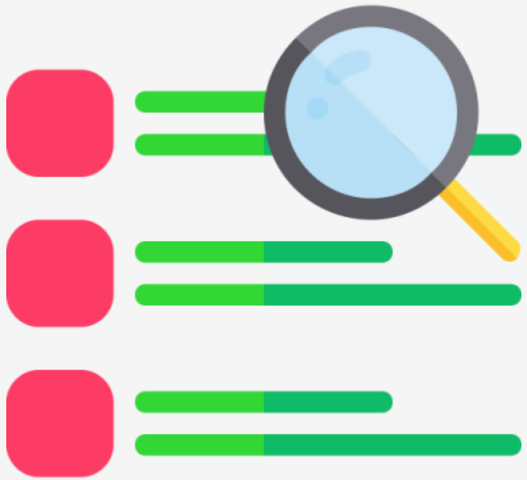
2022 한국소프트웨어종합학술대회 (KSC2022)

대조 학습에 기반한 한국어 단어 군집화

윤준호*, 나승훈

hoho5702@jbnu.ac.kr
nash@jbnu.ac.kr



목차

01

개요

군집화란?

레이블 없이 유사한 데이터들을 그룹화하는 비지도 학습 태스크

대표적인 알고리즘: GMM, K-Means, DBSCAN

고차원 임베딩의 군집화 ⟹ 높은 계산 복잡도, 낮은 성능

이를 위한 임베딩 차원 축소 알고리즘

1. 행렬 인수 분해 계열: PCA
2. 인접 그래프 계열: t-SNE, UMAP

UMAP – 고차원 공간에서의 데이터를 그래프로 만들고 저차원으로 그래프 투영
장점 – 빠른 속도, 임베딩 차원 크기에 대한 제한 X

대조 학습이란?

비슷한 데이터들의 representation 은 가깝게,

비슷하지 않은 데이터들은 멀어지도록 학습

결과적으로 잘 분포된 임베딩 공간을 형성

출처: https://ai.googleblog.com/2020/04/advancing-self-supervised-and-semi.html

최근 군집화와 심층 표현 학습을 결합하여 임베딩 공간안에서 최적화된 군집화

심층 신경망을 이용함에도 낮은 순도의 데이터를 이용할 경우 군집화가 어려움
(군집들 간에 비슷한 데이터를 가지는 경우)

Supporting Clustering with Contrastive Learning: SCCL

군집화 목적함수와 인스턴스별 대조 학습 목적 함수를 함께 최적화

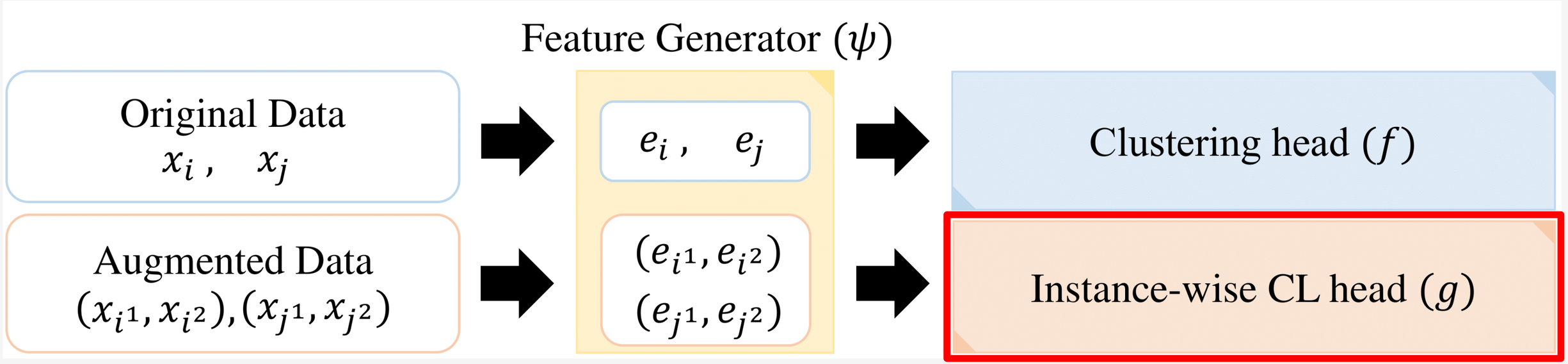
⇨ 짧은 문장 군집화 태스크에서 성능 개선

+ 한국어 데이터셋 적용
+ 효과적인 데이터 증강 기법 확인

02

방법

: SCCL

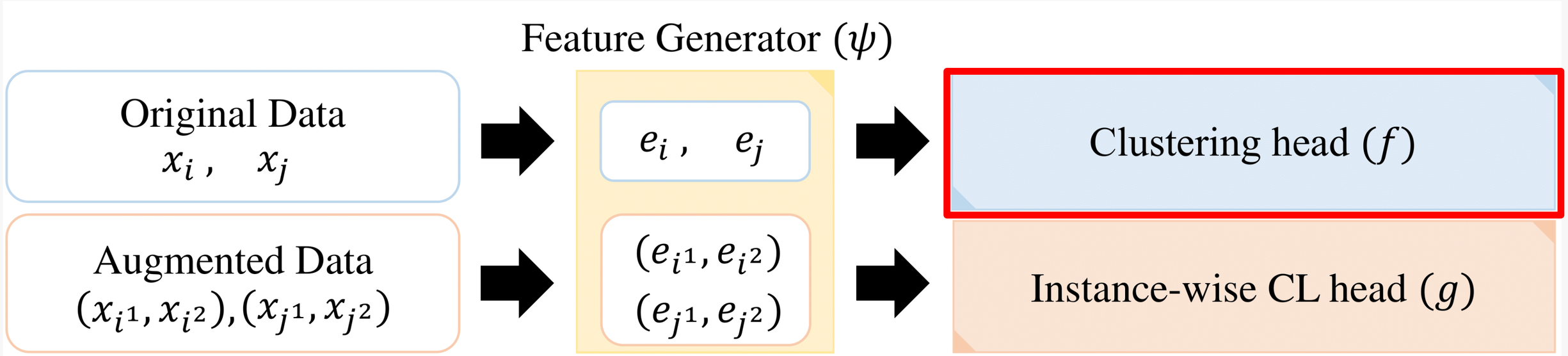


입력 배치 내의 각 인스턴스에 대해 무작위로 쌍을 증강시켜 증강배치 생성

동일한 데이터에서 증강된 쌍을 positive 쌍으로, 그 외는 negative 쌍으로 간주

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

유사도 sim 의 경우 두 데이터의 내적을 이용, 최종 $\mathcal{L}_{\mathrm{Instance-CL}} = \sum_{i=1}^{2M} \ell_i^I / 2M$



각 군집의 중심 점을 μ_k

배치 내의 인스턴스 x_j 의 임베딩 $e_j = \psi(x_j)$ 라고 할 때, x_j 가 k 번째 군집에 속할 확률

스튜던트 t분포 $$q_{jk} = \frac{(1 + \| e_j - \mu_k \|_2^2 / \alpha)^{-\frac{\alpha+1}{2}}}{\sum_{k'=1}^{K} (1 + \| e_j - \mu_{k'} \|_2^2 / \alpha)^{-\frac{\alpha+1}{2}}}$$

보조 분포 p_{jk} 를 활용하여 군집의 중심점들을 점진적으로 정제

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

$f_k = \sum_{j=1}^{M} q_{jk}\,, \quad k = 1, \ldots, K$ ⟹ 미니배치 내에서 근사한 약한 군집 빈도

보조 분포는 먼저 약한 군집 할당 확률 q_{jk} 를 제곱해 힘을 실어주고 관련된 군집 빈도로 정규화

그로 인해 신뢰도 높은 군집 할당으로부터의 학습을 장려하고

동시에 불균형 군집들로 인한 편향을 방지

군집 할당 확률과 보조 분포 사이의 KL 발산을 최적화하여 군집 할당 확률을 보조 분포에 가깝게

$$\ell_j^C = \mathrm{KL}\left[p_j \parallel q_j\right]$$

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

03

실험

데이터셋 - 3i4k 데이터셋

자주 사용되는 한국어 단어(서울대 음성 언어 처리 연구소에서 제공하는 말뭉치)와 짧은 발화

대조 학습을 위한 데이터 증강

KorEDA 의 임의 교환과 임의 삭제를 활용하여 명시적 데이터 증강

드롭아웃만을 노이즈로 이용하는 데이터 증강의 경우 특별한 방법 없이 하나의 문장을 사전 학습된 인코더에 두 번 통과

언어 모델

카카오브레인의 두 가지 KorNLU 데이터셋인 KorNLI 와 KorSTS 데이터셋을 모두 활용하여 멀티태스크로 학습시킨 ko-sroberta-multitask 를 기본 뼈대

768×K(군집 수) 크기의 선형 군집화 헤드 (f)

768 크기의 하나의 은닉층, 출력 벡터의 크기는 128인 대조 학습을 위한 다층 신경망 (g)

f 와 g 를 각각 이어 붙임

비교 실험

베이스라인 – 언어 모델을 통해 얻은 임베딩에 K-means 를 적용하여 얻은 결과

SCCL 모델을 이용한 결과와 비교

추가적으로 명시적 데이터 증강과 명시적 데이터 증강의 성능 비교

성능 평가

평가를 위해서는 SCCL 과 마찬가지로 정확도(ACC) 와 정규화된 상호정보(NMI) 를 사용

04

결과

표 1: 실험 결과

증강 기법	모델	ACC	NMI
-	베이스라인	30.50	19.92
	드롭아웃	**33.17**	**20.19**
임의 교환	1개 교환	32.97	19.89
	2개 교환	32.93	19.84
	3개 교환	32.95	19.94
임의 삭제	10% 확률	33.08	20.10
	20% 확률	32.71	20.03
	30% 확률	32.27	19.98

단순히 언어 모델을 통해 얻은 임베딩에 K-means 를 적용한 것보다
SCCL 을 적용한 결과들이 더 높은 정확도를 보이는 것을 알 수 있다.

명시적 데이터 증강 기법 중 가장 높은 성능을 보인 임의 삭제의 경우 점점 성능이 저하되는 경향
이는 문장들의 길이가 짧아 삭제 확률을 높일 수록 많은 의미를 상실하기 때문으로 보인다.

명시적 데이터 증강보다 드롭아웃만을 이용한 데이터 증강이 더 높은 성능을 보이는 것을 확인

05

결론

본 논문에서는 SCCL 에서 제시한 모델을 바탕으로 한국어 데이터셋에 적용해 성능을 비교
단순히 K-means 를 적용한 것보다 ACC 와 NMI 에서 각각 2.67 그리고 0.27 씩 향상

또한 드롭아웃만을 이용한 데이터 증강을 통해서 명시적 데이터 증강보다 쉽고
더 효과적인 대조 학습을 진행할 수 있음을 확인하였다.

발표 들어주셔서

감사합니다