Zero-Shot Learning Through Cross-Modal Transfer

> Socher, Richard, et al. "Zero-shot learning through cross-modal transfer." *arXiv preprint arXiv:1301.3666* (2013).

Abstract

1. Introduction

기존의 지도학습 모델은 한번도 보지 못한 데이터에 대해서 분류를 한다면 이미 학습된 데이터에 기반하여 분류하게 된다.(강아지,고양이로 학습한 모델은 자동차를 보았을 때, 강아지나 고양이로 분류하게 됨)

하지만 사람은 처음 보는 것에 대해서 간단한 요약을 본 후에 분류가 가능하다.
(ex. 2개의 바퀴를 가진 차량, 스틱으로 조종, 그 위에 서서 움직인다. -> 세그웨이)

이를 활용하여 자연어를 활용하여 처음 보는 물체에 대하여 사람과 같이 물체를 분류하는 방법을 제안한다.

1. 이미지를 neural network에 학습된 단어의 의미론적 공간에 mapping한다.
2. classifier는 test image를 training sample에서 본 class에 할당하려 하기 때문에, 처음 보는 class인지 확인하기 위해 알고 있는 class의 manifold에 새로운 이미지가 있는지 결정할 수 있는 모델을 추가로 활용한다.

manifold 간단한 설명 영상

2. Related Work

Zero-Shot Learning

Palatucci의 Zero-Shot Learning with Semantic Output Codes라는 논문에서는 특정 단어를 생각할 때 사람들의 fMRI 스캔을 직접 feature map에 mapping한 후 이를 활용하여 분류한다. 그러나 새로운 test instance에 대해서 seen or unseen class로 분류하지 않는다. 이와 반대로 본 논문에서는 outlier detection을 활용하여 seen or unseen class로 분류하도록 확장시킨다.

One-Shot Learning

One-Shot Learning은 training example이 적을 때 이를 인식하고 분류하는 방법이다.
이는 일반적으로 특징 표현, parameter 공유, context 유사도 등을 활용해 학습된다.

이 논문에서는 이와 유사하게 deep learning을 활용하여 low-level의 image feature를 학습하고 이를 확률 모델을 활용하여 자연어 기반의 지식으로 전달하는 방법을 사용하였다.

Knowledge and Visual Attribute Transfer

이전 연구에서는 잘 설계된 시각적 attribute를 사용하여 unseen class에 대해서 분류하였다.

이 논문과의 차이점은 비지도학습으로 학습된 단어에 대한 분포적 특성을 활용하여 분류하기 때문에 training image의 개수에 큰 영향을 받지 않는다.

Domain Adaption

Domain Adaption은 특정 domain에선 data가 많지만, 다른 domain에선 data가 적을 때 유용한 방법이다.

예를 들어 영화 리뷰에 대한 감정분석을 책 리뷰에 대해 적용시킨다면, 작업 방법이 달라지고, 각 domain에 대해서 feature들이 달라지게 된다.

Multimodal Embeddings

Multimodal Embedding은 소리나 영상, 텍스트와 같은 정보들을 연관짓는 것을 말한다.

이전 연구에선 단어와 image 영역을 canonical correlation analysis(변수들의 관계를 분석)를 통해 한 공간에 투영하여 annotation과 segmentation에서 sota를 달성했었다.

3. Word and Image Representations

nlp task에서 동시 발생 빈도는 단어의 특성을 나타내는데 효율적이라고 한다.
이를 표현하는 단어 벡터는 50차원으로 초기화 된다.

각 이미지는 F 차원으로 표현된다.

4. Projecting Images into Semantic Word Spaces

이미지의 특징을 50차원의 단어 공간에 project하는 방법을 학습하기 위한 objective function은 아래와 같다.

$$J(\theta) = \sum_{y \in Y_s} \sum_{x^i \in X_y} ||w_y - \theta^{(2)} tanh(\theta^{(1)} x^{(i)})||^2 \qquad [eq\ 1.]$$

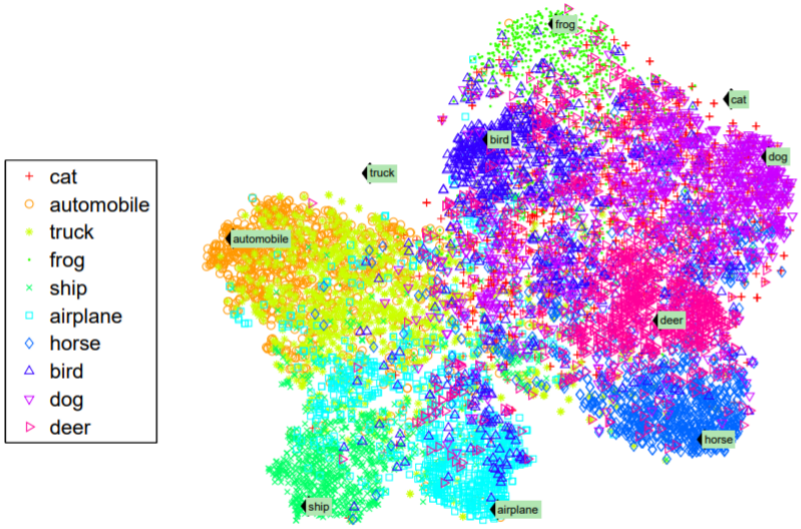
$$Y_s, Y_u = seen\ class, unseen\ class$$

$$w_y = class\ word\ vector$$

$$x^{(i)} \in X_y = training\ images$$

$$\theta^{(1)} \in \mathbb{R}^{h \times i}, \theta^{(2)} \in \mathbb{R}^{d \times h} = weights$$

이는 seen class에 대한 image의 feature vector와 가중치의 곱을 통해 단어벡터 W간의 l2 distance를 의미한다. 사용된 neural network는 two-layer neural network를 사용하였다고 한다.



[Figure 2] unseen class를 cat, truck으로 하여 학습한 모델을 T-SNE를 사용하여 semantic word space를 시각화한 것이다.

fig 2에서 seen class의 경우 word vector 주위로 clustering 되지만, unseen class의 경우는 그렇지 않다는 것을 확인할 수 있다. 하지만, unseen class를 살펴보면 cat은 의미적으로 유사한 dog, horse에 가깝게 mapping되며, automobile, ship과는 떨어져 있는 것을 확인할 수 있다. 이를 통해서 outlier를 찾은 다음, zero-shot word vector(unseen class vector)로 분류하는 아이디어에 동기부여했다고 한다.

5. Zero-shot Learning Model

논문에서 제시한 모델의 목적은 다음과 같다.

$$p(y|x)$$
$$y \in Y_s \cup Y_u \;\; given\ image\ x$$

이처럼 이미지 x가 주어졌을 때, unseen class와 seen class로 이루어진 class를 분류하는 것이다.
하지만 일반적인 분류기로는 seen class인지 unseen class인지 분류하기 어렵기 때문에 binary novelty random variable(V)와 semantic vector(f)를 제시하였다.

$$V \in \{s, u\}$$
$$f \in F_t$$

최종적으로 test image에 대한 class의 예측은 다음과 같다.

$$p(y|x, X_s, W, \theta) = \sum_{V \in \{s,u\}} P(y|V, x, X_s, F_s, W, \theta) P(V|x, X_s, F_s, W, \theta) \qquad [eq\ 2.]$$

5.1 Strategies for Novelty Detection

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

위는 eq2에서의 우측항에서 image가 unseen image일 확률을 의미한다. unseen image의 경우 기존의 seen image에서의 class와 가까울순 있지만 seen image의 class만큼은 가깝지 않다. 예를 들어 unseen image의 class가 실제로 고양이인 경우 seen image의 class에선 의미적으로 가까운 개와 가깝지만, 실제 개만큼은 가깝지 않다. 따라서 outlier detection을 통해 seen class인지 unseen class인지 결정할 수 있다.

outlier detection method

1. iso-metric, class-specific Gaussioans

클래스 별 mixture of gaussians(가우시안 분포의 밀도)이내의 thresholds를 사용한다.

$$y \in Y_s,$$
$$P(x|X_y, w_y, F_y, \theta) = P(f|F_y, w_y) = \mathcal{N}(f|w_y, \sum\nolimits_y)$$

위 수식은 X_y(seen class에서 image의 특징벡터) 와 theta(neural net의 가중치)를 통해 semantic vector인 f를 추정하고, 이 f를 평균이 w_y, y의 공분산을 가진 gaussian distribution으로 변환을 의미한다. 참고
이 때 overfitting을 막기 위하여 각 class 마다 같은 거리를 갖도록 변환시킨다.

이를 통해 새로운 image에서의 적용은 아래와 같다.

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

위 식은 unseen class를 모든 seen class에서의 mixture of gaussian이 임계치 T_y 이내이면 1을 할당하는 것을 의미한다.

위의 T_y를 설정하는 방법으로는 training image의 fraction을 임계값보다 높게 만드는 T_y를 결정합니다.

이 방법은 outlier에 대한 실제 확률을 제공하지 못한다는 단점이 있다.

loOP(local Outlier Probabilities)

Outlier에 대한 실제 확률을 제공하기 위해 lof를 변형한 loOP를 사용하였다고 한다.
fig 2를 보면 unseen class 는 data manifold에서 outlier가 아닌 것으로 보이며 novelty를 정하는데 보수적이다. 그렇기 때문에 test 데이터를 사용하지 않는다. 자세한 방법은 아래와 같다.

각 f 와 C(f)에 대한 probabilistic set distance를 계산한다. 이때 distance는 euclidean distance를 사용한다.

> f = training data의 sematic vector에 포함된 semantic vector
> C(f) = f에 가장 가까운 k개의 semantic vector

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

pdist를 활용하여 local outlier factor(lof, outlier정도와 비례)를 계산한다.

$$lof_\lambda(f) = \frac{pdist_\lambda(f, C(f))}{\mathbb{E}_{q\sim C(f)}\left[pdist_\lambda(f, C(q))\right]} - 1$$

outlier에 대한 확률을 구하기 위해서 normarlization factor Z를 정의 한다.

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

최종적으로 local outlier probability는 다음과 같다.

$$LoOP(f) = max\{0, erf(\frac{lof_\lambda(f)}{Z_\lambda(F_s)})\}$$

$$erf(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt$$

Classification

V = seen 인 경우 아무 classifier를 사용할 수 있지만, 논문에서는 softmax classifier를 사용하였다.

V = unseen 인 경우 isometric gaussian distribution에서 likelihood를 통해 분류하였다.

6. Experiments

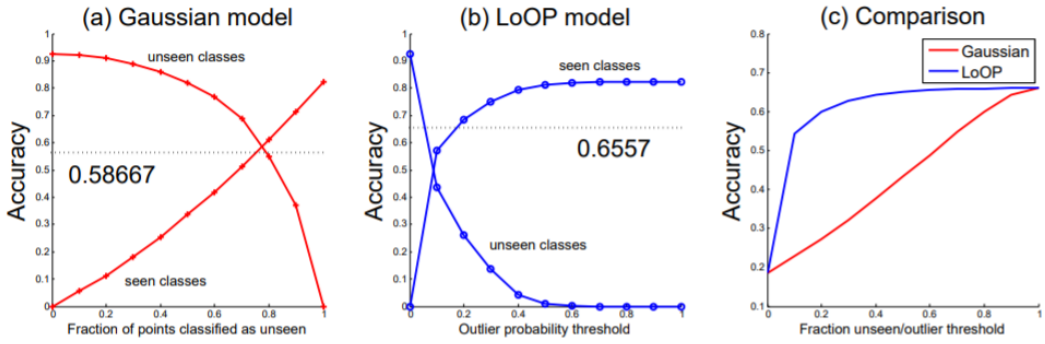


Figure 4: Comparison of accuracies for images from previously seen and unseen categories when unseen images are detected under the (a) Gaussian threshold model, (b) LoOP model. The average accuracy on all images is shown in (c) for both models. We also show a line corresponding to the single accuracy achieved in the Bayesian pipeline. In these examples, the zero-shot categories are "cat" and "truck".