Zero-Shot Learning Through Cross-Modal Transfer

Socher et al., 2013

Presented by Kang

Zero-shot learning? Cross-modal transfer?

Zero-shot Learning:

Training data 없이도, 사전에 관측되지 않은(unseen) 객체와 관측된 적 있는(seen) 객체를 모두 예측할 수 있는 학습 모델

Cross-modal transfer:

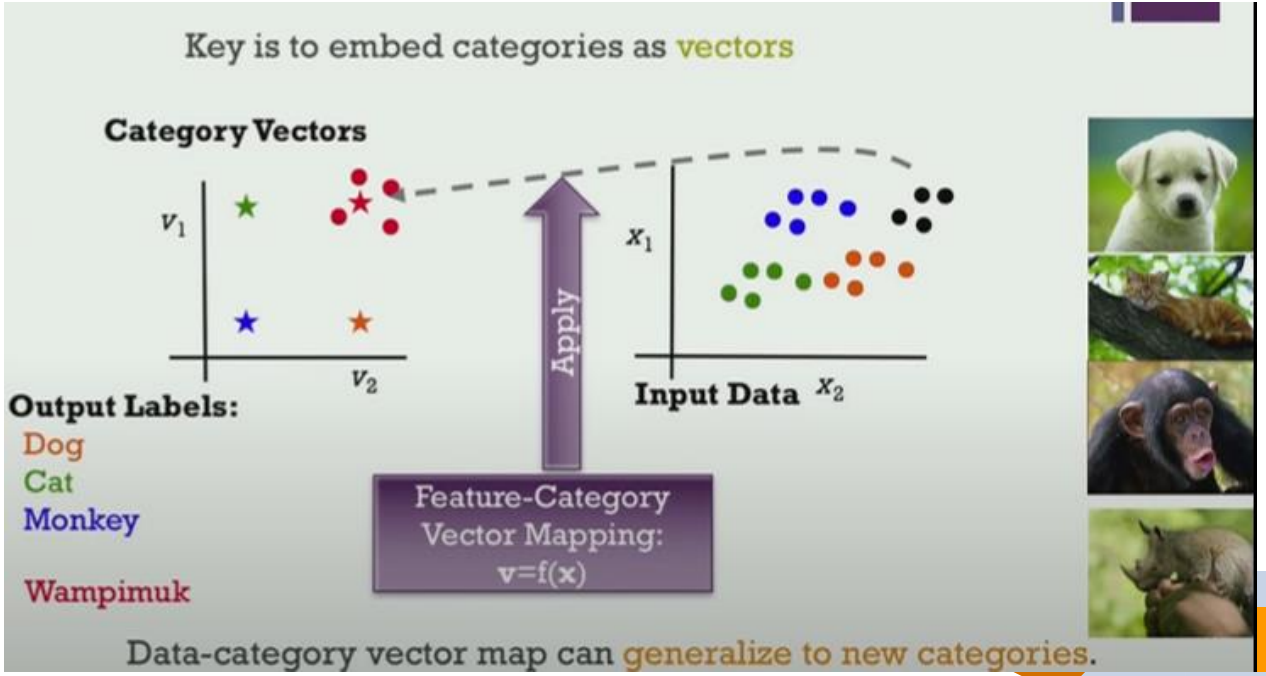
any kind of learning that involves information obtained from more than one modality(감각의 양상). A stimulus modality provides information obtained from a particular sensorial input, for example, visual, auditory, olfactory, or kinesthetic information.

이미지들을 신경망 모델로 학습된 단어의 의미적 공간에 위치시키는 것.

Zero-shot learning?

? New class ?

Word vector

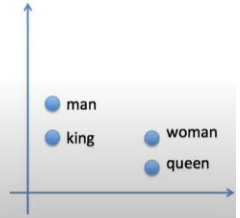


Word to Vector

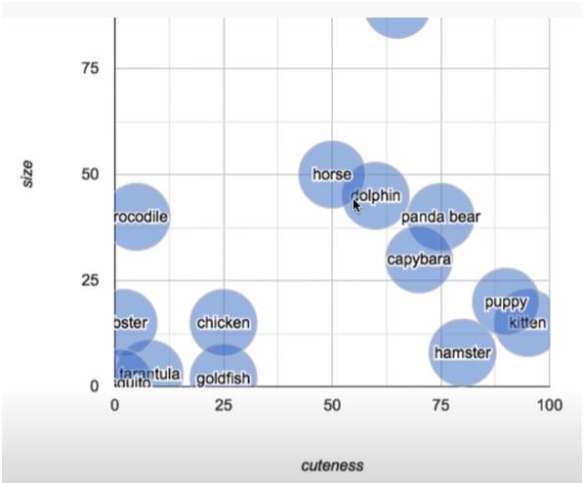
Embedding

Embedding is dense vector with similarity

unique word	encoding	embedding
king	[1, 0, 0, 0]	[1, 2]
man	[0, 1, 0, 0]	[1, 3]
queen	[0, 0, 1, 0]	[5, 1]
woman	[0, 0, 0, 0]	[5, 3]



	cuteness (0–100)	size (0–100)
kitten	95	15
hamster	80	8
tarantula	8	3
puppy	90	20
crocodile	5	40
dolphin	60	45
panda bear	75	40
lobster	2	15
capybara	70	30
elephant	65	90
mosquito	1	1
goldfish	25	2
horse	50	50
chicken	25	15

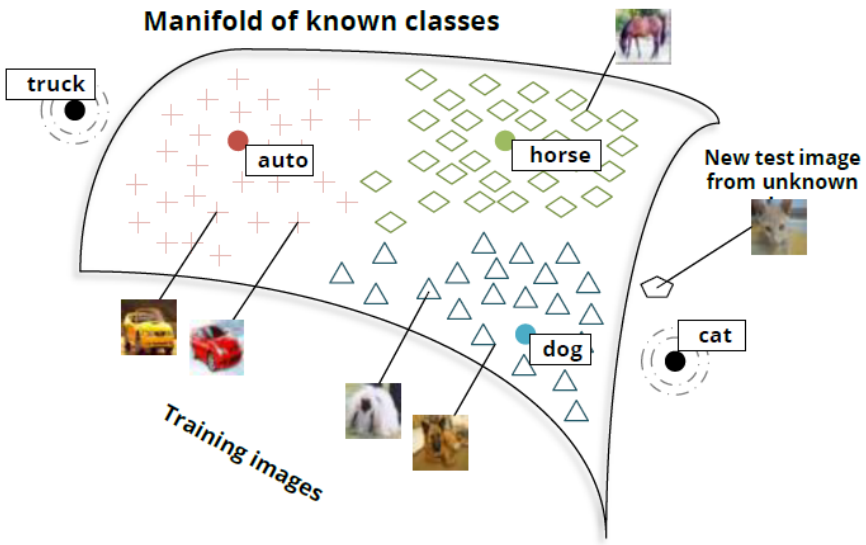


Why Zero-shot learning?

Vast and growing number of categories:

- Collecting and annotating examples is impossible
  - Especially deep learning scale
  - New categories always emerging

Zero-shot Learning?



[두 가지 main idea에 기초]

1) 이미지들을 신경망 모델로 학습된 단어의 의미적 공간에 매핑 (cross-modal transfer)
2) 기존의 분류기는 test 이미지를 seen class로 분류하는 것을 선호하기 때문에 모델은 새로운 이미지가 seen class에 해당하는지 아닌지를 결정하는 novelty detection을 포함

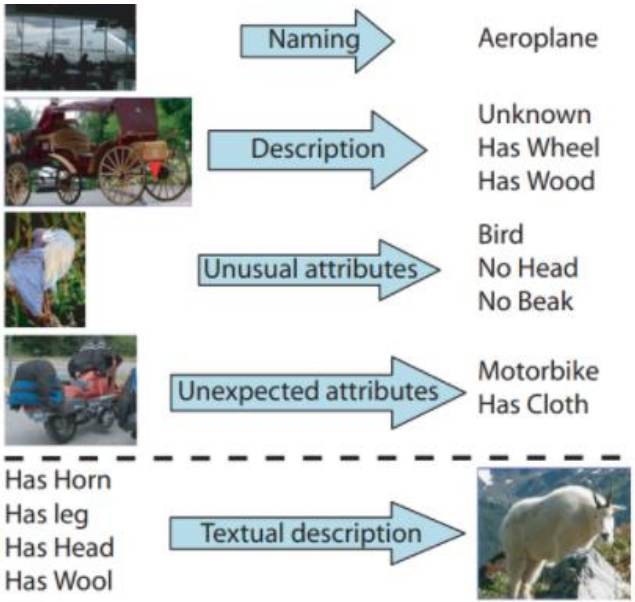
Cat과 truck는 unseen class여서 알려진 클래스여서 평면 안에 위치하지 않음.

시각적 이미지를 사용했기 때문에, 비지도학습 가능

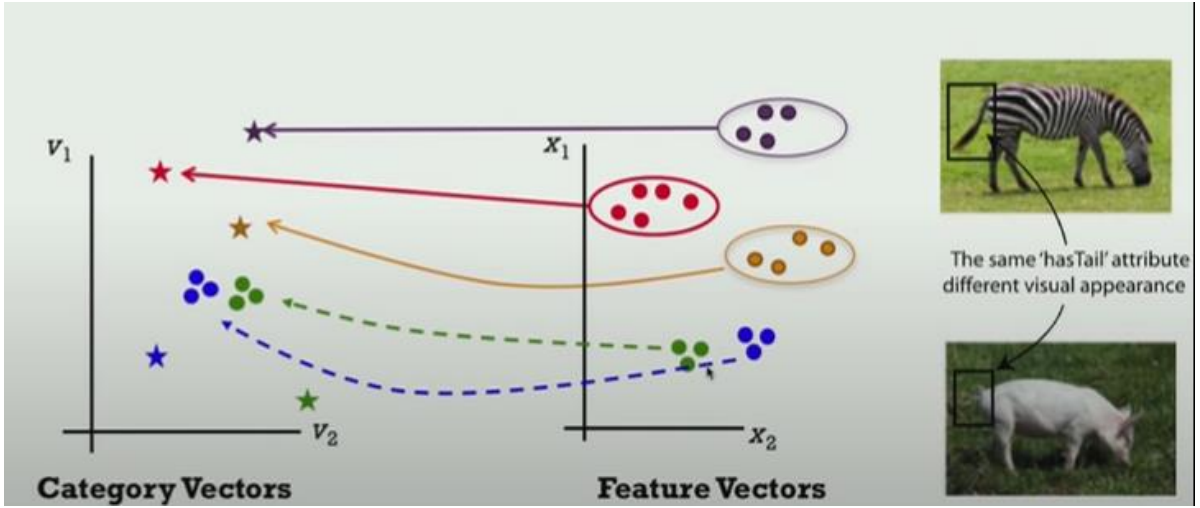
Related work

Zero-shot Learning	One-shot Learning	Knowledge and Visual Attribute Transfer	Domain Adaption	Multimodal Embeddings
- 본 연구와 가장 유사 - Palatucci et al., : fMRI로 특정 단어에 대한 사람들의 생각을 수동으로 특성을 분류 → Seen, unseen 간의 새로운 테스트 인스턴스는 분류 X → 본 연구에서는 분류	- 소수의 훈련 데이터로 시각 정보 학습 - 유사한 맥락(context), 공통된 특징을 공유할 때 사용됨. → 자연어로부터 cross-modal knowledge transfer 덕택에, 훈련 데이터가 필요 X	- 기존의 연구는 잘 정리된 시각적 특성을 이용 - 그러나 본 연구는 단어의 분포적 특성만 이용	- 한 분야(domain)에서는 이미 훈련된 데이터가 많지만, 다른 분야에서는 없을 때 사용. - 예) 영화 리뷰 → 책 리뷰에서 사용 가능	- Multimodal embeddings: 소리, 영상, 이미지, 글과 같이 다양한 소스로 이루어져 있는 정보와 연관 - 각 class 별로 소량의 훈련 데이터 필요

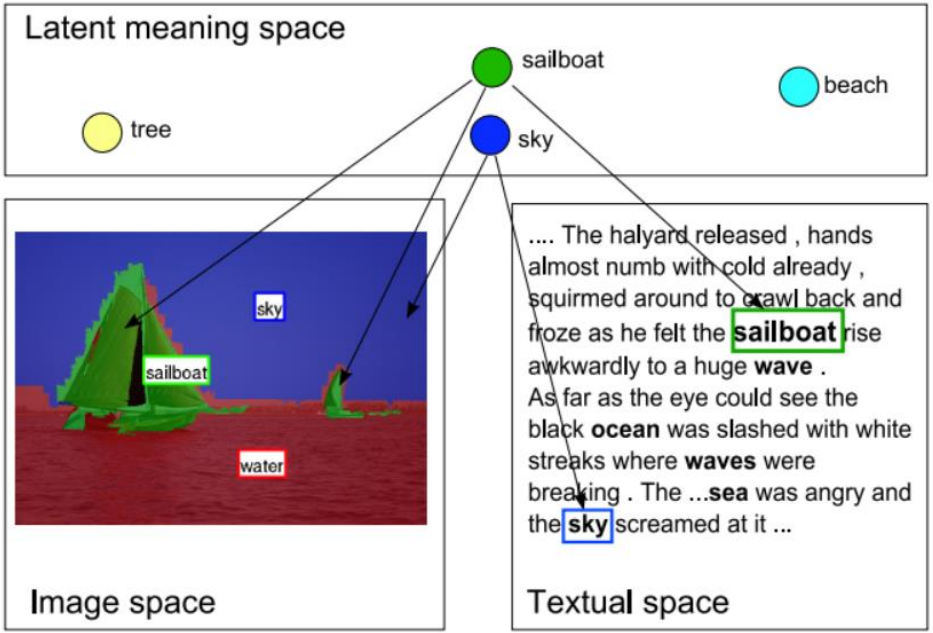
Related work - Knowledge and Visual Attribute Transfer



Related work - Domain Adaption



Related work - Multimodal Embeddings



Zero-shot learning model

Projecting Images into Semantic Word Spaces

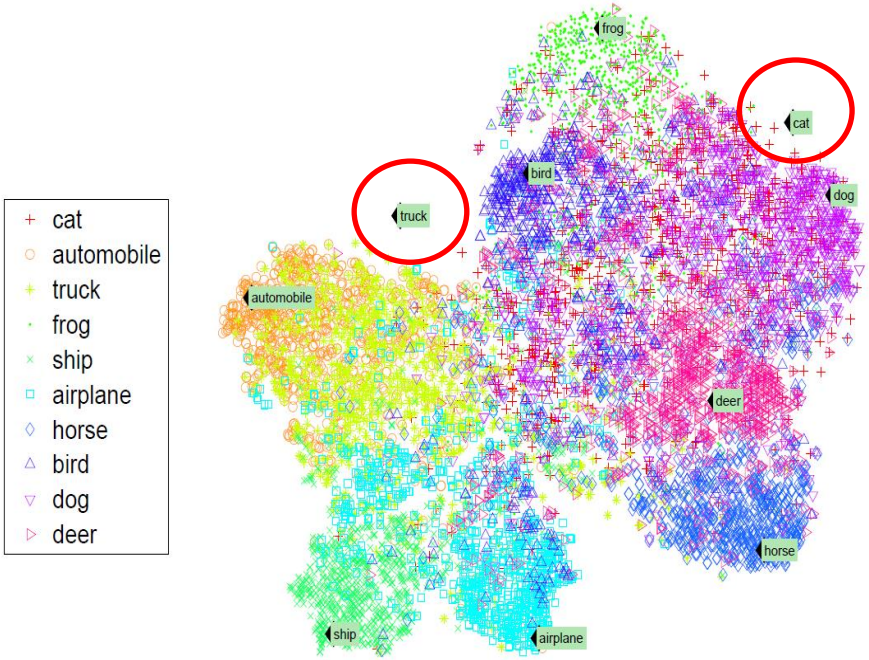
$$J(\Theta) = \sum_{y \in Y_s} \sum_{x^{(i)} \in X_y} \left|\left| w_y - \theta^{(2)} f\left(\theta^{(1)} x^{(i)}\right) \right|\right|^2$$

Y_s: seen class
Y_u: unseen class
W_y: class 이름에 대응되는 word vector
$x^{(i)} \in X_y$: 학습 이미지
$\theta^{(1)}, \theta^{(2)}$: neural network 가중치

이미지의 semantic relationship과 class membership을 학습시키기 위해 image feature vector를 d차원의 semantic word space인 f에 투영함

이미지와 단어 간의 매핑을 훈련시키기 위해 위의 식을 최소화함.

T-SNE visualization of the semantic word space



word vector와 image로 seen class와 unseen class의 semantic space를 시각화

10개의 데이터 셋 중, truck과 cat을 zero-shot learning을 위해 떼어놓음.

시각화 결과, truck과 cat만 unseen class로 분류됨.

그러나 자신이 속한 범주와 유사한 특성 근처에 위치해 있는 것을 발견.
(예: cat은 dog 근처에 위치, automobile로부터는 떨어져 있음)

Two Strategies for Novelty Detection

Gaussian model

- **Unseen class** 분류에 강함
- **Unseen** 이미지로 분류될 확률

$$P(V = u|x, X_s, F_s, \bar{W}, \theta)$$

- 가우시안 등간척도를 통해, 과적합 방지
- T_y를 통해 한계점(**threshold**)를 정해, **seen class**로 분류될 **margin** 확률을 계산

$$P(V = u|f, X_s, W, \theta) := \mathbb{1}\{\forall y \in Y_s : P(f|F_y, w_y) < T_y\}$$

- 임계값이 작을수록 unseen class로 판별되는 데이터가 적어짐
- 한계점: 이상치에 대해 실제 확률값을 도출하지 못함.

Local Outlier Probability(LoOP) model

- 가우시안 모델의 한계점 보완한 모델
- **K = 20,** $\lambda = 3$
- 확률적으로, **seen** 범주로 분류될 확률

$$\text{pdist}_\lambda(f, C(f)) = \lambda\sqrt{\frac{\sum_{q\in C(f)} d(f,q)^2}{|C(f)|}}$$

- **Euclidean** 거리계산

$$\text{lof}_\lambda(f) = \frac{\text{pdist}_\lambda(f, C(f))}{\mathbb{E}_{q\sim C(f)}[\text{pdist}_\lambda(f, C(q))]} - 1.$$

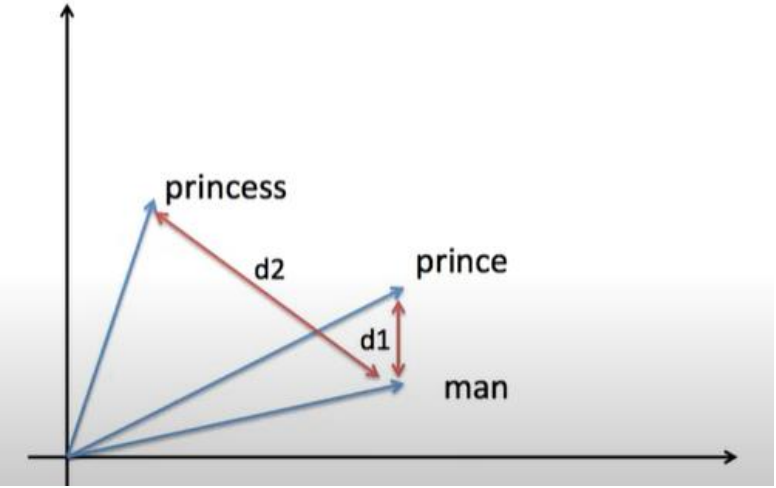
- **Seen data** 표준화

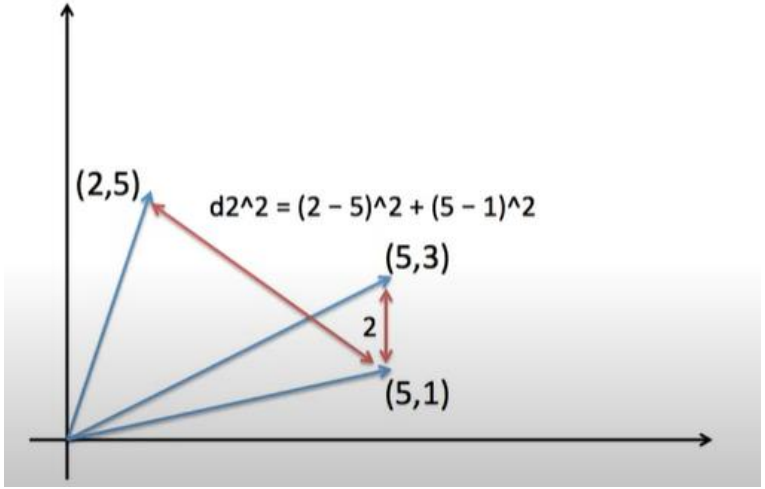
$$Z_\lambda(F_s) = \lambda\sqrt{\mathbb{E}_{q\sim F_s}[(\text{lof}(q))^2]}.$$

- 이상치일 확률

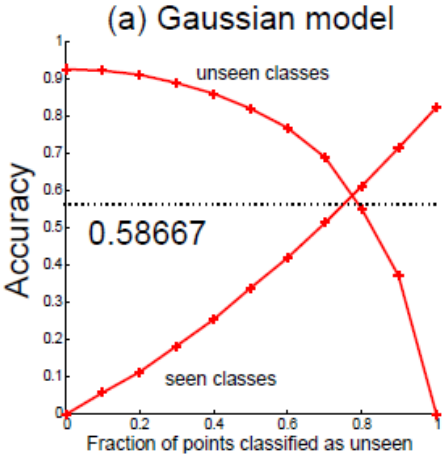
$$LoOP(f) = \max\left\{0, \text{erf}\left(\frac{\text{lof}_\lambda(f)}{Z_\lambda(F_s)}\right)\right\}$$

Two Strategies for Novelty Detection
- Euclidean distance

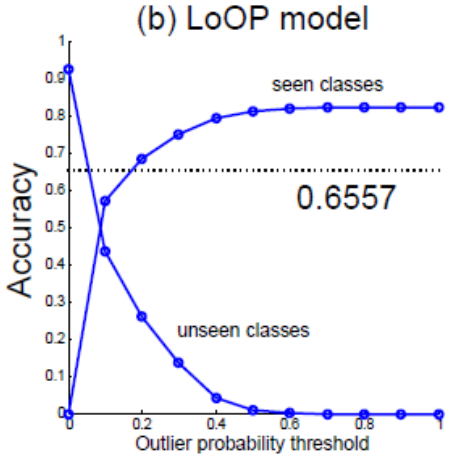




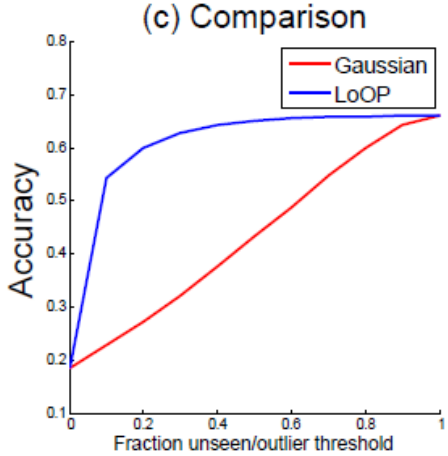
Two Strategies for Novelty Detection



Fraction이 **0**일수록, **unseen**으로 분류됨. 비교적 **unseen**을 잘 분류함.

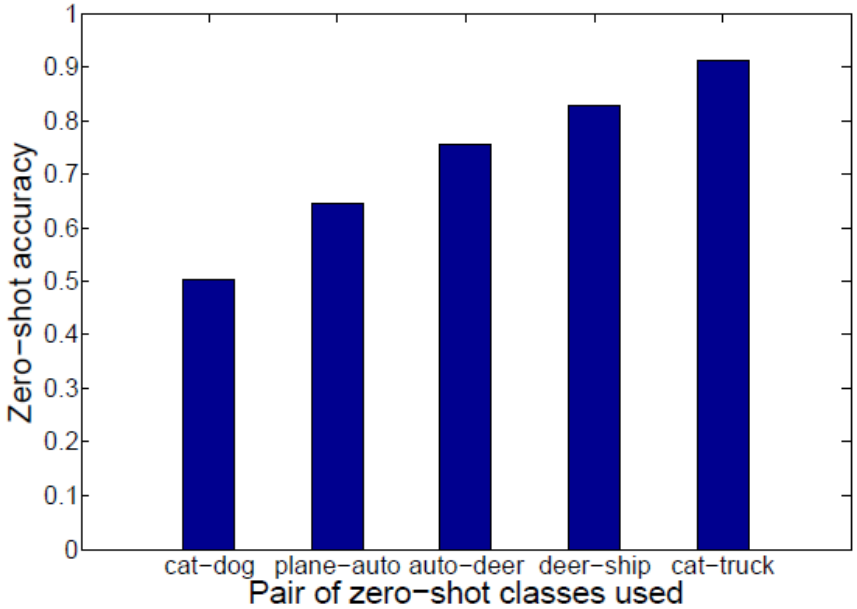


이상치 분류모델이라, **seen class**를 더 잘 분류함. 더 보수적.



정확도는 **LoOP**가 더 높음.

Zero-shot Accuracy



의미가 유사한 **class**가 **zero-shot**으로 선정되었을 수록, 정확도가 떨어짐.

Cat-truck의 경우, **90%** 정확도 달성

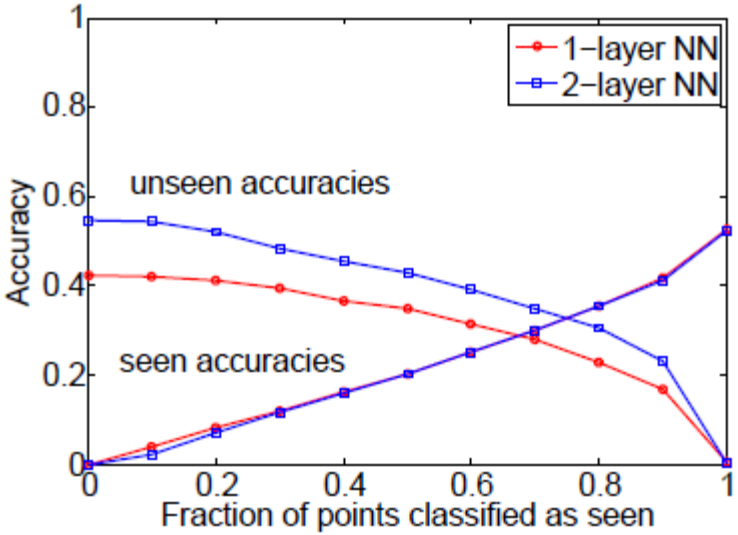
Zero-shot Accuracy

Bayesian pipeline (Gaussian)	74.25%
Bayesian pipeline (LoOP)	65.31%
Attribute-based (Lampert et al.)	45.25%

Attribute-based 보다 더 분류 정확도가 높았음.

Attributued-based:
동물 기반, 탈 것 기반의 특성(귀 모양, 깃털 유무, 바퀴 갯수 등등) 을 구분하는 분류기를 각각 만든 후(이 사진이 귀 모양이 0이냐 1이냐로 분류하는 모델), Unseen 데이터에 대해 labeling 하도록 한 것.

Extension to CIFAR 100

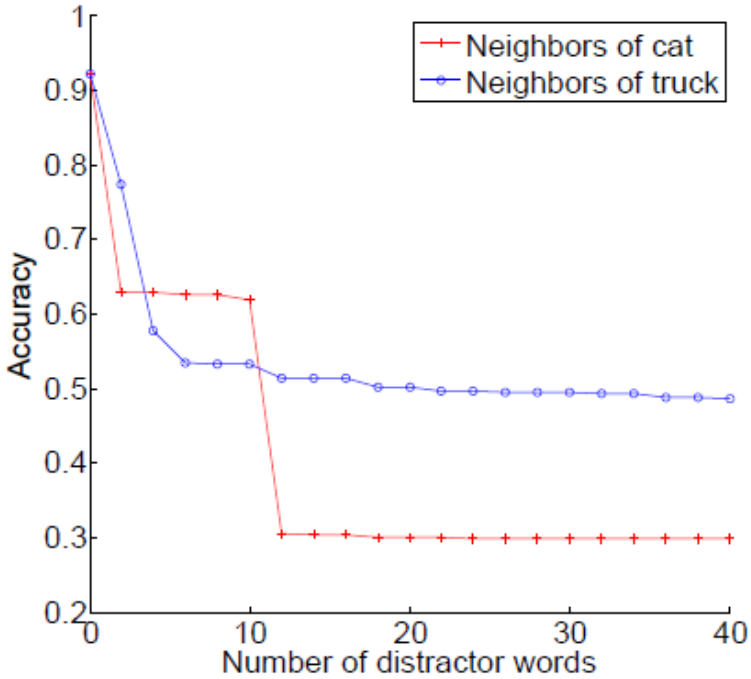


[10개 데이터에서 100개로 확장]

- Seen 최대분류 정확도: 52.7%
- unseen 최대분류 정확도: 52.7%

- 2-layer NN에서 더 정확도 높음

Zero-shot classes with Distractor words

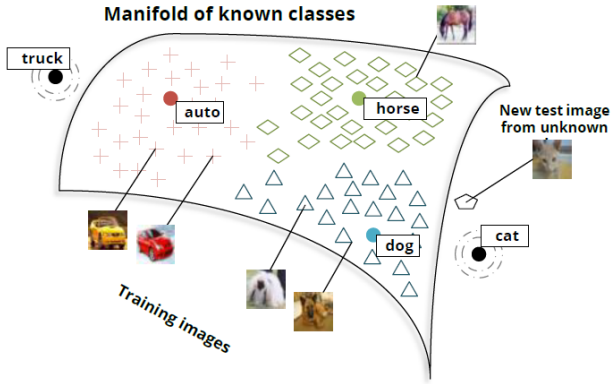


Zero-shot class: cat, truck
Distractor words: rabbit, kitten, mouse

→ Distractor 단어가 truck보다는 cat과 유사함.
→ 따라서 고양이보다Truck의 unseen 분류 정확도가 높았음.
→ 그러나 일정수준 아래로 정확도가 떨어지지는 않음

Conclusion

- 표준 분류(**softmax**)와 **zero-shot** 분류를 결합한 새로운 모델 제안
- **word vector**를 통해 비지도 학습임에도 다른 **modalities**간 정보 전달이 가능
- **Bayesian framework**를 통해, **seen**과 **zero-shot** 분류를 하나의 프레임워크로 통합



THANKS!