Are the Relationships of Class Representations in Vision and Language Similar?

Mar 15, 2021

Presenter: Sungguk Cha

MINDs Lab

Abstract

- Multi-modal zero-shot learning field is hot.
- We show that domain shift causes changes in representation of CNNs.
- We show that language-based zero-shot learning in computer vision has limitation in generalization intrinsically.

MINDs Lab

Contents

MINDs Lab

Advertisement

It is my fourth presentation in Algorithm Session!

- ➔ Zero-Shot Semantic Segmentation
  via Spatial and Multi-Scale Aware Visual Class Embedding
  Nov 23, 2020

- ➔ How Can We Correlate Inter-Domain Knowledge?
  Reviewing Consistent Structural Relation Learning for Generalized Zero-Shot Segmentation
  Dec 21, 2020

- ➔ Representation Learning: How Should Feature Be Learned in Vision?
  Introducing CLIP and an Unsupervised Semantic Segmentation Approach
  Feb 22, 2021

- ➔ **Are the Relationships of Class Representations in Vision and Language Similar?**
  Introducing an Experiment in SM-VCENet
  Mar 15, 2021

MINDs Lab

Introduction

➔ **Why** am I interested in class representations?

➔ **What** can we learn from this presentation?

MINDs Lab

Introduction:

Why am I interested in class representations?

➔ Class representation?

➔ Multi-modal approaches in computer vision

➔ Limitations of the multi-modal approaches

MINDs Lab

Introduction:

Why am I interested in class representations?

In this presentation,
class representation in language model == word embedding.

Language models: Word2Vec, BERT, GPT, …

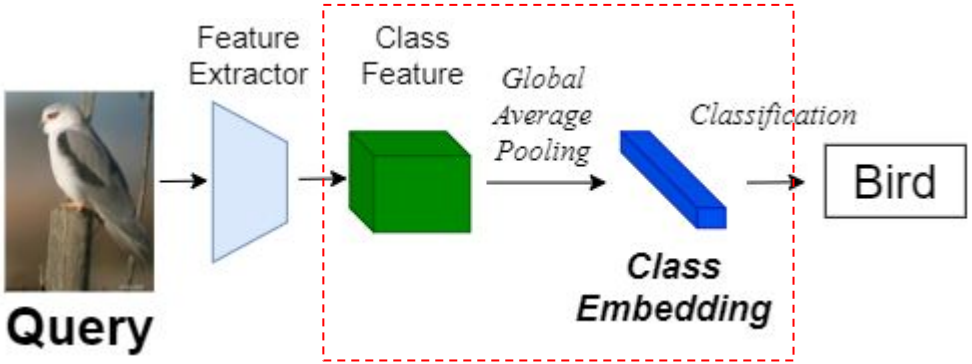
word embedding: *d* dimensional vector expression

MINDs Lab

Introduction:

Why am I interested in class representations?

Class representation in vision model



Introduction:

Why am I interested in class representations?

Why multi-modal approaches in computer vision?

MINDs Lab

Introduction:

Why am I interested in class representations?

Why multi-modal approaches in computer vision?

Prevalent approaches in computer vision field

- requires expensive dataset,
- predicts only learned categories.

MINDs Lab

Introduction:

Why am I interested in class representations?

Why multi-modal approaches in computer vision?

Prevalent approaches in computer vision field

- requires expensive dataset,
- predicts only learned categories.

To solve the problems, *zero-shot* learning is researched actively.

MINDs Lab

Introduction:

Why am I interested in class representations?

Why multi-modal approaches in computer vision?

Prevalent approaches in computer vision field

- requires expensive dataset,
- predicts only learned categories.

To solve the problems, *zero-shot* learning is researched actively.

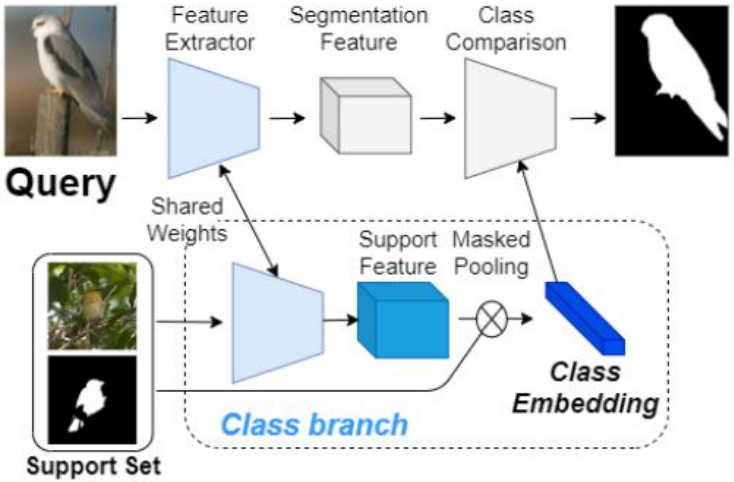
Zero-shot learning?

MINDs Lab

Introduction:

Why am I interested in class representations?

Few shot learning: recognize unseen categories with only a few support images



1. Generate class embedding from support set

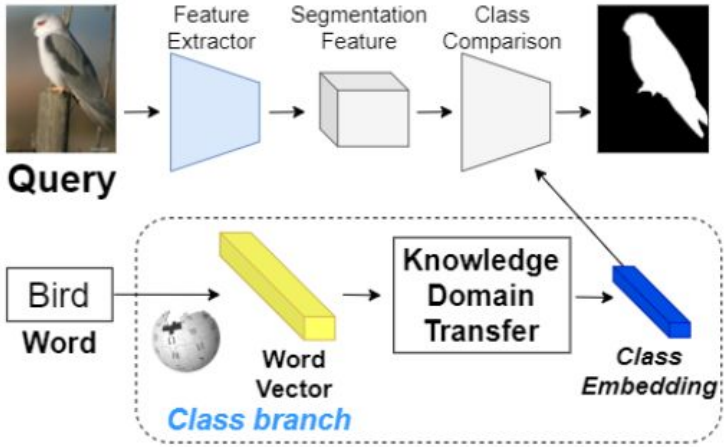
2. Compare *segmentation feature* and *class embedding*

Figure: Few-shot semantic segmentation

Introduction:

Why am I interested in class representations?

Zero shot learning: recognize unseen categories without a support image, **but with language model**



1. Generate class embedding from a word

2. Compare *segmentation feature* and *class embedding*

Figure: Zero-shot semantic segmentation

Introduction:

Why am I interested in class representations?

Multi-modal approaches in computer vision

- Image Classification: CLIP ~ 300M params
- Image Synthesis: DALL E ~ 13B params (GPT-3 small)
- Image Segmentation
- Zero-shot Sketch-based image retrieval

Usually, GPT-3 is not affordable.

Instead, zero-shot approaches use pretrained language model (word2vec).

MINDs Lab

Introduction:

Why am I interested in class representations?

Multi-modal approaches in computer vision examples

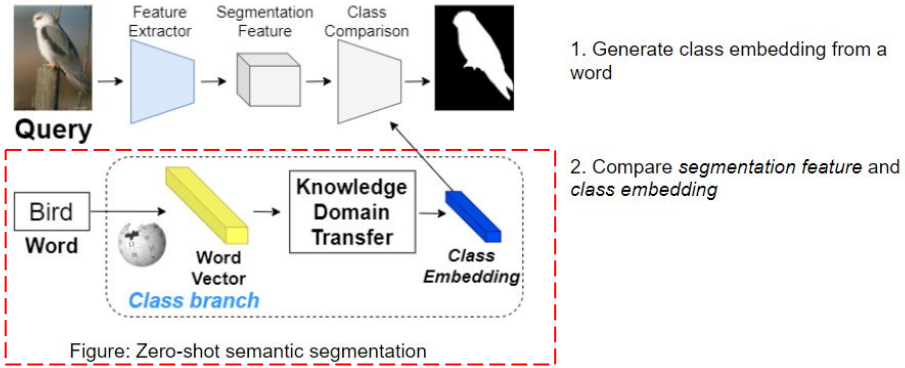


Figure: Zero-shot semantic segmentation

Image sources:
© Sungguk Cha 2021

Introduction:

Why am I interested in class representations?

Multi-modal approaches in computer vision examples

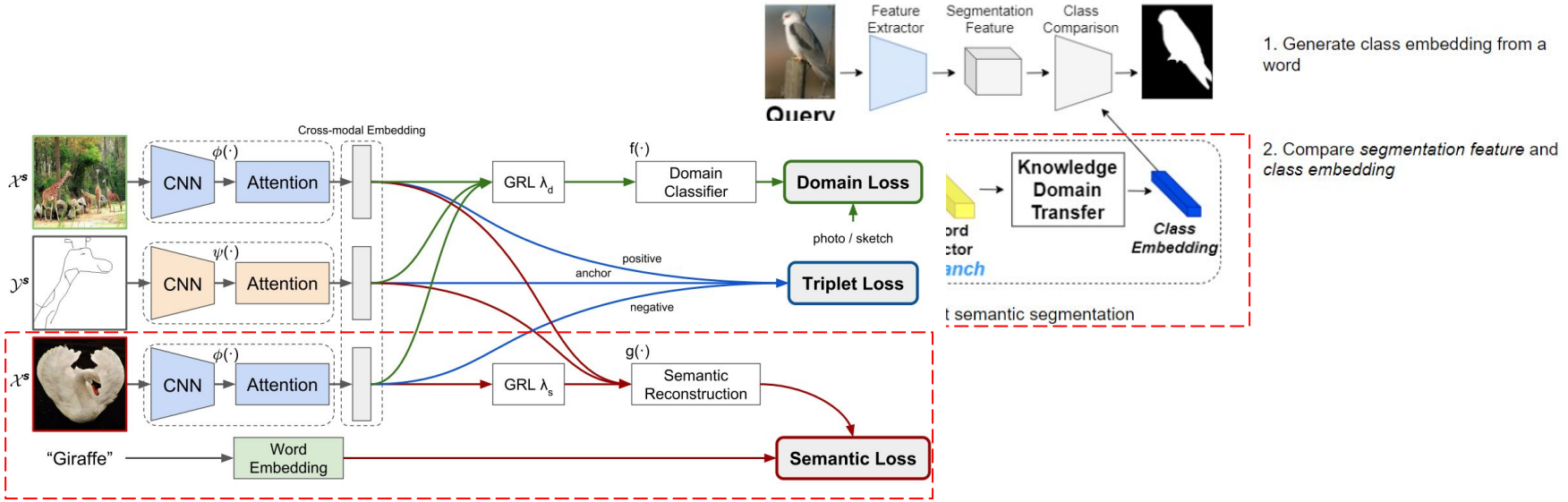


Image sources:
Doodle to Search: Practical Zero-Shot Sketch-based Image Retrieval, CVPR`19

MINDs Lab

Introduction:

Why am I interested in class representations?

Limitations of the multi-modal approaches

MINDs Lab

Introduction:

Why am I interested in class representations?

Limitations of the multi-modal approaches

The proposed zero-shot approaches use

only very little knowledge of language model.

MINDs Lab

Introduction:

Why am I interested in class representations?

Limitations of the multi-modal approaches

The proposed zero-shot approaches use
only very little knowledge of language model.

(e.g.,)

GloVe pretrained on Common Crawl with 2.2 M vocab. in 300 dimension

ZS-SBIR approaches use **125** word embeddings out of **2,200,000**. (Sketch)

ZSSS approaches use **20** word embeddings out of **2,200,000**. (PASCAL-5i)

MINDs Lab

Introduction:

Why am I interested in class representations?

(e.g.,)

GloVe pretrained on Common Crawl with 2.2 M vocab. in 300 dimension

ZS-SBIR approaches use 125 word embeddings out of 2,200,000. (Sketch)

ZSSS approaches use 20 word embeddings out of 2,200,000. (PASCAL-5i)

In my intuition,

it should not work in practice.

MINDs Lab

Introduction:

Why am I interested in class representations?

In my intuition,

it should not work in practice.

“I am very curious about

why it should not work in practice.”

MINDs Lab

Introduction:

Why am I interested in class representations?

“I am very curious about
why it should not work in practice.”

In this presentation, I address

“they have a limitation in class representations.”

“their class representation has weakness in ***generalization***.”

MINDs Lab

Introduction:

What can we learn from this presentation?

- Our experiment shows that
  domain shift causes change in class representation of CNN.

- Limitation of language model based zero-shot learning in computer vision:
  weakness in representation for generalization.

MINDs Lab

Experiment

We compare class representations

between vision models and language models

in zero-shot semantic segmentation (ZSSS) setting

20 categories of PASCAL VOC 2012

MINDs Lab

Experiment: Categories

```
PASCAL =
['aeroplane', 'bicycle', 'bird', 'boat', 'bottle',
 'bus', 'car', 'cat', 'chair', 'cow',
 'diningtable', 'dog', 'horse', 'motorbike', 'person',
 'plant', 'sheep', 'sofa', 'train', 'monitor']
```

MINDs Lab

Experiment: Correlation Metric

Cosine similarity

Class representation is expressed in *d* dimensional vector.

e.g., 2048d in ResNet, 300d in GloVe

MINDs Lab

Experiment:

Class representation in vision models

Given a dataset with 20 categories $X = \cup_i^{20} \cup_{j \in |C_i|} \{I_{i,j}\}$ and a model $f : I \rightarrow r$ that encode an image I into a d dimensional class representation r, the i-th category representation R_i is

$$R_i = \frac{1}{|C_i|} \sum_j^{|C_i|} f(I_{i,j}) \tag{1}$$

tl;dr

i-th class representation := mean of encoded vector from all samples of i-th class

MINDs Lab

Experiment:

Class representation in vision models

We used

ResNet 34, 50 and 101 pretrained on ImageNet 1k.

We obtained class representations from

PASCAL-5i test set

and COCO-20i test set.

MINDs Lab

Experiment:

Class representation in language models

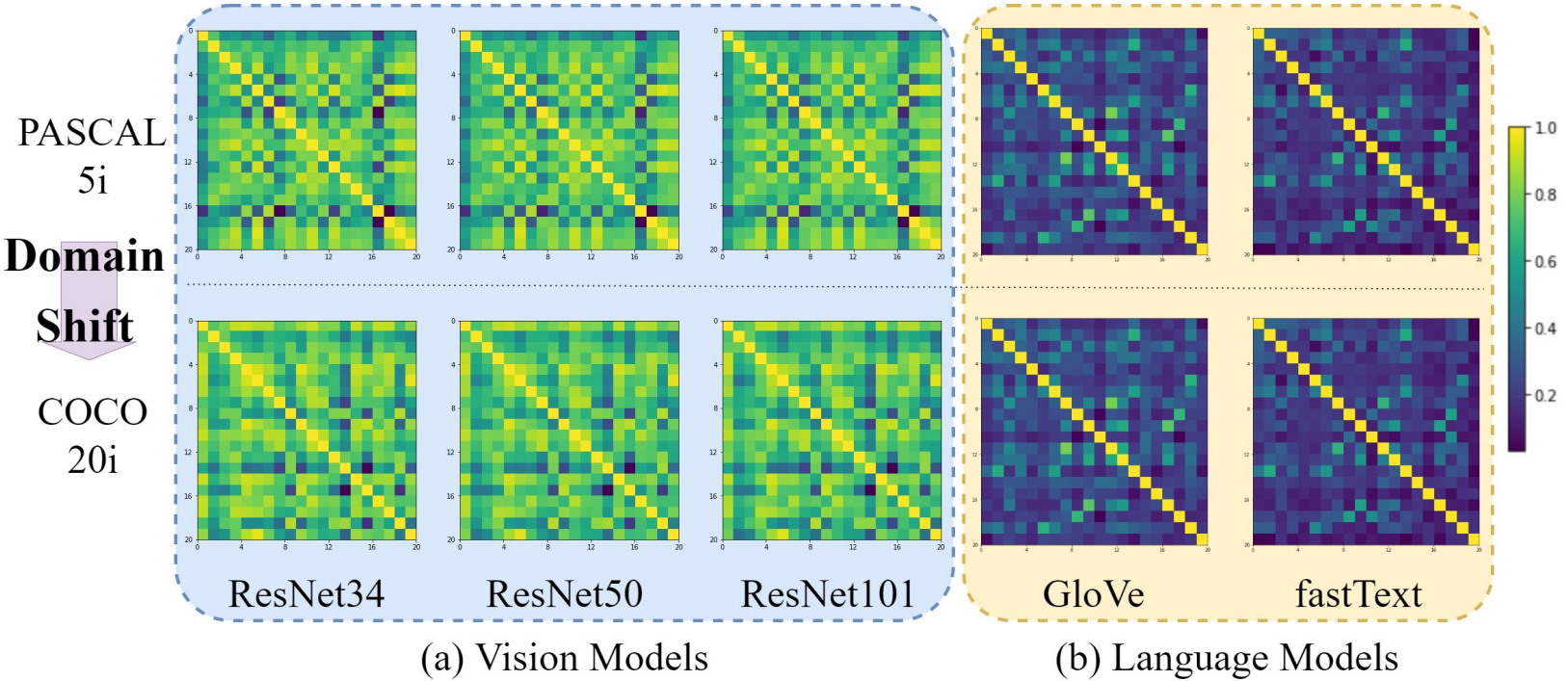
We used

GloVe pretrained on Common Crawl with 2.2M vocab in 300d,

fastText pretrained on Common Crawl with 2M vocab in 300d.

MINDs Lab

Experimental Results



MINDs Lab

Experimental Results

Representation correlation of **vision models** on the same domain show the same pattern and high similarity mean.

MINDs Lab

Experimental Results

Representation correlation of vision models on the same domain
show the same pattern and high similarity mean.

Representation correlation of **language models** trained on the same dataset
show a similar pattern and low similarity mean.

MINDs Lab

Experimental Results

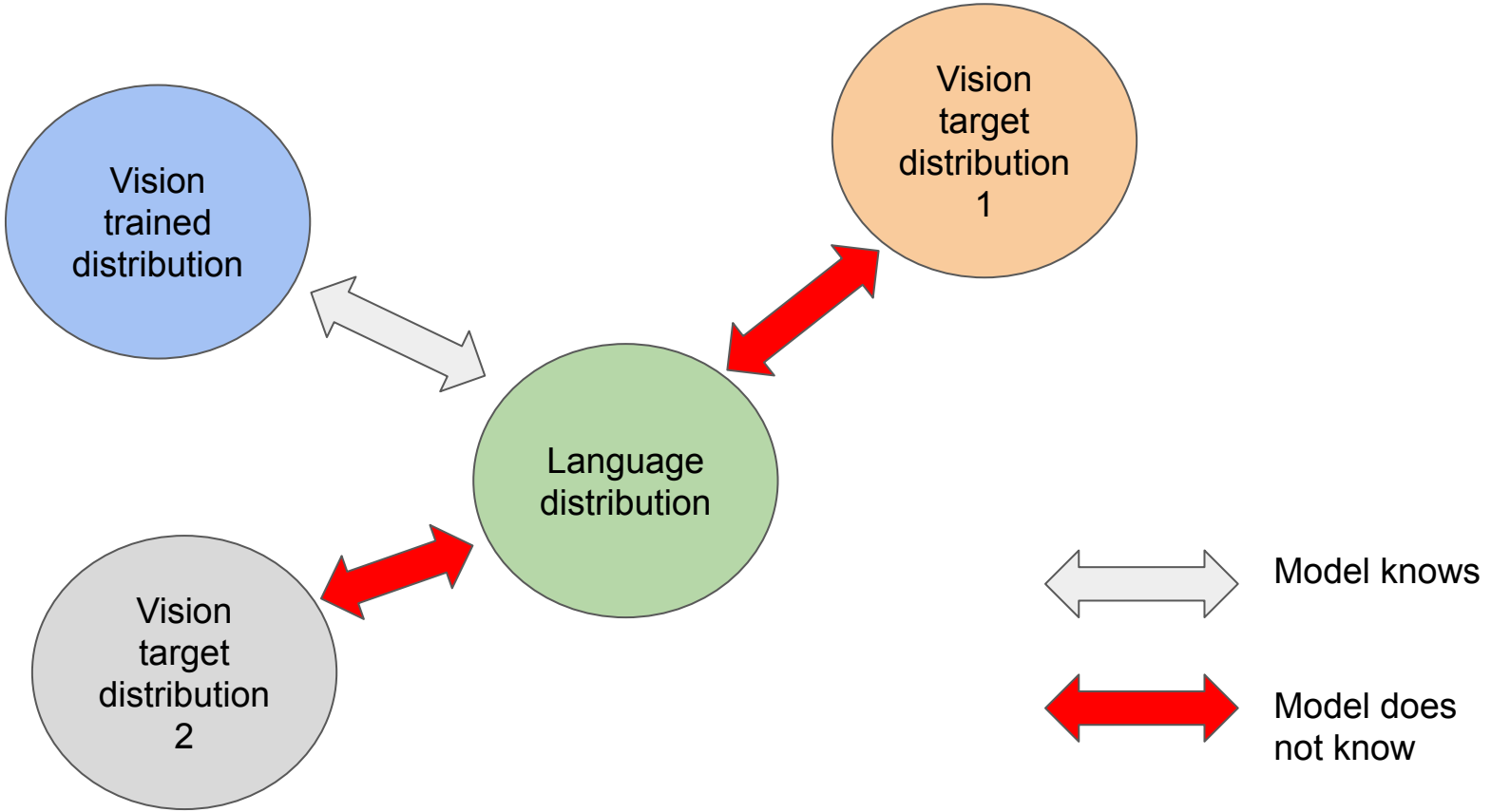
Representation correlation of vision models on the same domain
show the same pattern and high similarity mean.

Representation correlation of language models trained on the same dataset
show a similar pattern and low similarity mean.

Distribution gap exists between vision models and language models.

MINDs Lab

We showed that



MINDs Lab

Conclusion

We address that

“Language model based zero-shot learning approaches

have limitation in generalization.

In other words,

they work only on the trained domain.”

MINDs Lab