SCIENCES SORBONNE UNIVERSITÉ

GROUNDED REPRESENTATION LEARNING FOR NLP & APPLICATIONS TO ZERO-SHOT LEARNING

Workshop on Representation Learning for Complex Data,
University Lyon 2

Friday 24th May, 2019

Éloi Zablocki

LIP6

MLIA Machine Learning & Deep Learning for Information Access

The presented research involves:




- → Patrick Gallinari
- → Benjamin Piwowarski
- → Laure Soulier
- → Patrick Bordes

Introduction

Representation learning for NLP

Encoding word semantics —Applications

- → Language understanding (QA, summarization, NER, POS Tagging, sentiment analysis)
- → Machine translation
- → Statistical language modeling (speech recognition, dialog systems)
- → Zero-Shot Learning
- → ...

Representation learning for NLP

SCIENCES SORBONNE UNIVERSITÉ

Encoding word semantics —Applications

- → Language understanding (QA, summarization, NER, POS Tagging, sentiment analysis)
- → Machine translation
- → Statistical language modeling (speech recognition, dialog systems)
- → Zero-Shot Learning
- → ...

Distribution Semantic Models

- → Idea: Represent the *semantic* of a word in a vector encoding the distribution of the word contexts.
- → Hypothesis: Based on the *distributional hypothesis* (Harris 1954):

  "Words that occur in similar contexts should have similar meanings"
- → Property: Semantic similarity is often quantified by spatial proximity or related geometric measures

Why do we need multi-modal word embeddings?

SCIENCES SORBONNE UNIVERSITÉ

Context

→ Language is ambiguous, biased and lacks common-sense...

Why do we need multi-modal word embeddings?

SCIENCES SORBONNE UNIVERSITÉ

Context

→ Language is ambiguous, biased and lacks common-sense...

Example

→ When we say:

"Hand me the salt"

Why do we need multi-modal word embeddings?

SCIENCES SORBONNE UNIVERSITÉ

Context

→ Language is ambiguous, biased and lacks common-sense...

Example

→ When we say:

"Hand me the salt"

→ We mean:

"Hand me the salt ... or rather the receptacle that contains it. It has a cylindrical shape and is about 3 inches high. This object does not float in the air and lies on the table, in other words, in direct contact with it. If I ask you this favor, it means that it is closer to you than it is to me. Besides, when you give it to me, the aperture should be on the top so that the salt is not spoiled on the table because of gravity, which makes that dropped objects fall down. When you hand it to me, I expect that you give it to by making contact with my hand, located at the end of my arm, that I am going to bring closer to you. Salt is an ionic compound that can be formed by the neutralization reaction of an acid and a base. It is usually extracted from salted water, from seas and ocean. Salt is essential for life in general, and saltiness is one of the basic human tastes. Salt is one of the oldest and most ubiquitous food seasonings. The World Health Organization recommends that adults should consume less than 2000 mg of sodium, equivalent to 5 grams of salt per day. Edible salt is sold

Why do we need multi-modal word embeddings?

Context

→ Language is ambiguous, biased and lacks common-sense...

unlikely to be mentioned	likely to be mentioned
something expected	unusual facts
trivial facts	something with value

Langage ≠ Real world

Why do we need multi-modal word embeddings?

SCIENCES SORBONNE UNIVERSITÉ

Context

→ Language is ambiguous, biased and lacks common-sense...

unlikely to be mentioned	likely to be mentioned
something expected trivial facts	unusual facts something with value

$$\boxed{\text{Langage} \neq \text{Real world}}$$

→ ...while images are unequivocal depictions

Why do we need multi-modal word embeddings?

SCIENCES SORBONNE UNIVERSITÉ

Context

→ Language is ambiguous, biased and lacks common-sense...

unlikely to be mentioned	likely to be mentioned
something expected	unusual facts
trivial facts	something with value

$$\boxed{\text{Langage} \neq \text{Real world}}$$

→ ...while images are unequivocal depictions

Can language be *grounded* in the visual world?

Why do we need multi-modal word embeddings?

SCIENCES SORBONNE UNIVERSITÉ

Context

→ Language is ambiguous, biased and lacks common-sense...

unlikely to be mentioned	likely to be mentioned
something expected trivial facts	unusual facts something with value

$$\boxed{\text{Langage} \neq \text{Real world}}$$

→ ...while images are unequivocal depictions

Can language be *grounded* in the visual world?

Motivation

→ *Psychological studies*: meaning of words is grounded in perception

Why do we need multi-modal word embeddings?

SCIENCES SORBONNE UNIVERSITÉ

Context

→ Language is ambiguous, biased and lacks common-sense...

unlikely to be mentioned	likely to be mentioned
something expected trivial facts	unusual facts something with value

$$\boxed{\text{Langage} \neq \text{Real world}}$$

→ ...while images are unequivocal depictions

Can language be *grounded* in the visual world?

Motivation

→ *Psychological studies*: meaning of words is grounded in perception

→ *Human Reporting Bias* (van Durme 2013)

Why do we need multi-modal word embeddings?

SCIENCES SORBONNE UNIVERSITÉ

Context

→ Language is ambiguous, biased and lacks common-sense...

unlikely to be mentioned	likely to be mentioned
something expected trivial facts	unusual facts something with value

$$\boxed{\text{Langage} \neq \text{Real world}}$$

→ ...while images are unequivocal depictions

Can language be *grounded* in the visual world?

Motivation

→ *Psychological studies*: meaning of words is grounded in perception

→ *Human Reporting Bias* (van Durme 2013)

→ *Complementarity* between text and images

Why do we need multi-modal word embeddings?

Context

→ Language is ambiguous, biased and lacks common-sense...

unlikely to be mentioned	likely to be mentioned
something expected trivial facts	unusual facts something with value

Langage ≠ Real world

→ ...while images are unequivocal depictions

Can language be *grounded* in the visual world?

Motivation

→ *Psychological studies*: meaning of words is grounded in perception

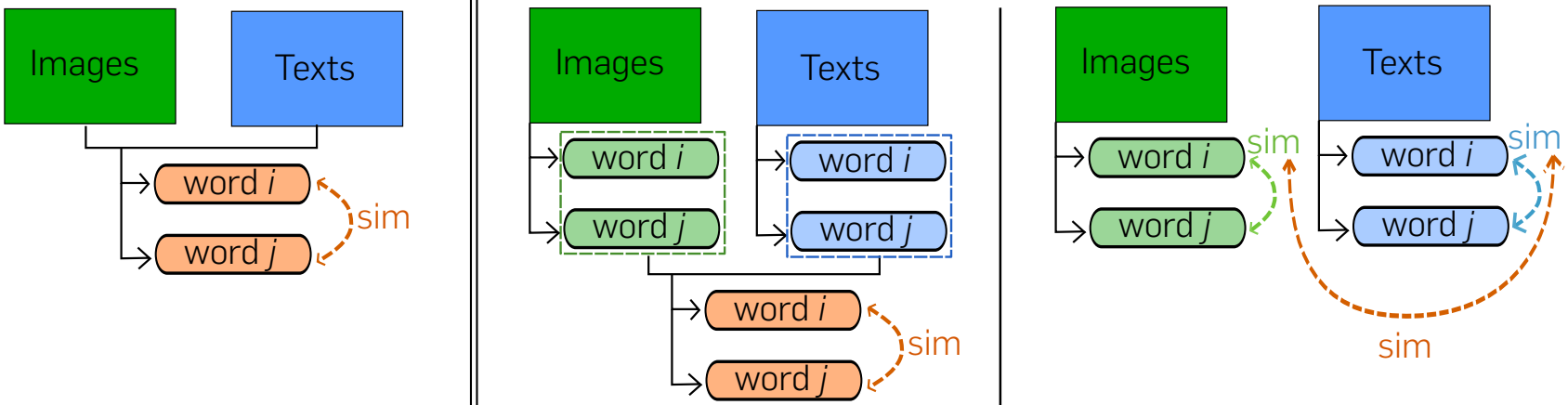
→ *Human Reporting Bias* (van Durme 2013)

→ *Complementarity* between text and images

→ Advances in *Computer Vision*

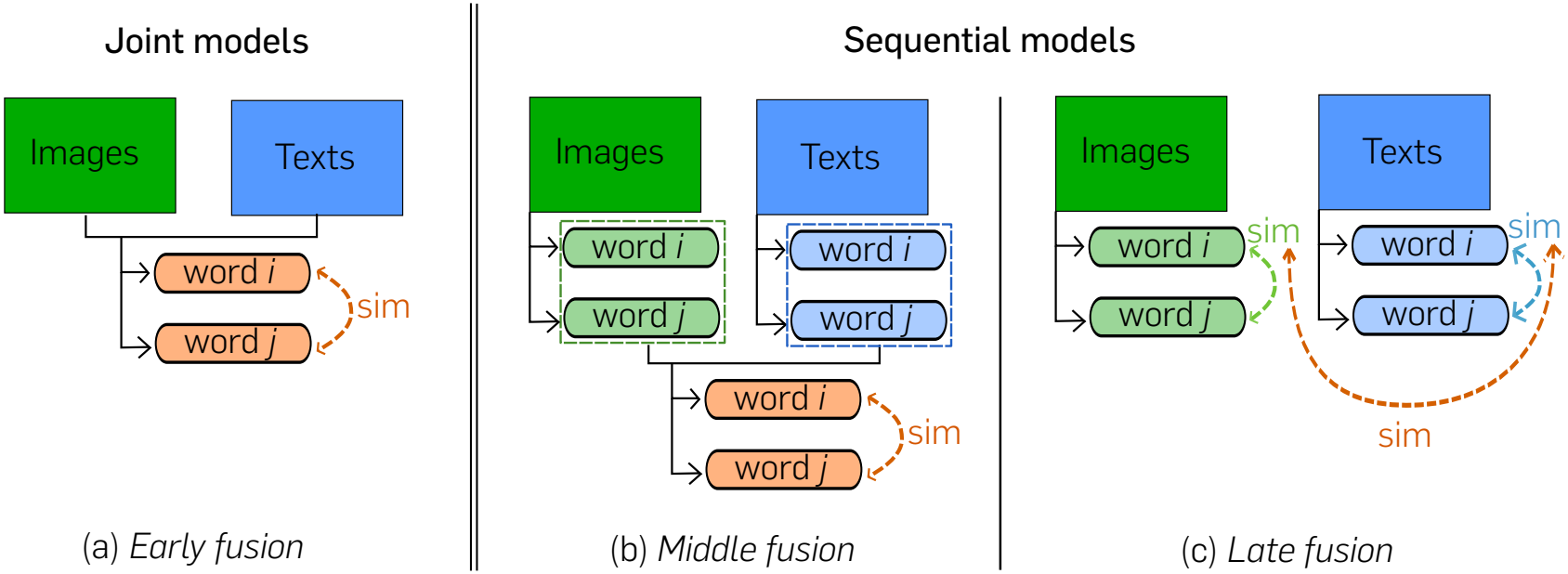
→ Focus on learning word representation

Multi-modal fusion techniques - Sequential models



(a) *Early fusion* (b) *Middle fusion* (c) *Late fusion*

Multi-modal fusion techniques - Sequential models

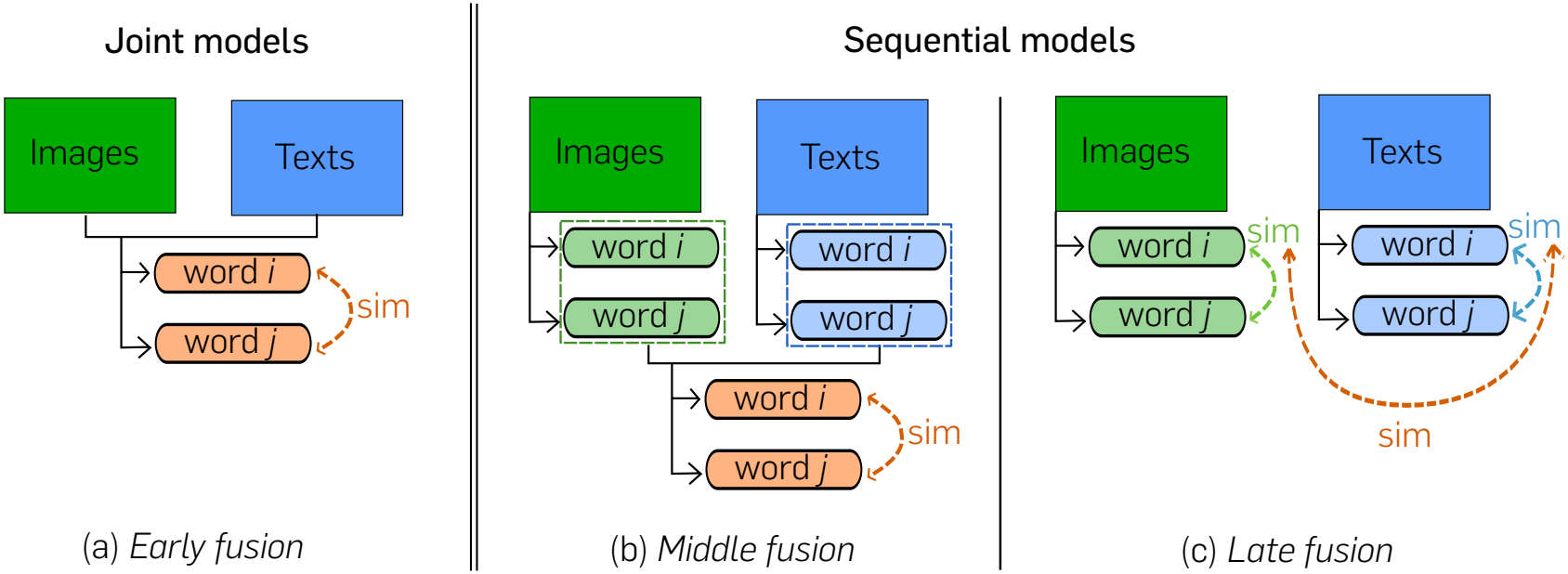


(a) *Early fusion* (b) *Middle fusion* (c) *Late fusion*

Sequential models

1. Separately construct representations for words in each modality
   - → **Text**: GloVe or Word2Vec
   - → **Image**: Aggregation of activations obtained from a pre-trained CNN forwarded on images

Multi-modal fusion techniques - Sequential models

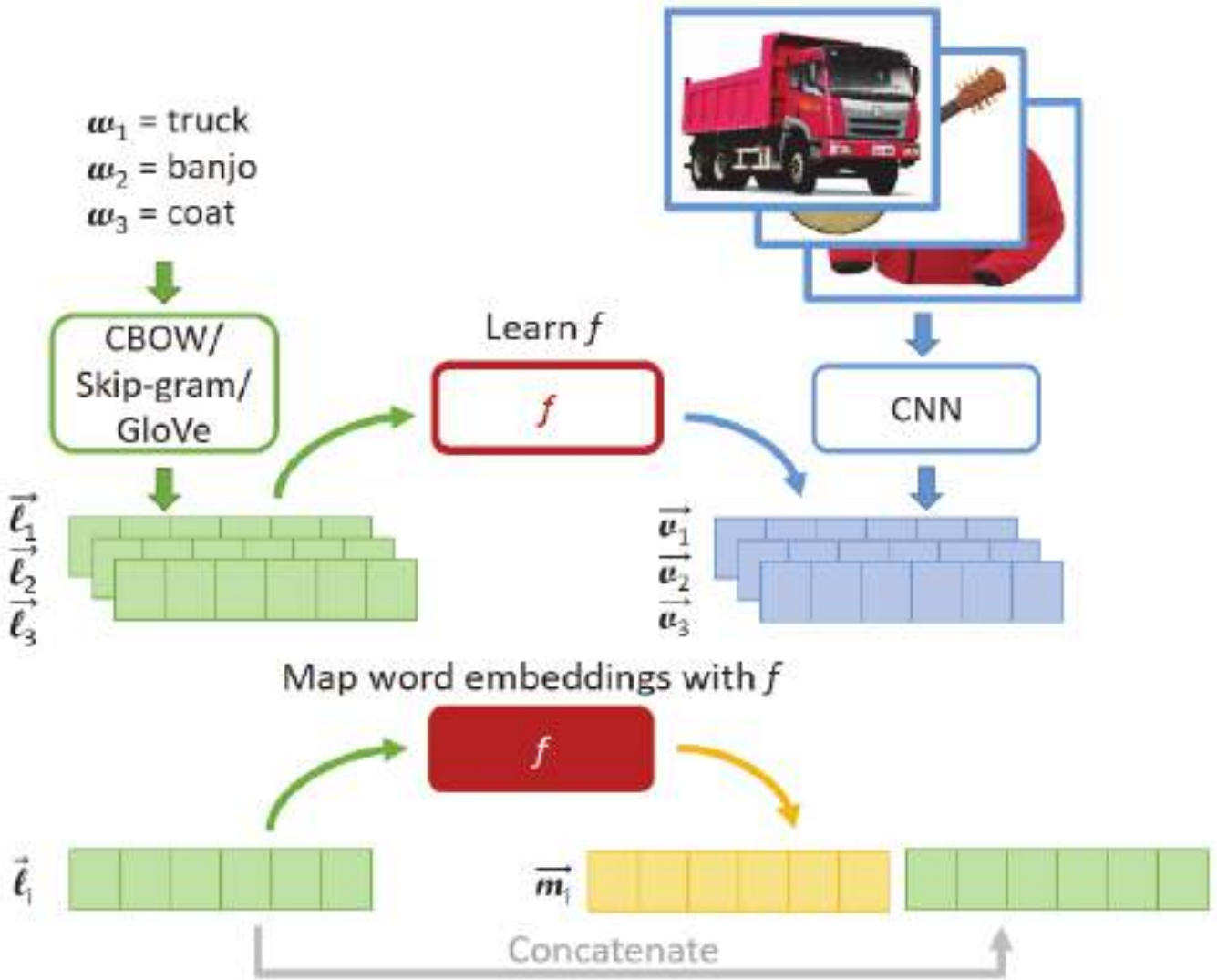


(a) *Early fusion* (b) *Middle fusion* (c) *Late fusion*

Sequential models

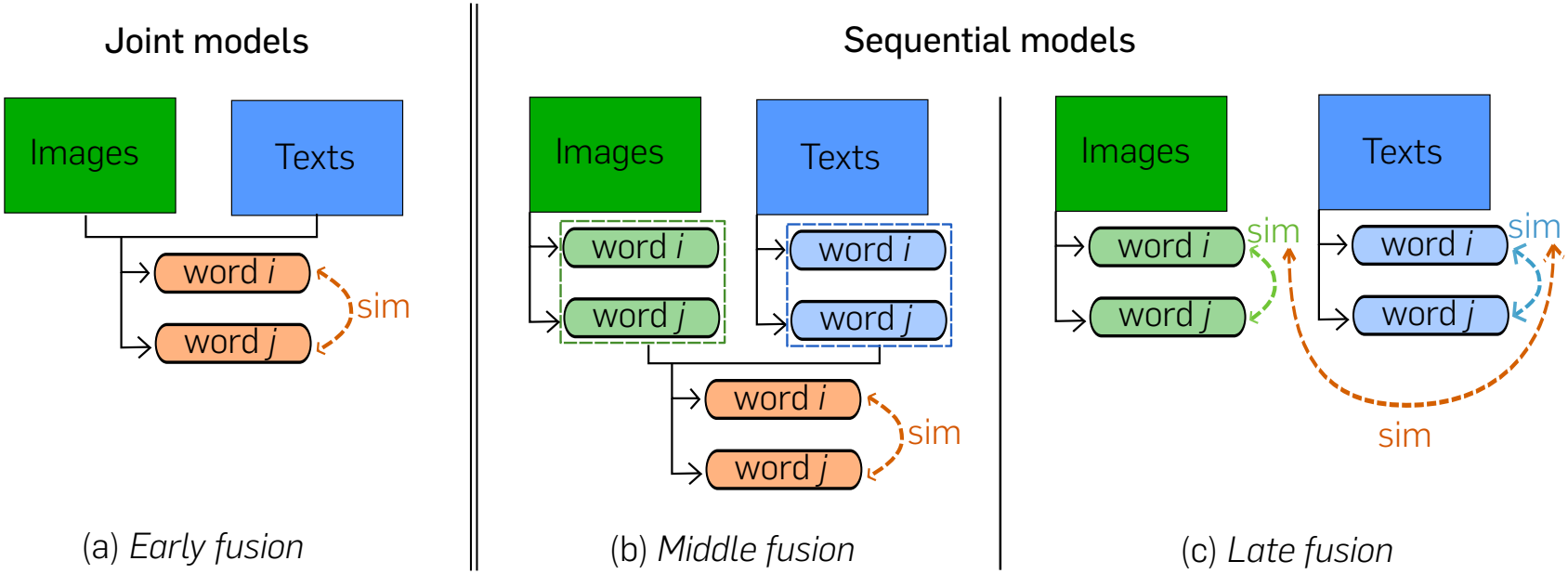
1. Separately construct representations for words in each modality
   - → **Text**: GloVe or Word2Vec
   - → **Image**: Aggregation of activations obtained from a pre-trained CNN forwarded on images
2. Combine them with different techniques
   - → **Middle fusion**: form multi-modal representations (e.g. concatenation, CCA)
   - → **Late fusion**: interaction in the downstream task (e.g. linear combination of similarity scores)

Example of sequential technique (middle fusion)



Sequential model (Collell et al. 2017)

Multi-modal fusion techniques - Joint models



(a) *Early fusion* (b) *Middle fusion* (c) *Late fusion*

Joint models (a.k.a. early fusion)

Jointly learn multi-modal representations from multiple modalities

→ With aligned text and images: bayesian models, grounded language models...

→ Without aligned text and images: skip-gram extension

Multi-modal fusion techniques

Joint models without aligned texts and images

Extensions to the *skip-gram* algorithm under the **distributional hypothesis**:

"Words that occur in similar contexts should have similar meanings"

Additional hypothesis in (Hill et al. 2014): "The frequency of appearance of concrete concepts in texts correlates with the likelihood of experiencing it in the real world."

Text corpus

Feature-Norm for concrete words

	has_legs	can_fly	has_teeth	is_green	is_animal	...
monkey	1	0	1	0	1	...
crocodile	1	0	1	1	1	...
plane	0	1	0	0	0	...
...	...	...	...			

"He saw a crocodile in his backyard!"

→ *Add fake sentence generated with feature-norms*

Fake sentence: "crocodile legs crocodile teeth crocodile green"

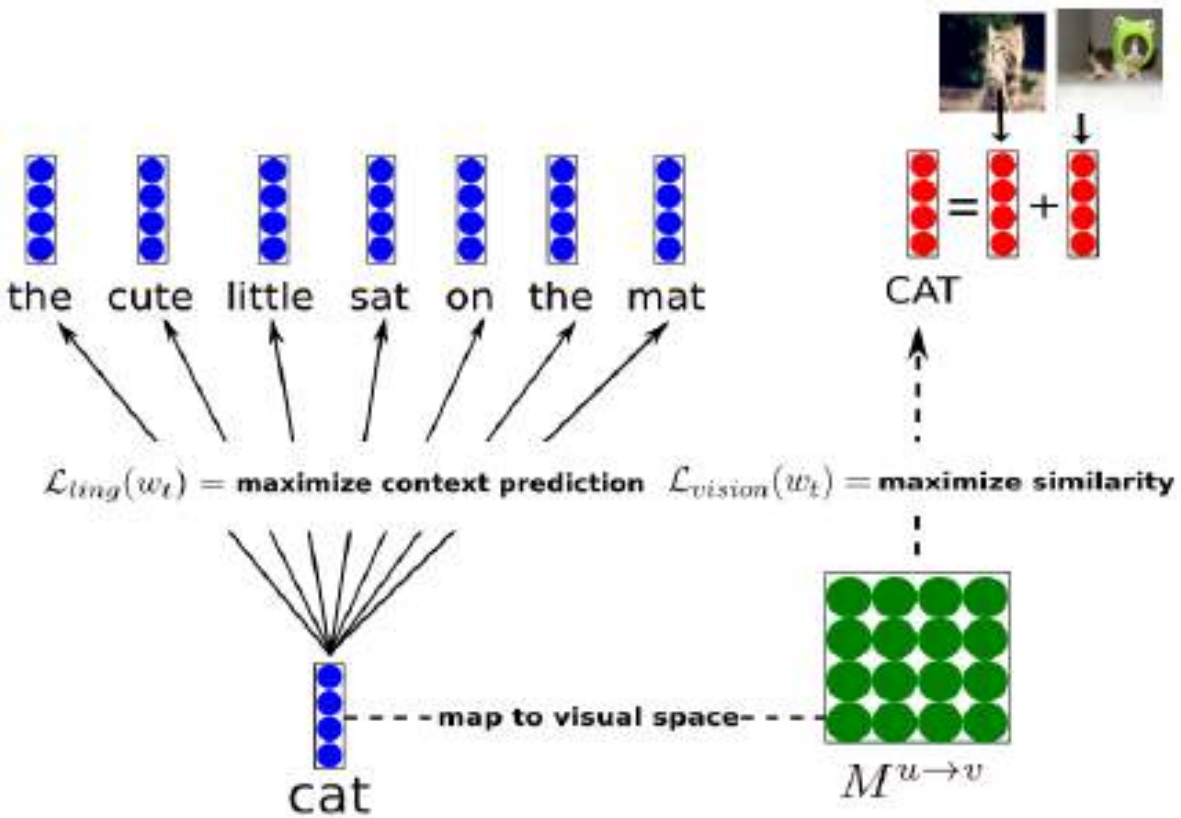
Multi-modal fusion techniques

Joint models without aligned texts and images

Extension of (Lazaridou et al. 2015)

→ Use deep visual features instead of feature norms

→ Use a triplet loss: $\sum_{v^-} \max\left(0, \gamma - \cos(u, v) + \cos(u, v^-)\right)$



Evaluating word representations (1/3)

SCIENCES SORBONNE UNIVERSITÉ

→How can we **evaluate** word representations?

Word Similarity

Spearman/Pearson correlation between cosine similarity and human judgement.
SimLex, *WordSim*, *SemSim*, *VisSim*, *MEN*, ...

word 1	word 2	human judgement	model (cosine)
cat	dog	0.76	0.86
stupid	dumb	0.91	0.52
cloud	book	0.13	-0.11
attach	join	0.56	0.17
advise	baseball	0.04	0.11
...	...	...	...

Evaluating word representations (2/3)

Feature-Norm prediction

- → Predicting feature-norms (e.g. *is red*, *can fly*) of concrete objects.
- → Features are regrouped in 9 categories: encyclopedic, function, sound, tactile, taste, taxonomic, color, form_and_surface, motion (Collell et al. 2016).
  - → Textual embedding better for: encyclopedic, function
  - → Visual embedding better for: motion, form_and_surface, color

	has_legs	can_fly	has_teeth	is_green	is_animal	...
monkey	1	0	1	0	1	...
crocodile	1	0	1	1	1	...
plane	0	1	0	0	0	...
...	...	...	...			

Evaluating word representations (3/3)

SCIENCES SORBONNE UNIVERSITÉ

Concreteness prediction

word	concreteness (human)
dog	0.97
cloud	0.72
war	0.32
hope	0.22

Analogy prediction

Predict analogies

Paris	France	London	England
man	woman	king	queen
do	doing	eat	eating
...	...	...	...

Several benchmarks:

https://github.com/EloiZ/embedding_evaluation

Learning Multi-Modal Word Representation Grounded in Visual Context

Research questions

Learning Multi-Modal Word Representation Grounded in Visual Context,
É. Zablocki, B. Piwowarski, L. Soulier, P. Gallinari, AAAI 2018.

Observation

→ Multimodal models have shown **complementarity** of text and language ...

→ ...but use **direct features** from objects and ignore **visual context**

Research questions

Learning Multi-Modal Word Representation Grounded in Visual Context,
É. Zablocki, B. Piwowarski, L. Soulier, P. Gallinari, AAAI 2018.

Observation

→ Multimodal models have shown **complementarity** of text and language ...

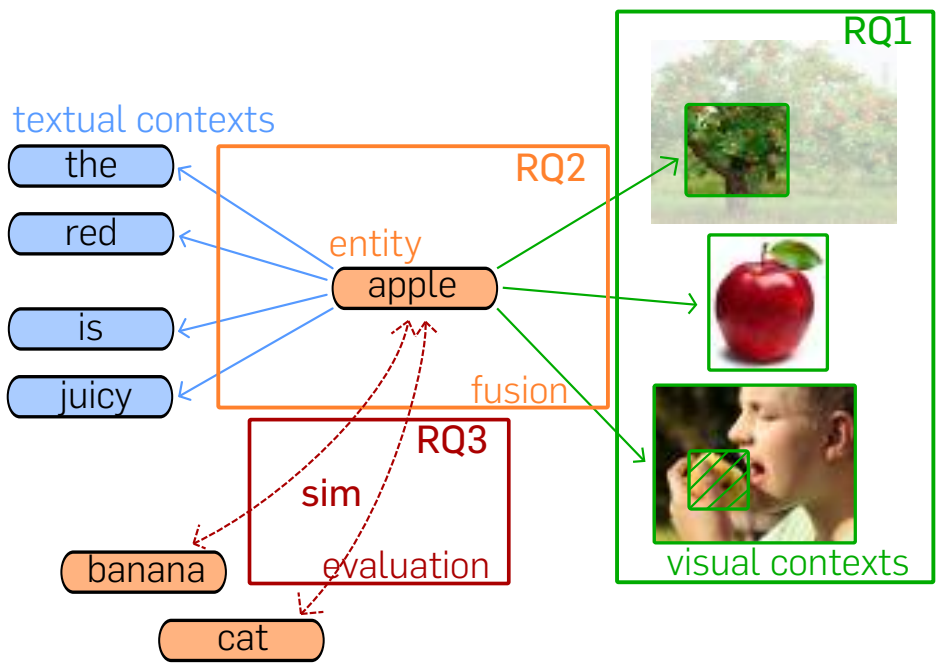
→ ...but use **direct features** from objects and ignore **visual context**

Research Questions

→ **RQ1**: What is a visual context and how can we model it?

→ **RQ2**: How can we learn representations jointly from texts and images using contexts?

→ **RQ3**: How can we evaluate the contribution of the visual modality to the final embeddings?



Visual skip-gram: model

Recall: Text skip-gram

$$\mathcal{L}_{\text{text}} = -\sum_{e \in \mathcal{D}} \sum_{c \in \mathcal{C}_e} \left[\log \sigma(t_e^\top \cdot u_c) + \sum_{c^-} \log \sigma(-t_e^\top \cdot u_{c^-}) \right]$$

T, U embedding tables, σ sigmoid function, $\mathcal{C}_e$ set of contexts of entity e

Visual skip-gram: model

Recall: Text skip-gram

$$\mathcal{L}_{\text{text}} = -\sum_{e \in \mathcal{D}} \sum_{c \in \mathcal{C}_e} \left[\log \sigma(t_e^\top \cdot u_c) + \sum_{c^-} \log \sigma(-t_e^\top \cdot u_{c^-}) \right]$$

T, U embedding tables, σ sigmoid function, $\mathcal{C}_e$ set of contexts of entity e

Idea: Use a skip-gram objective with visual contexts

$$\mathcal{L}_{\text{image}} = -\sum_{e \in \mathcal{D}} \sum_{c \in \mathcal{C}_e} \left[\log \sigma(t_e^\top \cdot f_\theta(c)) + \sum_{c^-} \log \sigma(-t_e^\top \cdot f_\theta(c^-)) \right]$$

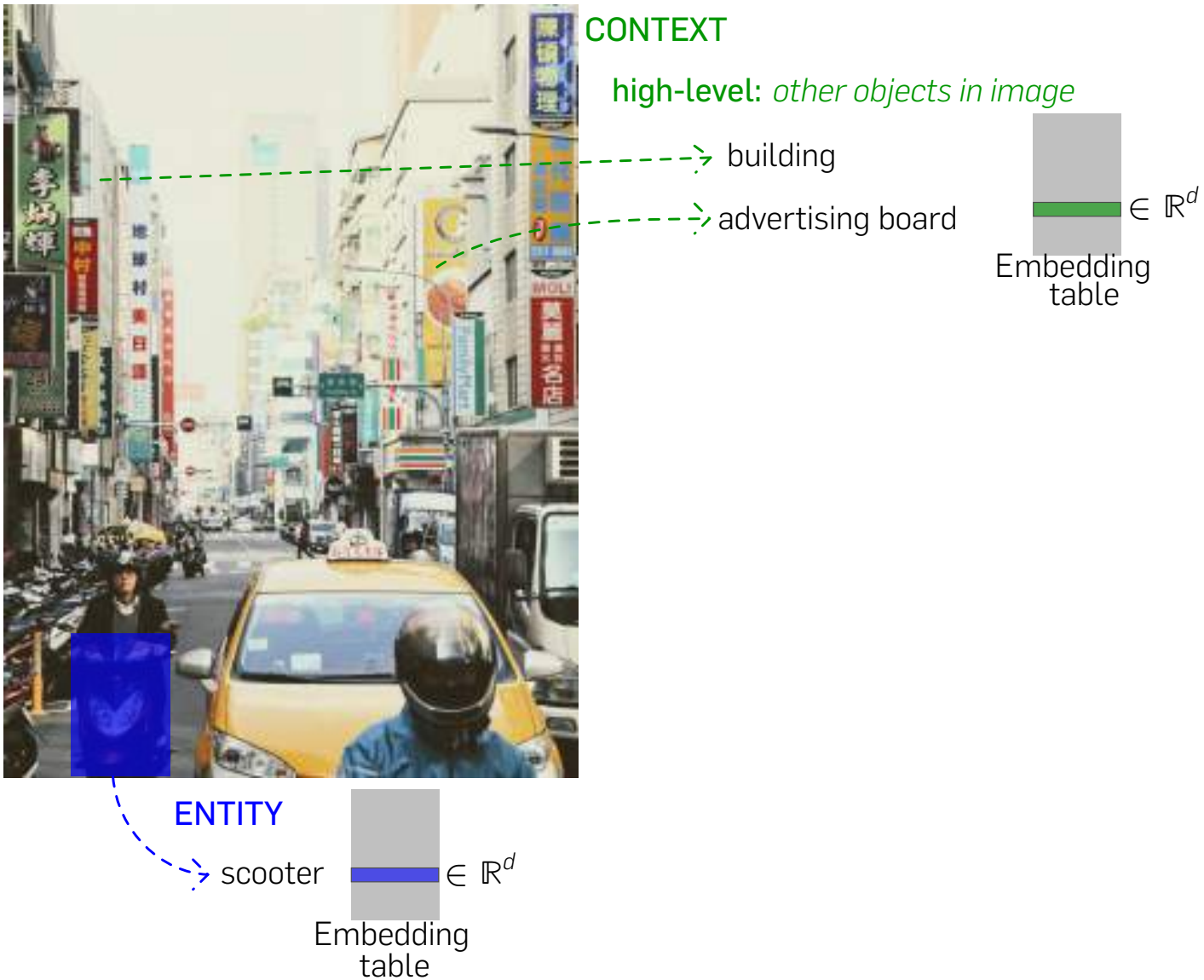
→ What is $\mathcal{C}_e$?

→ What is $c \in \mathcal{C}_e$?

→ What is $f_\theta(c)$?

Visual skip-gram: instantiation

Visual skip-gram: instantiation



Visual skip-gram: instantiation

CONTEXT

high-level: *other objects in image*

building

advertising board

$\in \mathbb{R}^d$

Embedding table

low-level:

- full-image where entity is masked out

- random patches around entity

Visual context

CNN

$\in \mathbb{R}^{2048}$

N

$\in \mathbb{R}^d$

projection

extract visual features

ENTITY

scooter

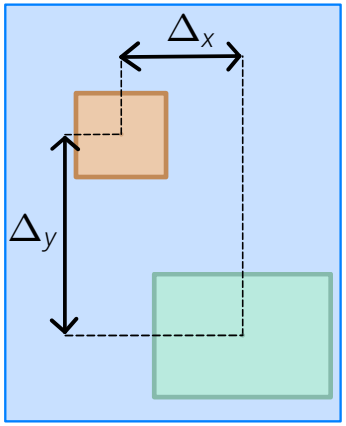
$\in \mathbb{R}^d$

Embedding table

Visual skip-gram: spatial information

Can we use **spatial information**?

$$\mathcal{L}_{\text{image}} = -\sum_{e \in \mathcal{D}} \sum_{c \in \mathcal{C}_e} \left[\log \sigma(t_e^\top \cdot f_\theta(c)) + \sum_{c^-} \log \sigma(-t_e^\top \cdot f_\theta(c^-)) \right]$$



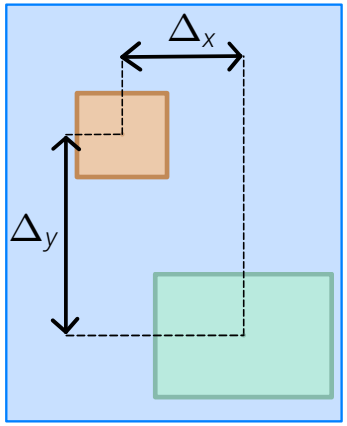
Can we enhance word representation with visual spatial information?

Form a spatial vector $s_{(e,c)}$ and integrate it in $f_\theta(c)$ to form $f_\theta^{sp}(c, s_{(e,c)})$

Visual skip-gram: spatial information

Can we use **spatial information**?

$$\mathcal{L}_{\text{image}} = -\sum_{e \in \mathcal{D}} \sum_{c \in \mathcal{C}_e} \left[\log \sigma(t_e^\top \cdot f_\theta(c)) + \sum_{c^-} \log \sigma(-t_e^\top \cdot f_\theta(c^-)) \right]$$



Can we enhance word representation with visual spatial information?

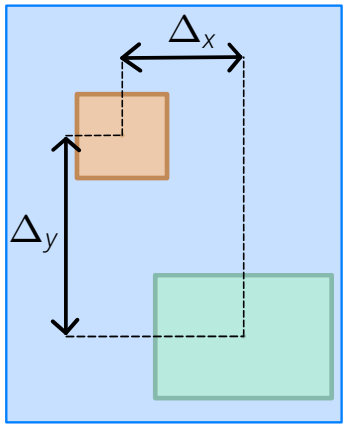
Form a spatial vector $s_{(e,c)}$ and integrate it in $f_\theta(c)$ to form $f_\theta^{sp}(c, s_{(e,c)})$

$s_{(e,c)}$	spatial information in a 4-d vector, two modelings $\rightarrow$ **Low-level**: $s_{(e,c)} = (\Delta_x, \Delta_y, \Delta_{\text{width}}, \Delta_{\text{height}})$ $\rightarrow$ **High-level**: $s_{(e,c)} = (\mathbb{1}_{\text{above}}, \mathbb{1}_{\text{below}}, \mathbb{1}_{\text{beside}}, \mathbb{1}_{\text{bigger}})$

Visual skip-gram: spatial information

Can we use **spatial information**?

$$\mathcal{L}_{\text{image}} = -\sum_{e\in\mathcal{D}}\sum_{c\in\mathcal{C}_e}\left[\log\sigma(t_e^\top\cdot f_\theta(c)) + \sum_{c^-}\log\sigma(-t_e^\top\cdot f_\theta(c^-))\right]$$



Can we enhance word representation with visual spatial information?

Form a spatial vector $s_{(e,c)}$ and integrate it in $f_\theta(c)$ to form $f_\theta^{sp}(c, s_{(e,c)})$

$s_{(e,c)}$	spatial information in a 4-d vector, two modelings → **Low-level**: $s_{(e,c)} = (\Delta_x, \Delta_y, \Delta_{\text{width}}, \Delta_{\text{height}})$ → **High-level**: $s_{(e,c)} = (\mathbb{1}_{\text{above}}, \mathbb{1}_{\text{below}}, \mathbb{1}_{\text{beside}}, \mathbb{1}_{\text{bigger}})$
f_θ^{sp}	integration of spatial feature, two integration methods → **Linear**: $f_\theta^{sp}(c, s_{(e,c)}) = M.(u_c \oplus s_{(e,c)})$, where $M \in \mathbb{R}^{d\times(d+4)}$ → **Bilinear**: $f_\theta^{sp}(c, s_{(e,c)}) = s_{(e,c)} M u_c$, where $M \in \mathbb{R}^{4\times d\times d}$

Multi-modal skip-gram

Model

$$\mathcal{L}(T, U, \theta) = \mathcal{L}_{\text{text}}(T, U) + \alpha \mathcal{L}_{\text{image}}(T, \theta), \text{ where } \alpha \in [0, 1]$$

- → T **shared** multi-modal word embeddings
- → U **textual** context parameters
- → θ **visual** context parameters
- → T, U and θ learned with SGD ; α found with cross-validation

Grounding words in visual context: Experiments

Evaluation

→ **Word Similarity**: correlation between cosine similarity and human judgement. e.g. *sim*('cat', 'dog') = 0.8 ; *sim*('cloud', 'book') = 0.1 ;

Grounding words in visual context: Experiments

Evaluation

- → **Word Similarity**: correlation between cosine similarity and human judgement. e.g. *sim*('cat', 'dog') = 0.8 ; *sim*('cloud', 'book') = 0.1 ;
- → **Feature-Norm Prediction**: predict objects' attributes from embedding with a linear SVM. e.g. *has_legs*('cat') = True, *is_red*('dog') = False

Grounding words in visual context: Experiments

Evaluation

- → **Word Similarity**: correlation between cosine similarity and human judgement. e.g. *sim*('cat', 'dog') = 0.8 ; *sim*('cloud', 'book') = 0.1 ;
- → **Feature-Norm Prediction**: predict objects' attributes from embedding with a linear SVM. e.g. *has_legs*('cat') = True, *is_red*('dog') = False
- → **Concreteness Prediction**: predict words' concreteness with a linear SVM. e.g. *conc*('dog')=0.9, *conc*('life')=0.1

Grounding words in visual context: Experiments

Evaluation

- → **Word Similarity**: correlation between cosine similarity and human judgement. e.g. *sim*('cat', 'dog') = 0.8 ; *sim*('cloud', 'book') = 0.1 ;
- → **Feature-Norm Prediction**: predict objects' attributes from embedding with a linear SVM. e.g. *has_legs*('cat') = True, *is_red*('dog') = False
- → **Concreteness Prediction**: predict words' concreteness with a linear SVM. e.g. *conc*('dog')=0.9, *conc*('life')=0.1

Data

- → Wikipedia (4.5 million articles)
- → Visual Genome (108k images, 30 objets per image)

Multi-modal skip-gram: results

Results

→ **Multi-modal embeddings** > text-only embeddings

→ 9% average improvement on all evaluation benchmarks

Multi-modal skip-gram: results

Results

- → **Multi-modal embeddings** > text-only embeddings
  - → 9% average improvement on all evaluation benchmarks
- → **Visual context** (objects surroundings) > Visual features from object
  - → 3.2% average improvement on word similarity tasks

Multi-modal skip-gram: results

Results

- → **Multi-modal embeddings** > text-only embeddings
  - → 9% average improvement on all evaluation benchmarks
- → **Visual context** (objects surroundings) > Visual features from object
  - → 3.2% average improvement on word similarity tasks
- → Visual context is **complementary** to visual features from objects
  - → ensemble model performs 6% better

Multi-modal skip-gram: results

Results

- → **Multi-modal embeddings** > text-only embeddings
  - → 9% average improvement on all evaluation benchmarks
- → **Visual context** (objects surroundings) > Visual features from object
  - → 3.2% average improvement on word similarity tasks
- → Visual context is **complementary** to visual features from objects
  - → ensemble model performs 6% better
- → **Spatial information** helps to improve word representations.
  - → and all spatial modelings and integration perform equally

Multi-modal skip-gram: results

Results

- → **Multi-modal embeddings** > text-only embeddings
  - → 9% average improvement on all evaluation benchmarks
- → **Visual context** (objects surroundings) > Visual features from object
  - → 3.2% average improvement on word similarity tasks
- → Visual context is **complementary** to visual features from objects
  - → ensemble model performs 6% better
- → **Spatial information** helps to improve word representations.
  - → and all spatial modelings and integration perform equally
- → **High-level context** > low-level context
  - → 1% average improvement on all tasks

Multi-modal skip-gram: results

Results

- → **Multi-modal embeddings** > text-only embeddings
  - → 9% average improvement on all evaluation benchmarks
- → **Visual context** (objects surroundings) > Visual features from object
  - → 3.2% average improvement on word similarity tasks
- → Visual context is **complementary** to visual features from objects
  - → ensemble model performs 6% better
- → **Spatial information** helps to improve word representations.
  - → and all spatial modelings and integration perform equally
- → **High-level context** > low-level context
  - → 1% average improvement on all tasks
- → **Joint model** > Sequential CCA
  - → 5% average improvement

Conclusion and Perspectives

Conclusion

→ Images can help NLP ...

→ ...for richer semantics to captured in word representations ...

→ with knowledge about visual contexts and spatial organization.

Perspectives

→ Evaluation on downstream tasks

→ Leverage other perceptual information sources (audio, KB, ...)

→ Grounded sentences representations (videos)

Changing perspectives

Two complementary approaches

→ Images can help for learning NLP representations

Changing perspectives

Two complementary approaches

→ Images can help for learning NLP representations

Conversely,

→ Using NLP representations can help Computer Vision

Changing perspectives

Two complementary approaches

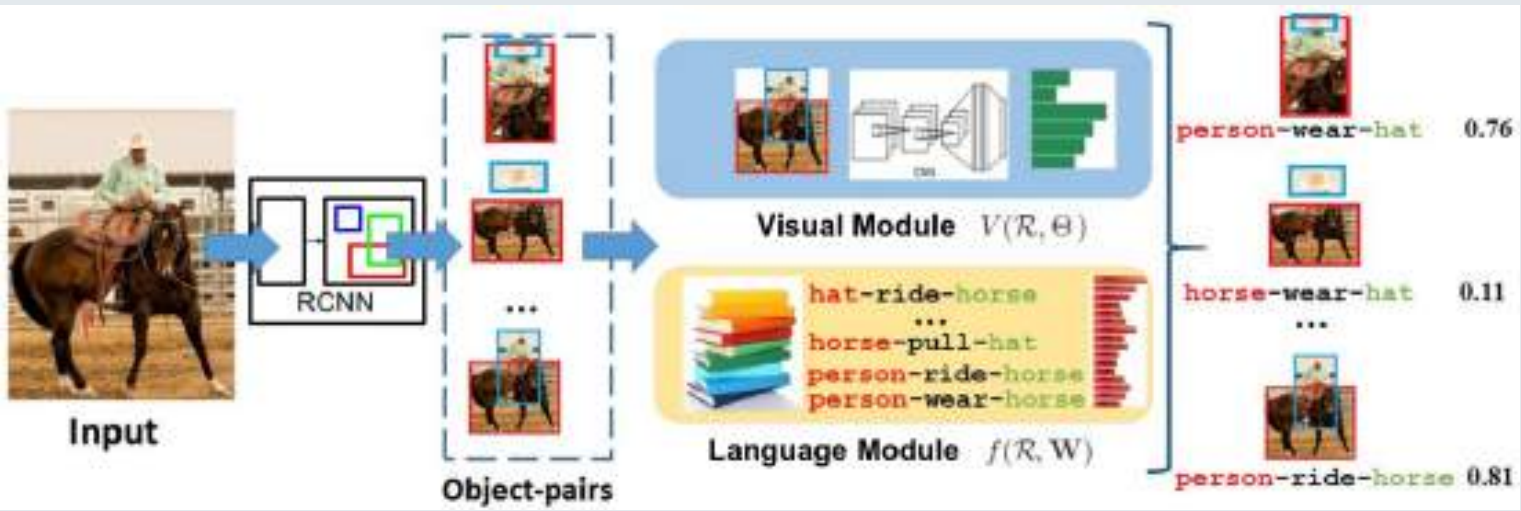
→ Images can help for learning NLP representations

Conversely,

→ Using NLP representations can help Computer Vision

NLP can assist computer vision, e.g. when no/few supervision available

→ Visual relationship detection



Visual Relationship Detection with Language Priors (from [Lu et al. 2016])

Changing perspectives

Two complementary approaches

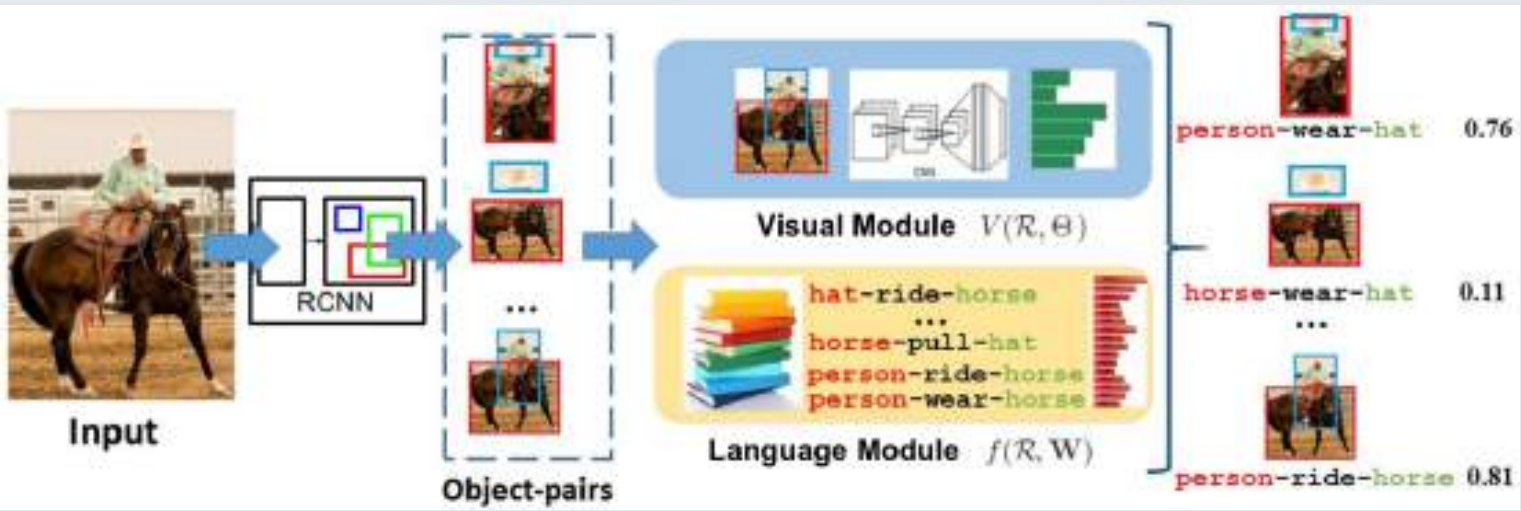
→ Images can help for learning NLP representations

Conversely,

→ Using NLP representations can help Computer Vision

NLP can assist computer vision, e.g. when no/few supervision available

→ Visual relationship detection



Visual Relationship Detection with Language Priors (from [Lu et al. 2016])

→ **Zero-shot learning**

Zero-Shot Recognition with Semantic Representations and Context

Zero-Shot Learning —Task & Hypothesis

Two domains ...

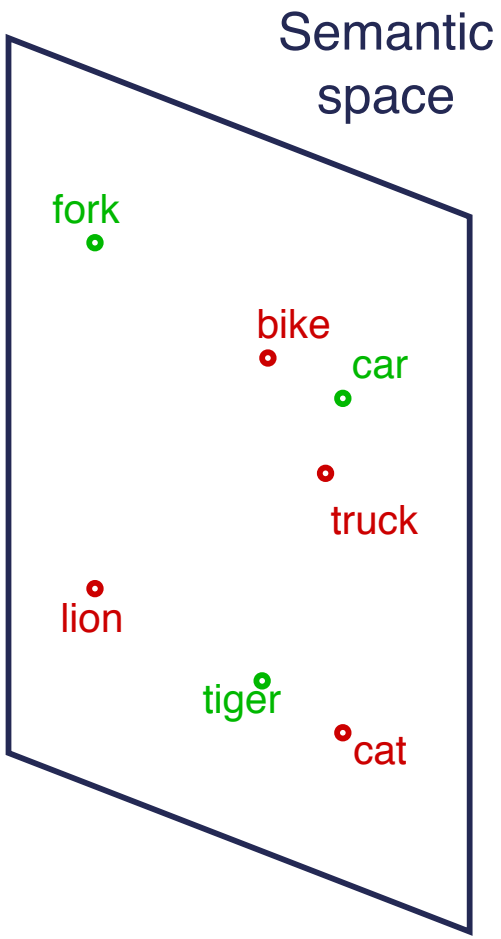
→ Source domain $\mathcal{S}$
Classes with training instances

→ Target domain $\mathcal{T}$
Zero-shot classes

...in a common space

→ Attributes

→ Semantic representation



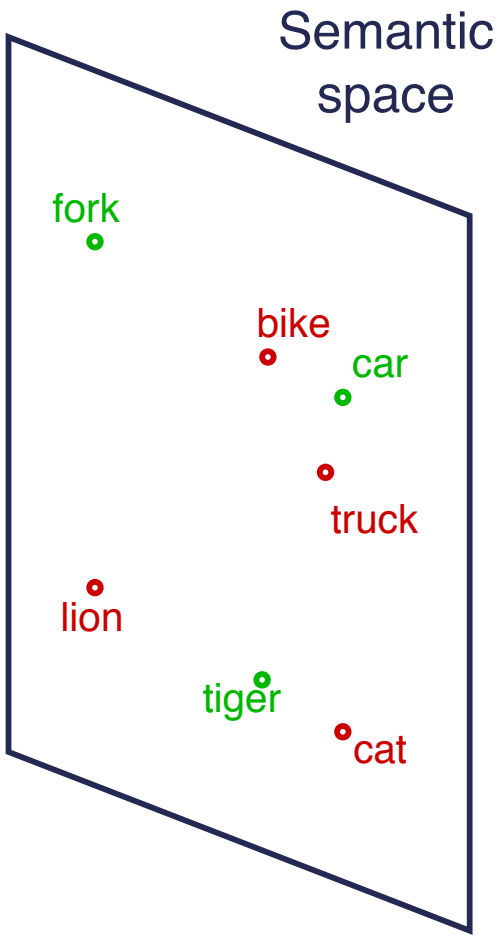
Zero-Shot Learning —Task & Hypothesis

Two domains ...

- → Source domain $\mathcal{S}$
  Classes with training instances
- → Target domain $\mathcal{T}$
  Zero-shot classes

...in a common space

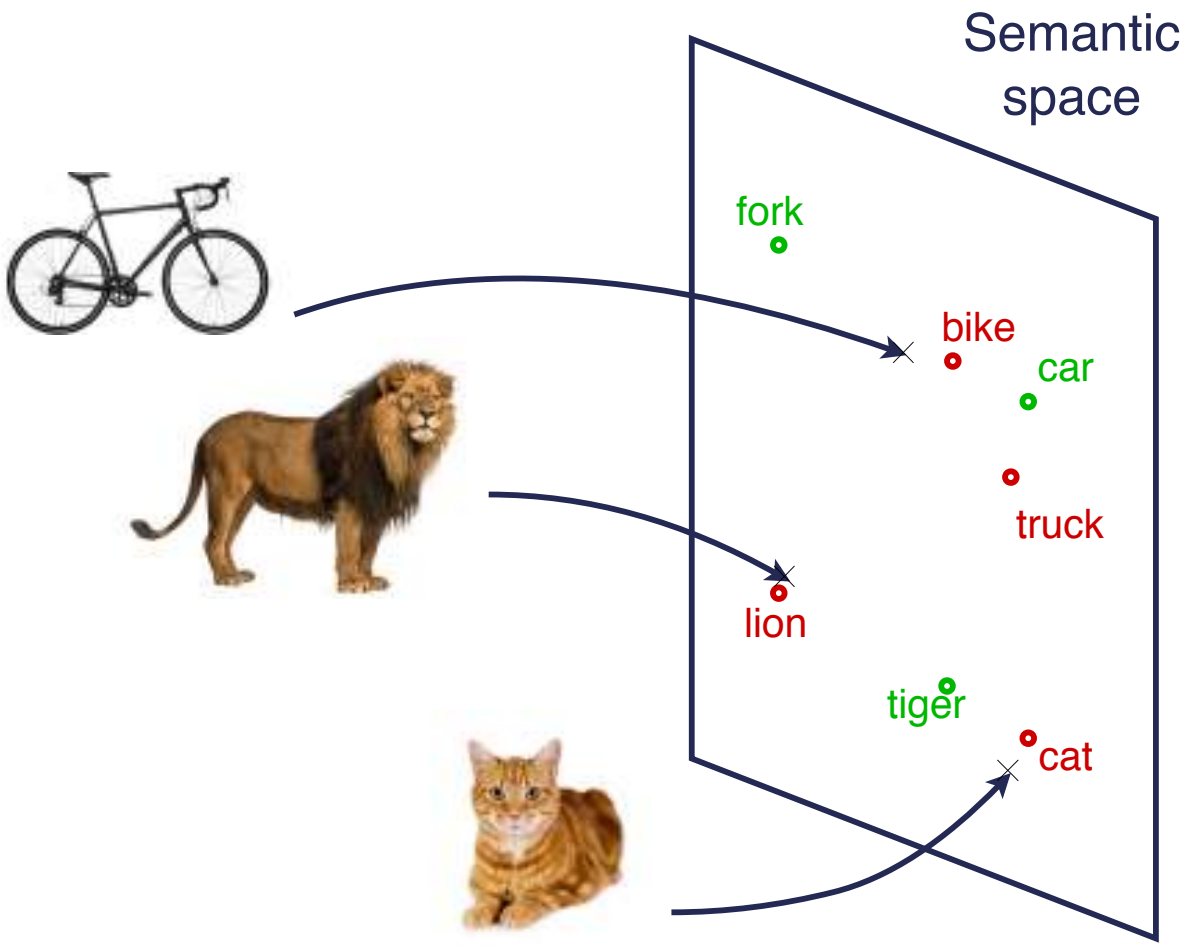
- → Attributes
- → Semantic representation



~~Maximize likelihood of $P(i, \mathcal{V})$ →null probability to classes of the target domain~~

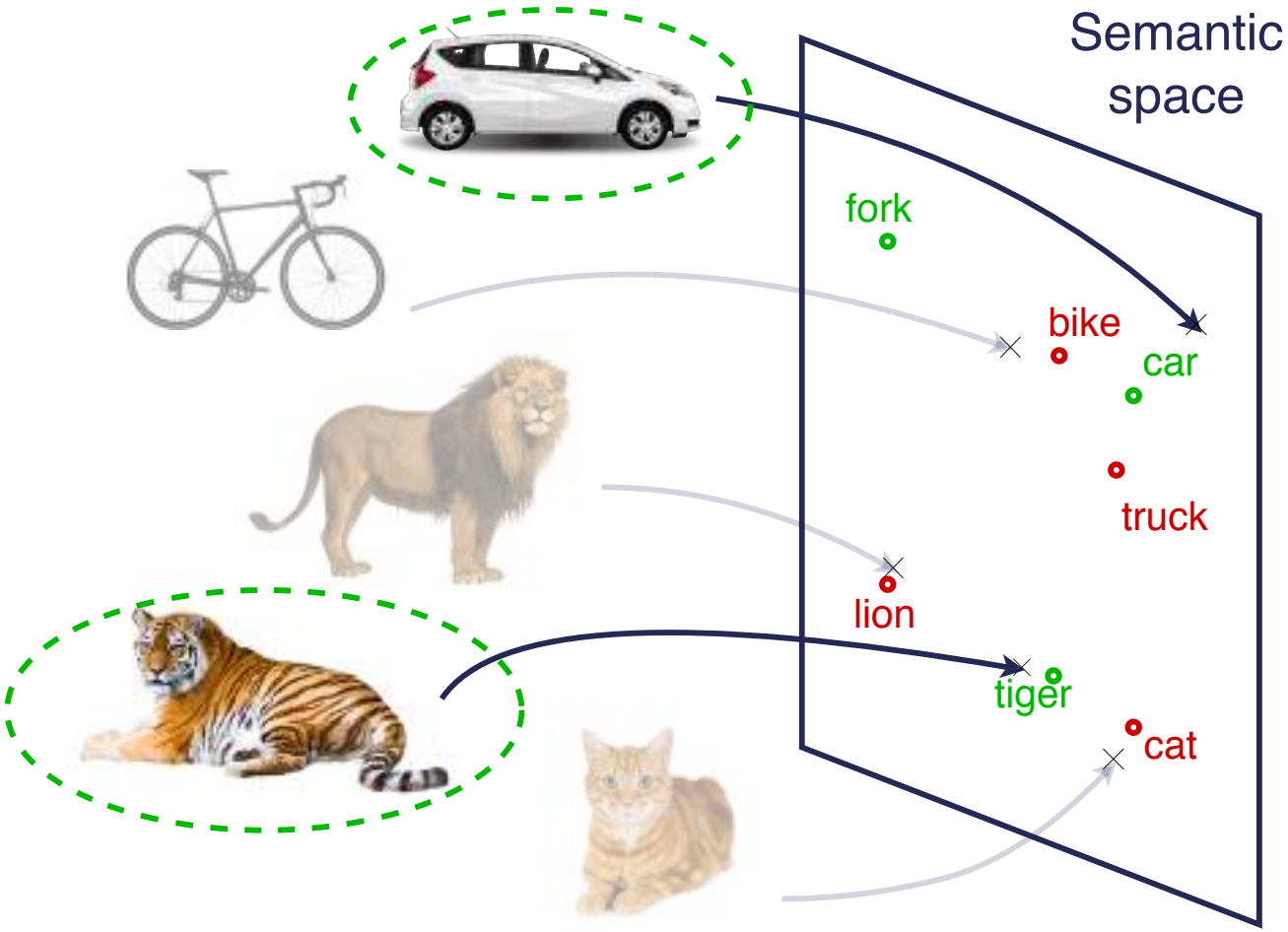
Idea: project images to the semantic space (Socher et al. 2013)

Zero-Shot Learning —Training



$$\mathcal{L}_V = \mathop{\mathbb{E}}_{i,\mathcal{V}\sim P^\star} \mathop{\mathbb{E}}_{j\sim\mathcal{U}} \left\lfloor \gamma - f(\mathcal{V})^\top w_i + f(\mathcal{V})^\top w_j \right\rfloor_+$$

Zero-Shot Learning —Inference



$$i^{\star} = \arg\min_{i \in \mathcal{T}} \cos(w_i, f(\mathcal{V}))$$

Motivation

→ Can language help visual understanding?

Motivation

→ Can language help visual understanding?

	Vision	Language
Region of Interest	Classification	Zero-Shot learning

Motivation

→ Can language help visual understanding?

→ Can visual context help visual understanding?

	Vision	Language
Region of Interest	Classification	Zero-Shot learning
Visual context	Context-aware classification	

Motivation

→ Can language help visual understanding?

→ Can visual context help visual understanding?

	Vision	Language
Region of Interest	Classification	Zero-Shot learning
Visual context	Context-aware classification	**????**

Research Question

Can **language** be leveraged along with **visual context** to help visual understanding?

→ Context-aware Zero-Shot Learning

Zero-Shot Learning —Adding context ?

→ "A **tiger** is most recognizable for its dark vertical stripes on reddish-orange fur with a lighter underside"

→ "I saw two **tigers** in the savanna, one was sleeping under a tree, the other one was hunting an antelope"

→ "We went to the zoo! An animal keeper was feeding a **tiger** with a dead boar. The **tiger** was shredding the poor animal with its huge teeth -_-"

Overview

Context-Aware Zero-Shot Learning for Object Recognition,
É. Zablocki, P. Bordes, B. Piwowarski, L. Soulier, P. Gallinari, ICML 2019.

Objective

Simultaneously use:

- → $\mathcal{V}$: The visual area containing an object of interest
- → $\mathcal{C}$: The visual context surrounding the object of interest

Hypothesis

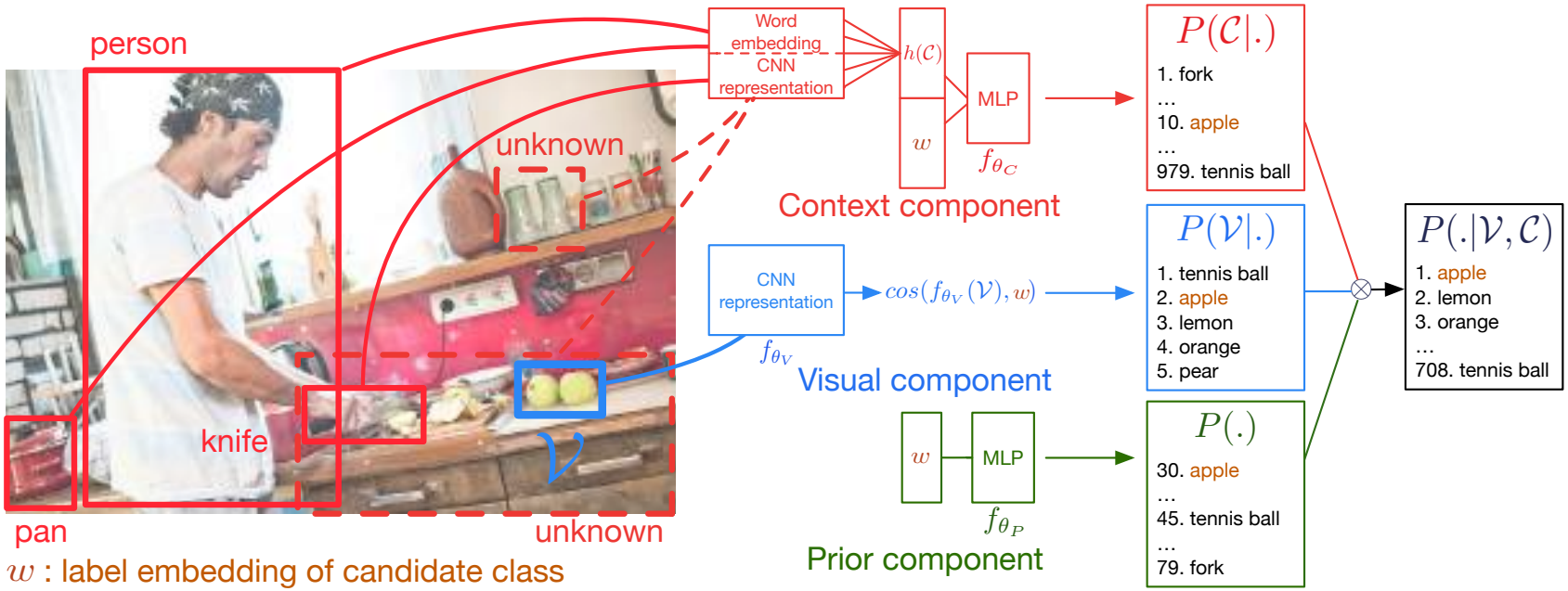
$\mathcal{V}$ and $\mathcal{C}$ conditionally independent given class i:

$$(\mathcal{V} \perp\!\!\!\perp \mathcal{C}) \mid i$$

Model

$$P(i \mid \mathcal{V}, \mathcal{C}) \propto P(\mathcal{V} \mid i) P(\mathcal{C} \mid i) P(i)$$

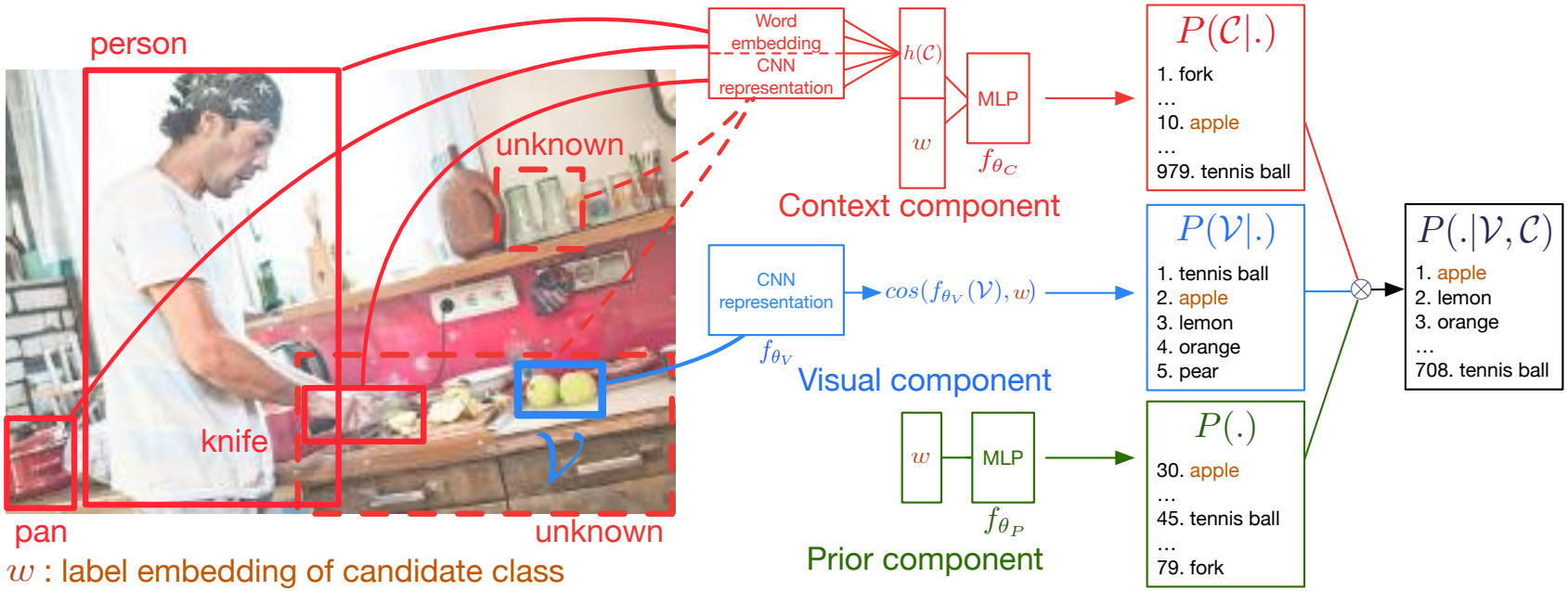
Overview



Model

$$P(i \mid \mathcal{V}, \mathcal{C}) \propto P(\mathcal{V} \mid i) P(\mathcal{C} \mid i) P(i)$$

Model's components



For a class *i*, and its semantic representation w_i:

Visual component

$$\log P(\mathcal{V} \mid i; \theta_V) \propto \cos(f_{\theta_V}(\mathcal{V}), w_i) := \log \widetilde{P}_{visual}$$

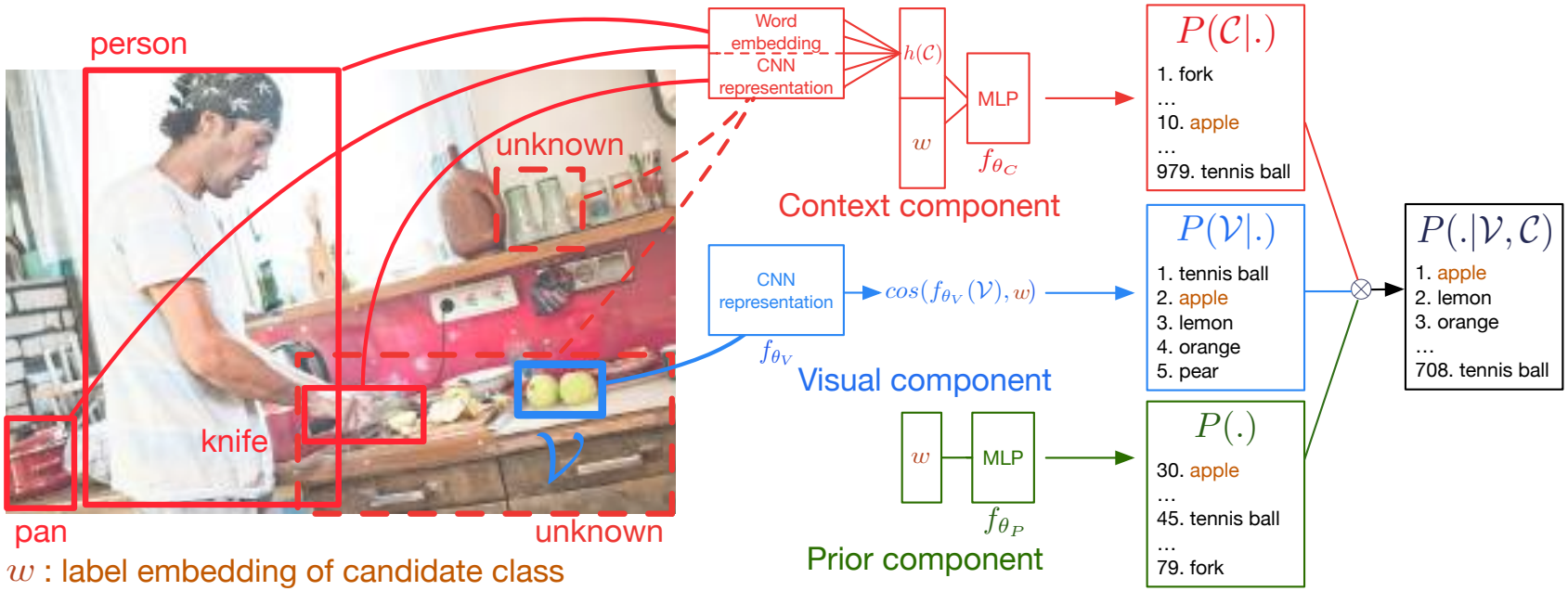
$$f_{\theta_V}(\mathcal{V}) = W_V.\texttt{CNN}(\mathcal{V}) + b_V$$

Prior component

$$\log P(i; \theta_P) \propto f_{\theta_P}(w_i) := \log \widetilde{P}_{prior}$$

$$f_{\theta_P}(w_i) = \texttt{MLP}_{\theta_P}(w_i)$$

Model's components

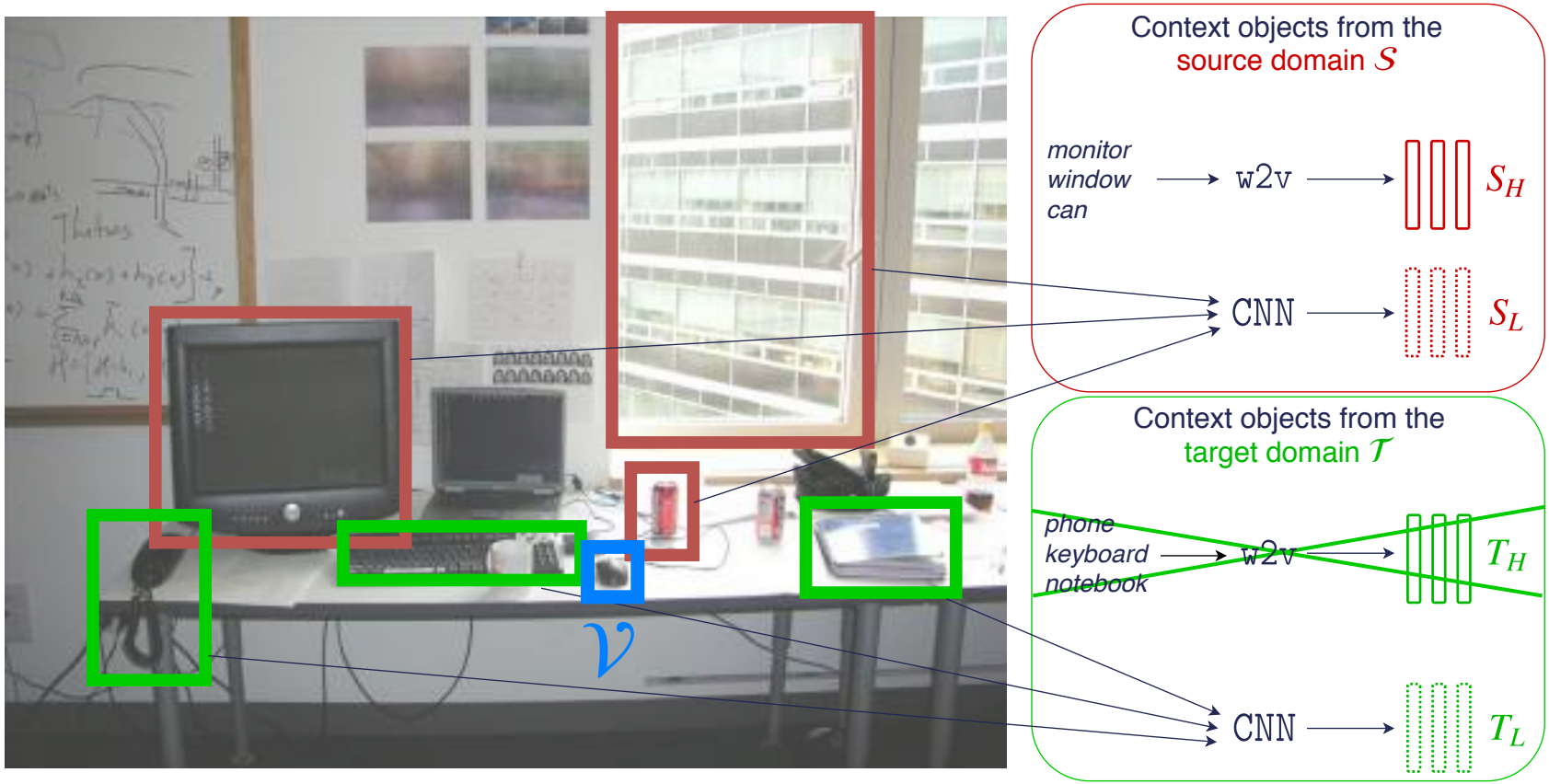


Context component

$$\log P(\mathcal{C} \mid i; \theta_C) \propto f_{\theta_C}\big(h(\mathcal{C}) \oplus w_i\big) := \log \widetilde{P}_{context}$$

$$f_{\theta_C} = \texttt{MLP}$$

Modeling contextual information



$$h(C_{S_H \cap T_L}) = \underbrace{S_H \; T_L}_{\text{Average}}$$

Learning

Positive examples sampled from true **data distribution** $P^\star$

Prior component

Negative sampled in the **uniform distribution** U

$$\mathcal{L}_P = \mathop{\mathbb{E}}_{i\sim P^\star} \mathop{\mathbb{E}}_{j\sim U} \left\lfloor \gamma_P - f_{\theta_P}(w_i) + f_{\theta_P}(w_j) \right\rfloor_+$$

Hence,

$$P^\star(i) > P^\star(j) \Rightarrow f_{\theta_P}(i) > f_{\theta_P}(j)$$

Learning

Positive examples sampled from true **data distribution** $P^\star$

Prior component

Negative sampled in the **uniform distribution** U

$$\mathcal{L}_P = \mathop{\mathbb{E}}_{i\sim P^\star} \mathop{\mathbb{E}}_{j\sim U} \left\lfloor \gamma_P - f_{\theta_P}(w_i) + f_{\theta_P}(w_j) \right\rfloor_+$$

Hence,

$$P^\star(i) > P^\star(j) \Rightarrow f_{\theta_P}(i) > f_{\theta_P}(j)$$

Context and Visual component

Negative sampled in the true **data distribution** $P^\star$

$$\mathcal{L}_V = \mathop{\mathbb{E}}_{i,\mathcal{V}\sim P^\star} \mathop{\mathbb{E}}_{j\sim P^\star} \left\lfloor \gamma_V - f_{\theta_V}(\mathcal{V})^\top w_i + f_{\theta_V}(\mathcal{V})^\top w_j \right\rfloor_+$$

$$\mathcal{L}_C = \mathop{\mathbb{E}}_{i,\mathcal{C}\sim P^\star} \mathop{\mathbb{E}}_{j\sim P^\star} \left\lfloor \gamma_C - f_{\theta_C}(\mathcal{C}, w_i) + f_{\theta_C}(\mathcal{C}, w_j) \right\rfloor_+$$

Hence,

$$P^\star(\mathcal{V} \mid i) > P^\star(\mathcal{V} \mid j) \Rightarrow f_{\theta_V}(\mathcal{V})^\top w_i > f_{\theta_V}(\mathcal{V})^\top w_j$$

$$P^\star(\mathcal{C} \mid i) > P^\star(\mathcal{C} \mid j) \Rightarrow f_{\theta_C}(\mathcal{C}, w_i) > f_{\theta_C}(\mathcal{C}, w_j)$$

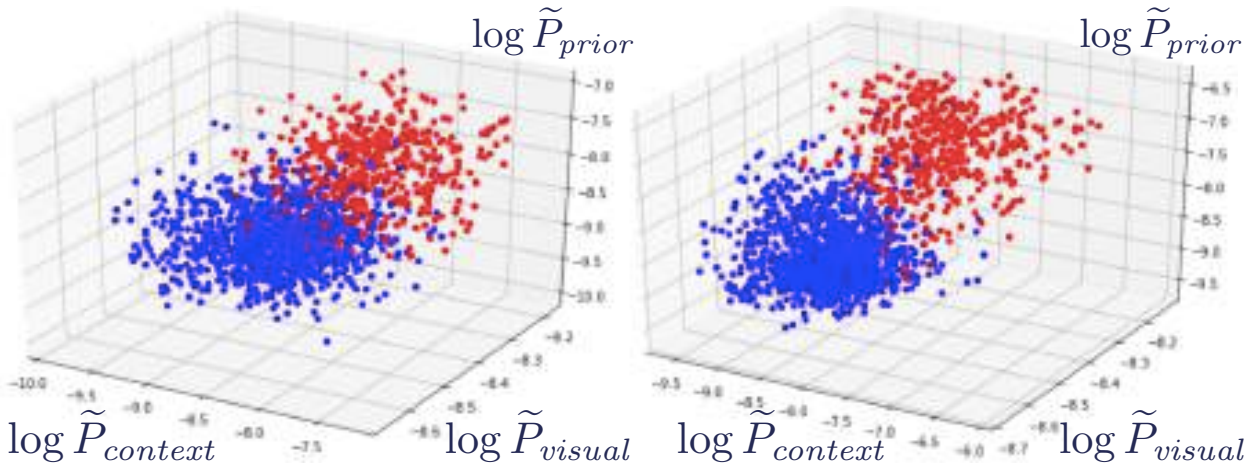
Inference

$$\log P(\mathcal{C} \mid i) \propto \log \widetilde{P}_{context} \qquad \log P(\mathcal{V} \mid i) \propto \log \widetilde{P}_{visual} \qquad \log P(i) \propto \log \widetilde{P}_{prior}$$

Hypothesis

The normalization constants are scalars, independent of the class i, hence:

$$P(i \mid \mathcal{V}, \mathcal{C}) = \underbrace{(\widetilde{P}_{context})^{\alpha_C}}_{P(\mathcal{C} \mid i)} \cdot \underbrace{(\widetilde{P}_{visual})^{\alpha_V}}_{P(\mathcal{V} \mid i)} \cdot \underbrace{(\widetilde{P}_{prior})^{\alpha_P}}_{P(i)}$$

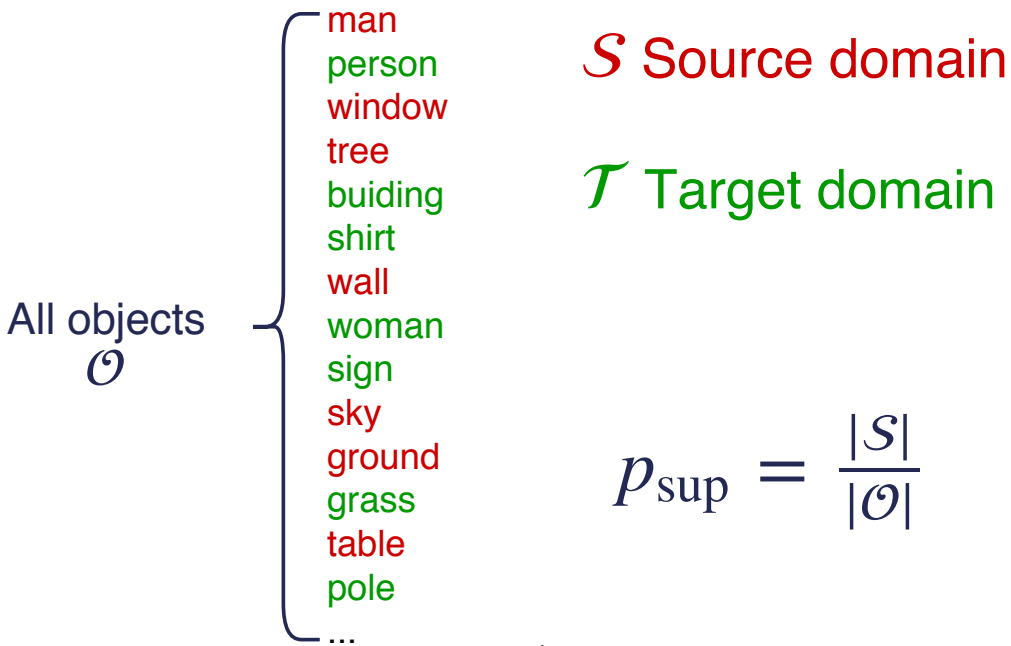


Target domain Source domain

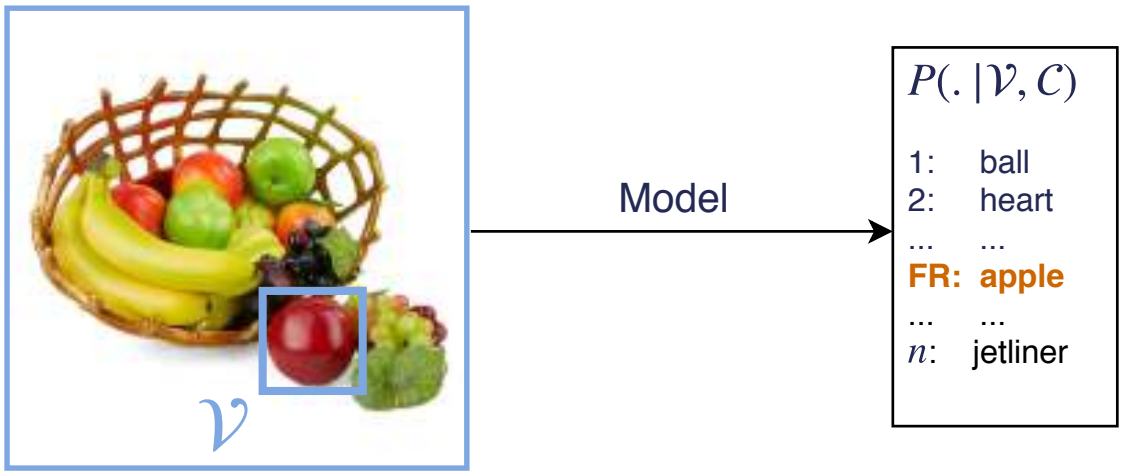
Data

Visual genome

- → 108k images
  - → 3.8M instances, 105K unique objects, 31 instances per image
- → Filtered (> 10 occurrences, need a *w2v* embedding)
  - → 3.4M instances, 4.8K unique objects
- → Highly imbalanced:
  - → 10% of most frequent classes ⟶ 84% of instances



Evaluation



Metric

For a collection of M test images:

$$\text{MFR} = \frac{1}{M} \sum_{i=1}^{M} \frac{2 * (\text{FR}_i - 1)}{n - 1}$$

→ Lower the better

→ Random: MFR = 100%

→ Perfect model: MFR = 0%

Importance of context

Model	p_{sup} Probability	Target domain $\mathcal{T}$ 10%	50%	90%	Source domain $\mathcal{S}$ 10%	50%	90%
Random	$\mathcal{U}$	*100*	*100*	*100*	*100*	*100*	*100*
M($\varnothing$)	$P(.)$	38.6	23.7	13.8	12.0	10.6	11.2
M($\mathcal{V}$)	$P(\mathcal{V}\|.)P(.)$	20.5	10.7	6.0	1.5	2.6	3.6
M($\mathcal{C}$)	$P(\mathcal{C}\|.)P(.)$	28.7	14.4	9.1	4.2	4.3	4.4
M($\mathcal{C}, \mathcal{V}$)	$P(\mathcal{C}\|.)P(\mathcal{V}\|.)P(.)$	**18.1**	**9.0**	**5.2**	**1.1**	**1.9**	**2.4**
$\delta_{\mathcal{C}}$ (%)		*11.6*	*16.4*	*12.1*	*23.7*	*27.3*	*31.5*

Table 1: MFR scores (%). δ_C is the relative improvement of M($\mathcal{C}, \mathcal{V}$) over M($\mathcal{V}$).

Results

- → Context is useful: M($\mathcal{C}$) $\overset{\text{MFR}}{<}$ *Random*
- → Visual frequency can be learned from textual semantics: M($\varnothing$) generalizes
- → Context is complementary to visual appearance: M($\mathcal{C}, \mathcal{V}$) $\overset{\text{MFR}}{<}$ M($\mathcal{V}$)

Modeling contextual information

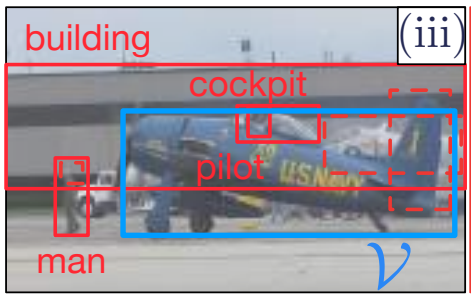
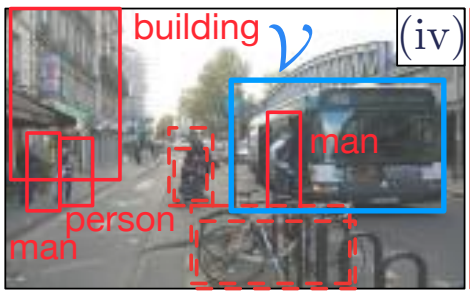
Context objects in $\mathcal{T}$	Context objects in $\mathcal{S}$	$\mathcal{T}$	$\mathcal{S}$
No context		10.72	2.64
Full image		9.19	2.13
$\varnothing$	image	9.01	2.05
image	$\varnothing$	9.00	2.13
$\varnothing$	label	8.96	1.92
image	image	8.60	1.93
image	label	8.52	1.86
image	image & label	**8.31**	**1.79**

Increasing supervision helps:

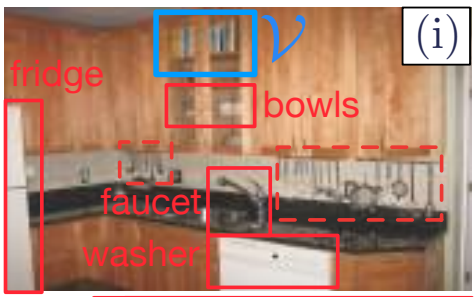
1. When bounding boxes are available...
2. When context objects are labeled...
3. When more context objects are used...
4. When low-level info. is used complementarily

Best model:
→ 22% relative improvement in the target domain $\mathcal{T}$ (32% in $\mathcal{S}$)

Qualitative results

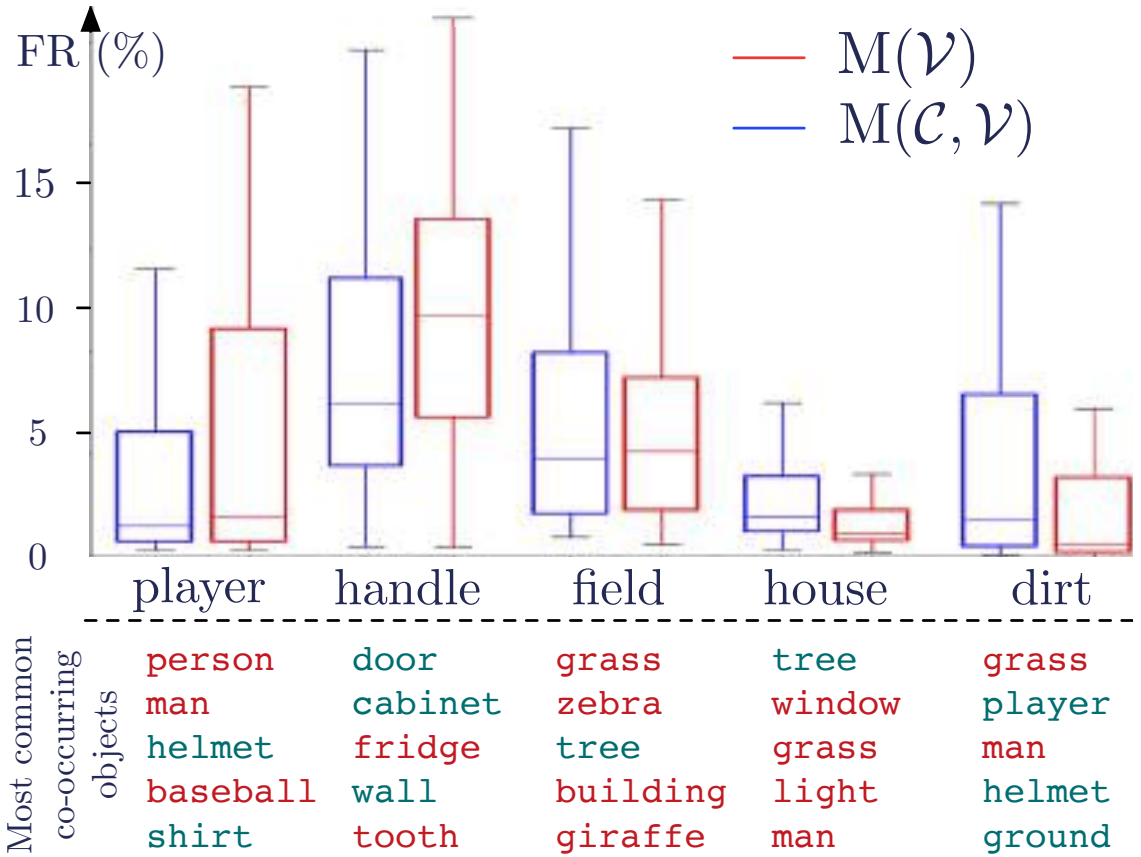
	$P(\mathcal{C}\|.)$	$P(\mathcal{V}\|.)$	$P(.)$	$P(.\|\mathcal{V},\mathcal{C})$
	1. lilies 2. flower 3. garden 4. carnations 5. orchids	1. needle 2. fingertip 3. kitten … 223. flower	1. tree 2. woman 3. car … 8. flower	1. flower 2. tip 3. hair … 29. needle
	1. water 2. river 3. sea … 9. boat	1. filters 2. connector 3. indicator … 1757. boat	1. tree 2. woman 3. car … 25. boat	1. boat … 31. water … 376. filters
	1. jetliner 2. engines 3. air 4. airplane 5. rotor	1. flight 2. KLM 3. jetliner … 11. airplane	1. tree … 256. airplane … 1717. jetliner	1. airplane 2. flight 3. runway 4. air 5. hand
	1. men 2. woman 3. widow … 752. bus	1. bus 2. business 3. parking 4. shops 5. downtown	1. tree 2. woman 3. car … 13. bus	1. bus 2. car 3. building 4. woman 5. vehicle

Negative results



	$P(\mathcal{C}\|.)$	$P(\mathcal{V}\|.)$	$P(.\|\mathcal{V})$	$P(.\|\mathcal{V},\mathcal{C})$
(i)	1. freezer 2. ovens 3. heater 4. … 981. books	1. shelving 2. books 3. bookshelf 4. cartons 5. papers	1. books 2. boxes 3. wall 4. shelving 5. table	1. table 2. room 3. boxes … 40. books
(ii)	1. specks 2. whiskers 3. paws 4. … 1242. leaves	1. sole 2. mold 3. scrape 4. leaves 5. branch	1. tree 2. canopy 3. hand 4. wall 5. leaves	1. hand 2. wall 3. nose … 74. leaves

Qualitative results

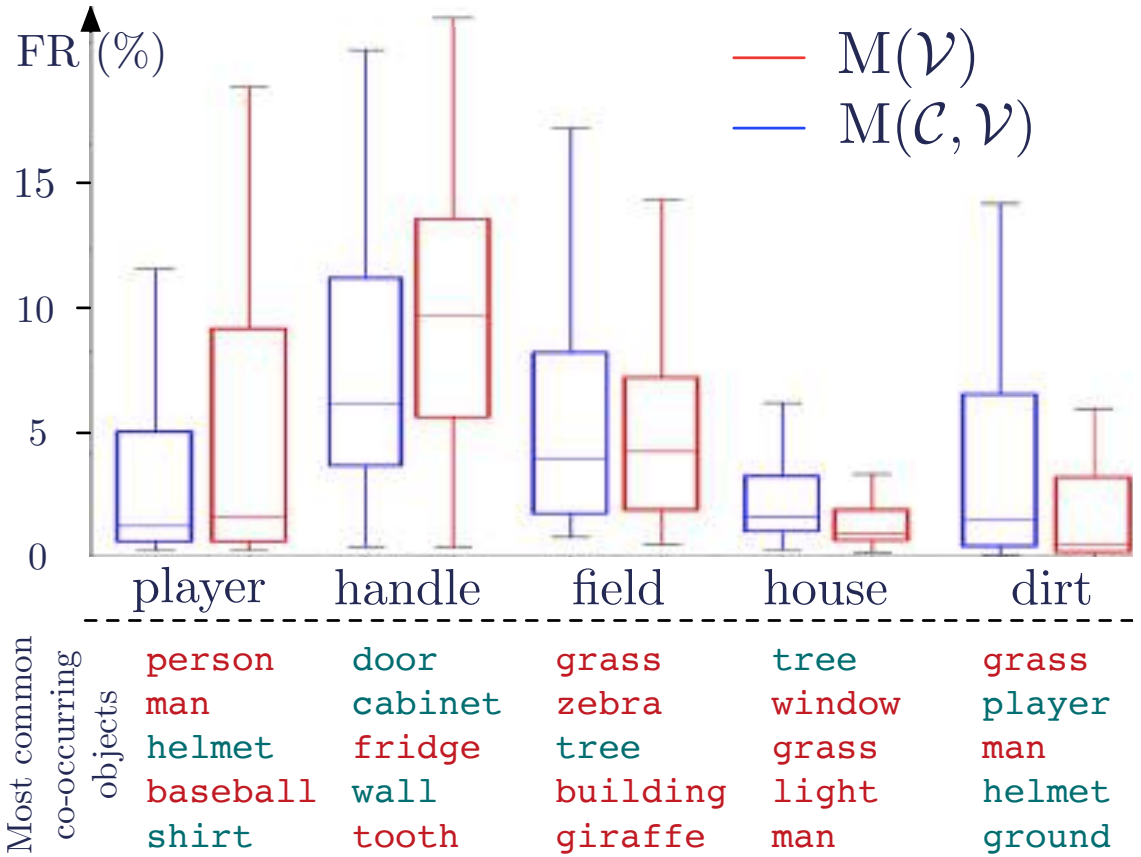


Context can help ...

→ refining predictions

→ disambiguating generic shapes

Qualitative results



Context can help ...

- → refining predictions
- → disambiguating generic shapes

...but not always

- → Objects with lots of possible context
- → $\rho(\#(\text{context}), \delta_C) = -0.33$

Conclusion and Perspectives

Conclusion

- → Language can help to bridge the gap of missing supervision in computer vision
- → Visual contexts and priors are contained withing word representation

Conclusion and Perspectives

Conclusion

- → Language can help to bridge the gap of missing supervision in computer vision
- → Visual contexts and priors are contained withing word representation

Perspectives

- → Learn spatial arrangement from language data
- → Zero-shot object detection

Conclusion

Images can help NLP

- → to learn grounded representations
- → model the semantic gap between language and images?
- → useful for downstream tasks? (open-domain QA, machine translation, ...)

Language can help Computer Vision

- → using a semantic space (build with NLP techniques)
- → evaluating NLP models (captioning, VQA)
- → low/few supervision (e.g. zero-shot learning)

Questions?

SCIENCES SORBONNE UNIVERSITÉ

—Thanks for your attention —

Questions ?