THE UNIVERSITY OF BRITISH COLUMBIA

Topics in AI (CPSC 532S):
Multimodal Learning with Vision, Language and Sound

Lecture 12: Coordinated Representations and Joint Embeddings (cont)

Logistics

- **Assignment 4** — Due **tomorrow**
  - BLUE4 scores
  - Self-attention

- **Project pitches** right after break (**register** on Google form!!!)

- Two more lectures after that from me, then paper readings

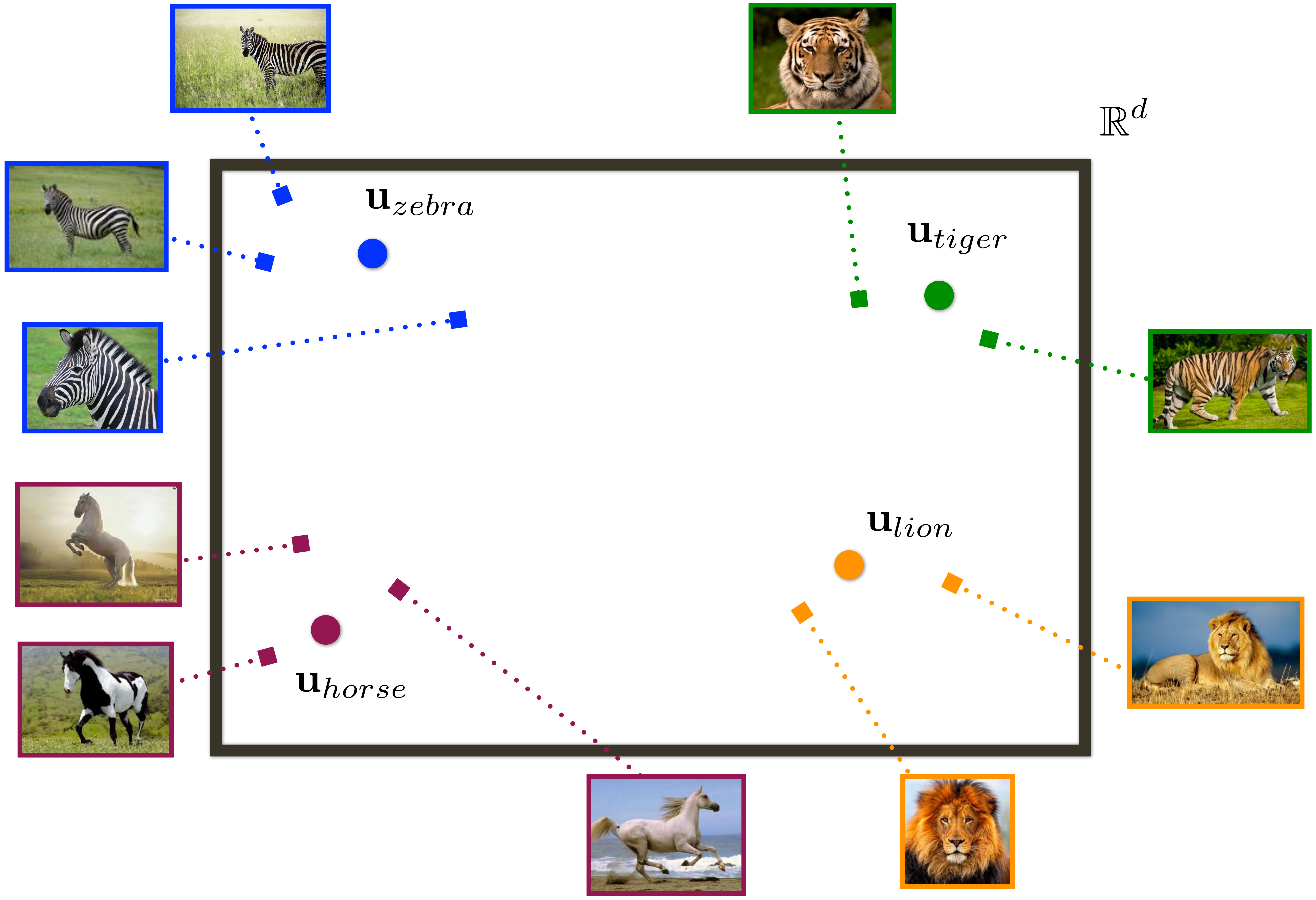
Semantic Embeddings

Image Embedding

$$\Psi(I_i) = \mathbf{W} \cdot CNN(I_i; \mathbf{\Theta}) \colon \mathbb{R}^D \rightarrow \mathbb{R}^d$$

Label Embedding

$$\Psi_L(word_i) = \mathbf{u}_i : \{1, ..., L\} \rightarrow \mathbb{R}^d$$



word2vec: Unsupervised Word Embedding

Distributional Semantics Hypothesis: words that are used and occur in the same context tend to have similar meaning

Label Embedding

$$\Psi_L(word_i) = \mathbf{u}_i : \{1, ..., L\} \rightarrow \mathbb{R}^d$$

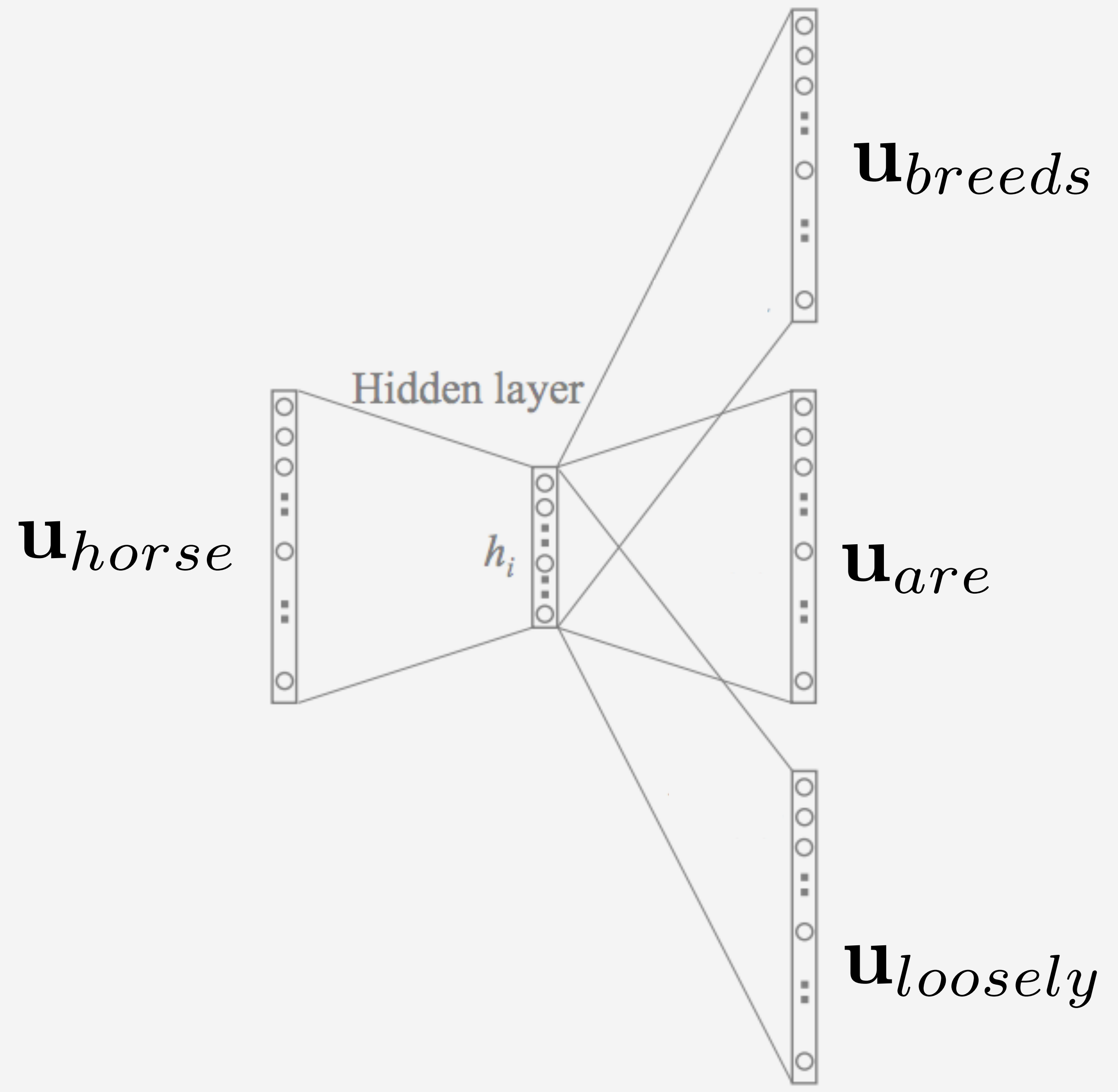
word2vec: Unsupervised Word Embedding

Distributional Semantics Hypothesis: words that are used and occur in the same context tend to have similar meaning

Label Embedding

$$\Psi_L(word_i) = \mathbf{u}_i : \{1, ..., L\} \rightarrow \mathbb{R}^d$$

e.g., Horse breeds are loosely divided into three categories

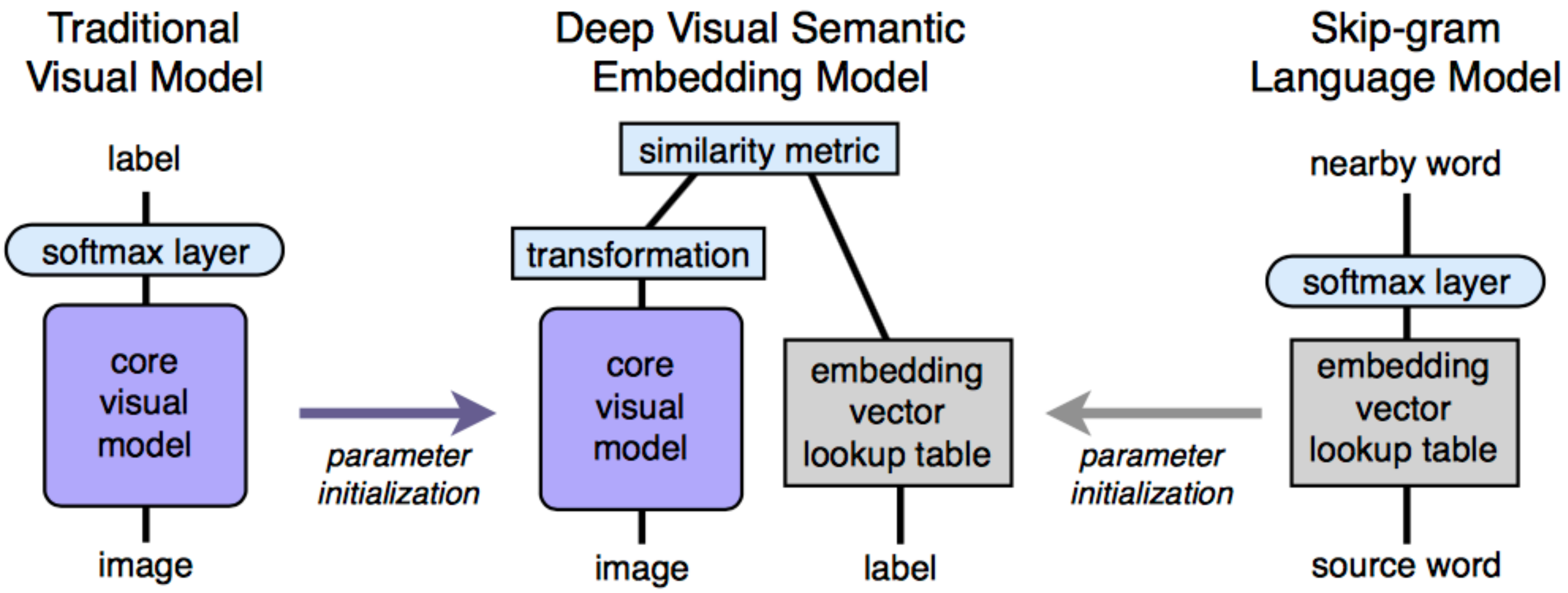


Skip-gram Model: unsupervised semantic representation for words

[Mikolov, Sutskever, Chen, Corrado, Dean, NIPS'13]

DeViSE: A Deep Visual-Semantic Embedding Model

[Frome et al., 2013]



$$loss(image, label) = \sum_{j \neq label} \max[0, margin - \vec{t}_{label} M \vec{v}(image) + \vec{t}_j M \vec{v}(image)]$$

DeViSE: A Deep Visual-Semantic Embedding Model

[Frome et al., 2013]

Supervised Results

Model type	dim	Flat hit@k (%)				Hierarchical precision@k			
		1	2	5	10	2	5	10	20
Softmax baseline	N/A	**55.6**	**67.4**	**78.5**	**85.0**	0.452	0.342	0.313	0.319
DeViSE	500	53.2	65.2	76.7	83.3	0.447	**0.352**	**0.331**	**0.341**
	1000	54.9	66.9	78.4	**85.0**	**0.454**	0.351	0.325	0.331
Random embeddings	500	52.4	63.9	74.8	80.6	0.428	0.315	0.271	0.248
	1000	50.5	62.2	74.2	81.5	0.418	0.318	0.290	0.292
Chance	N/A	0.1	0.2	0.5	1.0	0.007	0.013	0.022	0.042

Zero-shot Results

Model	200 labels	1000 labels
DeViSE	31.8%	9.0%
Mensink et al. 2012 [12]	35.7%	1.9%
Rohrbach et al. 2011 [17]	34.8%	-

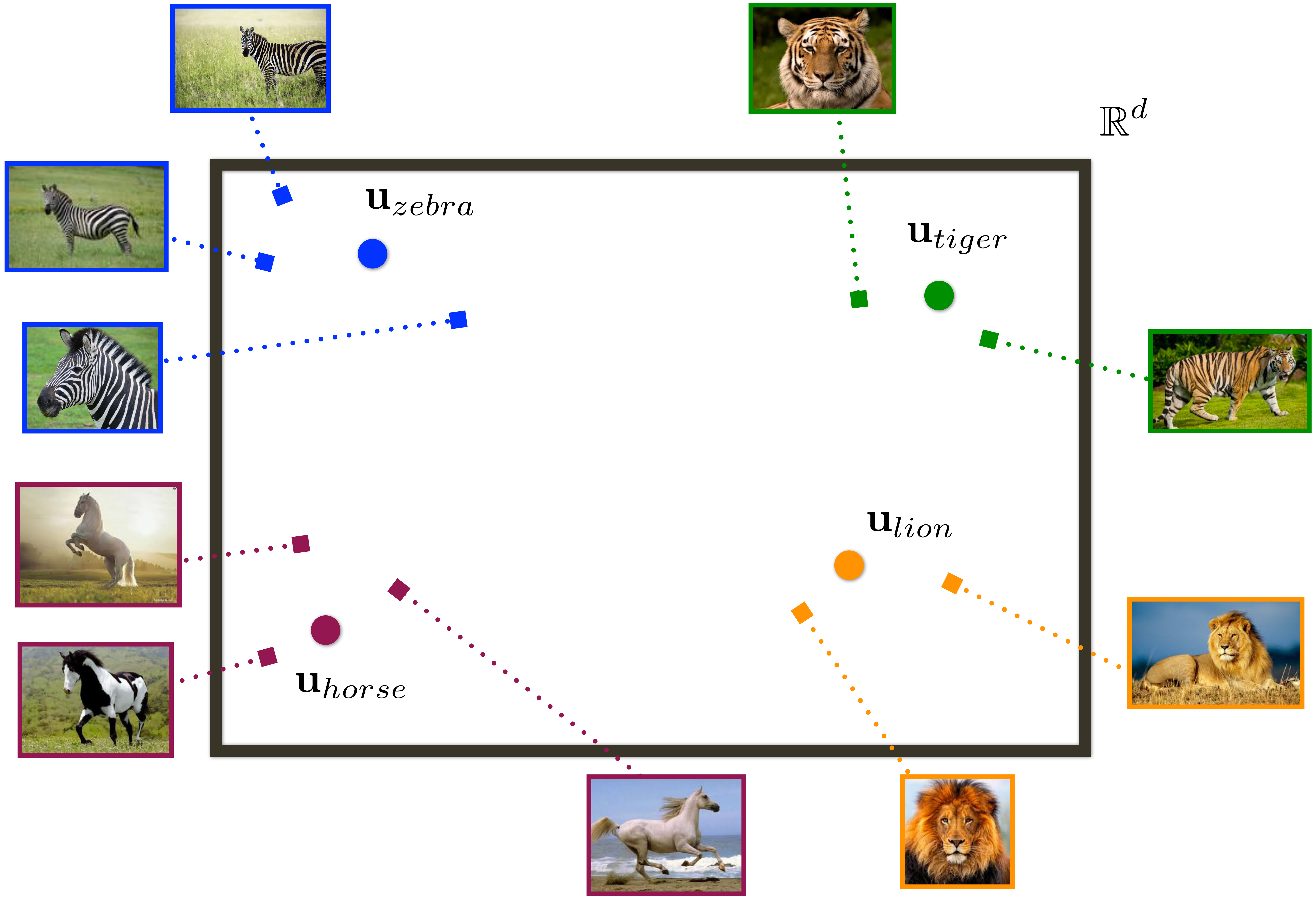
Semantic Embeddings

Image Embedding

$$\Psi(I_i) = \mathbf{W} \cdot CNN(I_i; \mathbf{\Theta}) \colon \mathbb{R}^D \rightarrow \mathbb{R}^d$$

Label Embedding

$$\Psi_L(word_i) = \mathbf{u}_i : \{1, ..., L\} \rightarrow \mathbb{R}^d$$



word2vec: Unsupervised Word Embedding

[Fu et al., 2016]

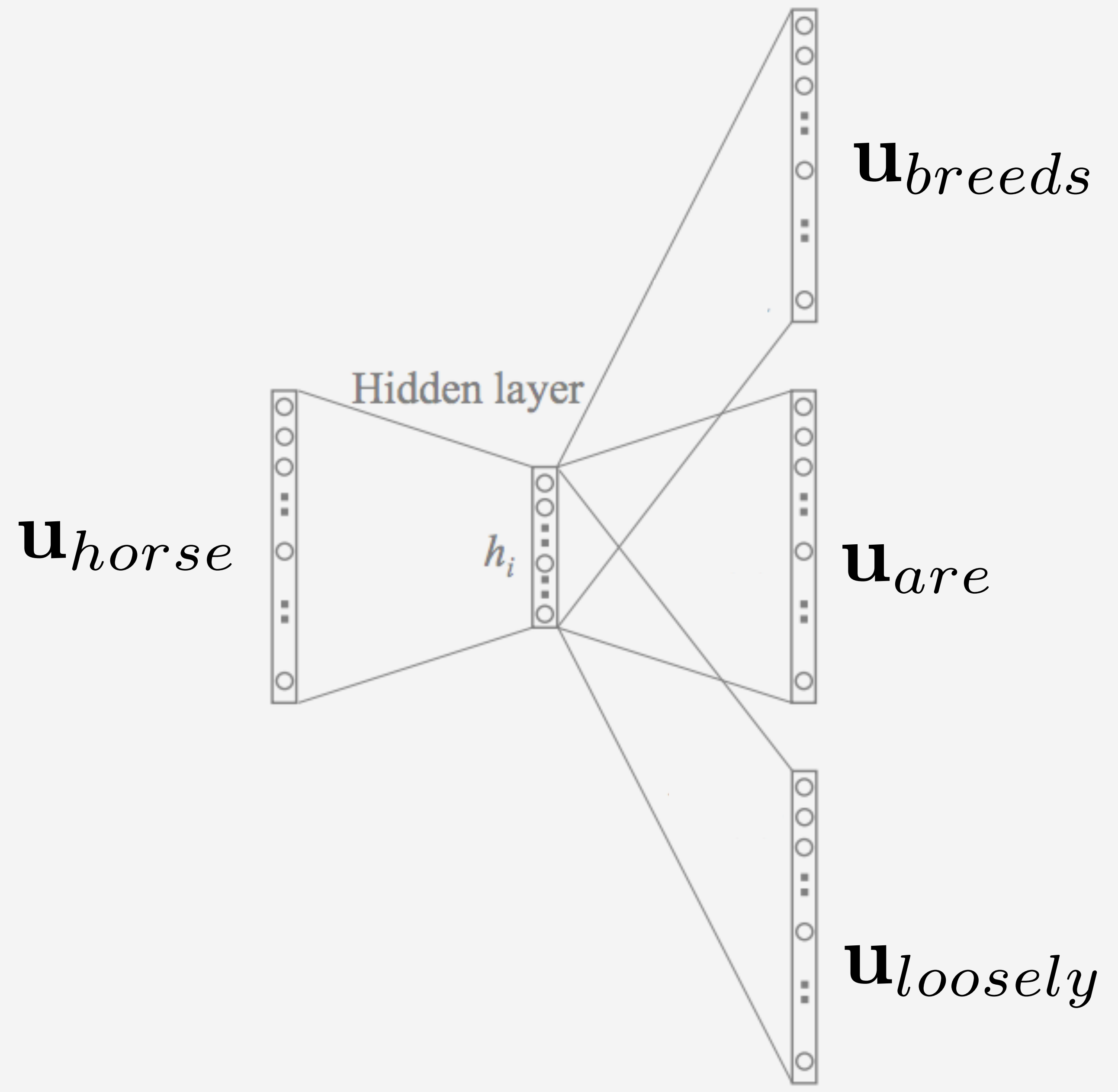
Distributional Semantics Hypothesis: words that are used and occur in the same context tend to have similar meaning

Label Embedding

$$\Psi_L(word_i) = \mathbf{u}_i : \{1, ..., L\} \rightarrow \mathbb{R}^d$$

$L = 310,000$

e.g., Horse breeds are loosely divided into three categories



Skip-gram Model: unsupervised semantic representation for words

(trained from 7 billion word linguistic corpus)

Semi-supervised **Vocabulary Informed** Learning

[Fu et al., 2016]

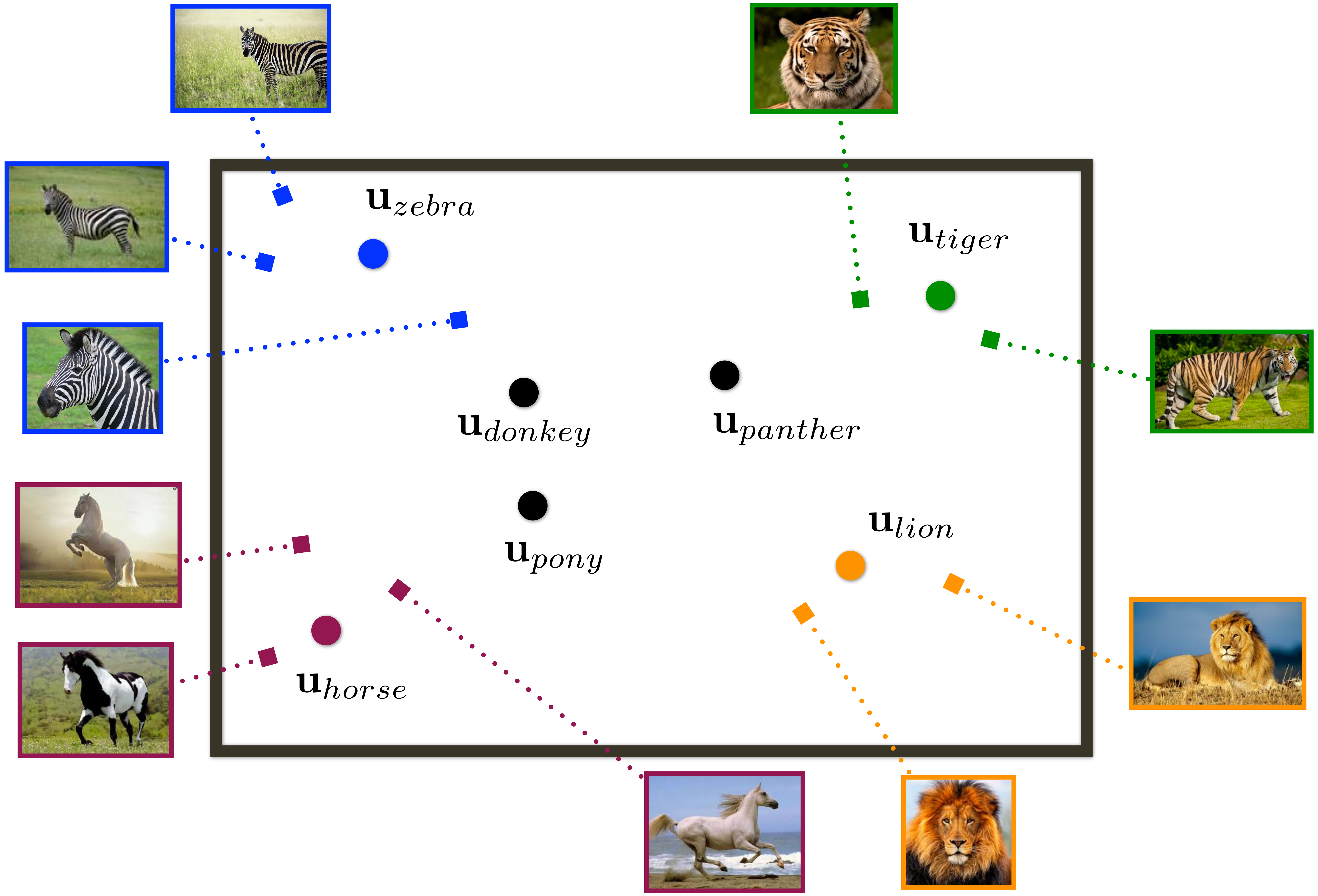
Image Embedding

$$\Psi(I_i) = \mathbf{W} \cdot CNN(I_i; \boldsymbol{\Theta}) : \mathbb{R}^D \rightarrow \mathbb{R}^d$$

Label Embedding

$$\Psi_L(word_i) = \mathbf{u}_i : \{1, ..., L\} \rightarrow \mathbb{R}^d$$

$L = 310,000$



Semi-supervised **Vocabulary Informed** Learning

[Fu et al., 2016]

Image Embedding

$$\Psi(I_i) = \mathbf{W} \cdot CNN(I_i; \mathbf{\Theta}) : \mathbb{R}^D \rightarrow \mathbb{R}^d$$

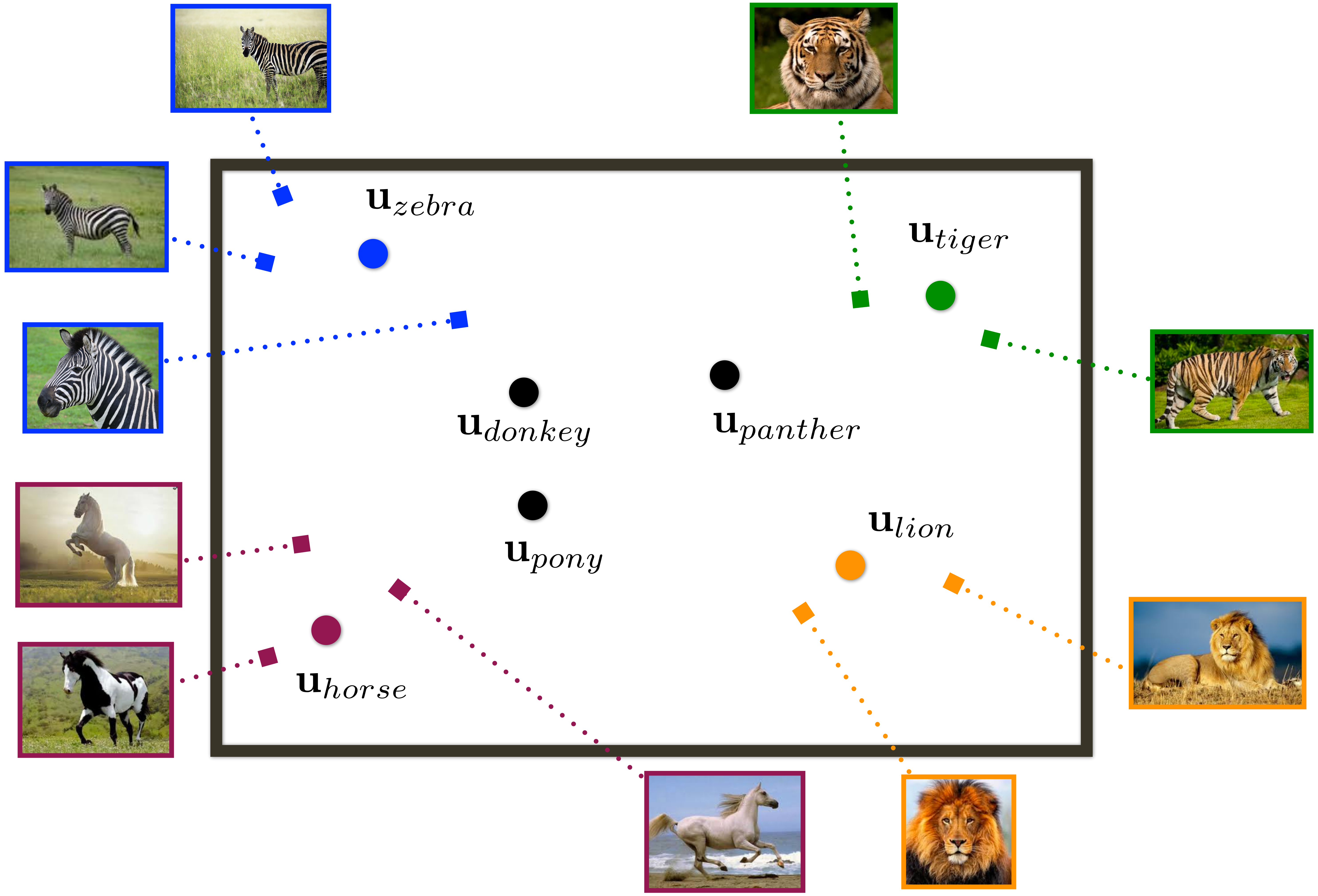
Label Embedding

$$\Psi_L(word_i) = \mathbf{u}_i : \{1, ..., L\} \rightarrow \mathbb{R}^d$$

$L = 310,000$

Similarity in Embedding Space

$$D(\mathbf{u}, \mathbf{u}') = ||\mathbf{u} - \mathbf{u}'||_2^2$$



Semi-supervised **Vocabulary Informed** Learning

[Fu et al., 2016]

Image Embedding

$$\Psi(I_i) = \mathbf{W} \cdot CNN(I_i; \mathbf{\Theta}) : \mathbb{R}^D \rightarrow \mathbb{R}^d$$

Label Embedding

$$\Psi_L(word_i) = \mathbf{u}_i : \{1, ..., L\} \rightarrow \mathbb{R}^d$$

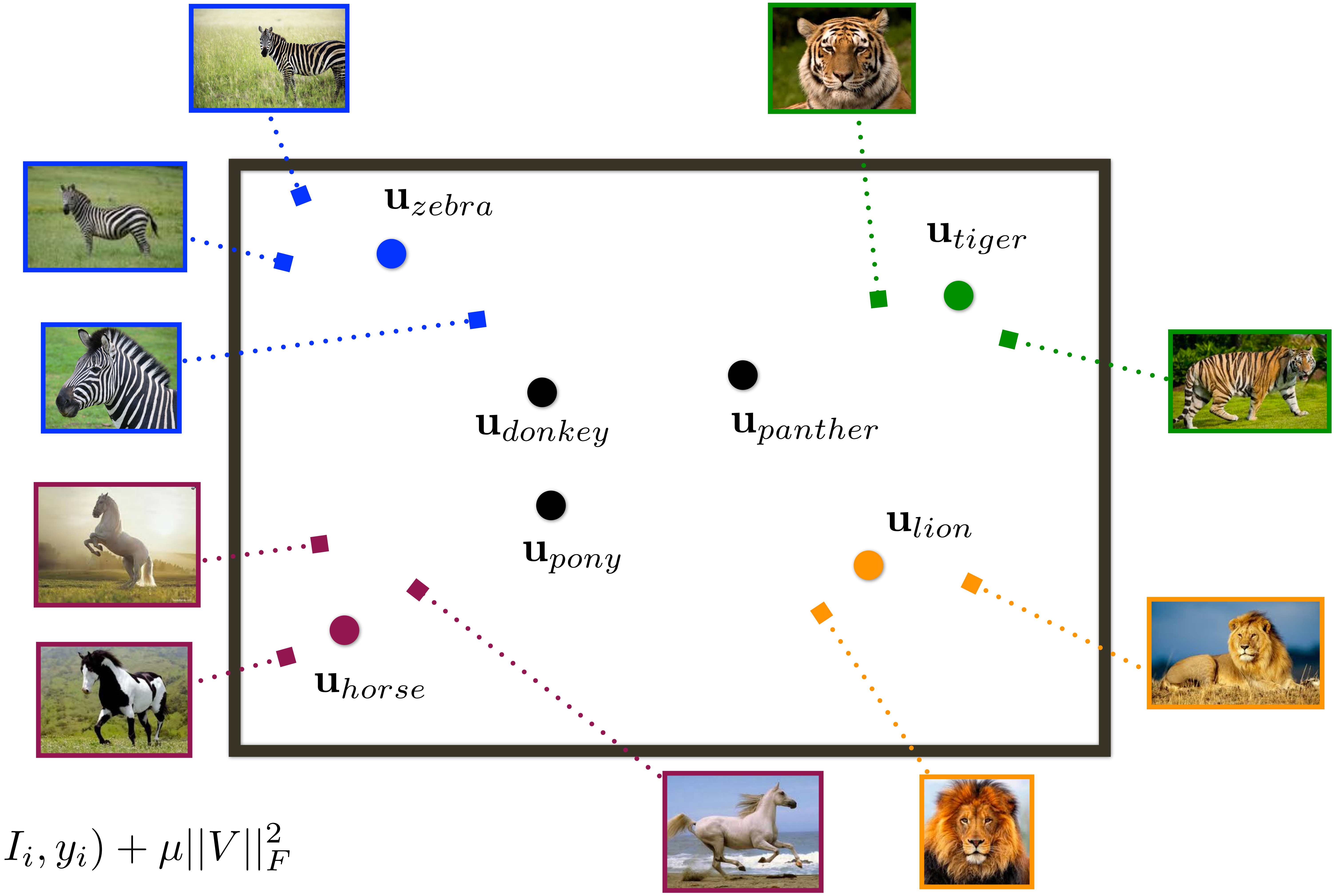
$L = 310,000$

Similarity in Embedding Space

$$D(\mathbf{u}, \mathbf{u}') = ||\mathbf{u} - \mathbf{u}'||_2^2$$

Objective Function:

$$\min_{\mathbf{W}} \sum_i^N \mathcal{L}_C(\mathbf{W}, \mathbf{V}, I_i, y_i) + \mathcal{L}_R(\mathbf{W}, \mathbf{V}, I_i, y_i) + \mu||V||_F^2$$



Semi-supervised **Vocabulary Informed** Learning [Fu et al., 2016]

$$\mathcal{L}_C(\mathbf{W}, \mathbf{U}, \mathbf{x}_i, y_i) = \sum [1 + \underline{D(\mathbf{W}\mathbf{x}_i, \mathbf{u}_{y_i})} - \underline{D(\mathbf{W}\mathbf{x}_i, \mathbf{u}_c)}]$$

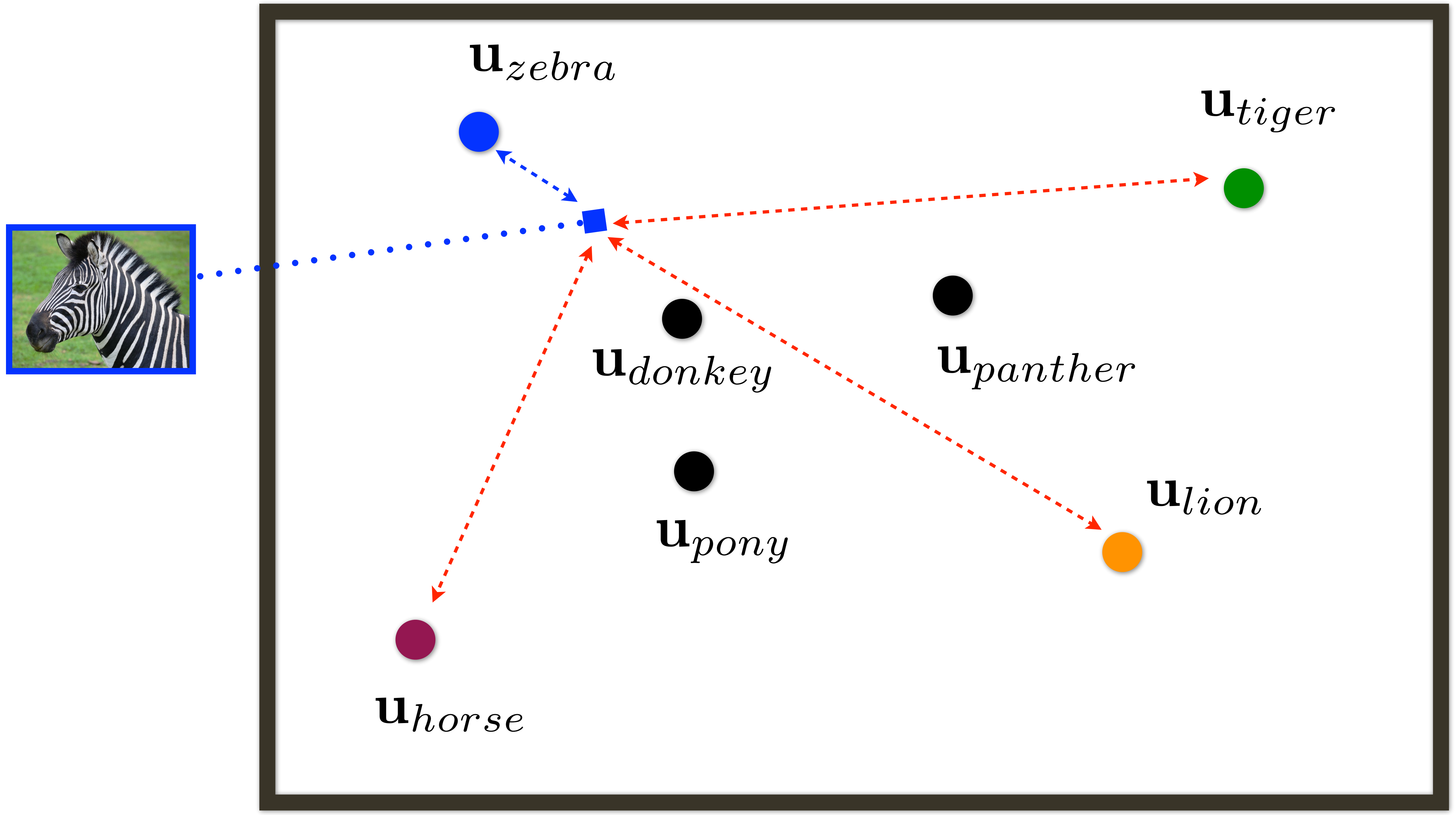
Image Embedding

$$\Psi(I_i) = \mathbf{W} \cdot CNN(I_i; \mathbf{\Theta}) : \mathbb{R}^D \to \mathbb{R}^d$$

Label Embedding

$$\Psi_L(word_i) = \mathbf{u}_i : \{1, ..., L\} \to \mathbb{R}^d$$

$L = 310,000$



Similarity in Embedding Space

$$D(\mathbf{u}, \mathbf{u}') = ||\mathbf{u} - \mathbf{u}'||_2^2$$

Objective Function:

$$\min_{\mathbf{W}} \sum_{i}^{N} \mathcal{L}_C(\mathbf{W}, \mathbf{V}, I_i, y_i) + \mathcal{L}_R(\mathbf{W}, \mathbf{V}, I_i, y_i) + \mu ||V||_F^2$$

Semi-supervised **Vocabulary Informed** Learning

[Fu et al., 2016]

$$\mathcal{L}_C(\mathbf{W}, \mathbf{U}, \mathbf{x}_i, y_i) = \sum [1 + D(\mathbf{W}\mathbf{x}_i, \mathbf{u}_{y_i}) - D(\mathbf{W}\mathbf{x}_i, \mathbf{u}_c)]$$

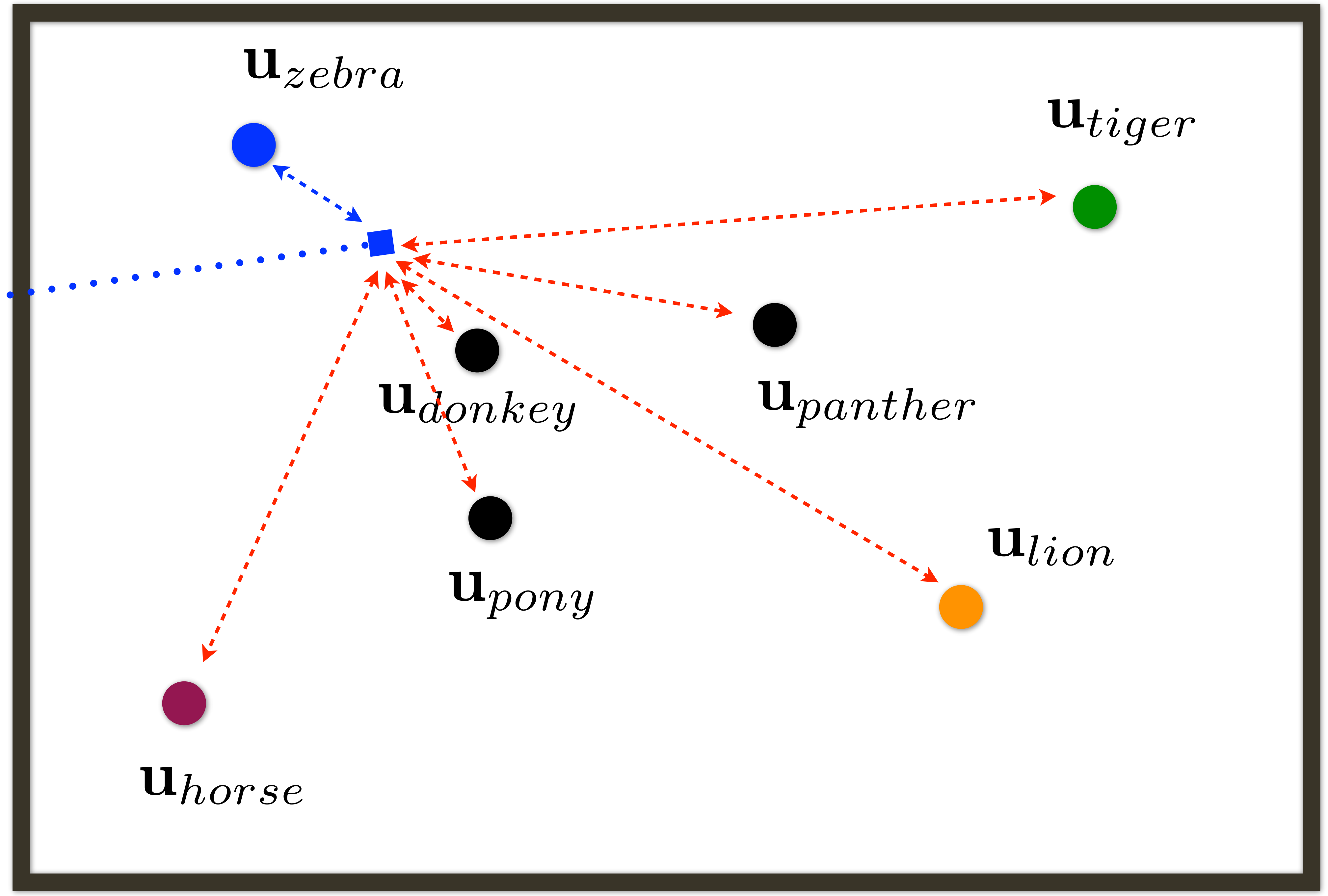
Image Embedding

$$\Psi(I_i) = \mathbf{W} \cdot CNN(I_i; \mathbf{\Theta}) \colon \mathbb{R}^D \to \mathbb{R}^d$$

Label Embedding

$$\Psi_L(word_i) = \mathbf{u}_i : \{1, ..., L\} \to \mathbb{R}^d$$

$L = 310,000$



Similarity in Embedding Space

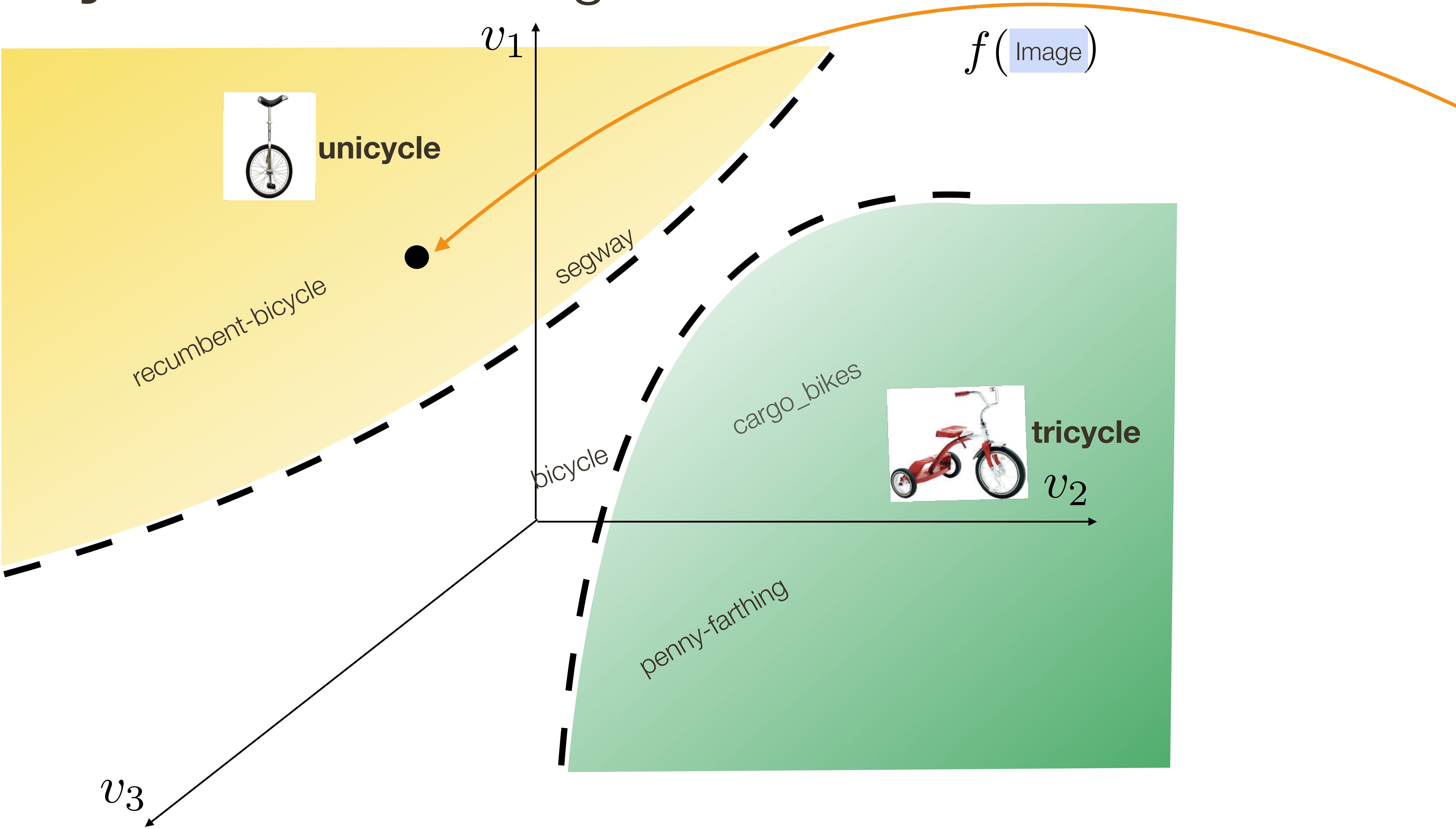
$$D(\mathbf{u}, \mathbf{u}') = ||\mathbf{u} - \mathbf{u}'||_2^2$$

Objective Function:

$$\min_{\mathbf{W}} \sum_i^N \mathcal{L}_C(\mathbf{W}, \mathbf{V}, I_i, y_i) + \mathcal{L}_R(\mathbf{W}, \mathbf{V}, I_i, y_i) + \mu ||V||_F^2$$

Vocabulary Informed Recognition

[Fu et al., 2016]



Experiments: Datasets

[Fu et al., 2016]

Animals with Attributes

Otter

Polar Bear

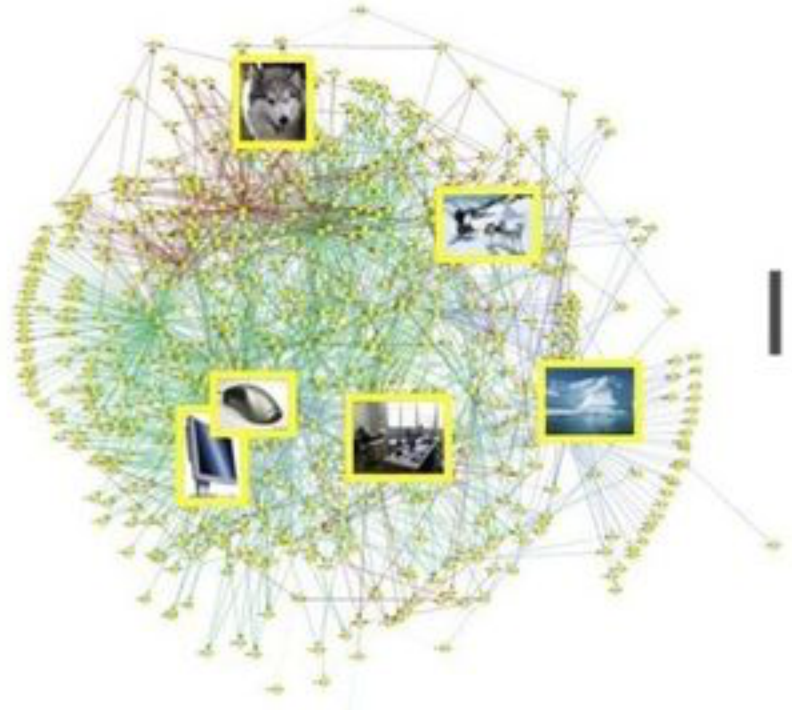
...

Auxiliary: 40 Animal Classes (annotated)

Target: 10 Animal Classes (**NO** annotation)

[Lampert, Nickisch, Harmeling CVPR'09]

ImageNet



IMAGENET

Auxiliary: 1,000 General Classes (annotated)

Target: 360 General Classes (**NO** annotation)

[Deng *et al.*, CVPR'09]

Experiments: Settings

[Fu et al., 2016]

AwA/ImageNet	No. Testing Classes			No. Testing Words	
	Auxiliary	Target	Total	Vocabulary	Chance(%)
SUPERVISED	✓		40/1000	40/1000	2.5/0.1
ZERO-SHOT		✓	10/360	10/360	10/0.28
OPEN-SET	✓	✓	50/1360	310K/310K	3.2E-04

The tasks are only separated in **evaluation**;
We train **one unified model** for all the settings

Experiments: Settings

[Fu et al., 2016]

Training

Otter

Polar Bear

Donkey

Poney

Panther

310, 000

Testing

Supervised

Zero-shot

Experiments: Settings

[Fu et al., 2016]

Training

Otter

Polar Bear

Donkey

Poney

Panther

$310,000$

Testing

Open-set

Experiments: Settings

[Fu et al., 2016]

AwA/ImageNet	No. Testing Classes			No. Testing Words	
	Auxiliary	Target	Total	Vocabulary	Chance(%)
SUPERVISED	✓		40/1000	40/1000	2.5/0.1
ZERO-SHOT		✓	**10**/360	**10**/360	**10**/0.28
OPEN-SET	✓	✓	50/1360	310K/310K	3.2E-04

The tasks are only separated in **evaluation**;
We train **one unified model** for all the settings

Zero-shot Results

[Fu et al., 2016]

Results with AWA

Method	Features	Accuracy
SS-Voc: full instances	$CNN_{OverFeat}$	78.3
Akata *et al*. CVPR 2015	$CNN_{GoogLeNet}$	73.9
TMV-BLP (Fu *et al*. ECCV 2014)	$CNN_{OverFeat}$	69.9
AMP (SR+SE) (Fu *et al*. CVPR 2015)	$CNN_{OverFeat}$	66.0
DAP (Lampert *et al*. TPAMI 2013)	CNN_{VGG19}	57.5
PST (Rohrbach *et al*. NIPS 2013)	$CNN_{OverFeat}$	53.2
DS (Rohrbach *et al*. CVPR 2010)	$CNN_{OverFeat}$	52.7
IAP (Lampert *et al*. TPAMI 2013)	$CNN_{OverFeat}$	44.5
HEX (Deng *et al*. ECCV 2014)	CNN_{DECAF}	44.2

+4.4%

Zero-shot Results

[Fu et al., 2016]

Results with AWA

3.3% of training data

Method	Features	Accuracy
SS-Voc: full instances	$\text{CNN}_{\text{OverFeat}}$	78.3
800 instances (20 inst*40 class);	$\text{CNN}_{\text{OverFeat}}$	74.4
Akata *et al*. CVPR 2015	$\text{CNN}_{\text{GoogLeNet}}$	73.9
TMV-BLP (Fu *et al*. ECCV 2014)	$\text{CNN}_{\text{OverFeat}}$	69.9
AMP (SR+SE) (Fu *et al*. CVPR 2015)	$\text{CNN}_{\text{OverFeat}}$	66.0
DAP (Lampert *et al*. TPAMI 2013)	$\text{CNN}_{\text{VGG19}}$	57.5
PST (Rohrbach *et al*. NIPS 2013)	$\text{CNN}_{\text{OverFeat}}$	53.2
DS (Rohrbach *et al*. CVPR 2010)	$\text{CNN}_{\text{OverFeat}}$	52.7
IAP (Lampert *et al.* TPAMI 2013)	$\text{CNN}_{\text{OverFeat}}$	44.5
HEX (Deng *et al*. ECCV 2014)	$\text{CNN}_{\text{DECAF}}$	44.2

+0.5%

Zero-shot Results

[Fu et al., 2016]

Results with AWA

0.82% of training data

Method	Features	Accuracy
SS-Voc: full instances	$\text{CNN}_{\text{OverFeat}}$	78.3
800 instances (20 inst*40 class);	$\text{CNN}_{\text{OverFeat}}$	74.4
200 instances (5 inst*40 class);	$\text{CNN}_{\text{OverFeat}}$	68.9
Akata *et al*. CVPR 2015	$\text{CNN}_{\text{GoogLeNet}}$	73.9
TMV-BLP (Fu *et al*. ECCV 2014)	$\text{CNN}_{\text{OverFeat}}$	69.9
AMP (SR+SE) (Fu *et al*. CVPR 2015)	$\text{CNN}_{\text{OverFeat}}$	66.0
DAP (Lampert *et al*. TPAMI 2013)	$\text{CNN}_{\text{VGG19}}$	57.5
PST (Rohrbach *et al*. NIPS 2013)	$\text{CNN}_{\text{OverFeat}}$	53.2
DS (Rohrbach *et al*. CVPR 2010)	$\text{CNN}_{\text{OverFeat}}$	52.7
IAP (Lampert *et al.* TPAMI 2013)	$\text{CNN}_{\text{OverFeat}}$	44.5
HEX (Deng *et al*. ECCV 2014)	$\text{CNN}_{\text{DECAF}}$	44.2

Weakly-supervised **Visual Grounding** of Phrases

[Xiao et al., 2017]

Given **image-sentence pairs** learn how to **localize** arbitrary language phrase or sentence in new images

The man at bat readies to swing at the pitch while the umpire looks on.

A large bus sitting next to a very tall building.

[Lin, Maire, Belongie, Hays, Perona, Ramanan, Dollar, Zitnick, ECCV'14]

Weakly-supervised **Visual Grounding** of Phrases

[Xiao et al., 2017]

Given **image-sentence pairs** learn how to **localize** arbitrary language phrase or sentence in new images

The man at bat readies to swing at the pitch while the umpire looks on.

A large bus sitting next to a very tall building.

a man

[Lin, Maire, Belongie, Hays, Perona, Ramanan, Dollar, Zitnick, ECCV'14]

Weakly-supervised **Visual Grounding** of Phrases

[Xiao et al., 2017]

Given **image-sentence pairs** learn how to **localize** arbitrary language phrase or sentence in new images

The man at bat readies to swing at the pitch while the umpire looks on.

A large bus sitting next to a very tall building.

[Lin, Maire, Belongie, Hays, Perona, Ramanan, Dollar, Zitnick, ECCV'14]

a table

Weakly-supervised **Visual Grounding** of Phrases

[Xiao et al., 2017]

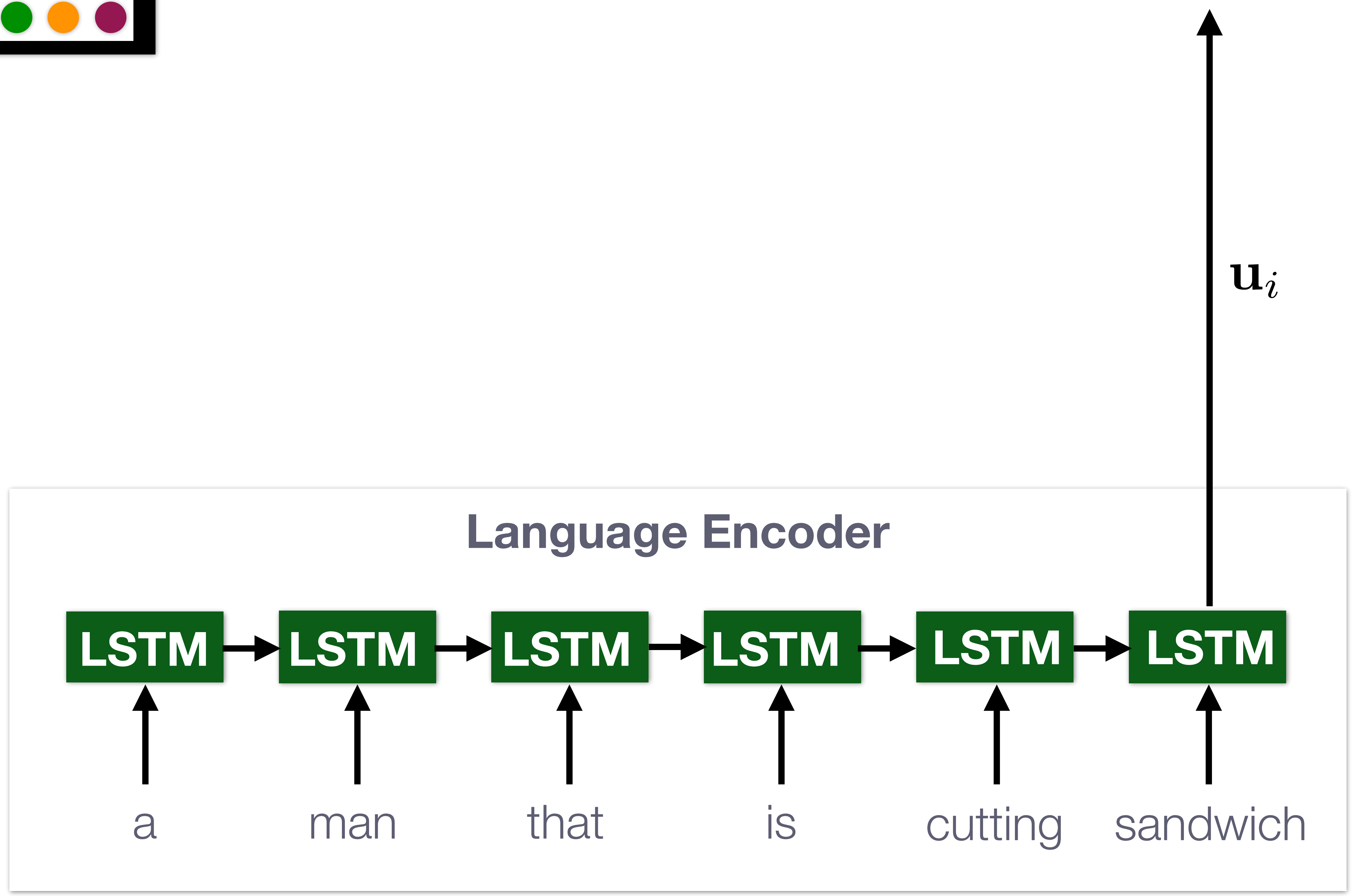
Label Embedding

$$\Psi_L(phrase_i) = \mathbf{u}_i$$

Weakly-supervised **Visual Grounding** of Phrases [Xiao et al., 2017]

Label Embedding

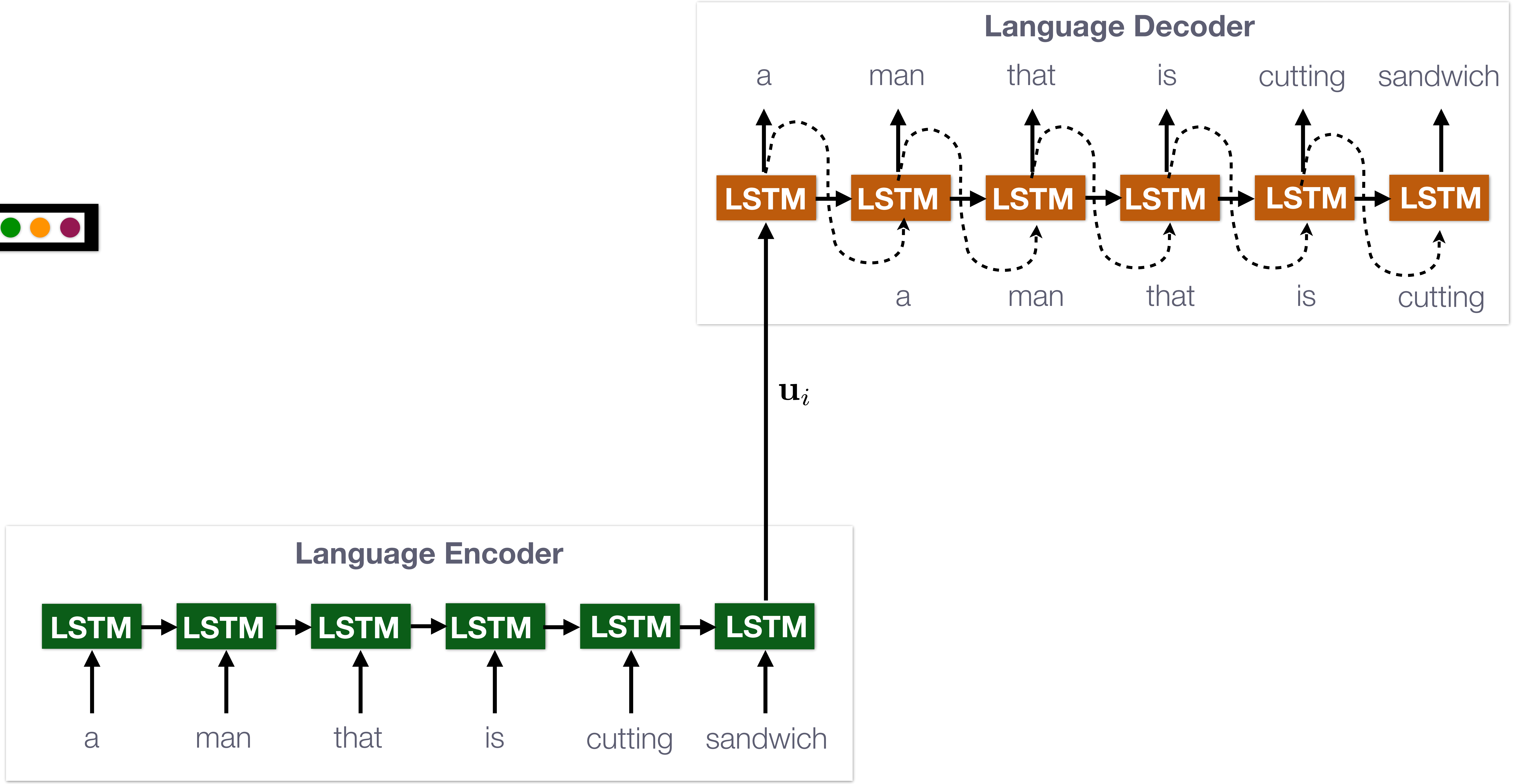
$$\Psi_L(phrase_i) = \mathbf{u}_i$$



Weakly-supervised **Visual Grounding** of Phrases [Xiao et al., 2017]

Label Embedding

$$\Psi_L(phrase_i) = \mathbf{u}_i$$

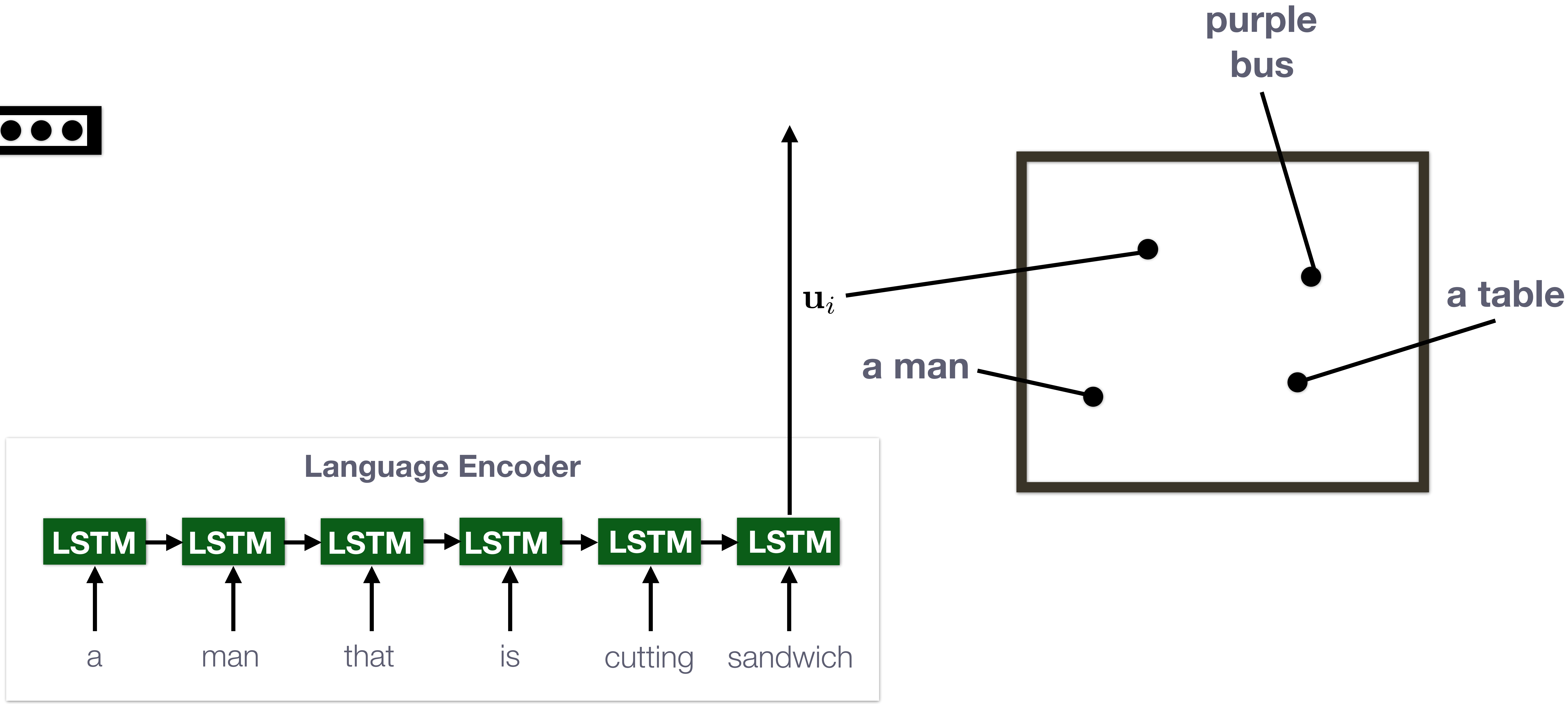


Weakly-supervised **Visual Grounding** of Phrases

[Xiao et al., 2017]

Label Embedding

$$\Psi_L(phrase_i) = \mathbf{u}_i$$

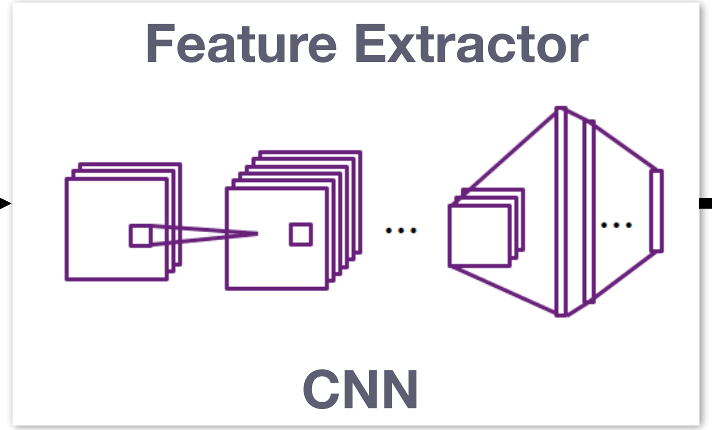


Weakly-supervised **Visual Grounding** of Phrases

[Xiao et al., 2017]

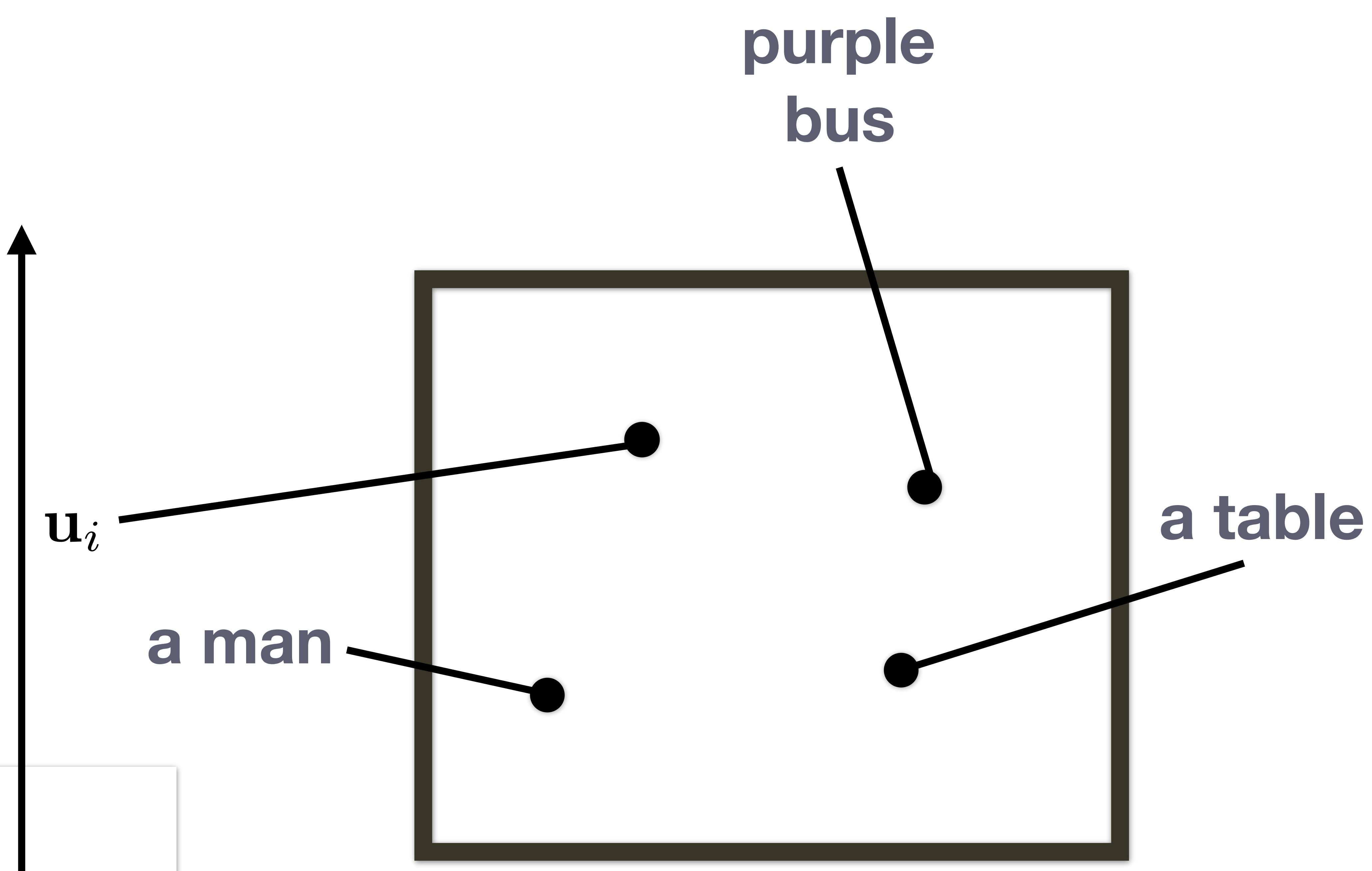
Image Embedding

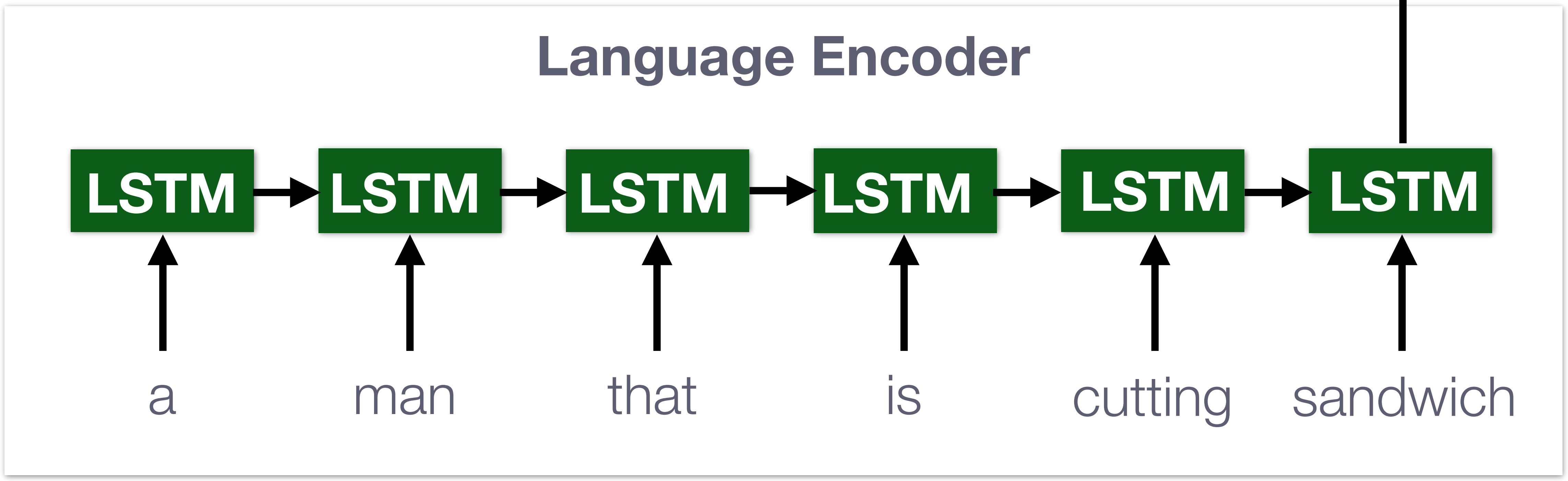
$$\Psi(I_i) = \mathbf{W} \cdot CNN(I_i; \mathbf{\Theta})$$



Label Embedding

$$\Psi_L(phrase_i) = \mathbf{u}_i$$





Weakly-supervised **Visual Grounding** of Phrases

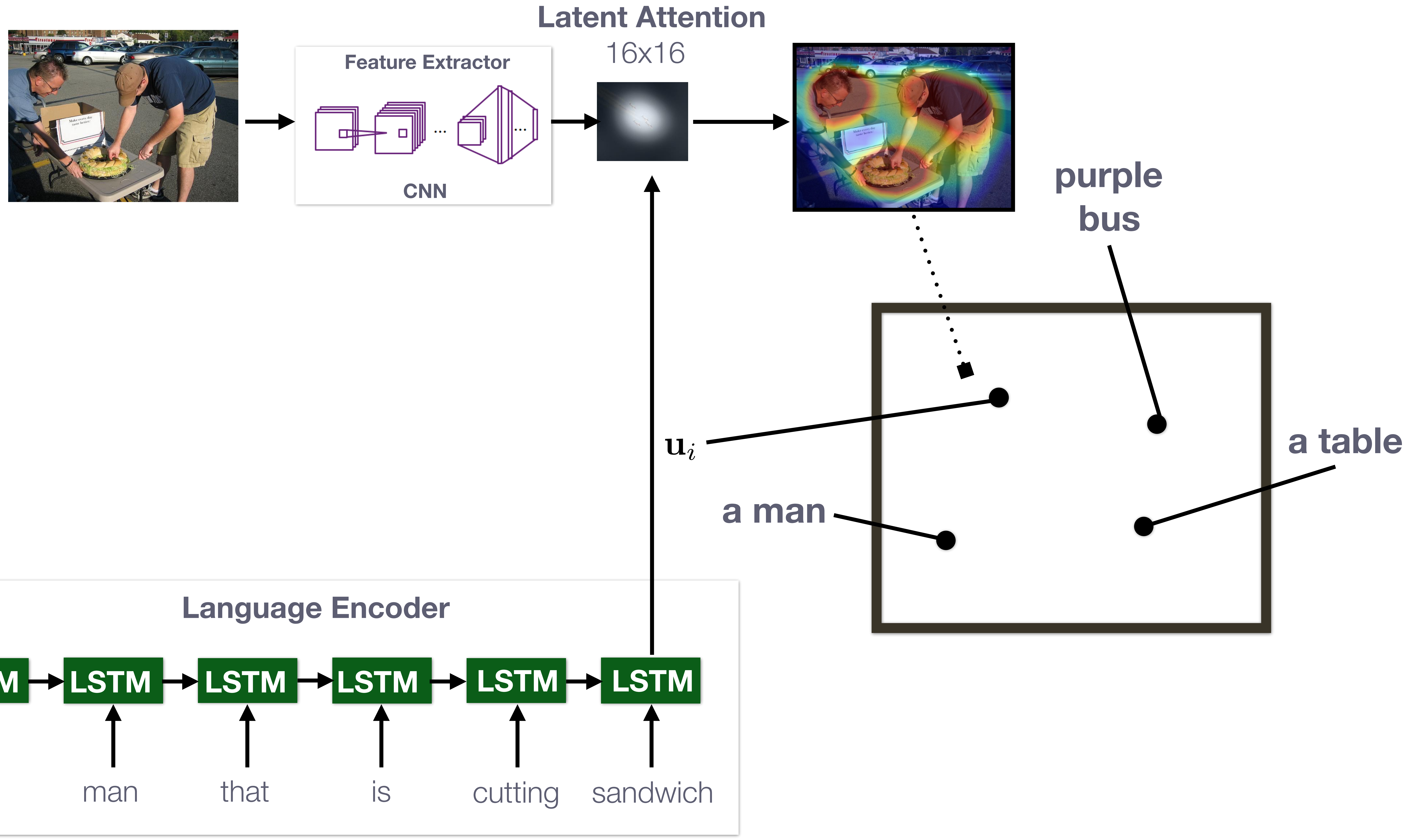
[Xiao et al., 2017]

Image Embedding

$$\Psi(I_i) = \mathbf{W} \cdot CNN(I_i; \mathbf{\Theta})$$

Label Embedding

$$\Psi_L(phrase_i) = \mathbf{u}_i$$



Weakly-supervised **Visual Grounding** of Phrases

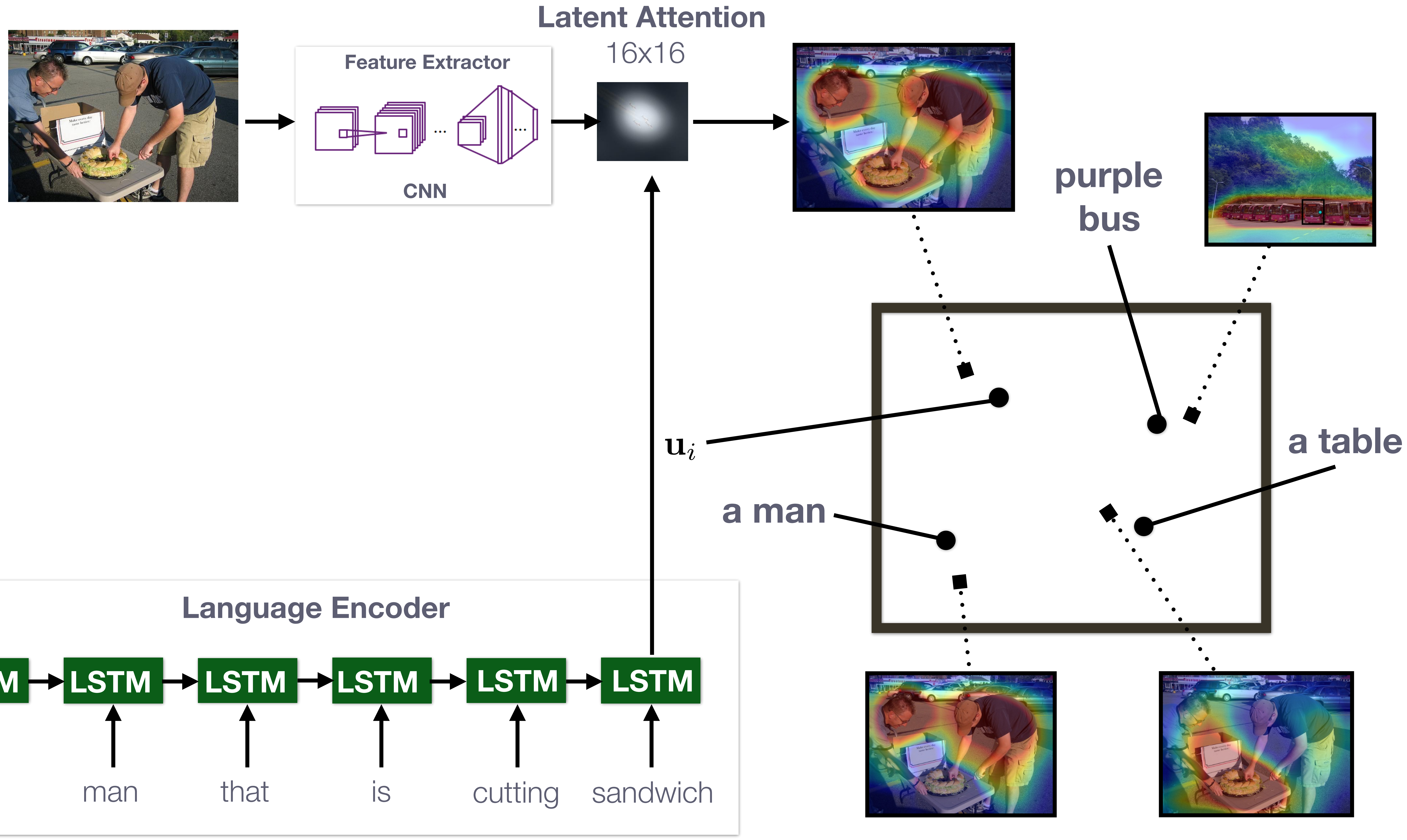
[Xiao et al., 2017]

Image Embedding

$$\Psi(I_i) = \mathbf{W} \cdot CNN(I_i; \mathbf{\Theta})$$

Label Embedding

$$\Psi_L(phrase_i) = \mathbf{u}_i$$



Weakly-supervised **Visual Grounding** of Phrases

[Xiao et al., 2017]

Image Embedding

$$\Psi(I_i) = \mathbf{W} \cdot CNN(I_i; \mathbf{\Theta})$$

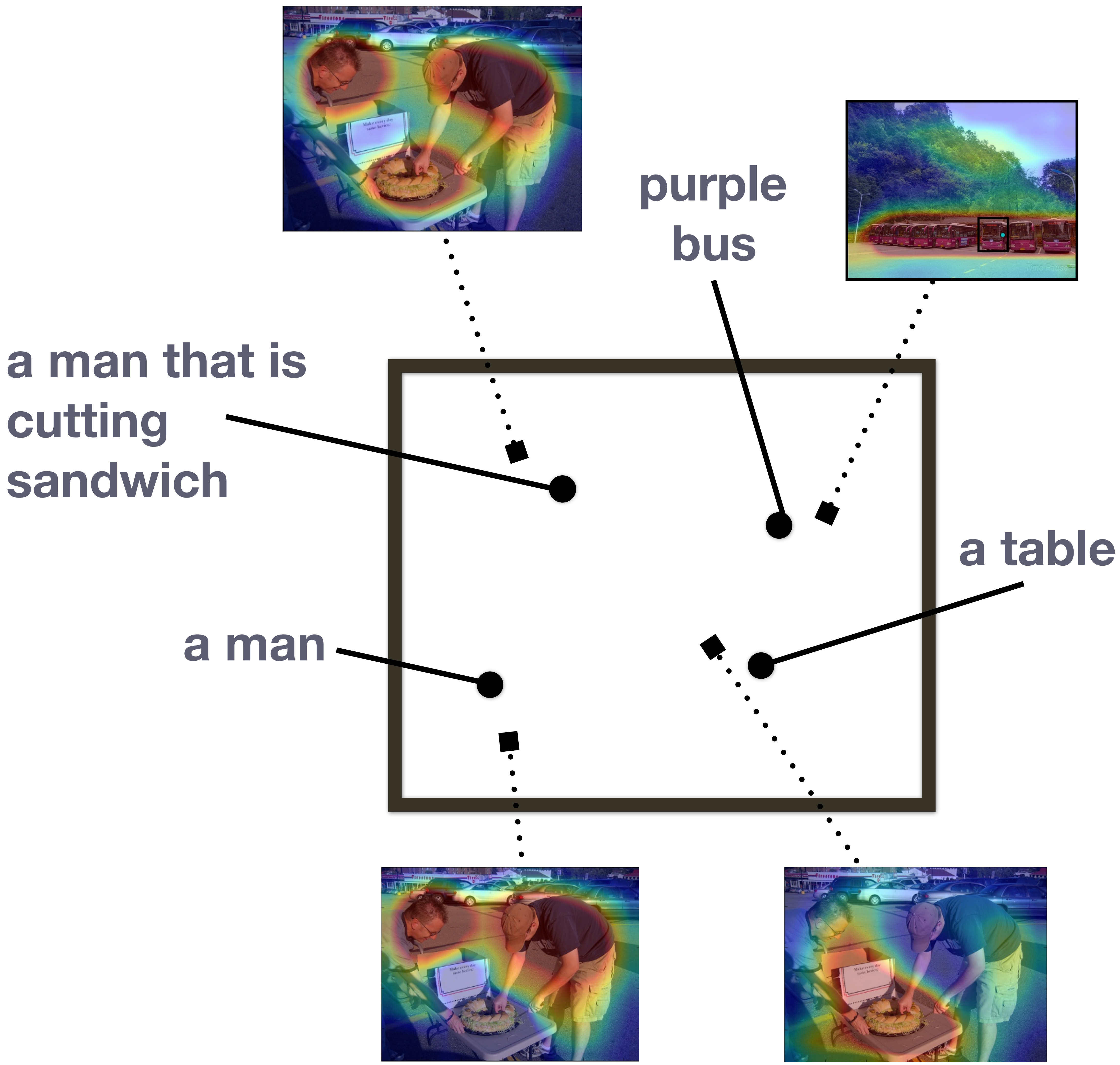
Label Embedding

$$\Psi_L(phrase_i) = \mathbf{u}_i$$

Similarity in Embedding Space

$$D(\mathbf{u}, \mathbf{u}') = ||\mathbf{u} - \mathbf{u}'||_2^2$$

Objective Function:



Combination of previous discriminative similarity and **linguistic regularization**

Weakly-supervised **Visual Grounding** of Phrases

[Xiao et al., 2017]

For **noun phrases**:

- **siblings** should have **disjoin**
- **parents** should be **union of**

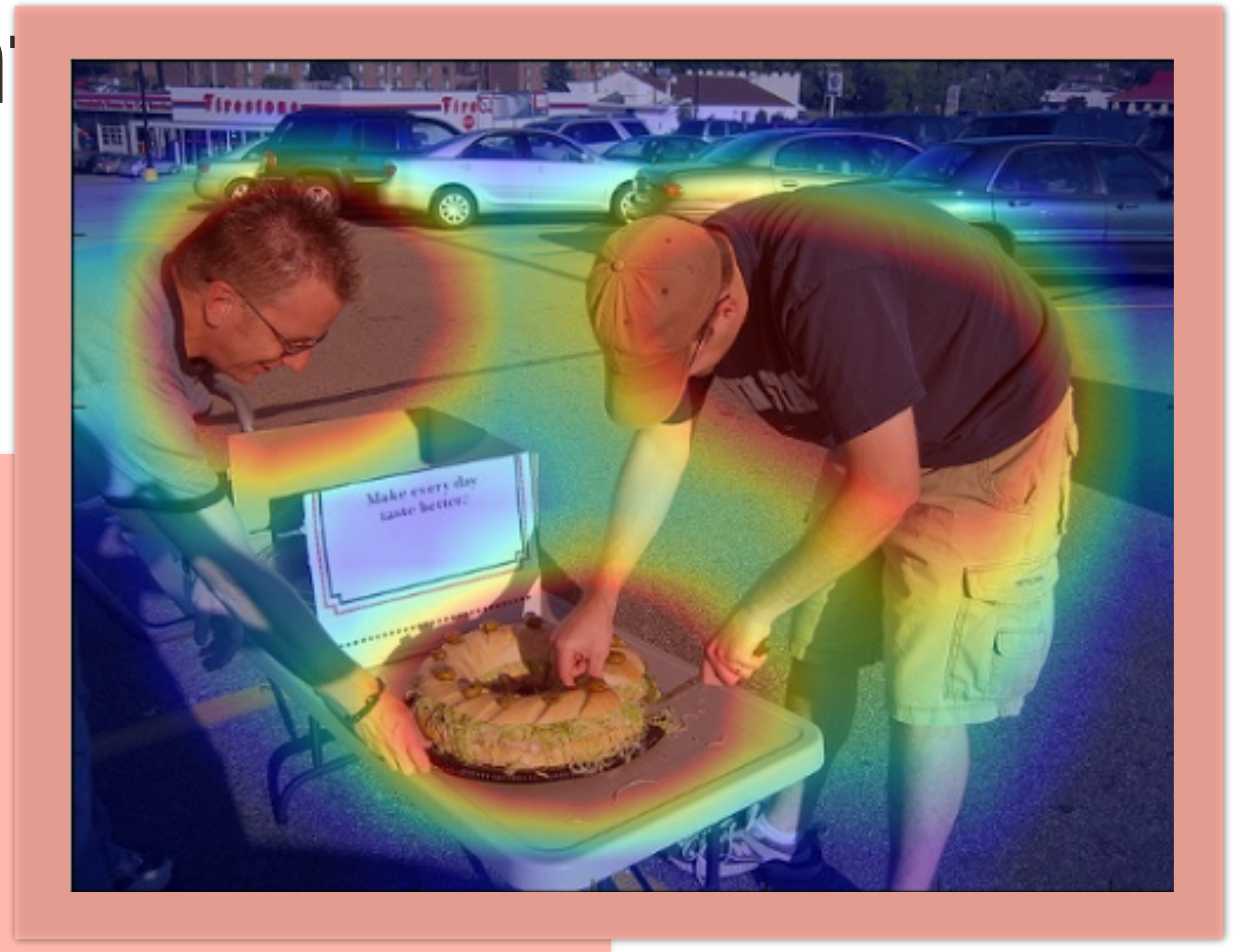
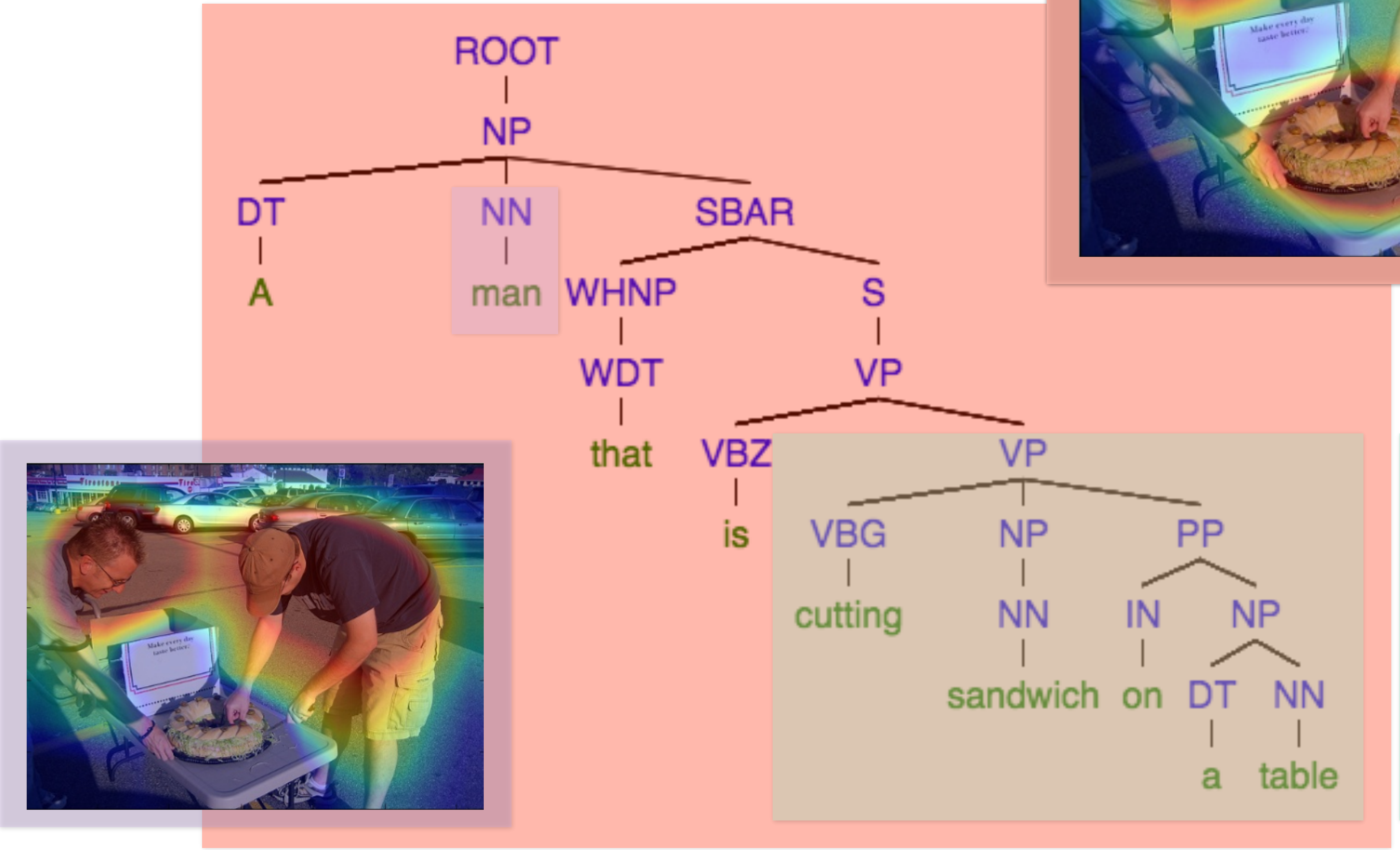


Image Embedding

$$\Psi(I_i) = \mathbf{W} \cdot CNN(I_i; \mathbf{\Theta})$$

Label Embedding

$$\Psi_L(phrase_i) = \mathbf{u}_i$$



Similarity in Embedding Space

$$D(\mathbf{u}, \mathbf{u}') = ||\mathbf{u} - \mathbf{u}'||_2^2$$

Objective Function:

Combination of previous discriminative similarity and **linguistic regularization**

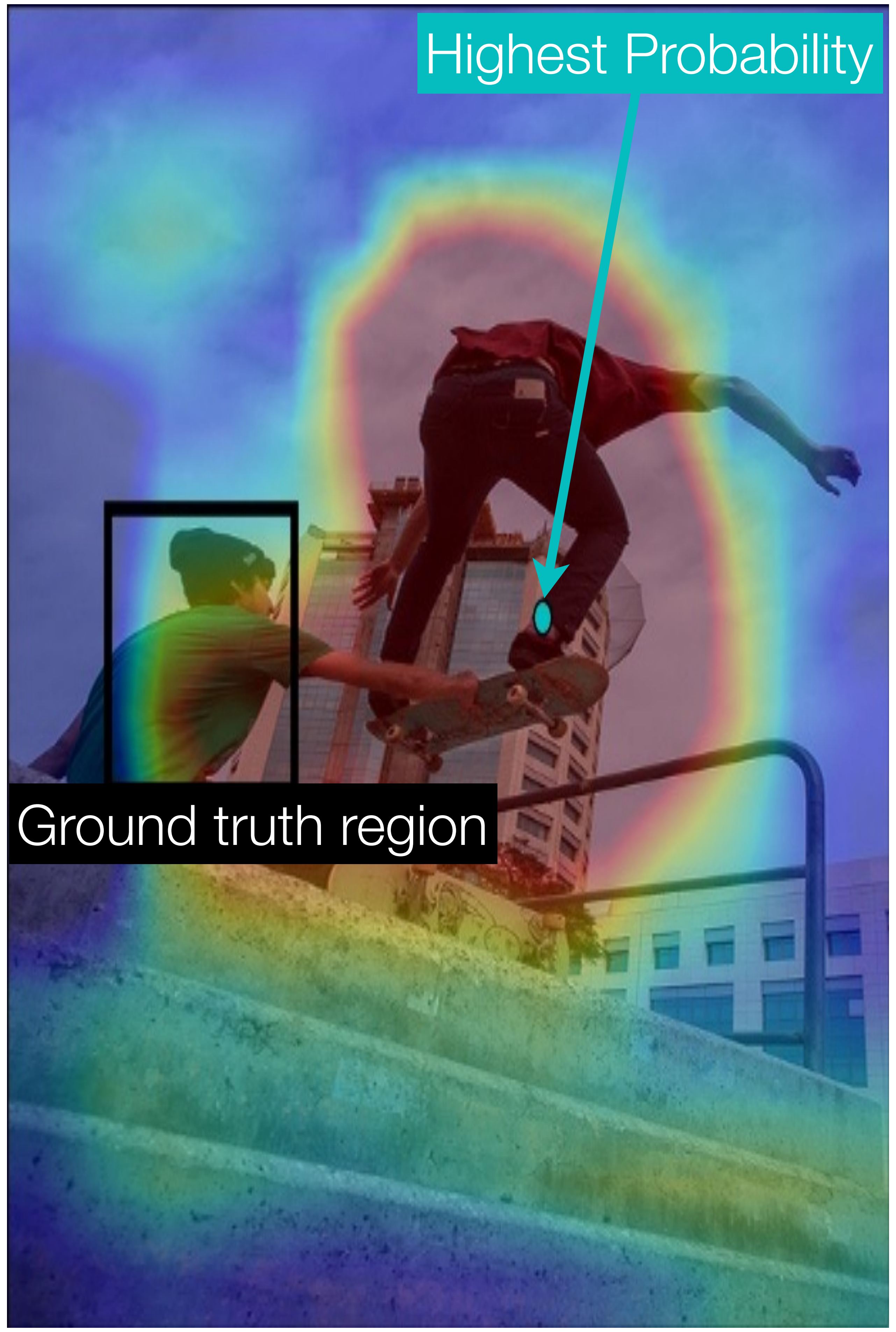
Qualitative **Results**

[Xiao et al., 2017]

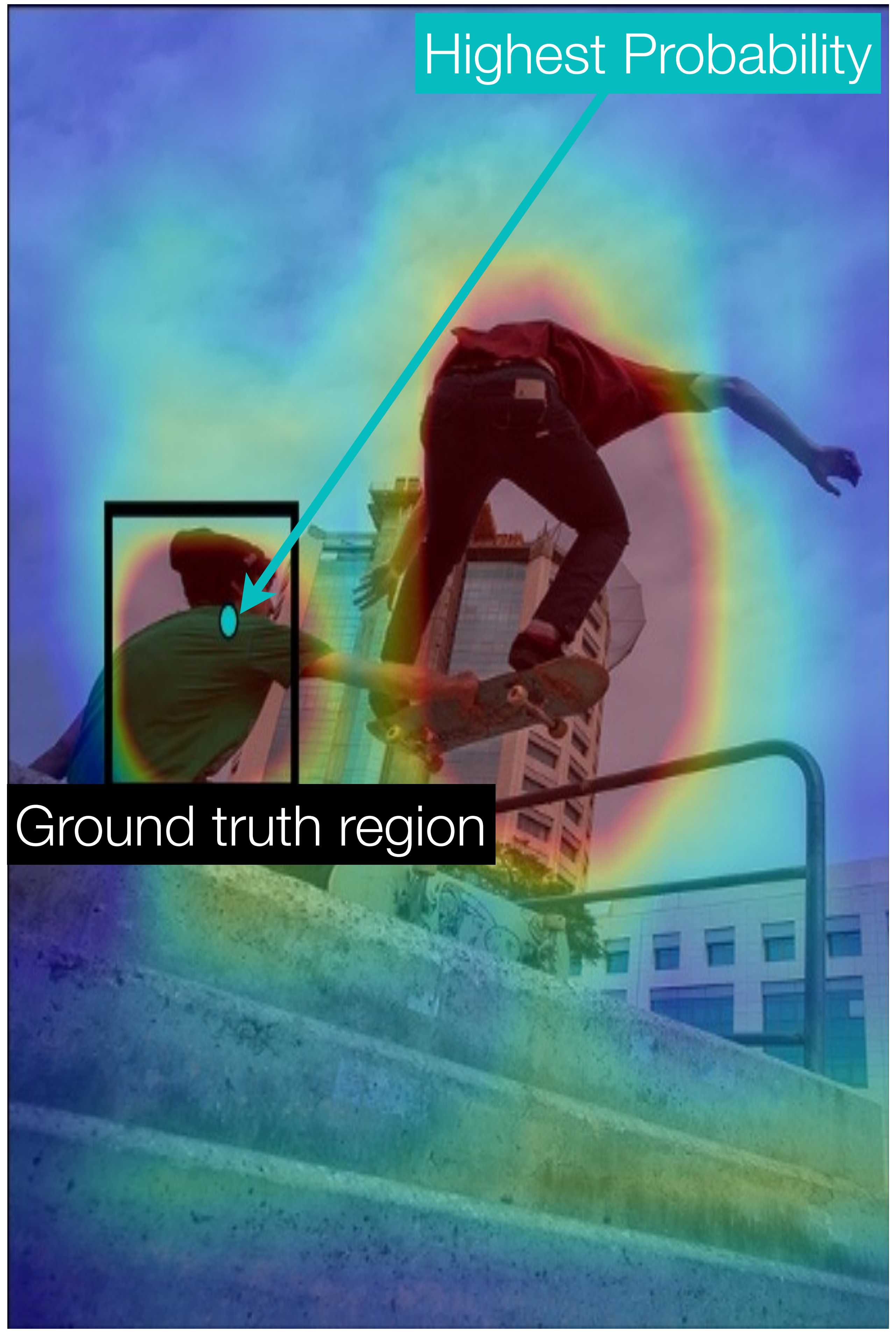
Input:

guy in green t-shirt holding skateboard

NO linguistic constraints



Our Model



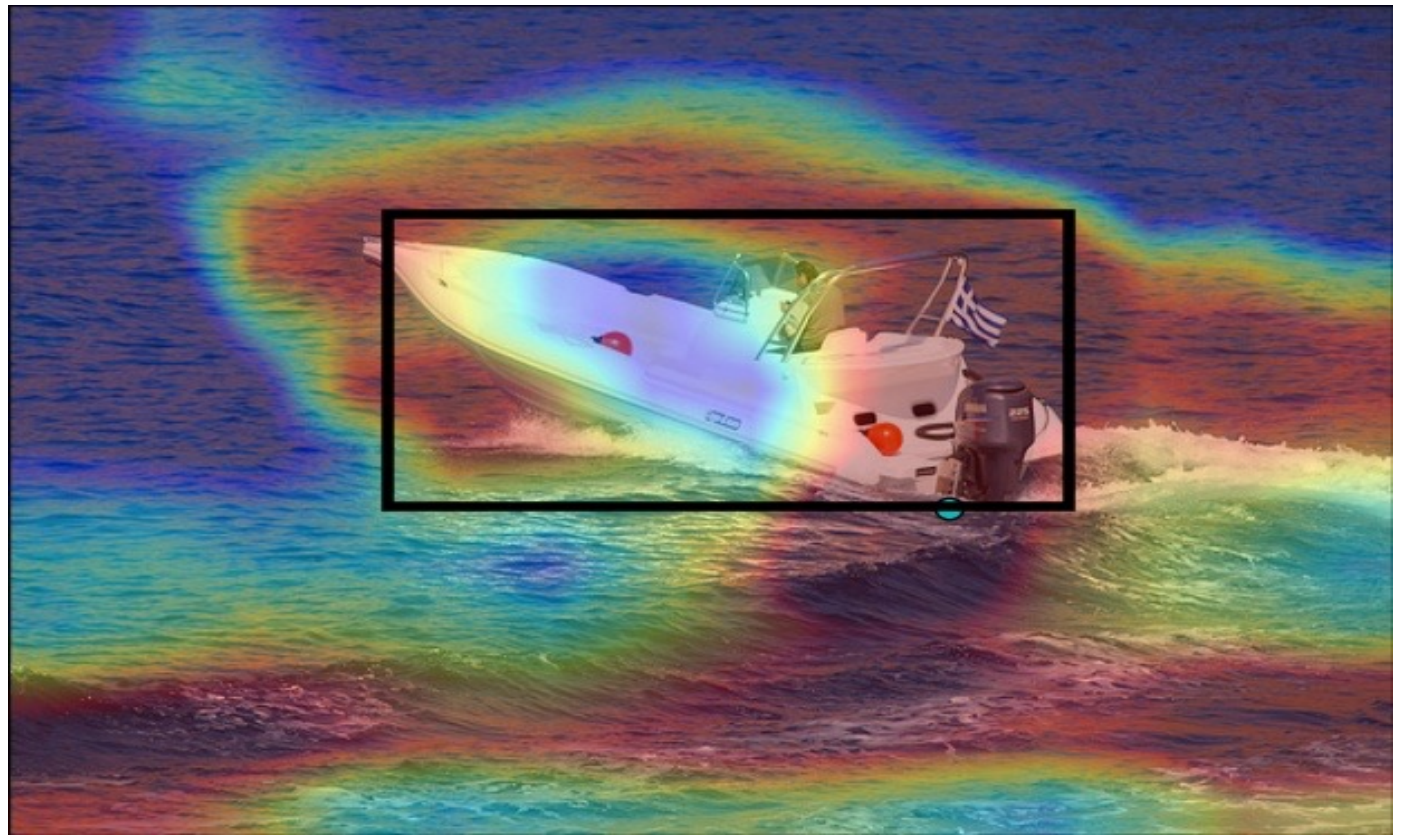
Qualitative **Results**

[Xiao et al., 2017]

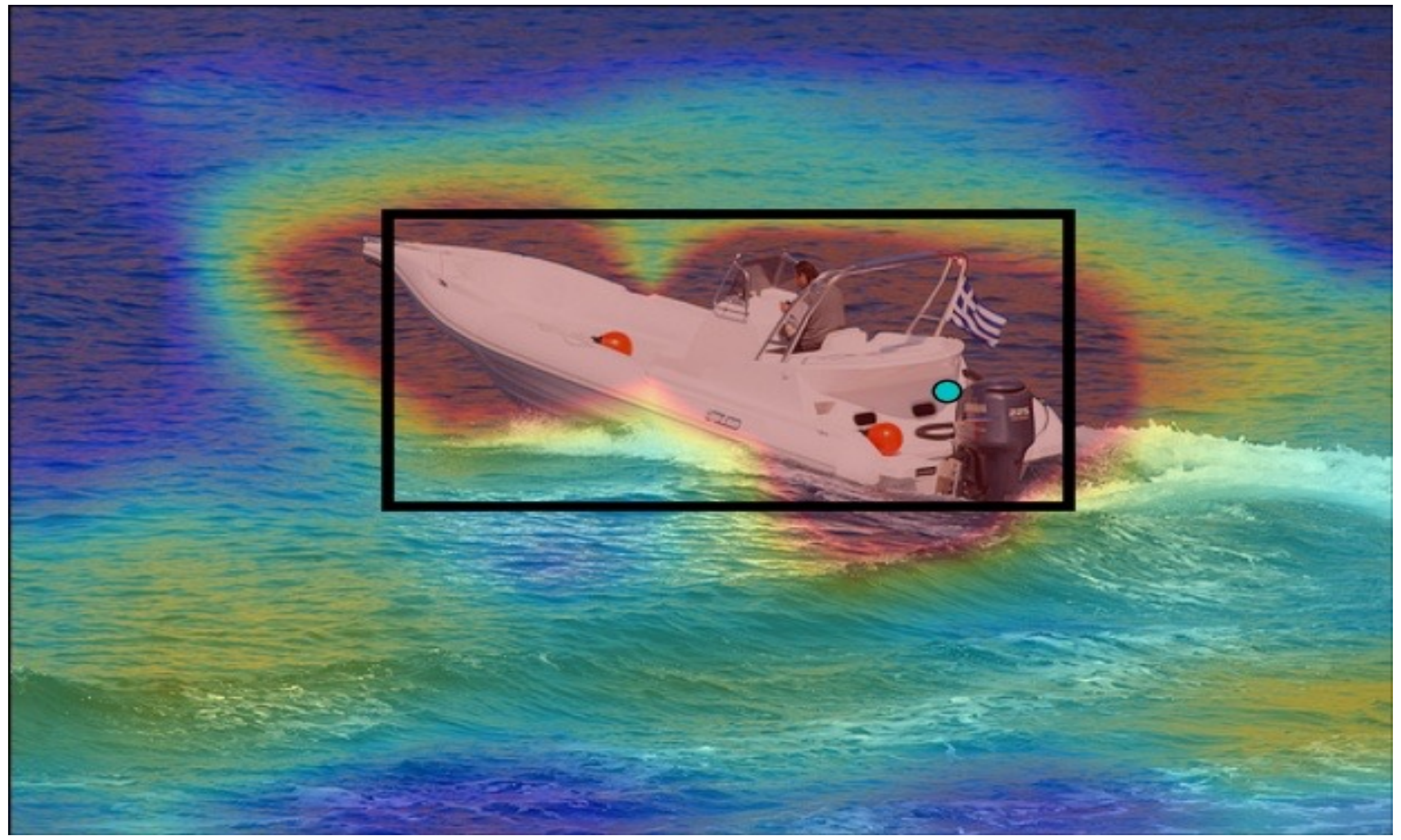
Input:

a person driving a boat

NO linguistic constraints



Our Model

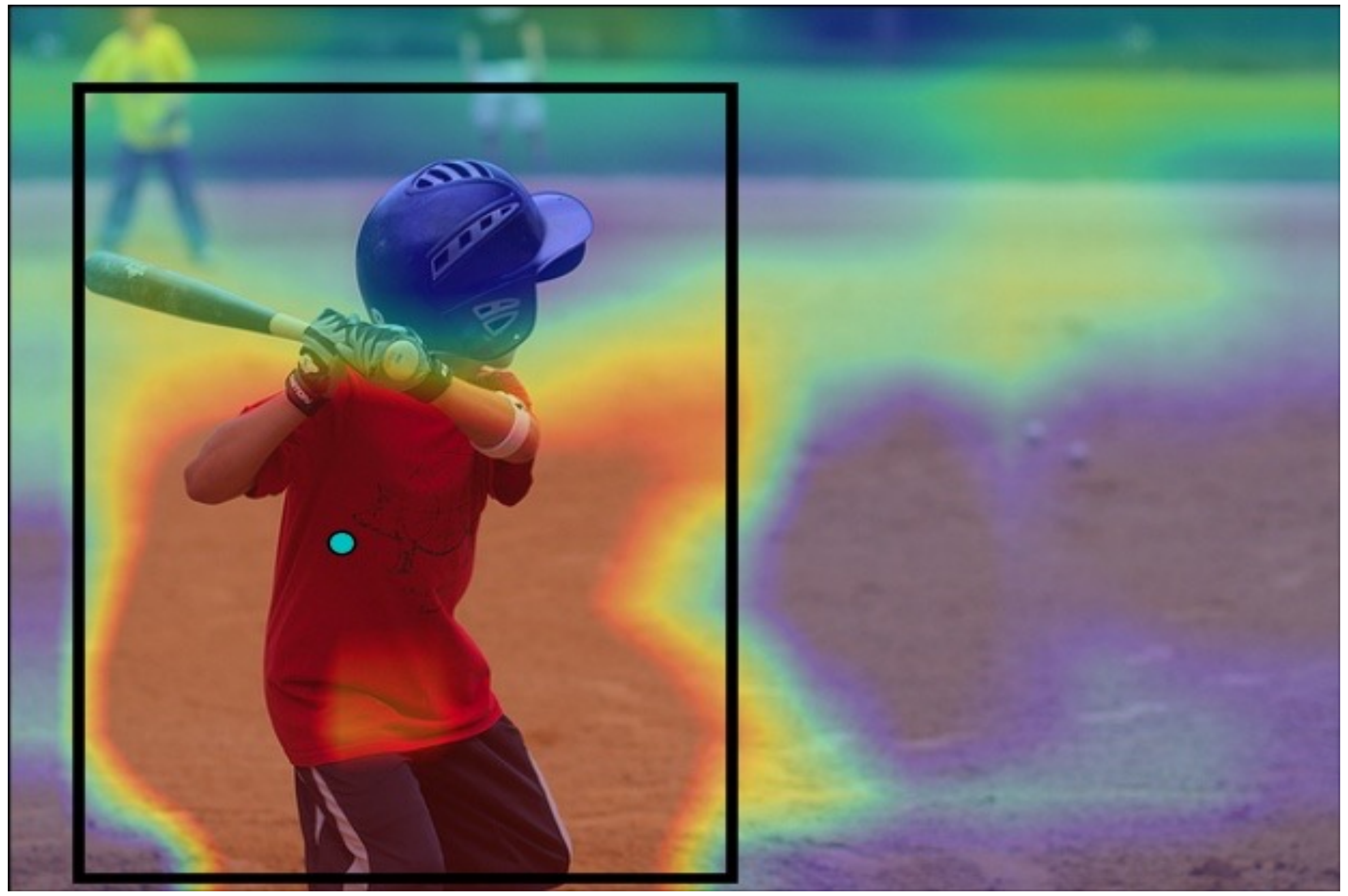


Qualitative **Results**

Input:

a child wearing black protective helmet

NO linguistic constraints [Xiao et al., 2017]



Our Model

Quantitative Results

[Xiao et al., 2017]

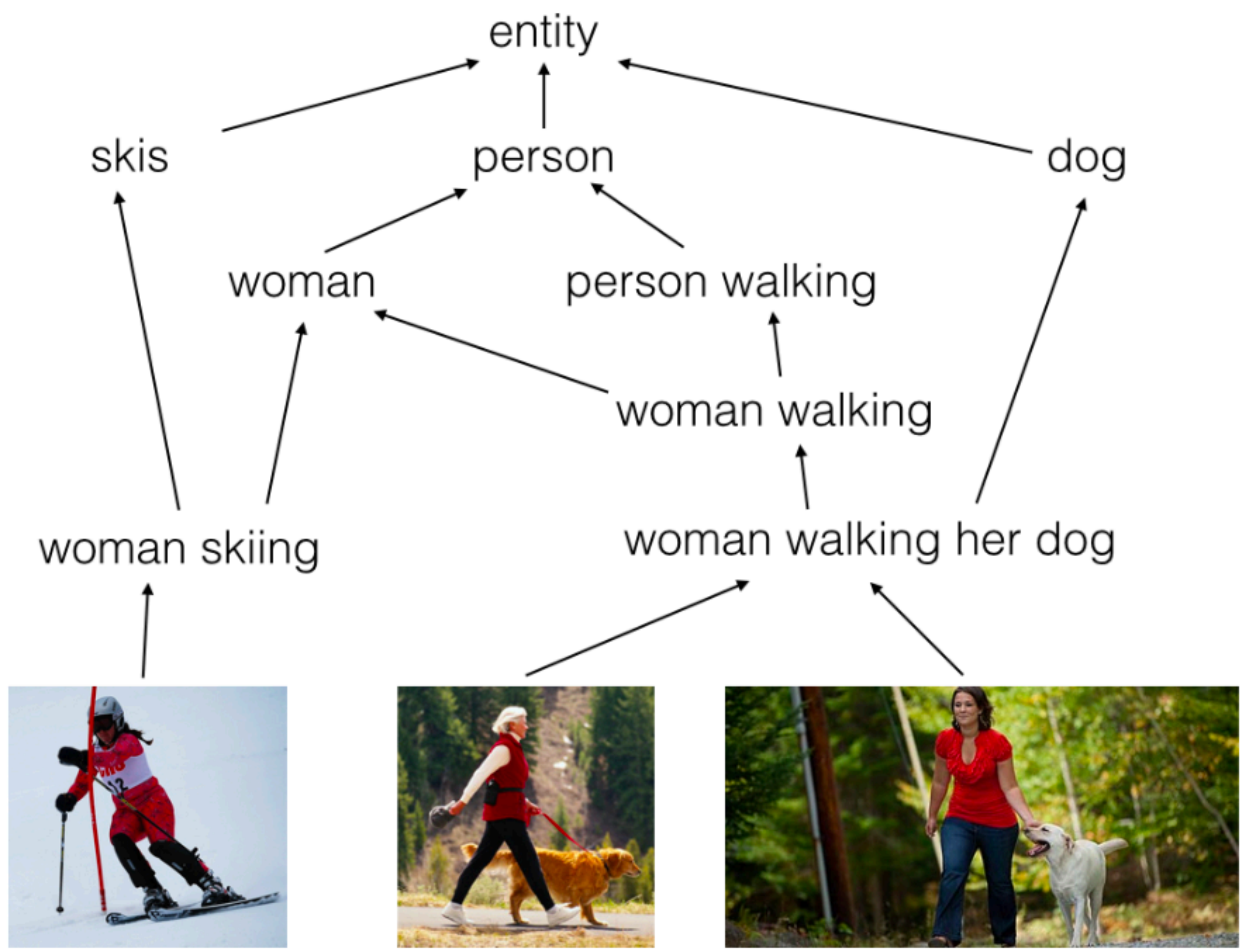
Segmentation performance on COCO dataset

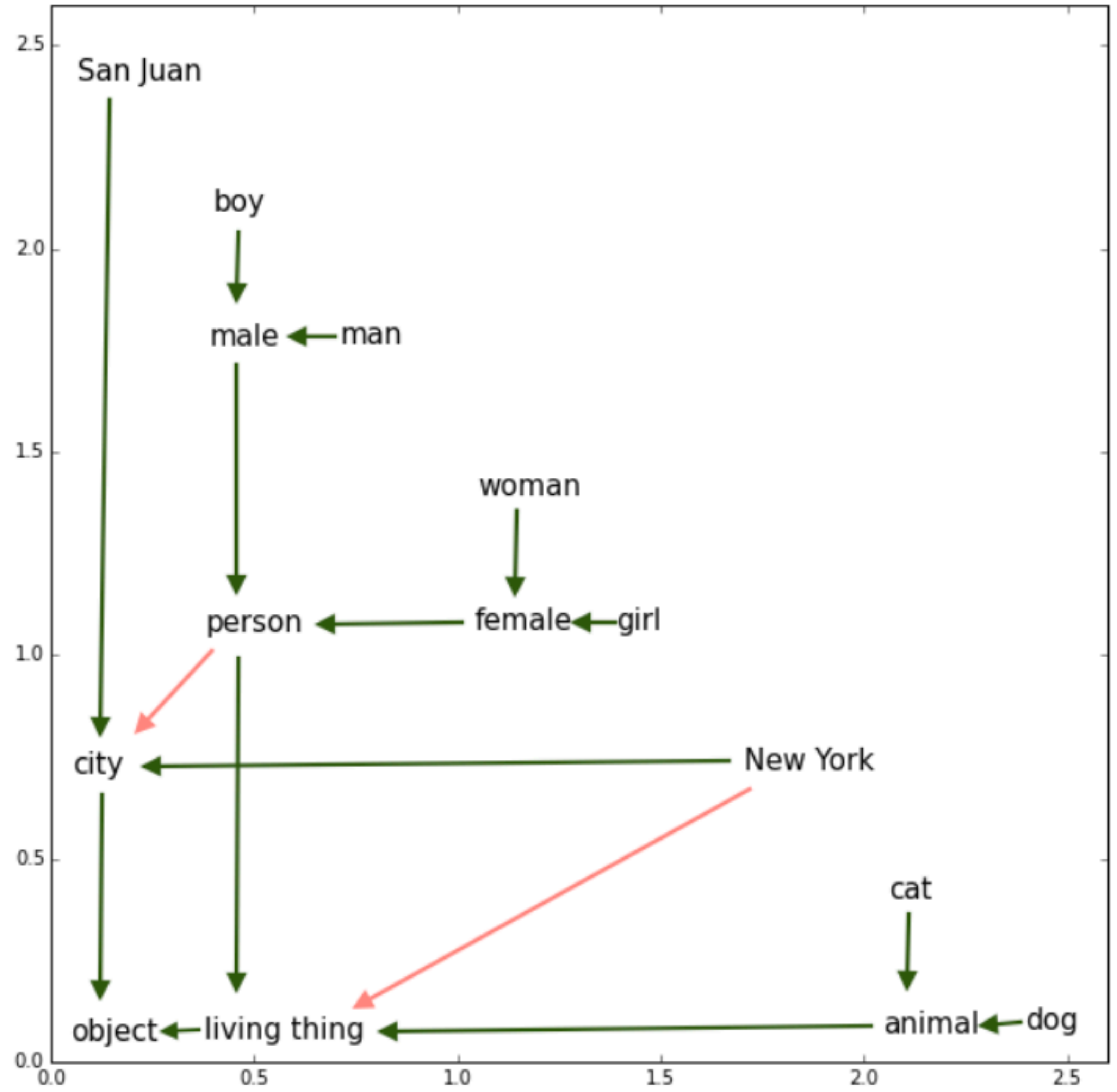
[Lin, Maire, Belongie, Hays, Perona, Ramanan, Dollar, Zitnick, ECCV'14]

	IoU@0.3	IoU@0.4	IoU@0.5	Avg mAP
Non-strcutred	0.302	0.199	0.110	0.203
Parent-Child	0.327	0.213	0.118	0.219
Sibling	0.316	0.203	0.114	0.211
Ours	**0.347**	**0.246**	**0.159**	**0.251**

Order Embeddings

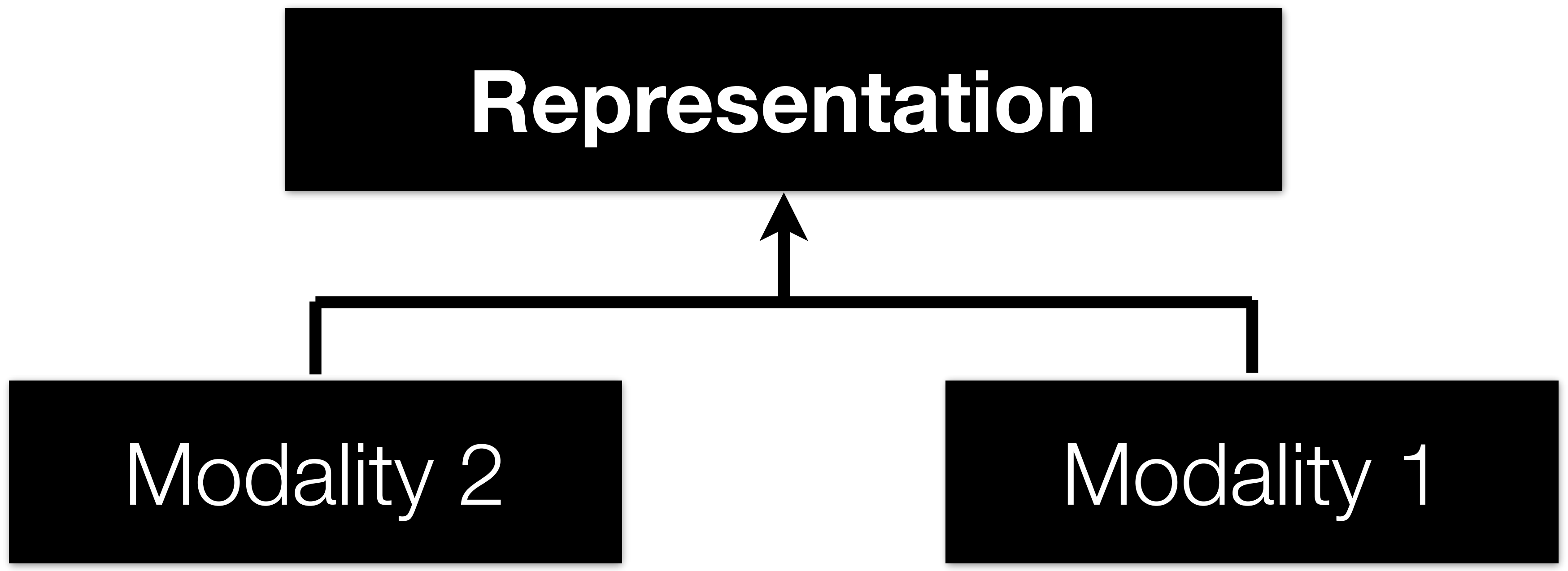
[Vendrov et al., 2016]





Multimodal Representation Types

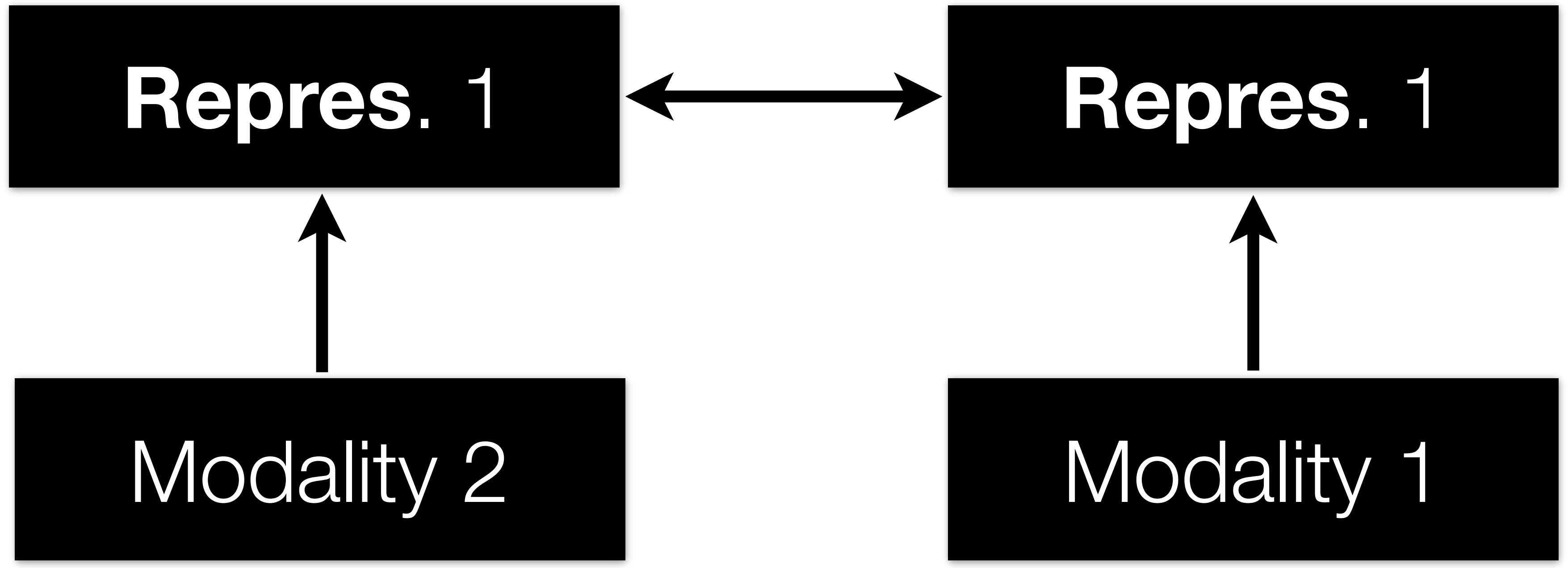
Joint representations:



— Simplest version: modality concatenation (early fusion)

— Can be learned supervised or unsupervised

Coordinated representations:



— Similarity-based methods (e.g., cosine distance)

— Structure constraints (e.g., orthogonality, sparseness)

— CCA (unsupervised), joint embeddings (supervised)

*slide from Louis-Philippe Morency

Final Words …

Joint representations

- Project modalities to the same space
- Use when all the modalities are present during test time
- Suitable for multi-model fusion

Coordinated representations

- Project modalities to their own coordinated spaces
- Use when only one of the modalities is present during test-time
- Suitable for multimodal translation
- Good for multimodal retrieval

*slide from Louis-Philippe Morency