How Can We Correlate Inter-Domain Knowledge?

Reviewing

Consistent Structural Relation Learning for Generalized Zero-Shot Segmentation

Peike *et al.*, NeurIPS2020

Presentor: **Sungguk Cha**

MINDs Lab

Abstract

Recently, in image recognition tasks, approaches **using language domain knowledge directly correlating with visual knowledge** are actively researched.

In this talk, I will review an approach in which **visual feature is learnt to be similar to language domain knowledge**.

Closing, we will discuss "**is it desirable to force visual feature to be like language embedding?**"

MINDs Lab

Contents

1. Introduction to Correlating Language-Vision Knowledge in Image Recognition

2. Consistent Structural Relation Learning

3. Discussion

MINDs Lab

Introduction to
Correlating Language-Vision Knowledge in Image Recognition

- Zero-Shot Learning

- Zero-Shot Classification Approaches

- Zero-Shot Segmentation Approaches

MINDs Lab

Introduction to
Correlating Language-Vision Knowledge in Image Recognition

- Zero-Shot Learning

- Zero-Shot Classification Approaches

- Zero-Shot Segmentation Approaches

MINDs Lab

Introduction to
Correlating Language-Vision Knowledge in Image Recognition

Zero-Shot Learning?

1. Supervised Learning

2. Few-Shot Learning

3. Zero-Shot Learning

MINDs Lab

Introduction to
Correlating Language-Vision Knowledge in Image Recognition

Zero-Shot Learning?

1. Supervised Learning

2. Few-Shot Learning

3. Zero-Shot Learning

MINDs Lab

Introduction to
Correlating Language-Vision Knowledge in Image Recognition

Zero-Shot Learning?

1. Supervised Learning

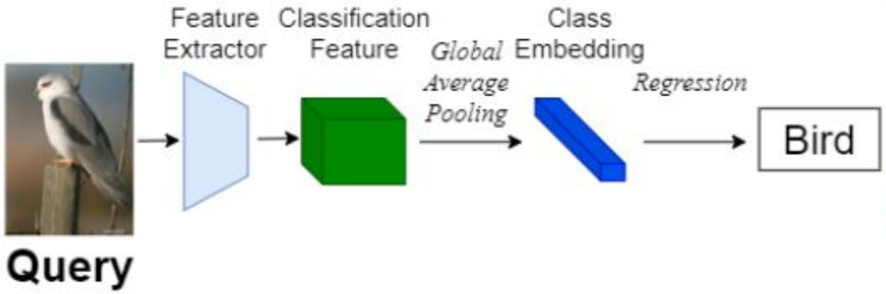


Figure: Image classification overview

Task definition:
Given an image, classify the query among ***N*** possible classes.

Method:
Given ***E***-dimension class embedding, solve ***N*** regression problems.

Challenge:
Cannot predict any class except the ***N*** classes.

MINDs Lab

Introduction to
Correlating Language-Vision Knowledge in Image Recognition

Zero-Shot Learning?

1. Supervised Learning

- Data hungry
- Cannot predict a novel class

MINDs Lab

Introduction to
Correlating Language-Vision Knowledge in Image Recognition

Zero-Shot Learning?

1. Supervised Learning

2. Few-Shot Learning

3. Zero-Shot Learning

MINDs Lab

Introduction to
Correlating Language-Vision Knowledge in Image Recognition

Zero-Shot Learning?

2. Few-Shot Learning

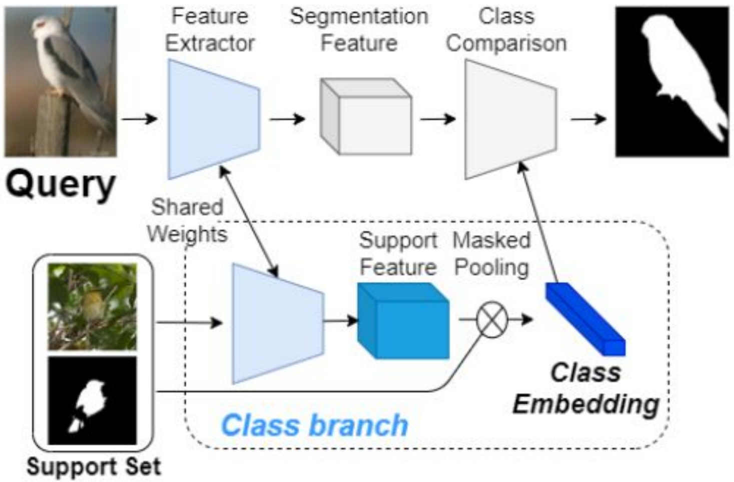


Figure: Few-shot semantic segmentation

1. Generate class embedding from support set

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

MINDs Lab

Introduction to
Correlating Language-Vision Knowledge in Image Recognition

Zero-Shot Learning?

2. Few-Shot Learning

- Learns to compare ‘support image’ and ‘query image’

Introduction to
Correlating Language-Vision Knowledge in Image Recognition

Zero-Shot Learning?

1. Supervised Learning
2. Few-Shot Learning
3. Zero-Shot Learning

MINDs Lab

Introduction to
Correlating Language-Vision Knowledge in Image Recognition

Zero-Shot Learning?

3. Zero-Shot Learning

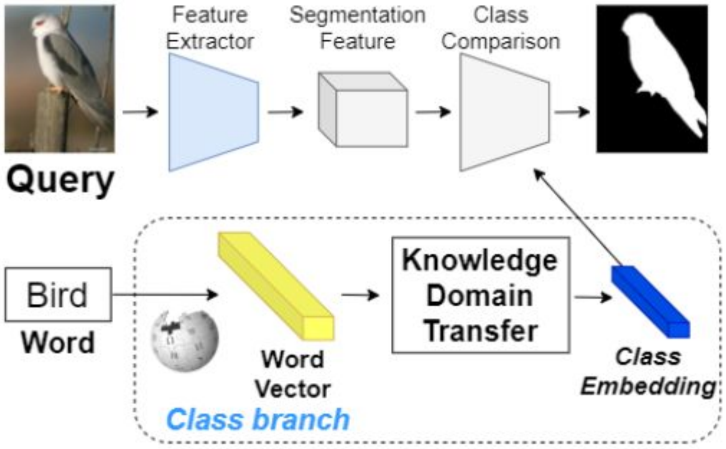


Figure: Zero-shot semantic segmentation

1. Generate class embedding from a word

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

MINDs Lab

Introduction to
Correlating Language-Vision Knowledge in Image Recognition

Zero-Shot Learning?

3. Zero-Shot Learning

- Learns to compare ‘**word vector** originated feature’ and ‘**query image**’

Introduction to
Correlating Language-Vision Knowledge in Image Recognition

- Zero-Shot Learning
- Zero-Shot Classification Approaches
- Zero-Shot Segmentation Approaches

MINDs Lab

Introduction to
Correlating Language-Vision Knowledge in Image Recognition

Zero-Shot Classification Approaches

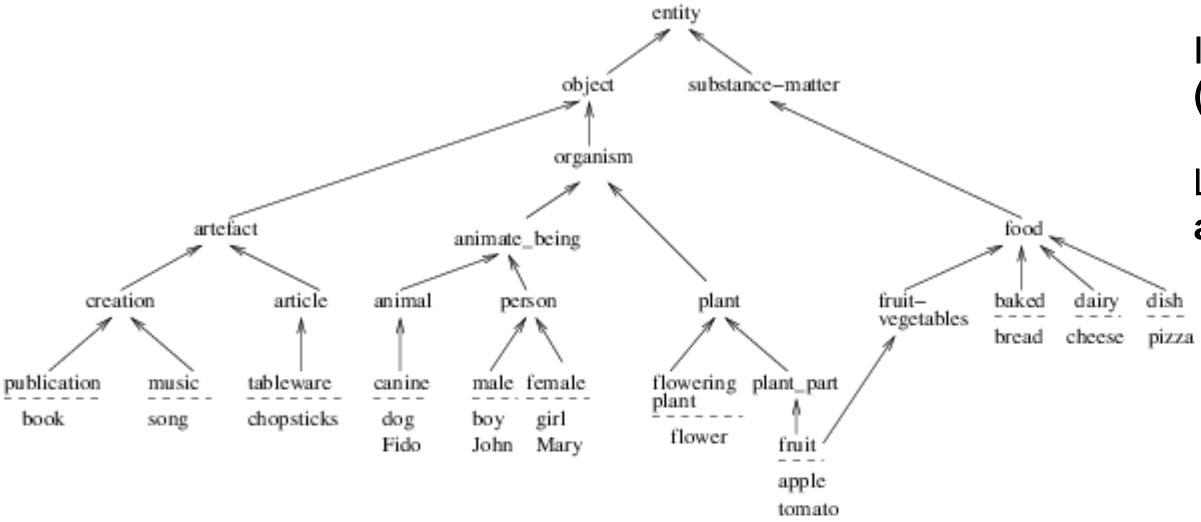
ImageNet 1k: 1,000 categories

ImageNet 23k: 23,000 categories

Imagenet source: https://syncedreview.com/2020/06/23/google-deepmind-researchers-revamp-imagenet/

Introduction to
Correlating Language-Vision Knowledge in Image Recognition

Zero-Shot Classification Approaches



ImageNet is based upon **WordNet (hierarchy)**.

Learn relationship from **word vector and hierarchy** with GNN.

Wordnet source: https://www.researchgate.net/figure/A-simplified-WordNet-hierarchy-of-synsets_fig2_228707772

Introduction to
Correlating Language-Vision Knowledge in Image Recognition

Zero-Shot Classification Approaches

- Wang *et al.,* Hyperbolic Visual Embedding Learning for Zero-Shot Recognition, CVPR 2020
- Kampffmeyer *et al.*, Rethinking knowledge graph propagation for zero-shot learning, CVPR 2019
- Liu *et al.,* Zero-shot recognition via semantic embeddings and knowledge graphs, CVPR 2018

Introduction to
Correlating Language-Vision Knowledge in Image Recognition

- Zero-Shot Learning
- Zero-Shot Classification Approaches
- Zero-Shot Segmentation Approaches

MINDs Lab

Introduction to
Correlating Language-Vision Knowledge in Image Recognition

Zero-Shot Segmentation Approaches

Low number of classes => Hard to utilize WordNet hierarchy

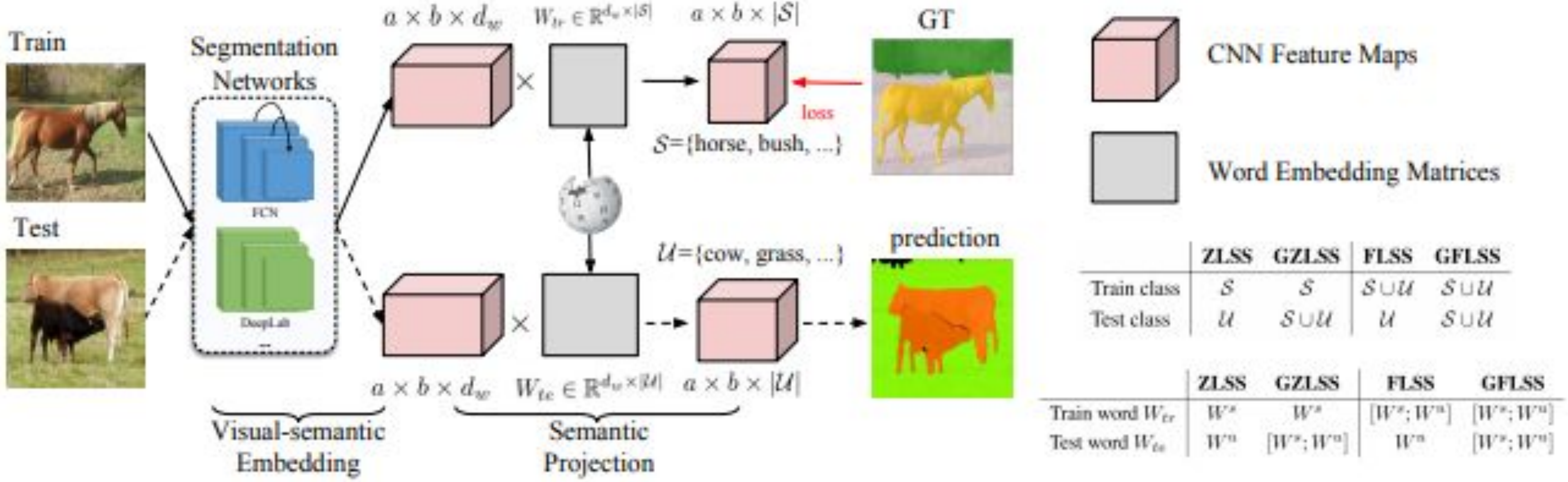
Task	Single-label Classification	Object Detection	Segmentation	Multi-label Classification
Number of Classes	ImageNet: **21,841** Open Images V5: **19,959**	COCO2017: **172** PASCAL VOC2007: **20** Open Images V5: **600**	Cityscapes: **19** PASCAL VOC2012: **20** ADE20k: **150*** PASCAL CONTEXT: **59*** Open Images V5: **350**	COCO2017: **172** PASCAL VOC2007: **20** NUS-WIDE: **128*** Open Images V5: **600**

MINDs Lab

Introduction to
Correlating Language-Vision Knowledge in Image Recognition

Zero-Shot Segmentation Approaches

Word vector as a classifier



	ZLSS	GZLSS	FLSS	GFLSS
Train class	$\mathcal{S}$	$\mathcal{S}$	$\mathcal{S} \cup \mathcal{U}$	$\mathcal{S} \cup \mathcal{U}$
Test class	$\mathcal{U}$	$\mathcal{S} \cup \mathcal{U}$	$\mathcal{U}$	$\mathcal{S} \cup \mathcal{U}$

	ZLSS	GZLSS	FLSS	GFLSS
Train word W_{tr}	W^s	W^s	$[W^s; W^u]$	$[W^s; W^u]$
Test word W_{te}	W^u	$[W^s; W^u]$	W^u	$[W^s; W^u]$

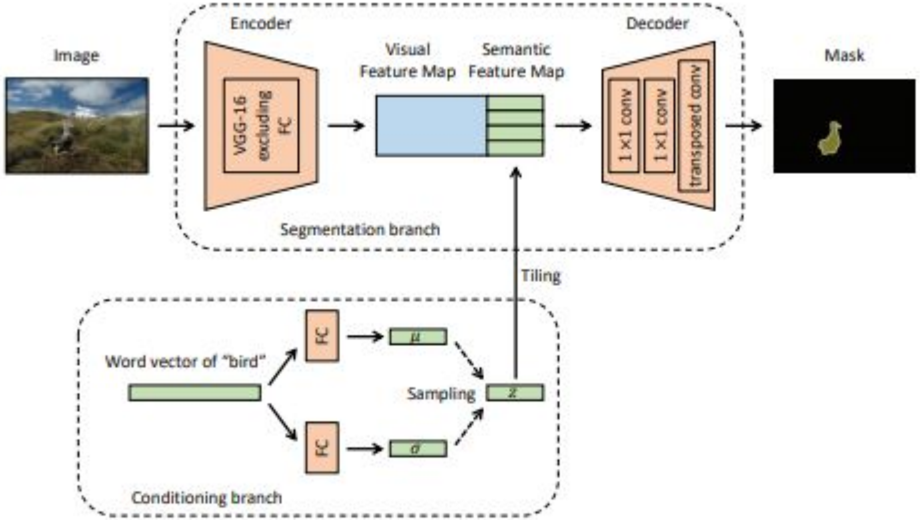
Source: Semantic Projection Network for Zero- and Few-Label Semantic Segmentation, Xian and Choudhury *et al.*

MINDs Lab

Introduction to
Correlating Language-Vision Knowledge in Image Recognition

Zero-Shot Segmentation Approaches

Word vector as a class embedding

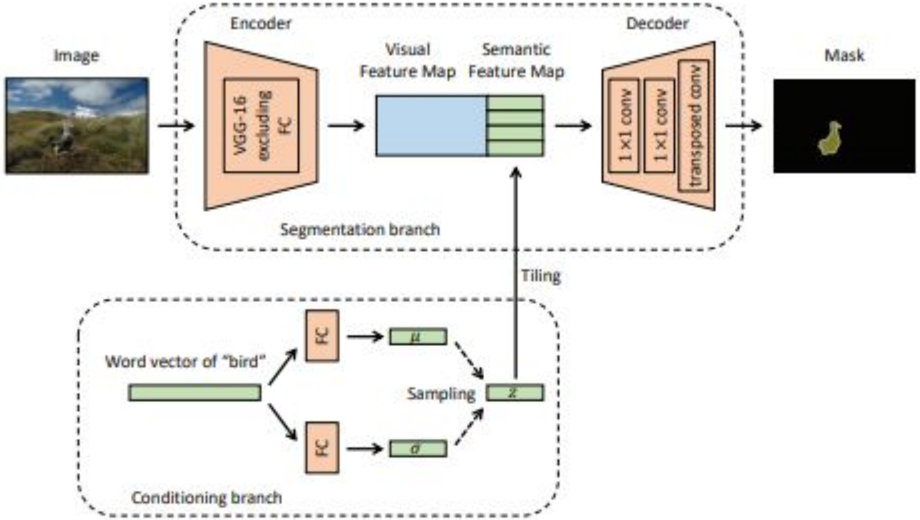


Source: Zero-Shot Semantic Segmentation via Variational Mapping, Kato *et al.*

Introduction to
Correlating Language-Vision Knowledge in Image Recognition

Zero-Shot Segmentation Approaches

Word vector as a class embedding



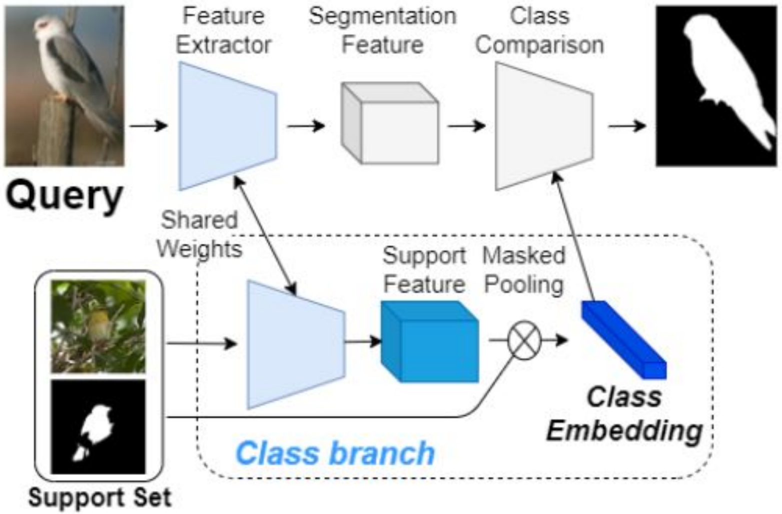


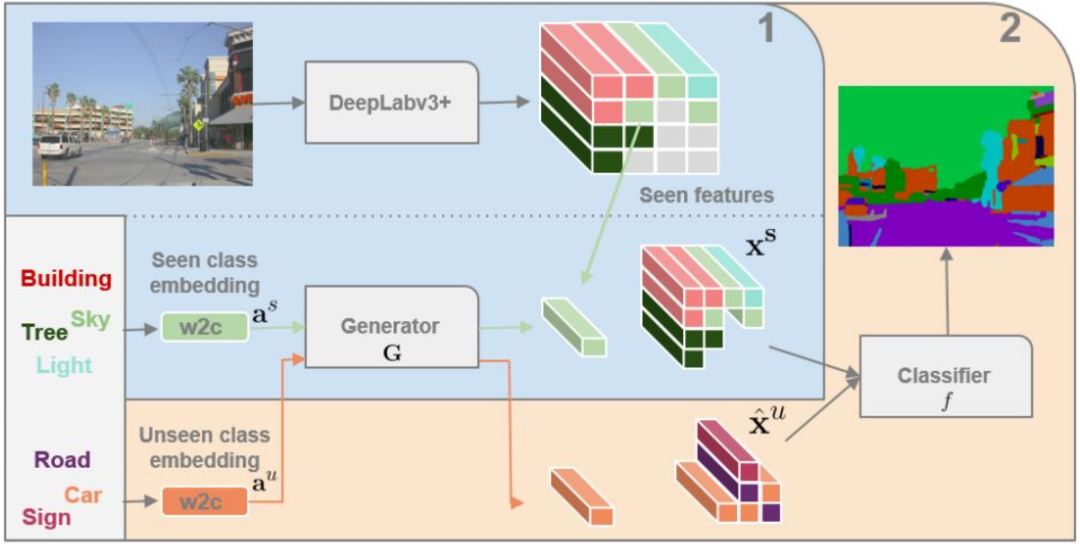
Figure: Few-shot semantic segmentation

MINDs Lab Source: Zero-Shot Semantic Segmentation via Variational Mapping, Kato *et al.*

Introduction to
Correlating Language-Vision Knowledge in Image Recognition

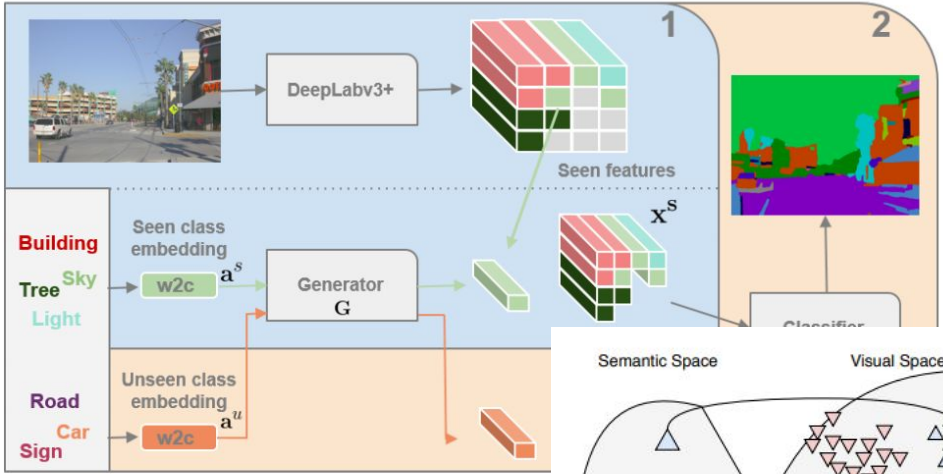
Zero-Shot Segmentation Approaches

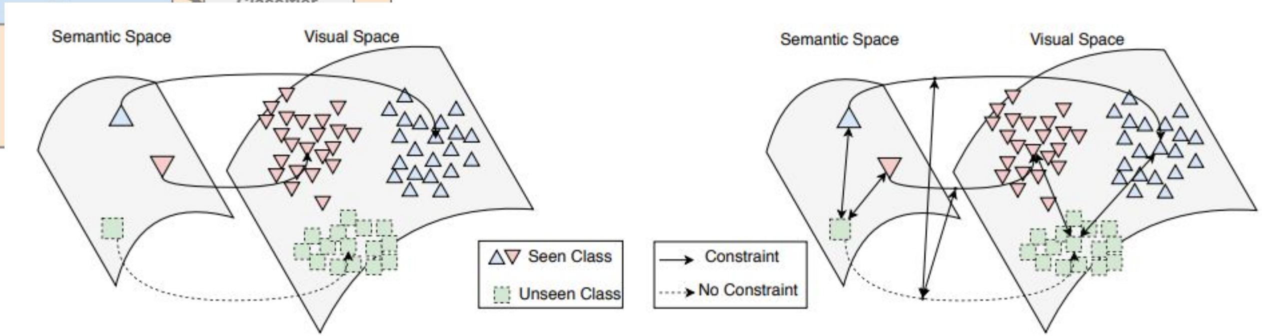
Synthesize visual feature from word vector



Source: Zero-Shot Semantic Segmentation, Bucher *et al.*

Consistent Structural Relation Learning



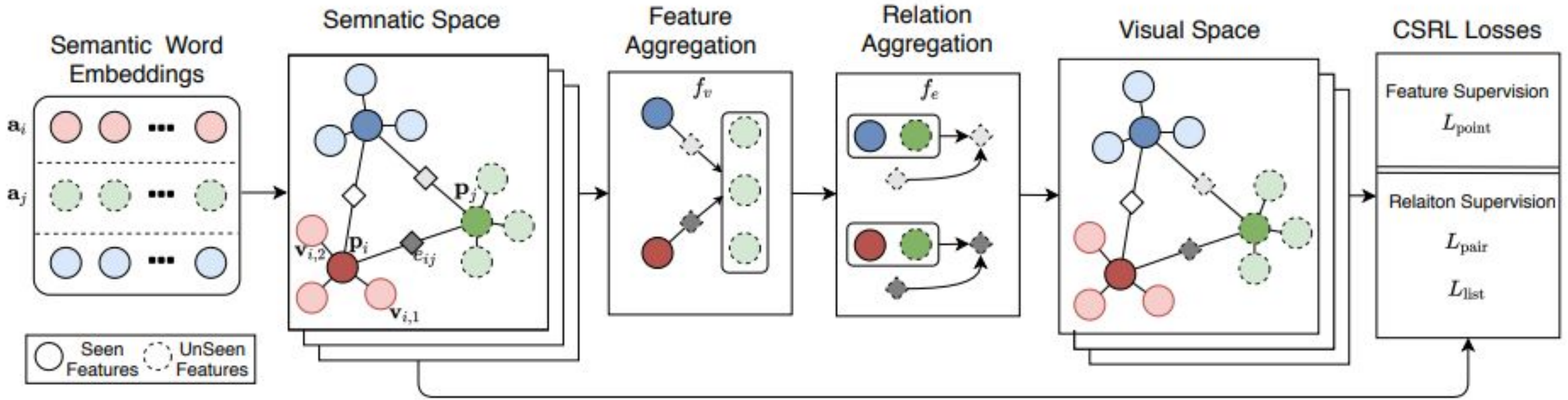


(a) Node-to-node generator

(b) Structural generator

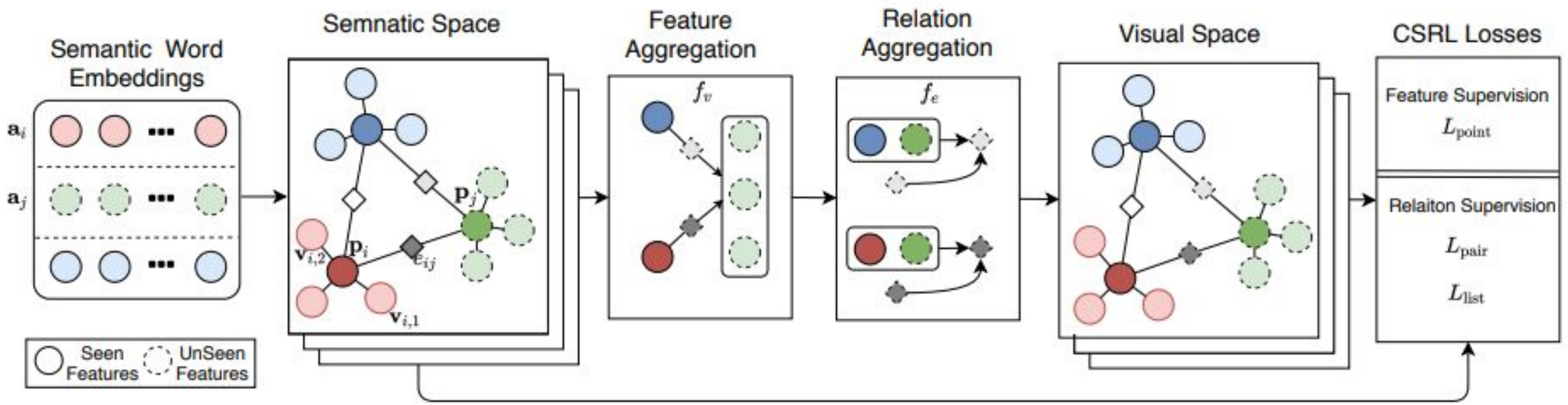
Source:
Zero-Shot Semantic Segmentation, Bucher *et al.*
Consistent Structural Relation Learning for Generalized Zero-Shot Segmentation, Li *et al.*

Consistent Structural Relation Learning



Source: Consistent Structural Relation Learning for Generalized Zero-Shot Segmentation, Li *et al.*

Consistent Structural Relation Learning



$$\mathcal{G} = (\mathcal{V}, \mathcal{E})$$
$$\mathcal{V} := \{\mathbf{v}_{i,n} | \forall i \in [1, |\mathcal{S} \cup \mathcal{U}|], n \in [1, N]\}$$
$$\mathcal{E} := \{e_{ij} | \forall i, j \in [1, |\mathcal{S} \cup \mathcal{U}|]\}$$

$\{\mathbf{a}_j | \mathbf{a}_j \in \mathcal{A}\}_{j=1}^{|\mathcal{S} \cup \mathcal{U}|}$ word vector

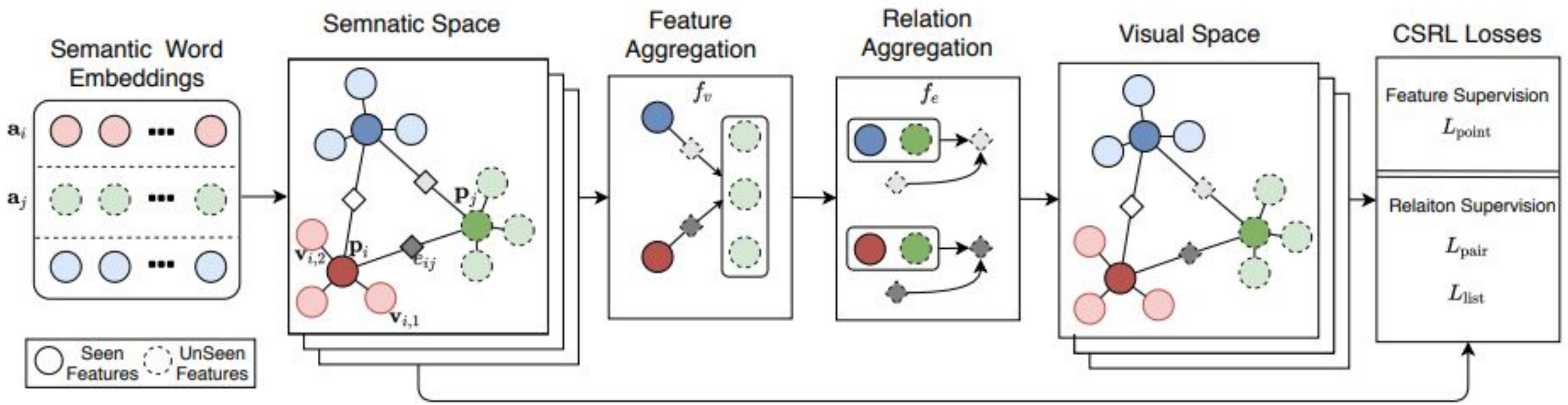
$\mathbf{v}^0_{i,n} = [\mathbf{a}_i \oplus \mathbf{z}_{i,n}]$ concat word vector and z ~ N(0, 1)

$e^0_{ij} = \mathbf{a_i} \cdot \mathbf{a_j} / \|\mathbf{a_i}\|_2 \|\mathbf{a_j}\|_2$ cosine similarity

MINDs Lab

Source: Consistent Structural Relation Learning for Generalized Zero-Shot Segmentation, Li *et al.*

Consistent Structural Relation Learning



Feature Aggregation

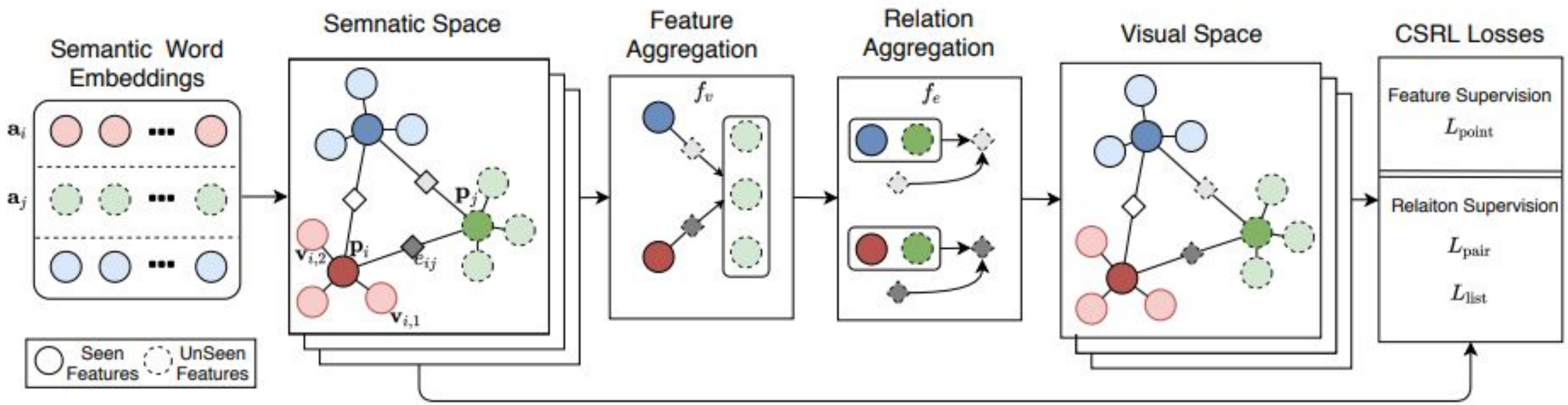
$$\mathbf{p}_i^{\ell-1} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{v}_{i,n}^{\ell-1}$$

node representative feature (prototype)

$$\mathbf{v}_{i,n}^{\ell} = f_v^{\ell}([\mathbf{v}_{i,n}^{\ell-1} \oplus \sum_{j=1, j\neq i}^{|\mathcal{S}\cup\mathcal{U}|} e_{ij}^{\ell-1}\mathbf{p}_i^{\ell-1}]; \phi_v^{\ell})$$

Source: Consistent Structural Relation Learning for Generalized Zero-Shot Segmentation, Li *et al.*

Consistent Structural Relation Learning



Relation Aggregation

$$e_{ij}^{\ell} = f_e^{\ell}(|\mathbf{p}_i^{\ell} - \mathbf{p}_j^{\ell}|; \phi_e^{\ell}) e_{ij}^{\ell-1}$$

aggregation

$$\tilde{e}_{ij}^{\ell} = \frac{\exp(e_{ij}/\gamma)}{\sum_{j'=1}^{|\mathcal{S}|} \exp(e_{ij'}/\gamma)}$$

normalization

Source: Consistent Structural Relation Learning for Generalized Zero-Shot Segmentation, Li *et al.*

MINDs Lab

Consistent Structural Relation Learning

Supervision on seen class

Make generator to synthesize feature similar to visual feature

x := visual feature
\hat{x} := **generated** visual feature

$$\mathcal{L}_{\text{point}} = \frac{1}{|\mathcal{S}|}\sum_{c=1}^{|\mathcal{S}|}[\mathbb{E}_{\mathbf{x},\mathbf{x}'\sim\mathcal{X}^c}K(\mathbf{x},\mathbf{x}') + \mathbb{E}_{\hat{\mathbf{x}},\hat{\mathbf{x}}'\sim\hat{\mathcal{X}}^c}K(\hat{\mathbf{x}},\hat{\mathbf{x}}') - 2\mathbb{E}_{\mathbf{x}\sim\mathcal{X}^c,\hat{\mathbf{x}}\sim\hat{\mathcal{X}}^c}K(\mathbf{x},\hat{\mathbf{x}})]$$

이유성 3 days ago
K(x,y)를 두 feature vector f(x)와 f(y)의 내적으로 생각하고(실제로 가능합니다),
저 MMD 식은 X에서 f(x)의 기댓값 mu_x와 hat(X)에서의 f(\hat(x))의 기댓값 mu_Y에 대해 mu_X-mu_Y의 L2 norm을 구한 것으로 해석할 수 있습니다.

이유성 3 days ago
대략 [(x1+...+xn)/n - (y1+...+ym)/m]^2을 전개한 형태로 보아도 좋아요 (edited)

Consistent Structural Relation Learning

Supervision on relationships (pair-wise consistency)

Constrain final relation to be similar to the initial relation

$$\mathcal{L}_{\text{pair}}(\mathbf{M}^{\mathcal{A}}, \mathbf{M}^{\hat{\mathcal{X}}}) = \frac{1}{|\mathcal{U}|} \sum_{i=1}^{|\mathcal{U}|} D_{\text{KL}}[\mathbf{M}_i^{\mathcal{A}} || \mathbf{M}_i^{\hat{\mathcal{X}}}]$$

$\mathbf{M} = \{e_{ij}^{\ell} | \forall i \in [1, |\mathcal{U}|], j \in [1, |\mathcal{S}|]\} \in \mathbb{R}^{|\mathcal{U}| \times |\mathcal{S}|}$ Relation matrix

$\mathbf{M}^{\mathcal{A}}$ Initial relation matrix

$\mathbf{M}^{\hat{\mathcal{X}}}$ Final relation matrix

MINDs Lab Source: Consistent Structural Relation Learning for Generalized Zero-Shot Segmentation, Li *et al.*

Consistent Structural Relation Learning

Supervision on relationships
(list-wise consistency)

Learn to preserve correlation ranking

$\pi(i)$ i-th permutation

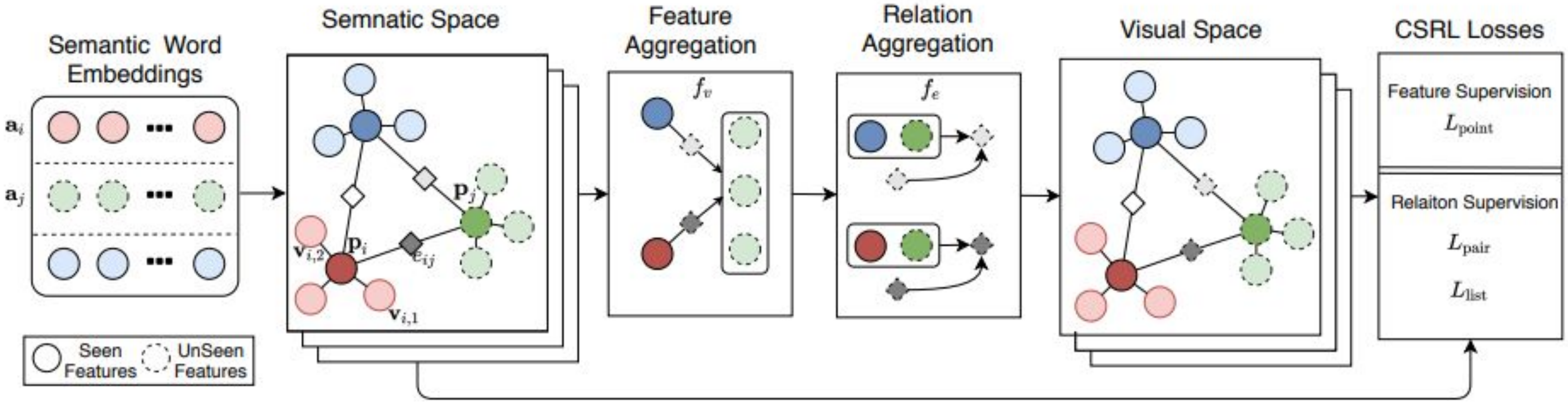
$$P(\pi|\mathbf{M}_i) = \prod_{j=1}^{|\mathcal{S}|} \frac{\exp(e_{i\pi(j)}/\gamma)}{\sum_{k=j}^{|\mathcal{S}|} \exp(e_{i\pi(k)}/\gamma)}$$

$$\mathcal{L}_{\text{list}}(\mathbf{M}^{\mathcal{A}}, \mathbf{M}^{\hat{\mathcal{X}}}) = \frac{1}{|\mathcal{U}|} \sum_{i=1}^{|\mathcal{U}|} D_{\text{KL}}[P(\pi \in \mathcal{P}|\mathbf{M}_i^{\mathcal{A}}) \| P(\pi \in \mathcal{P}|\mathbf{M}_i^{\hat{\mathcal{X}}})]$$

MINDs Lab

Source: Consistent Structural Relation Learning for Generalized Zero-Shot Segmentation, Li *et al.*

Consistent Structural Relation Learning



$$\mathcal{L}(\phi) = \mathcal{L}_{\text{point}} + \mathcal{L}_{\text{pair}} + \mathcal{L}_{\text{list}}$$

Source: Consistent Structural Relation Learning for Generalized Zero-Shot Segmentation, Li *et al.*

Consistent Structural Relation Learning

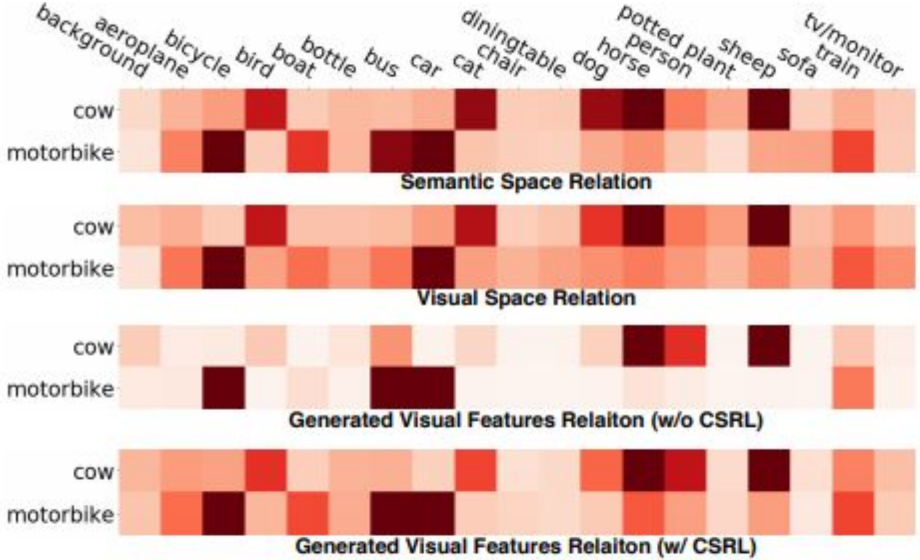


Figure 4: Relations between unseen (cow and motorbike) and seen categories.

Source: Consistent Structural Relation Learning for Generalized Zero-Shot Segmentation, Li *et al.*

MINDs Lab

Discussion

1. Is it **desirable** letting visual feature space be similar to language feature space?

2. We have no explicit intuition how visual feature space is looking like. Then, which form of the space can be **desirable**?

Source: Consistent Structural Relation Learning for Generalized Zero-Shot Segmentation, Li *et al.*