지능 증강 대상 탐색을 위한
멀티모달 데이터간 유사도 탐색

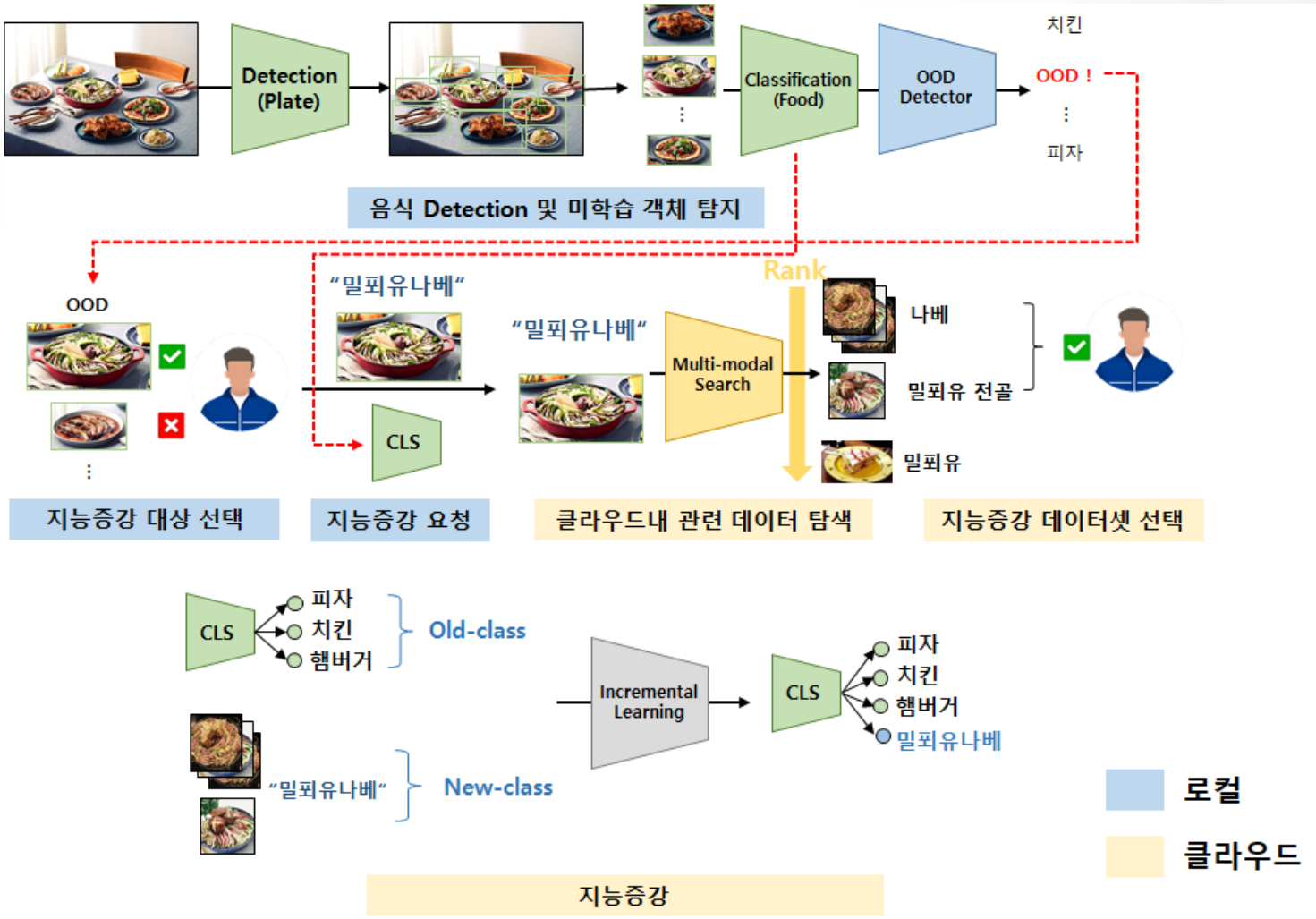
2021. 06. 24.

광주과학기술원 (GIST)

인공지능 연구실

발표자: 김종원 (통합과정, 지도교수: 이규빈)

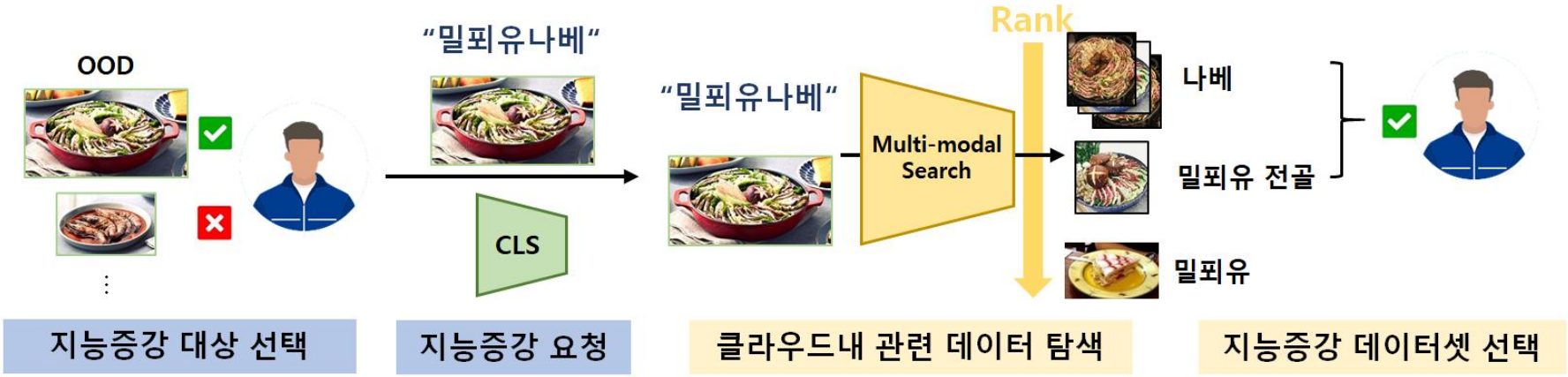
지능 증강 시스템 전체 개요



지능 증강 시스템 전체 개요

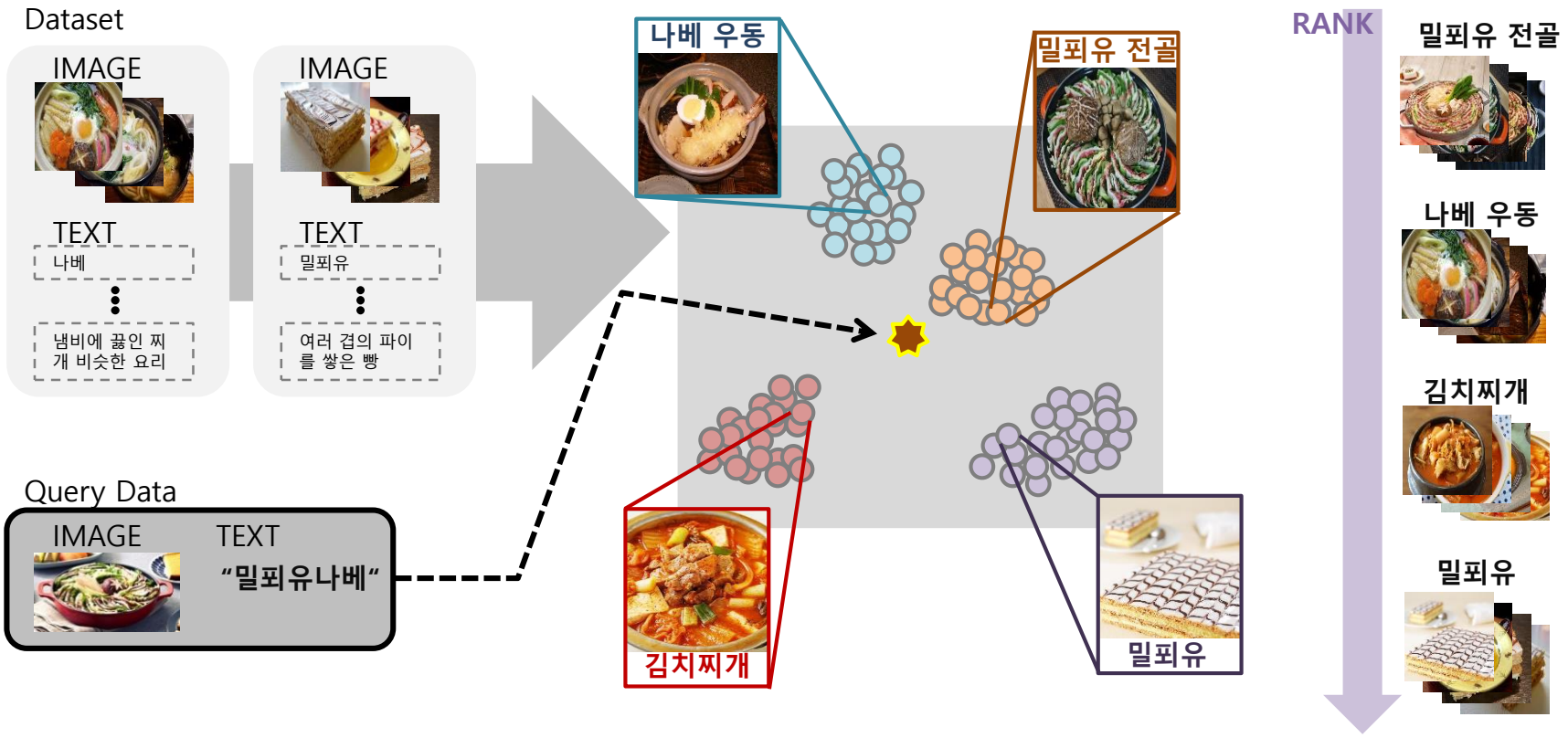
연구 목표

- 멀티모달 데이터를 활용한 지능 증강 데이터 탐색 알고리즘 개발
  - 사용자로부터의 지능증강요청을 처리하기 위해서는 적합한 데이터 탐색필요
  - 적합한 데이터를 탐색하여 증강된 지능을 사용자에게 제공



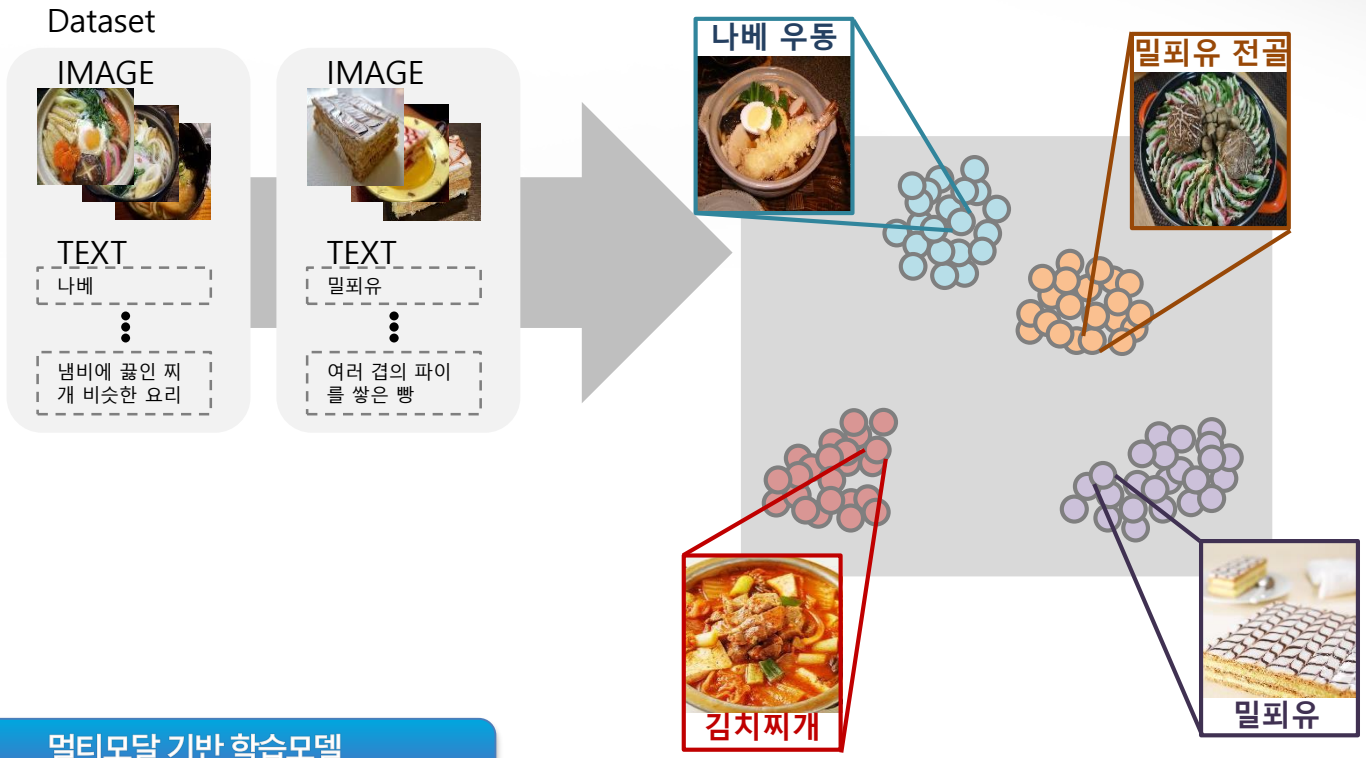
연구 내용

- 멀티모달 데이터를 활용한 지능 증강 데이터 탐색 알고리즘 개발
  - 멀티모달 데이터를 입력으로 받아 fusion feature를 생성
  - 생성된 fusion feature를 유사성을 고려하여 embedding
  - Embedding된 feature를 거리 기반으로 유사한 데이터를 탐색



연구 내용

- **Multi-modal fusion representation**

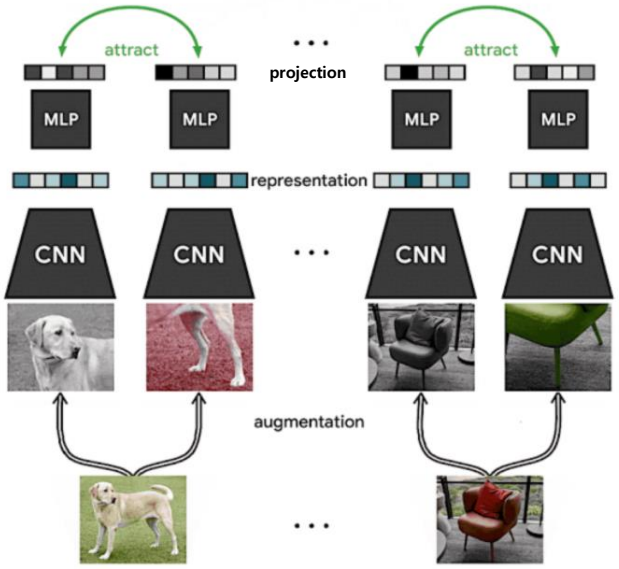


멀티모달 기반 학습모델

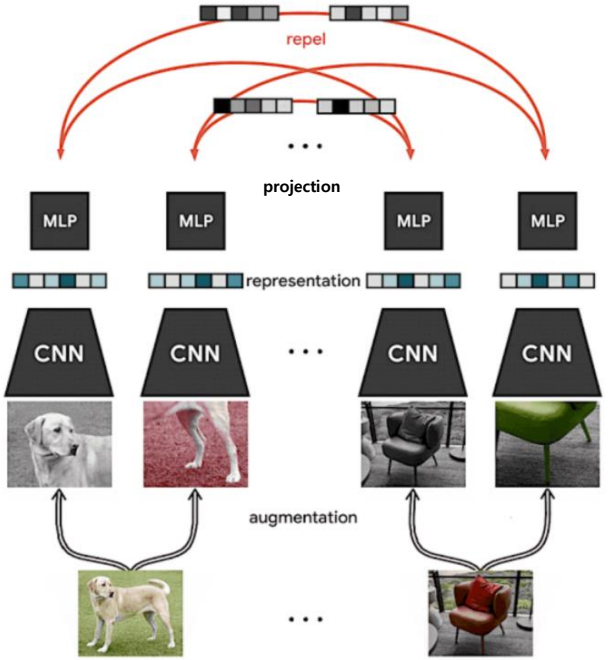
- 멀티모달 모델은 데이터를 embedding space 내에서 각 데이터 간의 유사성을 고려하여 mapping 하는 역할
- 멀티 모달 데이터를 기반으로 유사도 기반의 손실 함수를 사용하여 학습을 한다.

참고 문헌

- **A Simple Framework for Contrastive Learning of Visual Representations (SimCLR), 2020, Google**



Training scheme for positive samples



Training scheme for negative samples

상세 사항

- 같은 이미지에서 얻어진 Positive sample 쌍을 입력 받은 모델의 projection이 유사하게 생성되도록 훈련
- 다른 이미지에서 얻어진 Negative sample 쌍을 입력 받은 모델의 projection이 달라지게 생성되도록 훈련
- Contrastive learning에 적합한 multi-view를 생성하기 위한 Data augmentations 조합과 많은 양의 negative samples를 생성하기 위한 Batch size를 확인함으로써 Self-supervised contrastive learning의 표준을 정립

참고 문헌

- **Self-supervised multimodal versatile networks, 2020**

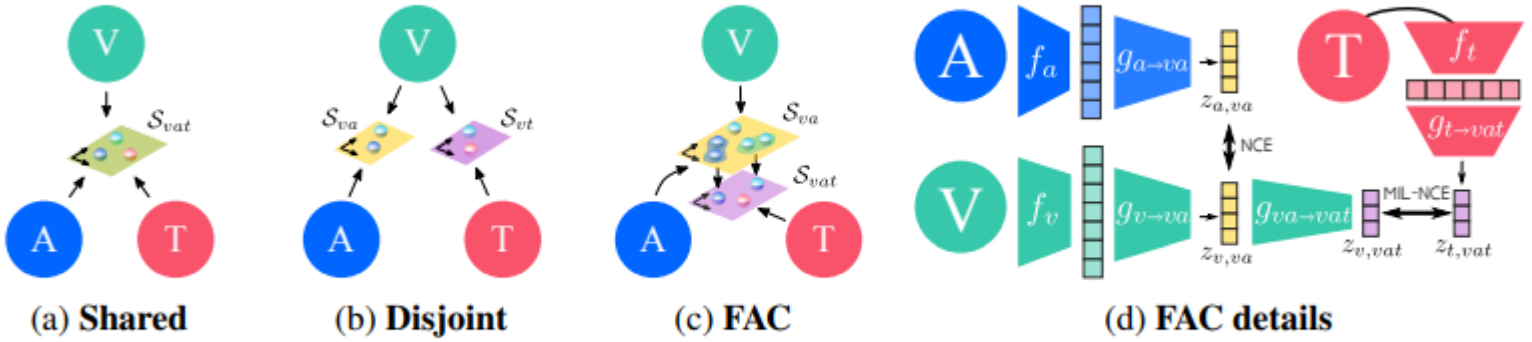


Figure 1: (a)-(c) Modality Embedding Graphs, (d) Projection heads and losses for the FAC graph. V=Vision, A=Audio, T=Text.

상세 사항

- Multi modal 데이터(영상, 음성, 텍스트)를 활용한 self-supervised learning 기법을 제시
- 학습은 HowTo100M 데이터 셋을 활용
- UCF101등의 데이터를 활용하여 평가
- 전체 3개 모달을 활용하지 않아도 동작을 하며 (최소 2개) 우수한 성능을 보여주었음 (SOTA 달성)

참고 문헌

- **Multimodal Clustering Networks for Self-supervised Learning from Unlabeled Videos, 2021**

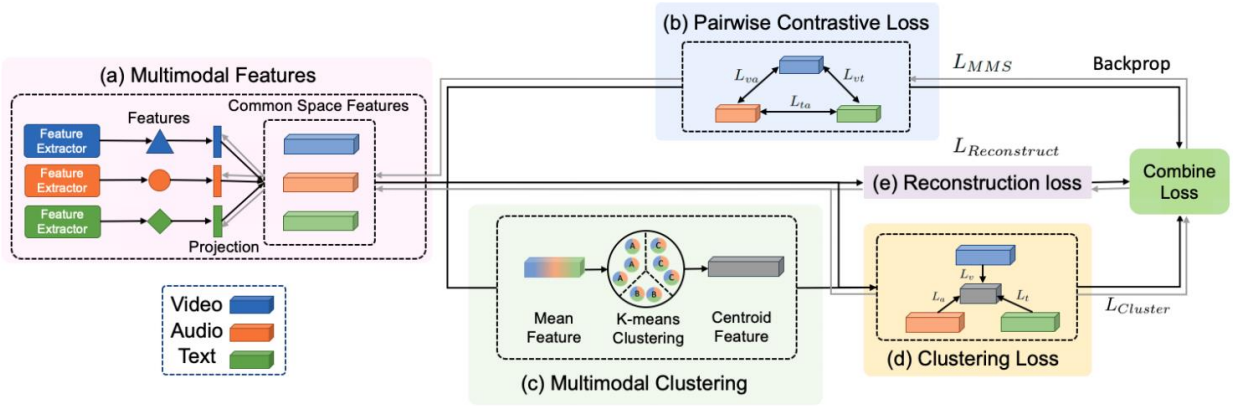


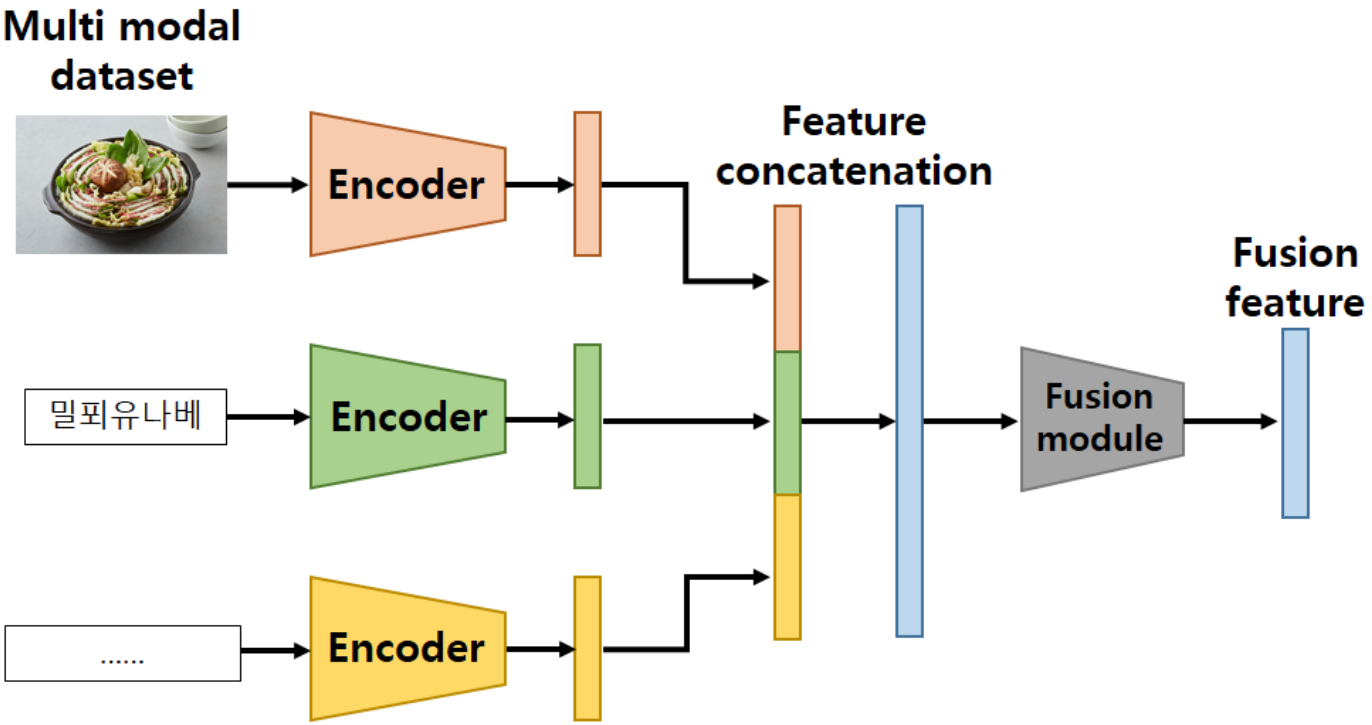
Figure 3: **Illustration of our proposed framework.** Our framework comprises four parts: (a) Extracting features from several modalities and projecting them into joint space. (b) Calculating contrastive loss pairwise to pull the features close across modalities. (c) Performing multimodal clustering across features from different domains in a batch. (d) Performing joint prediction across features to multimodal centroids to bring together semantically similar embeddings. (e) Reconstruction loss for regularization. Best viewed in color.

상세 사항

- Multi modal 데이터(영상, 음성, 텍스트)를 분석하기 위해서 "Pairwise Constrastive loss", "Reconstruction loss", "Clustering loss"를 활용하는 방법을 제시
- Pairwise Constrastive loss는 세가지 feature간에 서로 비교
- Clustering loss는 3가지 데이터의 feature의 평균을 내어 구한 "Mean feature"를 통해 얻은 "Centroid feature"와 각 원래 feature를 비교

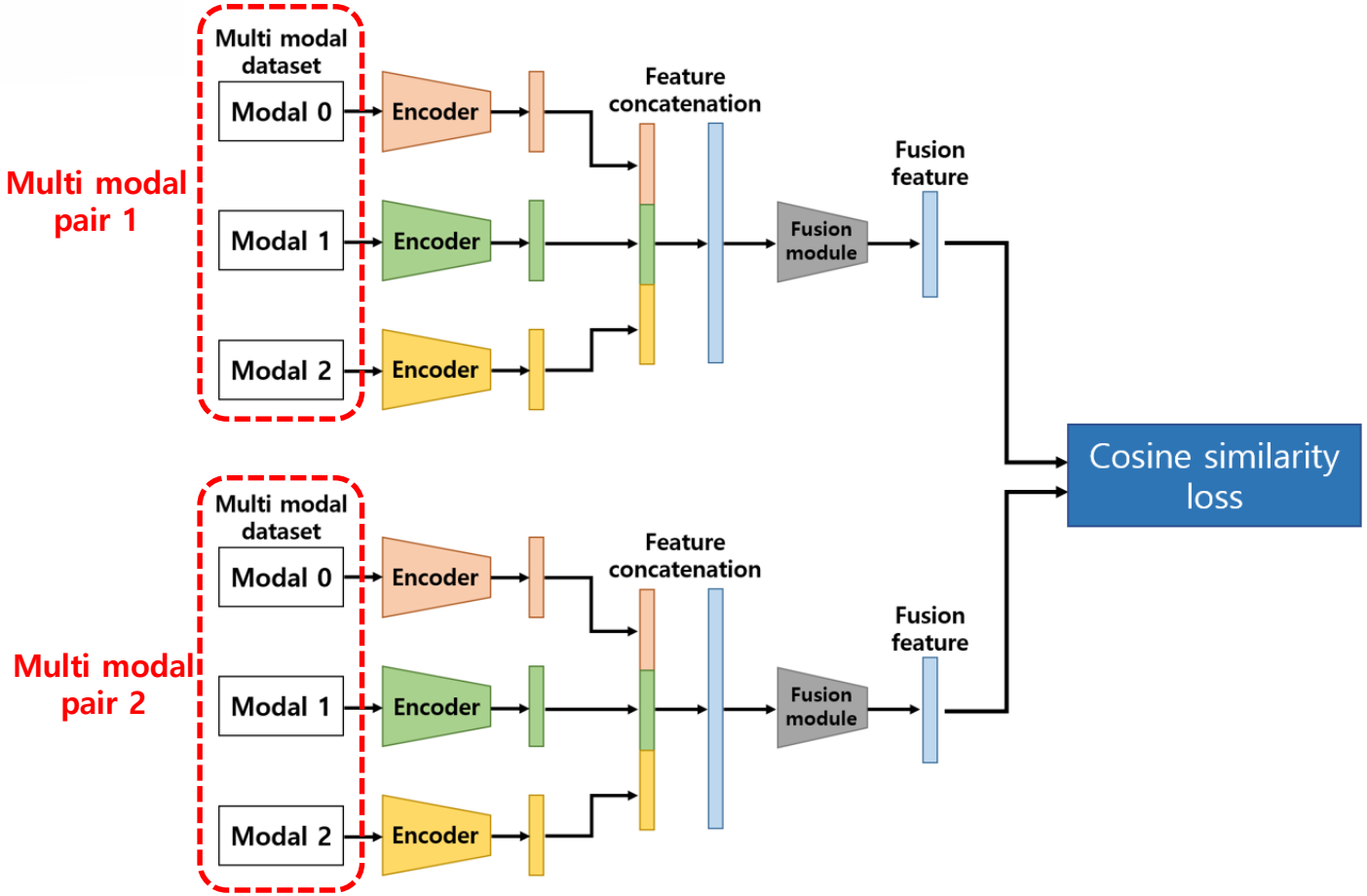
연구 내용

- Multi modal feature fusion
  - 학습 모델은 각 모달용 Encoder를 통해 modality feature를 생성
  - 각 모달별 modality feature은 fusion module을 거쳐 fusion feature로 변환



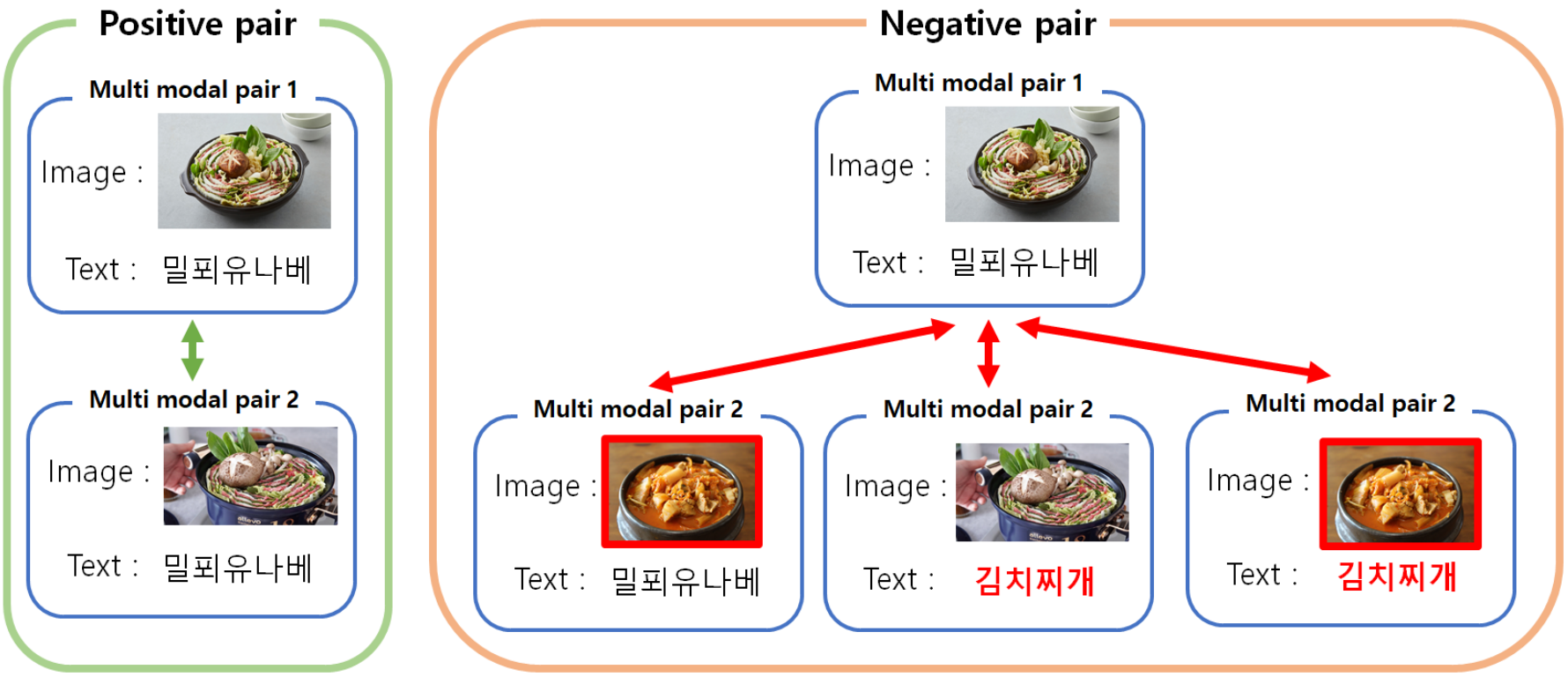
연구 내용

- Contrastive learning
  - fusion feature는 다른 fusion feature와 cosine similarity으로 비교
  - 두 feature가 유사하다면 1, 다르다면 0을 출력하도록 학습



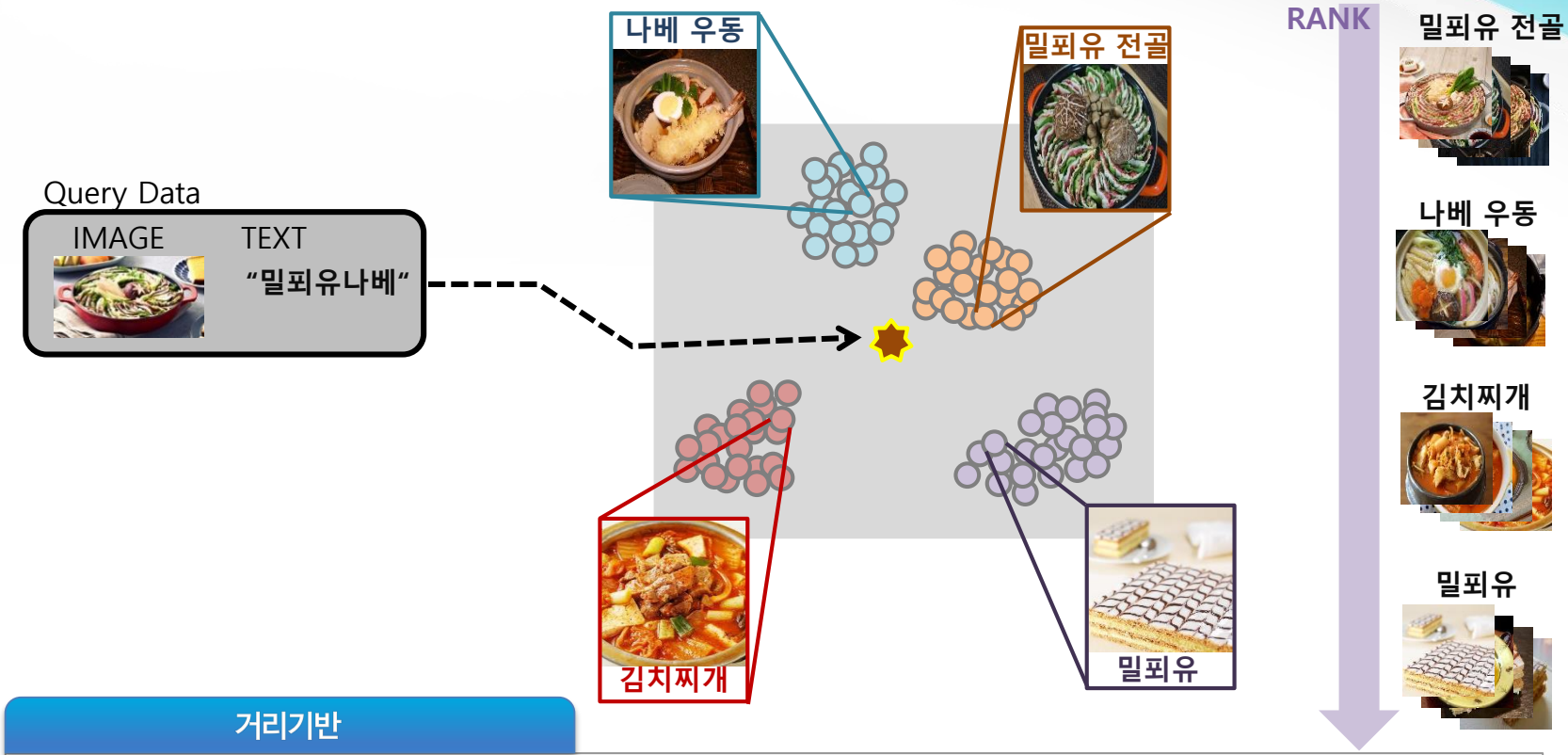
연구 내용

- Contrastive learning
  - Contrastive learning을 위한 데이터 pair를 각각 생성
  - Positive pair와 Negative pair로 나뉘어짐



연구 내용

- **Data search via feature distance**

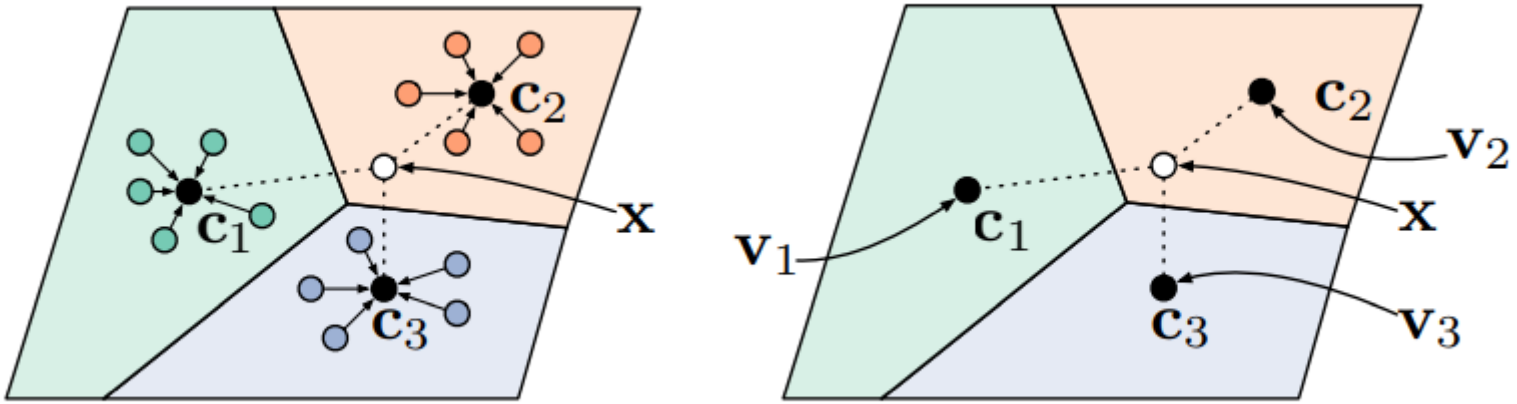


거리기반

- 멀티모달 기반 학습모델을 통해서 전체 데이터와 query data를 embedding space에 mapping
- 전체 데이터와 query data의 feature를 활용하여 거리기반으로 유사도 탐색
- query data 기준으로 가까운 순서로 각 subset을 sorting 하여 유사도 순으로 사용자에게 제공

참고 문헌

- **Prototypical Networks for Few-shot Learning, 2017**



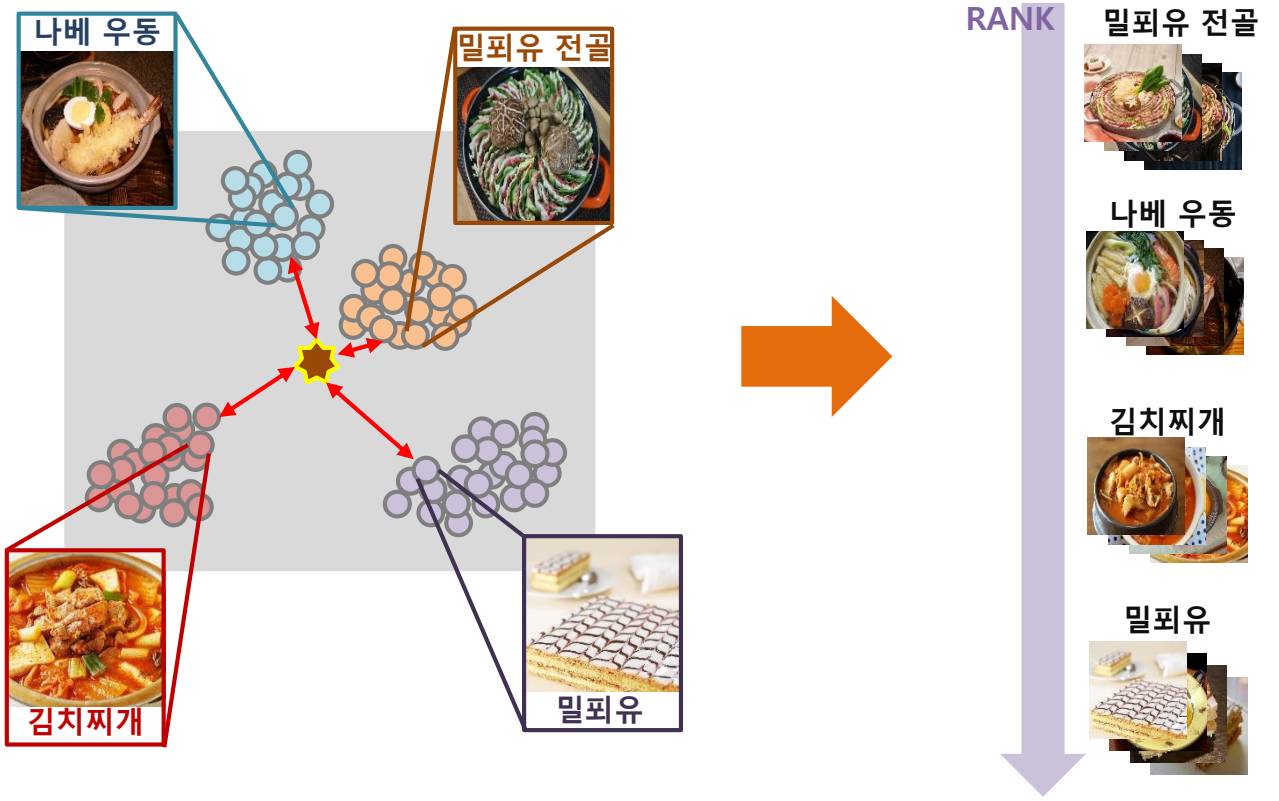
(a) Few-shot (b) Zero-shot

상세 사항

- 각 class 별 mean feature(= prototype)를 계산
- query feature와 mean feature 사이의 거리가 제일 짧은 class를 리턴
- distance는 euclidean distance, 각 class 별 prototype과 distance 계산

연구 내용

- Prototype base distance calculate
  - 전체 데이터의 각 class에 해당하는 prototype을 랜덤으로 선정
  - 선정된 prototype을 활용하여 입력된 query feature와 거리를 계산
  - 계산된 거리가 짧을수록 연관성이 높은 것으로 판단

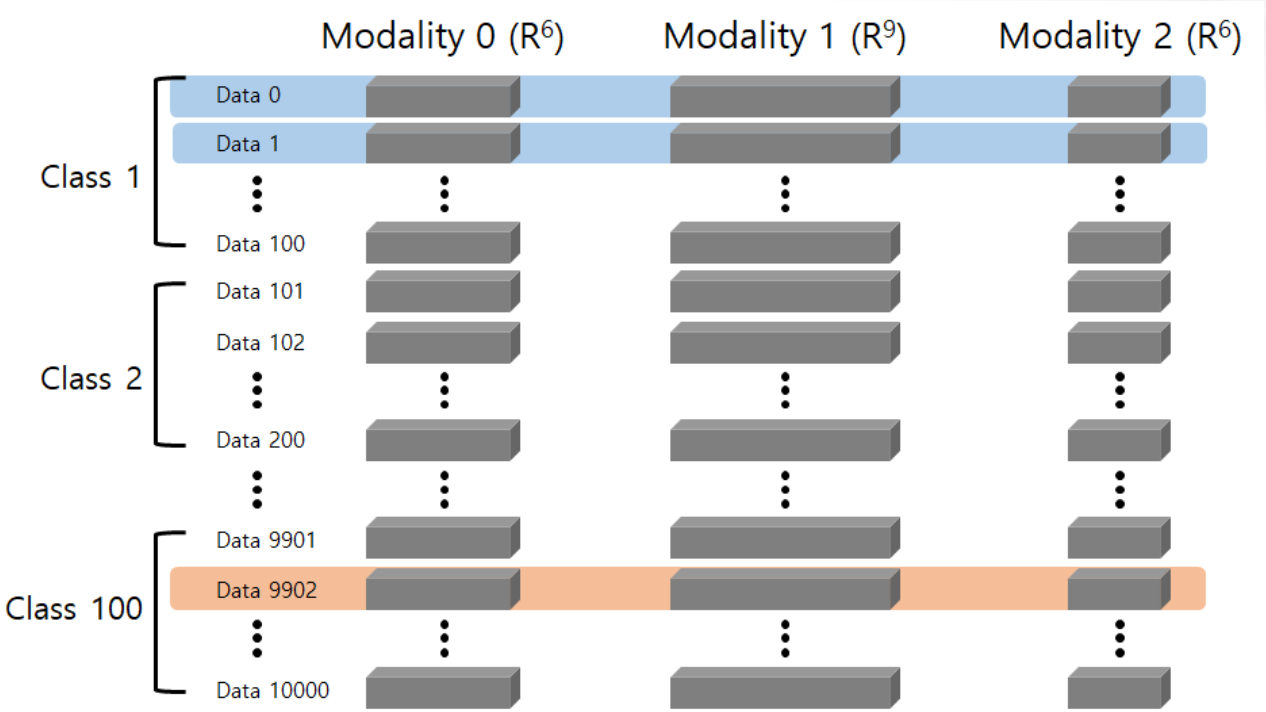


연구 결과

확률분포 기반 데이터 생성

- 같은 modality의 데이터는 (m, 1) 같은 크기의 벡터로 구성
- 각 modality 별로 m은 다른 값을 가짐, (modality = distribution)
- 같은 modality(dimension) 내에서 각 class는 각자 다른 distribution의 값을 의미
- Distribution type은 uniform, normal 두 종류
- 각 class 별 distribution parameter를 랜덤으로 설정하고 데이터를 샘플링

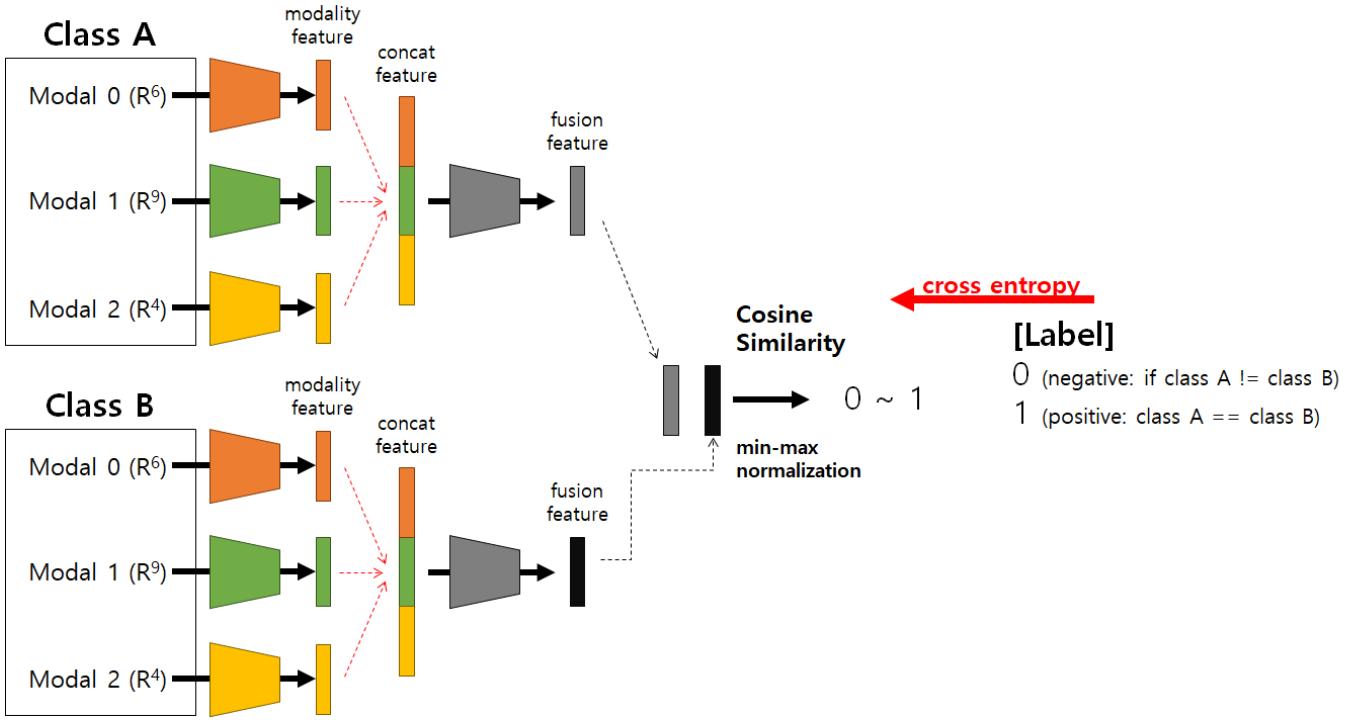
연구 결과



확률분포 기반 데이터 생성

- 각 class당 100개, 총 10000개 데이터 생성, Train set과 test set은 10:1
- Modality 0: (6 x 1) vector
- Modality 1: (9 x 1) vector
- Modality 2: (4 x 1) vector

연구 결과



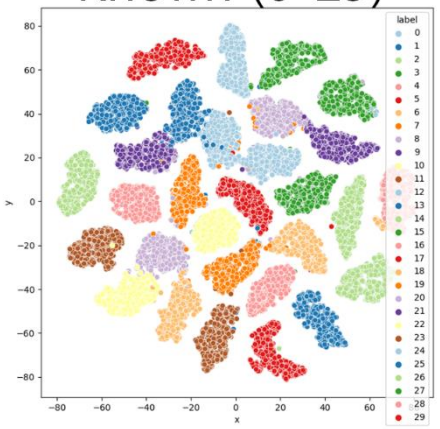
멀티모달 fusion model

- feature extractor = 3 FC layers ⇒ 모달리티 별 따로 존재
- feature fusion = 2 FC layers
- similarity = cosine similarity & min-max norm. (-1~1⇒0~1)

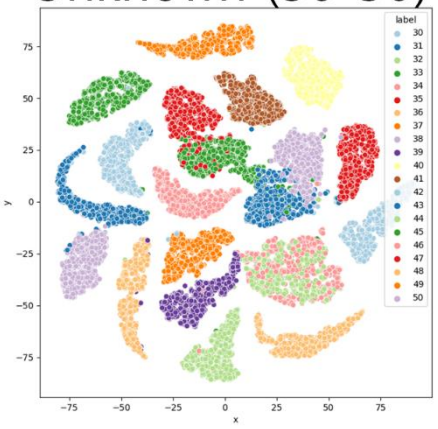
연구 결과

- **T-SNE 시각화**

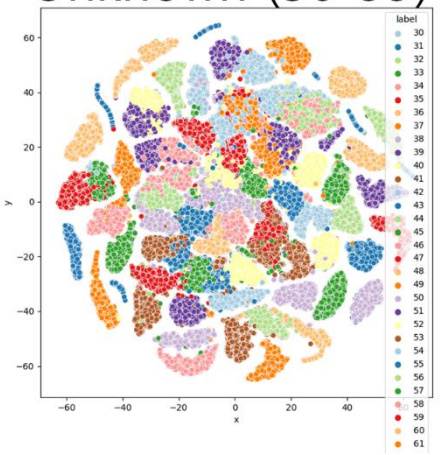
Known (0-29)



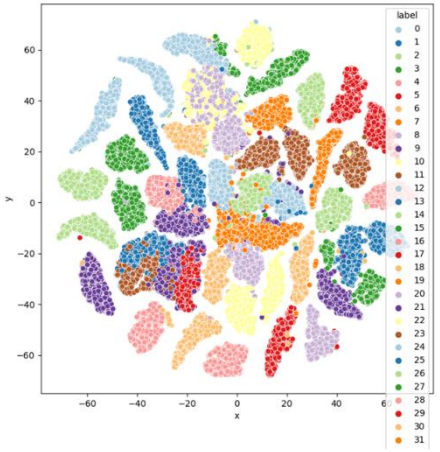
Unknown (30-50)



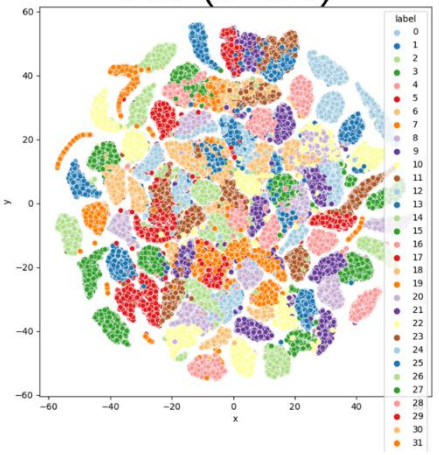
Unknown (30-99)



Known & Unknown (0-50)



All (0-99)



연구 결과

- **Euclidean distance (EUC)와 Cosine similarity (COS)를 loss로 학습 후 비교**
- **Cosine similarity가 좀더 높은 성능**

		EUC: Euclidean distance			COS: Cosine similarity								
		Known (0-29)				Knonw & Unknown (0-50)				All (0-99)			
Modality	num K	EUC top1	EUC top5	COS top1	COS top5	EUC top1	EUC top5	COS top1	COS top5	EUC top1	EUC top5	COS top1	COS top5
0	200	62.07%	91.13%	63.27%	95.53%	45.77%	83.26%	46.02%	86.08%	29.14%	64.07%	29.78%	66.11%
1	200	63.87%	92.10%	65.07%	95.67%	59.22%	90.71%	60.10%	92.65%	43.68	83.15	43.62	83.44
2	200	43.00%	80.37%	44.23%	85.03%	37.20%	72.43%	38.39%	76.31%	25.71	59.17	26.12	60.38
0, 1	200	88.27%	96.77%	93.73%	99.93%	78.65%	94.33%	83.20%	99.12%	66.75	88.97	71.85	93.98
0, 2	200	79.00%	95.30%	82.20%	99.37%	66.84%	89.98%	71.24%	95.65%	52.43	82.3	55.9	87.76
1, 2	200	89.40%	98.57%	90.10%	99.67%	80.80%	97.31%	81.00%	98.73%	68.85	93.44	69.2	**94.29**
0, 1, 2	5	95.73%	98.23%	**97.57%**	99.93%	80.63%	95.75%	82.47%	97.98%	69.65%	90.61%	68.81%	92.21%
	20	95.87%	98.57%	97.50%	**99.97%**	83.49%	96.71%	83.78%	97.16%	72.01%	93.15%	71.79%	93.40%
	200	**96.03%**	**98.83%**	**97.53%**	**99.97%**	**84.59%**	96.77%	85.16%	**98.04%**	**73.20%**	**93.35%**	73.56%	94.11%
	2000	95.83%	98.77%	97.53%	**99.97%**	84.49%	**96.80%**	**85.43%**	98.02%	72.82%	93.21%	**73.70%**	**94.21%**

연구 계획

- 음식 데이터 적용
  - 실제 환경에서 활용할 음식 데이터 기반 유사도 탐색 실험
  - 영어 데이터는 Recipe1M 활용
  - 한글 데이터는 만개의레시피 사이트의 정보를 가공하여 활용

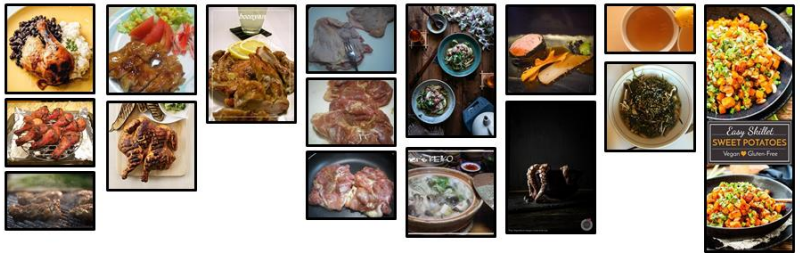
Dataset: Recipe1M+ ➔ 1M recipes & 13M images

"title": "Kombu Tea Grilled Chicken Thigh"

"**ingredients**": [
{"text": "2 Chicken thighs"},
{"text": "2 tsp Kombu tea"},
{"text": "1 White pepper"}],

"**instructions**": [
{"text": "Pierce the skin of the chicken with a fork or knife."},
{"text": "Sprinkle with kombu tea evenly on both sides of the chicken, about 1 teaspoon per chicken thigh."},
{"text": "Brown the skin side of the chicken first over high heat until golden brown."},
{"text": "Sprinkle some pepper on the meat just before flipping over."},
{"text": "Then brown the other side until golden brown."}]},

"**images**": [{"id": String, "ulr": String}]



Thank You