대조학습을 활용한 새로운 의도 카테고리 발견

서승연°[1], 이근배[1,2]
포항공과대학교 인공지능대학원[1], 포항공과대학교 컴퓨터공학과[2]
{ssy319,gblee}@postech.ac.kr

Novel Intent Category Discovery using Contrastive Learning

Seungyeon Seo°[1], Gary Geunbae Lee[1,2]
Graduate School of Artificial Intelligence, Pohang University of Science and Technology[1]
Computer Science and Engineering, Pohang University of Science and Technology[2]

요 약

라벨 데이터 수집의 어려움에 따라 라벨이 없는 데이터로 학습하는 준지도학습, 비지도학습에 대한 연구가 활발하게 진행되고 있다. 본 논문에서는 그의 일환으로 Novel Intent Category Discovery(NICD) 문제를 제안하고 NICD 연구의 베이스라인이 될 모델을 소개한다. NICD 문제는 라벨이 있는 데이터와 라벨이 없는 데이터의 클래스 셋이 겹치지 않는다는 점에서 기존 준지도학습의 문제들과 차이가 있다. 제안 모델은 RoBERTa를 기반으로 두 개의 분류기를 추가하여 구성되며 라벨이 있는 데이터셋과 라벨이 없는 데이터셋에서 각각 다른 분류기를 사용하여 라벨을 예측한다. 학습방법은 2단계로 먼저 라벨이 있는 데이터셋으로 요인표현을 학습한다. 두 번째 단계에서는 교차 엔트로피, 이항 교차 엔트로피, 평균제곱오차, 지도 대조 손실함수를 NICD 문제에 맞게 변형하여 학습에 사용한다. 논문에서 제안된 모델은 라벨이 없는 데이터셋에 대해 이미지 최고성능 모델보다 24.74 더 높은 정확도를 기록했다.

주제어: 대화 시스템, 사용자 발화 의도 예측, 준 지도학습, 지도대조학습

1. 서론

딥러닝 모델의 발전으로 많은 모델들이 분류 문제에서 좋은 성능을 보였지만 이러한 모델들은 라벨이 있는 데이터가 필요하다. 많은 양의 데이터를 라벨링 하는 것은 큰 비용과 노력을 요구하기 때문에 라벨 없이 학습하도록 하는 연구가 주목받고 있다. 이러한 맥락에서 이미지 분야에서 활발히 연구되고 있는 Novel Category Discovery(NCD) 문제는 라벨이 있는 데이터와 라벨이 없는 데이터를 모두 활용하기 위해 제안되었다[1].

NCD 문제에서는 라벨이 있는 데이터셋과 라벨이 없는 데이터셋이 주어진 상황에서 각 입력에 해당하는 클래스를 예측해야 한다. 이때 라벨이 있는 데이터셋의 라벨 개수와 라벨이 없는 데이터셋의 라벨 개수는 알고 있다고 가정한다. 일반적인 준 지도학습 문제와 가장 구별되는 점은 라벨이 있는 데이터의 클래스 셋과 라벨이 없는 데이터의 클래스 셋이 겹치지 않는다는 것이다. 따라서 NCD 문제에서는 두 데이터셋이 서로소 집합 관계라는 특성과 각 데이터셋의 라벨 개수를 학습 전략에 활용할 수 있다.

기존에 제시된 방법론은 이미지 데이터에서는 올바르게 작동하지만, 시퀀스 데이터인 텍스트에서는 곧바로 적용되지 않아 추가적인 연구가 필요하다. 이 논문에서는 NCD의 가정을 사용자 의도 찾기 문제에 적용함으로써 Novel Intent Catecory Discovery(NICD) 문제를 제시한다. 더불어 데이터셋 구성의 특징을 지도 대조 학습에 적용한 손실함수를 제안하여 NICD 방법론의 베이스라인을 제안한다.

2. 관련 연구

2.1 Novel Category Discovery

NCD 문제 해결법 중 가장 기본적인 방법은 D^l에서 데이터 임베딩을 학습하고, 학습된 모델을 베이스라인으로 D^u의 군집화 알고리즘을 학습하는 2단계 방법론이다. [2]는 첫 번째 단계에서 얻은 모델을 이용해 만든 수도라벨의 KL-divergence를 이용한 학습법을 소개했다. 이 저자는 내적 곱을 이항 교차 엔트로피 손실함수에 적용하여 다중 클래스 분류 문제로 해결하는 방법도 제안했다[3]. [4]는 잠재 공간을 기반으로 군집의 중심을 정한 후 가장 가까운 군집에 인스턴스를 할당했다.

2단계 방법론과는 달리 단일 단계 방법은 자기 지도학습 방법을 사용하여 두 데이터셋을 동시에 학습한다[5]. 기본적인 이미지 데이터 증강 방법인 회전, 자르기, 대칭이동 등을 적용하여 모델을 강건화할 수 있다. [6]은 널리 쓰이는 이미지 데이터 증강 기법인 MixUp[7]을 이용해서 새로운 샘플을 만들어 학습에 사용했다. NCD 문제에서 최고 성능을 낸 논문들은 대조 손실함수에 임베딩 유사성을 수도 라벨로 활용하거나[8], 교차 엔트로피 손실함수에 다중 뷰 자가 라벨링 전략을 적용했다[9].

NCD문제를 사용자 의도 찾기에 적용한다면 시스템에 새로운 도메인을 추가해야 할 때 추가된 의도의 개수만 안다면 라벨링 작업없이 성공적으로 시스템을 확장할 수 있다.

2.2 New Intent Discovery

New Intent Discovery(NID) 문제의 목표는 학습 과정에서 보지 못했던 발화도 예측할 수 있는 모델 설계다. NID 문제는

학습 과정에서 라벨이 없는 데이터셋도 사용한다는 점이 NCD 문제와 유사하지만 데이터셋 세부 구성에서 명확하게 다르다. NCD 문제의 경우 라벨이 있는 데이터셋 D^l_{NCD}과 라벨이 없는 데이터셋 D^u_{NCD}의 라벨 분포가 겹치지 않는다. 하지만 NID 문제는 라벨이 없는 데이터셋 $D^u_{NID} = \{x_i | y_i \in \{C_l, C_u\}\}$의 라벨 중 일부가 라벨이 있는 데이터셋 $D^l_{NID} = \{(x_i, y_i) | y_i \in C_l\}$에도 존재한다.

NID 방법론에서도 NCD와 마찬가지로 데이터 임베딩을 학습하고 그 결과를 이용해 대조 학습을 수행하는 2단계 학습 방법이 널리 사용된다[10, 11].

2.3 대조 학습(Contrastive Learning)

대조 학습[12]은 자기 지도학습 방법론 중 하나로 이미지와 언어 분야에서 좋은 성과를 거뒀다[10, 13, 14]. 대조 학습을 활용하면 정답 라벨이 없는 데이터 샘플을 증강하여 배치안의 다른 뷰들과 구별하도록 모델을 학습시킴으로써 강건한 요인 표현(feature representation)을 얻을 수 있다. [15]은 정답 라벨이 있는 데이터에 대조 학습을 적용한 지도 대조 학습(Supervised Contrastive Learning) 방법을 제안했다. 같은 클래스를 가진 샘플들은 임베딩 공간에서 가깝게 위치하도록 하고, 다른 클래스의 샘플들은 멀어지도록 학습시켰다. [16, 17]은 클래스 분류 문제에서 대조학습 또는 지도대조학습을 활용한 모델을 제안했고 새로운 클래스를 찾아내는 데 효과적임을 입증했다.

3. 방법론

3.1 Novel Intent Category Discovery

NICD 문제는 NCD의 기본 구성과 같은 구성을 가진다. 라벨이 있는 데이터셋 $D^l = \{(x^l_i, y^l_i), i = 1, ..., N\}$과 라벨이 없는 데이터셋 $D^u = \{x^u_i, i = 1, ..., M\}$을 이용해야 하며 라벨이 있는 데이터셋의 클래스 라벨 $y^l_i \in \{1, ..., L\}$의 개수 L과 라벨이 없는 데이터셋의 클래스 라벨 $y^u_i \in \{1, ..., K\}$ 개수 K가 주어진다. 이때 x_i와 y_i는 각각 목적 지향 대화 시스템의 사용자 발화(e.g. Can you arrange travel for 7 people for TR1389 on Sunday?)와 해당 발화에 해당하는 의도(e.g. book train)이며 D^l의 클래스셋 C_l과 D^u의 클래스셋 C_u는 겹치지 않는다.

3.2 모델 구조

앞선 연구들에 의해 성공적이라고 증명된 2단계 학습 전략을 사용했다. 첫 번째 단계에서는 D^l에 교차 엔트로피를 이용해 좋은 요인 표현을 학습하도록 했고, 두 번째 단계에서는 두 데이터셋 D^l, D^u에 4개의 손실함수를 적용해 라벨을 예측할 수 있도록 학습했다.

BERT 기반의 사전학습 언어모델인 RoBERTa[18]에 2개의 분류기를 추가하여 모델을 구성했다. 그림 1에서 볼 수 있듯이

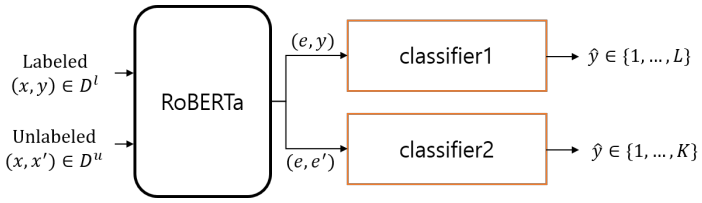

그림 1. 모델 구조

D^l과 D^u는 각각 다른 분류기(classifier)를 사용하며 분류기는 각 데이터셋의 라벨집합에서 입력 토큰의 라벨을 예측한다. 각 발화 x_i를 입력으로 하여 RoBERTa에서 해당 발화의 표현 e_i을 얻은 후, 첫 번째 [CLS] 토큰을 선형 레이어로 구성된 분류기에 넣어 라벨을 예측한다. D^u는 텍스트 증강 기법으로 증강하여 사용한다.

3.3 1단계: 좋은 요인표현 모델

베이스 모델은 기본적인 텍스트 분류 모델의 학습 방법을 따라 교차 엔트로피(Cross Entropy) 손실함수를 사용해 학습했다.

$$L_{CE} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log \hat{y}_i \tag{1}$$

교차 엔트로피 손실함수는 다중 분류 문제에서 많이 활용되는 손실함수[19]로 실제 라벨 y_i의 분포와 모델이 예측한 라벨 $\hat{y}_i$ 분포 사이의 차이를 나타낸다. 2단계에서 사용되는 손실함수 중 일부는 비슷한 라벨을 가진 발화들이 비슷한 임베딩 표현을 가지고 있다는 전제가 있기 때문에 1단계 학습이 필수적이다.

3.4 2단계: 다양한 손실함수를 이용한 라벨 예측 학습

D^l은 1단계에서 사용한 교차 엔트로피와 지도 대조(Supervised Contrastive) 손실함수를 사용해서 학습했고, D^u은 이항 교차 엔트로피(Binary Cross Entropy)와 수도라벨을 이용한 지도 대조 손실함수를 활용했다. 두 데이터셋 모두에 평균제곱오차(Mean Squared Error) 손실함수를 활용해 강건한 모델을 얻을 수 있도록 했다.

3.4.1 이항 교차 엔트로피

이항 교차 엔트로피 함수는 이진 분류에 사용되는 손실함수로 두 입력이 같은 라벨을 가지는지를 예측하기 위해 고안되었다.

$$L_{BCE} = -\frac{1}{M^2}\sum_{i=1}^{M}\sum_{j=1}^{M}[s_{ij} \log \hat{y}_i^\top \hat{y}_j + (1 - s_{ij}) \log(1 - \hat{y}_i^\top \hat{y}_j)] \tag{2}$$

우리가 가진 데이터에는 라벨이 없기 때문에 두 임베딩 사이의 유사도 s_{ij}를 계산하여 이항 교차 엔트로피 손실함수에 활용했다. s_{ij}는 유클리드 공간에서 두 벡터 e_i와 e_j의 거리가 임곗값보다 작으면 1, 크면 0으로 계산했다.

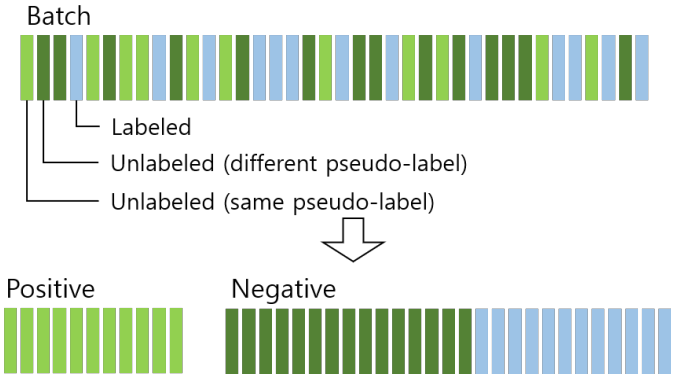

그림 2. 라벨이 없는 데이터셋의 SCL 학습 시 세트 구성

3.4.2 평균제곱오차

비슷한 발화들은 같은 라벨을 예측하는 강건성을 보장하기 위해 [5]가 사용한 손실함수를 도입했다.

$$L_{MSE} = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i^l - \hat{y}_i^{l\prime})^2 + \frac{1}{M}\sum_{i=1}^{M}(\hat{y}_i^u - \hat{y}_i^{u\prime})^2 \quad (3)$$

입력 발화 x_i와 발화를 증강하여 만든 x_i'의 각 예측 라벨 $\hat{y}_i$와 $\hat{y}_i'$가 같아지도록 학습된다.

3.4.3 지도 대조

지도 대조 학습은 같은 라벨을 가진 발화들은 가깝게, 다른 라벨을 가진 발화들은 멀게 표현되도록 임베딩을 학습한다[15]. NICD 문제에서 $C_l \cap C_u = \emptyset$이므로 e_i^u 발화의 라벨은 C_l에 반드시 존재하지 않는다. 따라서 다른 데이터셋의 데이터들은 서로 Negative 샘플로 설정할 수 있다. 그림 2에서 지도 대조 학습 시 설정되는 Positive/Negative 샘플을 확인할 수 있다. D^l에서도 같은 규칙을 적용하여 같은 라벨을 가진 데이터들은 Positive, 다른 라벨을 가진 데이터와 D^u의 데이터는 Negative로 설정한다.

$$L_{SCL} = \sum_{i=1}^{N}\mathcal{L}_i^{scl} \quad (4)$$

$$L_i^{scl} = -\frac{1}{N-1}\sum_{j=1}^{N}\mathbf{1}_{i\neq j}\cdot s_{ij}\cdot\log\frac{\exp(z_i\cdot z_j/\tau)}{\sum_{k=1}^{N}\mathbf{1}_{i\neq k}\exp(z_i\cdot z_k/\tau)} \quad (5)$$

z_i는 i번째 발화의 임베딩을 의미하고 s_{ij}는 L_{BCE}에서 사용한 거리 기반의 유사도 값을 사용했다. 배치 안에 있는 N개의 발화에 대해 L_i^{scl}을 계산하여 더해줬다.

3.4.4 전체 손실함수

$$L = L_{CE} + L_{BCE} + \omega(t)L_{MSE} + \alpha L_{SCL} \quad (6)$$

위에서 서술한 손실함수들을 결합한 손실함수 L로 1단계에서 학습한 베이스 모델을 재학습했다. $\omega(t)$는 ramp-up 함수로 시간이 지날수록 커져 평균제곱오차 손실함수의 기여를 높이고 α는 초매개변수로 실험적으로 설정했다.

4. 실험

4.1 실험환경

	Train	Dev	Test
Labeled	20962	2489	2410
Unlabeled	18940	2812	2907
Total	39902	5301	5317

표 1. 데이터 분포

Labeled	Unlabeled
find restaurant	find attraction
book restaurant	book hotel
find hotel	book train
find taxi	find train

표 2. 각 데이터셋의 의도라벨 종류

데이터셋 MultiWOZ 2.2 [20]는 대표적인 작업 지향 대화 데이터셋으로 8개 도메인의 대화로 이루어져 있다. 대화의 턴마다 사용자의 의도와 슬롯-밸류가 태깅되어 있어 대화 시스템 연구에 활발히 활용되고 있다. 이 연구의 목표는 사용자의 발화에서 의도를 추출하는 것이므로 주어진 데이터셋에서 의도가 나타난 발화만 추출하여 학습에 사용했다. 학습 데이터셋이나 평가 데이터셋에 나타나지 않은 의도를 가진 발화는 학습과 평가에서 제외했다. NICD 문제에서는 데이터셋을 라벨이 있는 데이터셋과 라벨이 없는 데이터셋으로 구분하여 학습에 활용하는데, 이미지 분야의 선행 연구를 따라 클래스 수를 5:5 비율로 데이터셋을 구성했다 2. 표 1에서 사용한 데이터의 분포를 알 수 있다. 각 발화는 영어-독일어 Back translation 기법[1]으로 증강해 사용했다.

세부구현 Hugging Face에서 제공하는 사전 훈련 모델 roberta-base[2]와 토크나이저를 활용했고 SGD 옵티마이저를 사용했다. roberta-base의 max sequence lenght는 32로 설정했으며 학습 시 learning rate, batch size, epoch는 각각 1e-6, 64, 40으로 설정했다. 손실함수에 사용되는 초매개변수 $\omega(t)$의 계수는 5.0, α는 10으로 설정해 학습했다. 발화 유사도 s_{ij} 계산 시 사용된 임곗값은 0.7이다. 학습에는 NVIDIA RTX A5000 4대를 사용했다.

[1]https://github.com/makcedward/nlpaug
[2]https://huggingface.co/roberta-base

Model	Unlabeled Dataset			Labeled Dataset		
	ACC	NMI	ARI	ACC	NMI	ARI
Baseline (Stage 1)	58.65	0.2847	**0.3399**	**78.63**	**0.4686**	**0.5319**
NCL	37.39	0.084	0.0851	73.61	0.3951	0.4443
Ours	**62.13**	**0.3524**	0.3131	77.68	0.4596	0.5221

표 3. NICD 성능비교

Model	Unlabeled Dataset			Labeled Dataset		
	ACC	NMI	ARI	ACC	NMI	ARI
NCL w/o Stage1	34.95	0.0518	0.0453	62.41	0.2636	0.2515
Ours w/o Stage1	40.80	0.0603	0.0638	32.16	0.0117	0.0094
Ours	**62.13**	**0.3524**	**0.3131**	**77.68**	**0.4596**	**0.5221**

표 4. 1단계 학습의 효과

4.2 평가 기준

클래스 예측의 주된 평가 기준으로 정확도(ACC)를 사용했다. 라벨이 없는 데이터의 경우 헝가리안 알고리즘을 이용해 예측 라벨의 모든 순열에서 가장 높은 정확도를 선택하는 클러스터 정확도를 사용하여 평가했다.

$$\text{ClusterACC} = \max_m \frac{\sum_{i=1}^{n} \mathbf{1}\{y_i = m(\hat{y}_i)\}}{n} \tag{7}$$

y_i는 정답 라벨을, $\hat{y}_i$는 예측 라벨을 의미하며 m은 예측 라벨 리스트의 순열을 나타낸다. 보조 평가 기준으로 NMI(Normalized Mutual Information)과 ARI(Adjusted Rand Index)를 사용해 클러스터링 품질 평가를 시행했다.

$$\text{NMI} = \frac{MI(y, \hat{y})}{\frac{1}{2}(H(y) + H(\hat{y}))} \tag{8}$$

NMI는 예측 라벨 분포 $\hat{y}$와 정답 라벨 분포 y의 유사도를 측정하는 척도로 엔트로피가 낮을수록 1에 가까워 좋은 성능을 나타낸다. ARI 역시 예측 클러스터와 정답 클러스터 사이의 일치를 측정하며 0은 무작위 클러스터, 1은 완벽히 일치함을 의미한다.

$$\text{ARI} = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right]/\binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right]/\binom{n}{2}} \tag{9}$$

a는 각 군집 내에서 가능한 모든 데이터 쌍이며 b는 예측과 실제 라벨이 일치하지 않는 쌍을 나타낸다. 테스트셋을 라벨이 있는 데이터와 라벨이 없는 데이터로 나눠 각각 평가해 각 분류기의 성능 변화를 측정했다.

4.3 실험 결과

성능 비교를 위해 이미지 NCD의 최고성능 모델에서 사용한 기법 NCL(Neighborhood Contrastive Learning)[8]을 적용하여 MultiWOZ 데이터를 학습시켰다. 표 3에서 볼 수 있듯이 D^u와 D^l 모두 ACC가 하락했다. 특히 D^u의 경우 NMI, ARI가 크게 하락해 클러스터링 품질이 매우 떨어졌다. NCL의 경우 이전 배치의 데이터들을 queue에 저장하고 이를 모두 Negative 샘플로 사용하는 부분이 있어 학습이 잘못된 방향으로 흘러갈 가능성이 크다. 반면 우리 모델은 D^u에서 베이스 모델보다 ACC가 3.48 향상되었다. D^l의 경우 ACC, NMI, ARI 모두 베이스 모델의 성능과 유사한 성능을 기록했다.

표 4의 w/o Stage 1은 1단계 학습을 거치지 않고 곧바로 roberta-base에 학습한 결과다. NCL과 Ours 모두 성능이 크게 떨어졌는데 이는 두 모델이 수도 라벨 계산 시 좋은 요인 표현 능력이 필요하기 때문이다. 잘못된 요인 표현은 다른 라벨 발화들의 임베딩을 비슷하게 생성해 같은 수도라벨을 가지게 할 수 있다.

Augmentation Method	Unlabeled Dataset		
	ACC	NMI	ARI
Synonym substitution	58.72	0.3272	0.2492
Random insert	54.39	0.2844	0.2166
Ours (Back translation)	**62.13**	**0.3524**	**0.3131**

표 5. 데이터 증강 방법의 영향

데이터 증강 방법의 영향을 알아보기 위해 유의어 대체(Synonym subtitution)와 랜덤 삽입(Random insert)로 증강한 데이터에 대해 2단계 학습법을 시행했다. 표5에서 두 방법 모두 역변역(Back translation)으로 증강한 데이터를 사용한 모델 Ours보다 각각 3.41, 7.74 낮은 ACC를 기록했으며 특히 랜덤 삽입의 경우 베이스라인보다도 낮은 ACC를 보임을 확인할 수

있다. 따라서 제안된 방법론이 증강데이터의 품질에 큰 영향을 받음을 알 수 있다.

5. 결론

본 논문에서는 사용자 의도 예측 분야의 라벨 데이터 부족 문제를 해결하기 위해 Novel Intent Category Discovery 문제를 제안하고 그 연구의 기반이 될 수 있는 베이스 모델을 제시했다. NICD 문제는 라벨셋이 서로 서로소 관계인 라벨이 있는 데이터셋과 라벨이 없는 데이터셋을 이용해 라벨을 예측하는 문제다. 모델은 총 2단계의 학습 과정을 거쳤으며 첫 번째 단계에서는 라벨이 있는 데이터를 사용했고 두 번째 단계에서는 라벨이 없는 데이터도 함께 사용해서 학습했다. 학습 시에는 분류 문제의 기본 손실함수인 교차 엔트로피, 이항 교차 엔트로피 함수를 각 데이터에 적용했고 더불어 지도 대조 손실함수와 평균 제곱 오차 손실함수로 강건한 표현을 학습했다. 4가지의 손실함수로 학습된 모델은 MultiWOZ 2.2에 대해 NCD 최고 성능 모델보다 좋은 성능을 보였다.

감사의 글

본 연구는 문화체육관광부 및 한국콘텐츠진흥원의 2021년 문화콘텐츠 R&D 전문인력 양성(문화기술 선도 대학원) 사업의 연구결과로 수행되었음(인공지능 및 증가상현실 기반 콘텐츠 메타버스 구축을 통한 R&D 전문인력 양성, R2021040136)
본 연구는 과학기술정보통신부 및 정보통신기획평가원의 대학ICT연구센터육성지원사업의 연구결과로 수행되었음(IITP-2023-2020-0-01789)

참고문헌

[1] K. Han, A. Vedaldi, and A. Zisserman, “Learning to discover novel visual categories via deep transfer clustering,” *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8401–8409, 2019.

[2] Y.-C. Hsu, Z. Lv, and Z. Kira, “Learning to cluster in order to transfer across domains and tasks,” *arXiv preprint arXiv:1711.10125*, 2017.

[3] Y.-C. Hsu, Z. Lv, J. Schlosser, P. Odom, and Z. Kira, “Multi-class classification without multi-class labels,” *arXiv preprint arXiv:1901.00544*, 2019.

[4] K. Han, A. Vedaldi, and A. Zisserman, “Learning to discover novel visual categories via deep transfer clustering,” *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8401–8409, 2019.

[5] K. Han, S.-A. Rebuffi, S. Ehrhardt, A. Vedaldi, and A. Zisserman, “Autonovel: Automatically discovering and learning novel visual categories,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 44, No. 10, pp. 6767–6781, 2021.

[6] Z. Zhong, L. Zhu, Z. Luo, S. Li, Y. Yang, and N. Sebe, “Openmix: Reviving known knowledge for discovering novel visual categories in an open world,” *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9462–9470, 2021.

[7] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, “mixup: Beyond empirical risk minimization,” *arXiv preprint arXiv:1710.09412*, 2017.

[8] Z. Zhong, E. Fini, S. Roy, Z. Luo, E. Ricci, and N. Sebe, “Neighborhood contrastive learning for novel class discovery,” *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10 867–10 875, 2021.

[9] E. Fini, E. Sangineto, S. Lathuiliere, Z. Zhong, M. Nabi, and E. Ricci, “A unified objective for novel class discovery,” *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9284–9292, 2021.

[10] Y. Zhang, H. Zhang, L.-M. Zhan, X.-M. Wu, and A. Lam, “New intent discovery with pre-training and contrastive learning,” *arXiv preprint arXiv:2205.12914*, 2022.

[11] H. Zhang, H. Xu, X. Wang, F. Long, and K. Gao, “Usnid: A framework for unsupervised and semi-supervised new intent discovery,” *arXiv preprint arXiv:2304.07699*, 2023.

[12] R. Hadsell, S. Chopra, and Y. LeCun, “Dimensionality reduction by learning an invariant mapping,” *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR’06)*, Vol. 2, pp. 1735–1742, 2006.

[13] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” *International conference on machine learning*, pp. 1597–1607, 2020.

[14] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, “Unsupervised feature learning via non-parametric instance discrimination,” *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3733–3742, 2018.

[15] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, “Super-

vised contrastive learning," *Advances in neural information processing systems*, Vol. 33, pp. 18 661–18 673, 2020.

[16] J. Lee, S. Seo, Y. Kim, and G. G. Lee, "Doric: Domain robust fine-tuning for open intent clustering through dependency parsing," *arXiv preprint arXiv:2303.09827*, 2023.

[17] P. Wang, K. Han, X.-S. Wei, L. Zhang, and L. Wang, "Contrastive learning based hybrid networks for long-tailed image classification," *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 943–952, 2021.

[18] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[19] J. Liu, W.-C. Chang, Y. Wu, and Y. Yang, "Deep learning for extreme multi-label text classification," *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval*, pp. 115–124, 2017.

[20] X. Zang, A. Rastogi, S. Sunkara, R. Gupta, J. Zhang, and J. Chen, "Multiwoz 2.2: A dialogue dataset with additional annotation corrections and state tracking baselines," *arXiv preprint arXiv:2007.12720*, 2020.