DEEP SEMANTIC-VISUAL EMBEDDING WITH LOCALIZATION

Thursday 4th October, 2018

Martin Engilberge, Louis Chevallier,
Patrick Pérez, Matthieu Cord

Machine Learning &
Deep Learning for
Information Access

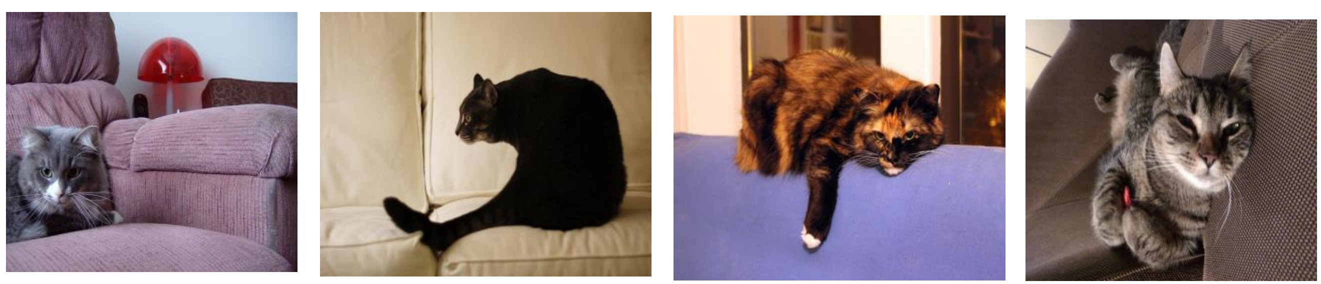
Tasks

Visual Grounding of phrases:

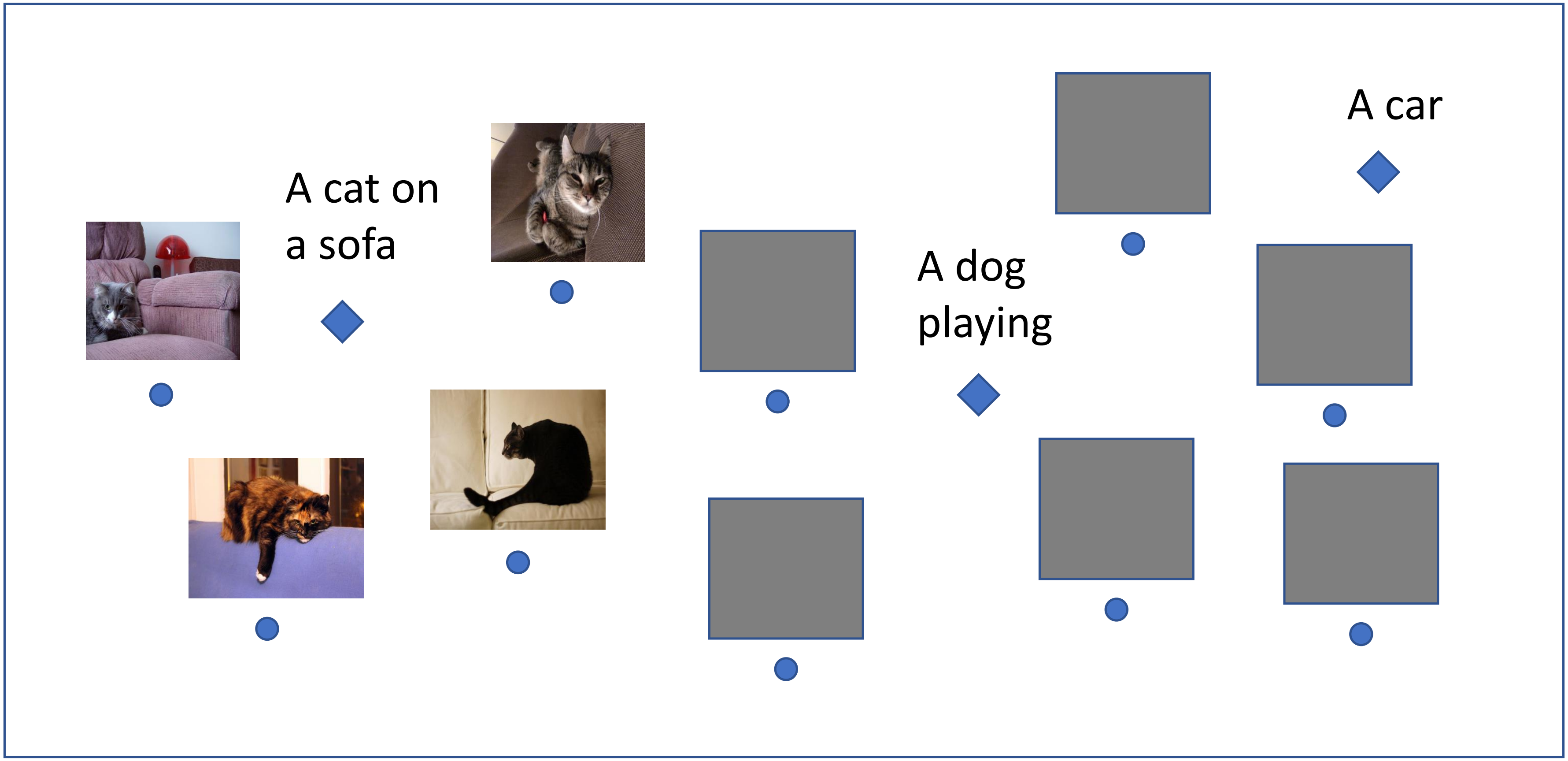
Localize any textual query into a given image.

Cross-modal retrieval:

Query: A cat on a sofa



Semantic visual embedding

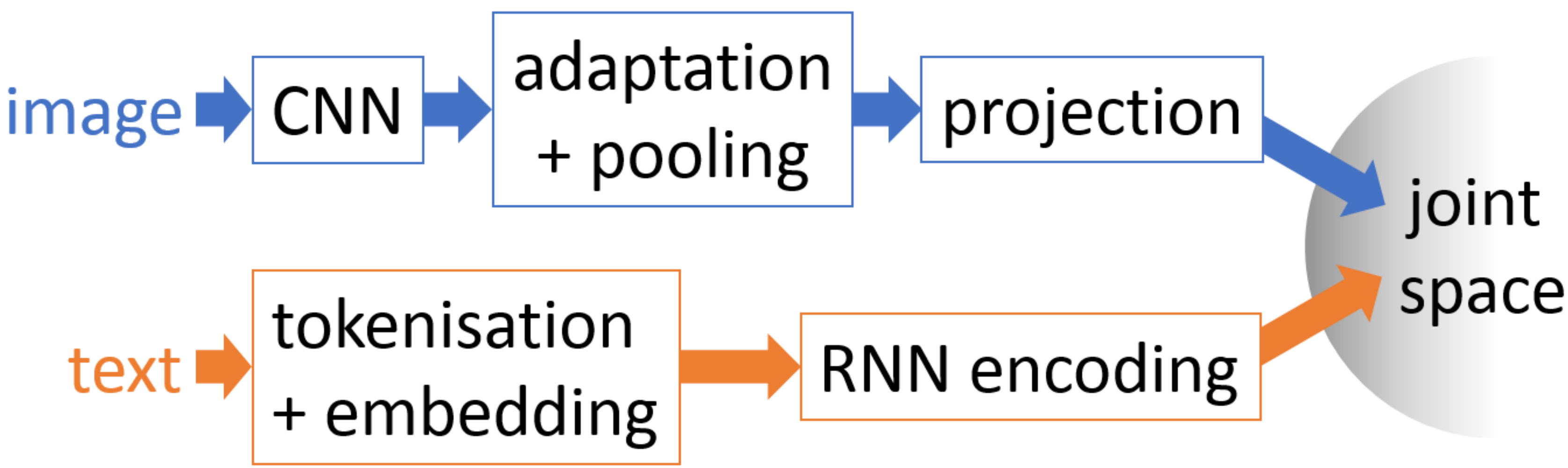


2D Semantic visual space example:

- Distance in the space has a semantic interpretation.
- Retrieval is done by finding nearest neighbors.

Approach

- Learning image and text joint embedding space.
- Visual grounding relying on the spatial-textual information modeling.
- Cross-modal retrieval leveraging the semantic space and the visual and textual alignment.



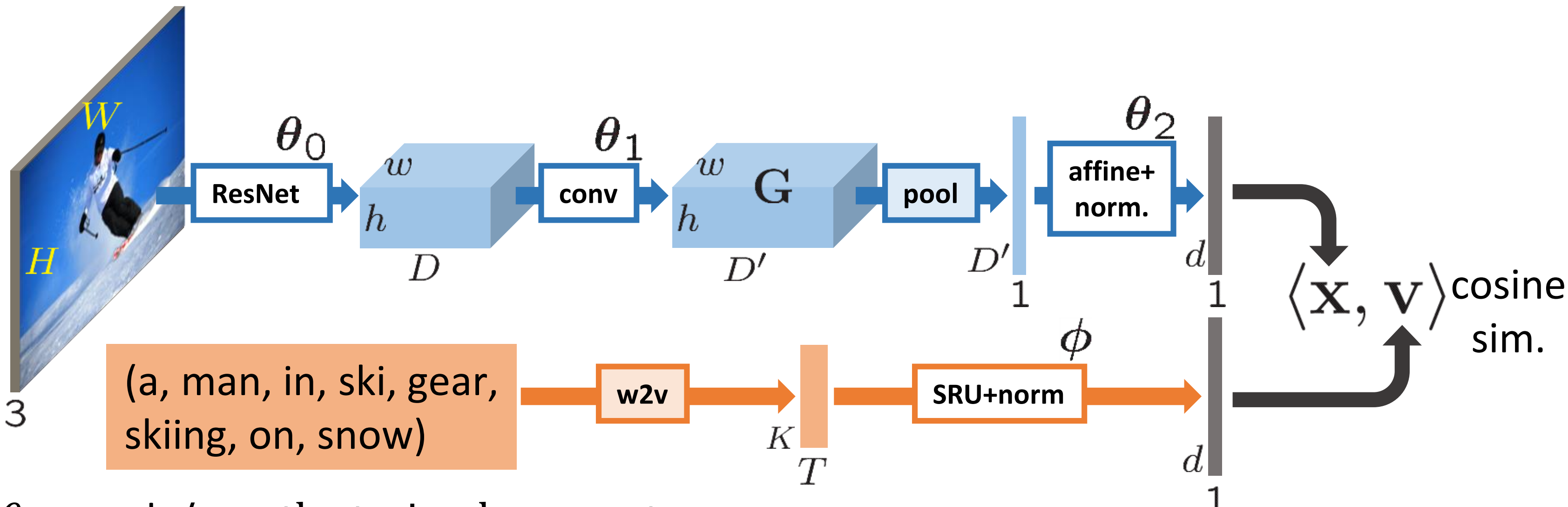
Semantic Embedding Model

Visual pipeline:

- ResNet-152 pretrained.
- Weldon spatial pooling.
- Affine projection
- normalization.

Textual pipeline:

- Pretrained word embedding.
- Simple Recurrent Unit (SRU).
- Normalization.



$\theta_{0:2}$ and ϕ are the trained parameters

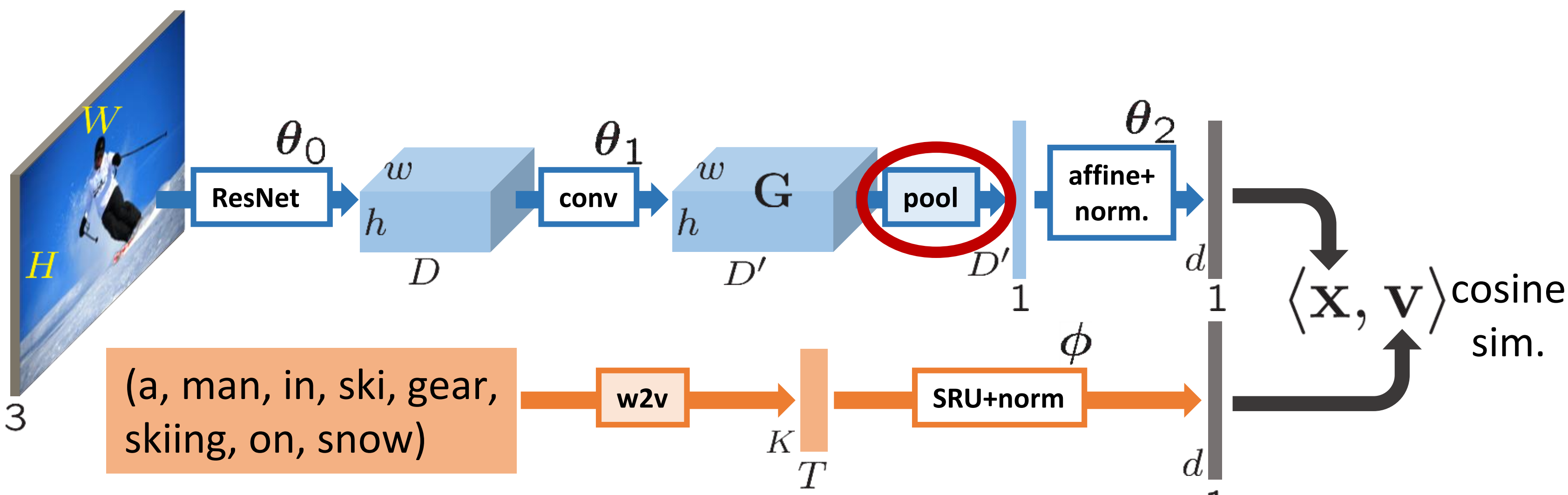
Semantic Embedding Model

Visual pipeline:

- ResNet-152 pretrained.
- Weldon spatial pooling.
- Affine projection
- normalization.

Textual pipeline:

- Pretrained word embedding.
- Simple Recurrent Unit (SRU).
- Normalization.



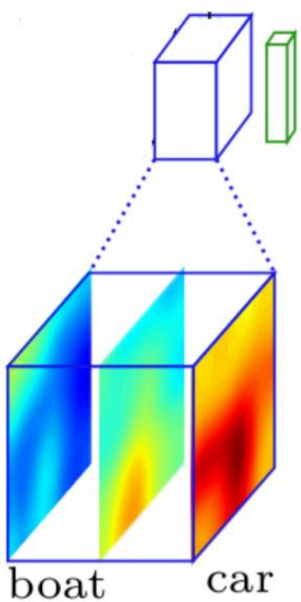
$\theta_{0:2}$ and ϕ are the trained parameters

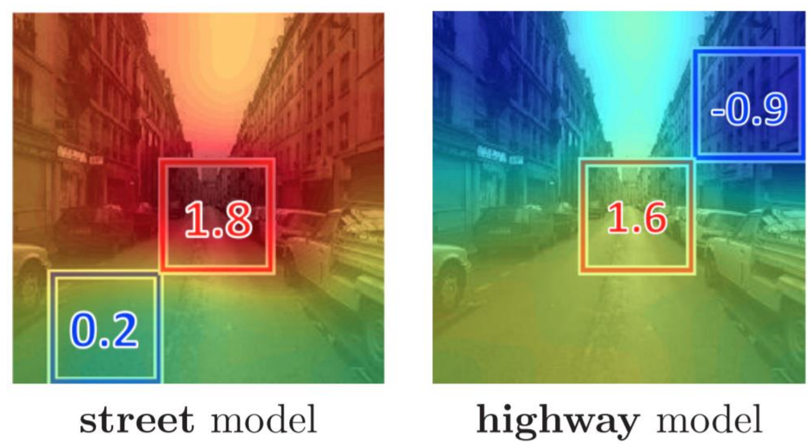
Pooling mechanisms

Weldon spatial pooling:

- Instead of global average/max pooling.
- Aggregate the min and max of each map.
- Produce activation map with finer localization.

i





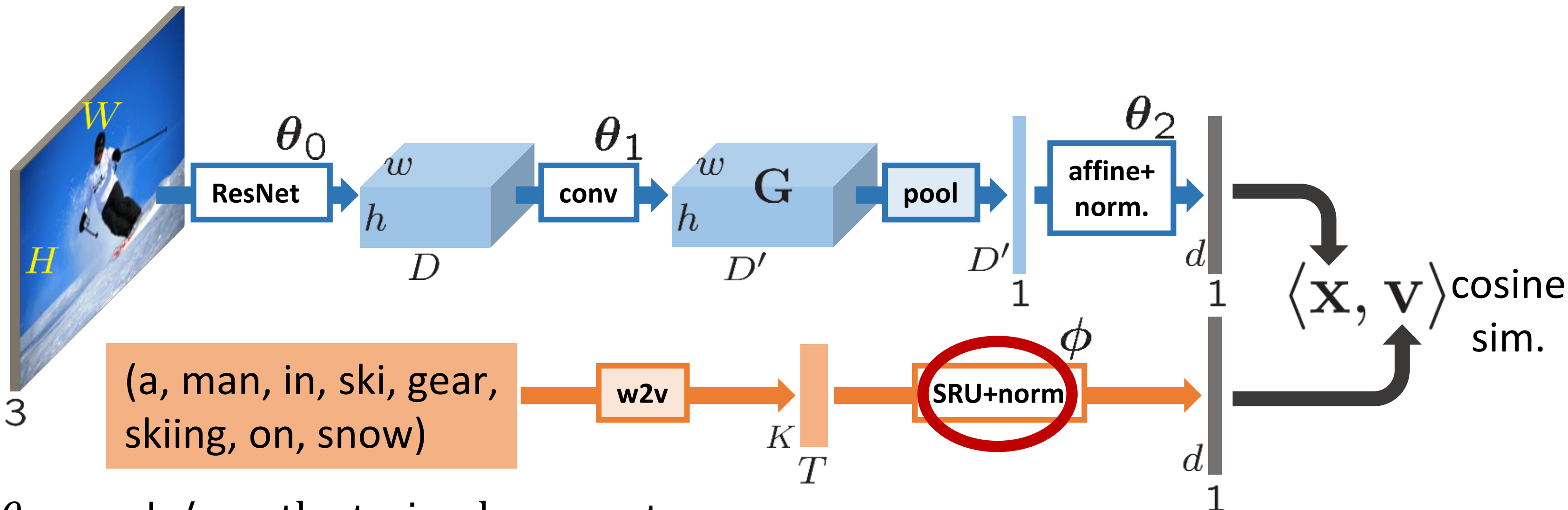
Semantic Embedding Model

Visual pipeline:

- ResNet-152 pretrained.
- Weldon spatial pooling.
- Affine projection
- normalization.

Textual pipeline:

- Pretrained word embedding.
- Simple Recurrent Unit (SRU).
- Normalization.



$\theta_{0:2}$ and ϕ are the trained parameters

Simple Recurrent Unit: SRU

Recurrent neural network:

- Fixed sized representation for variable length sequence.
- Able to capture long-term dependency between words.

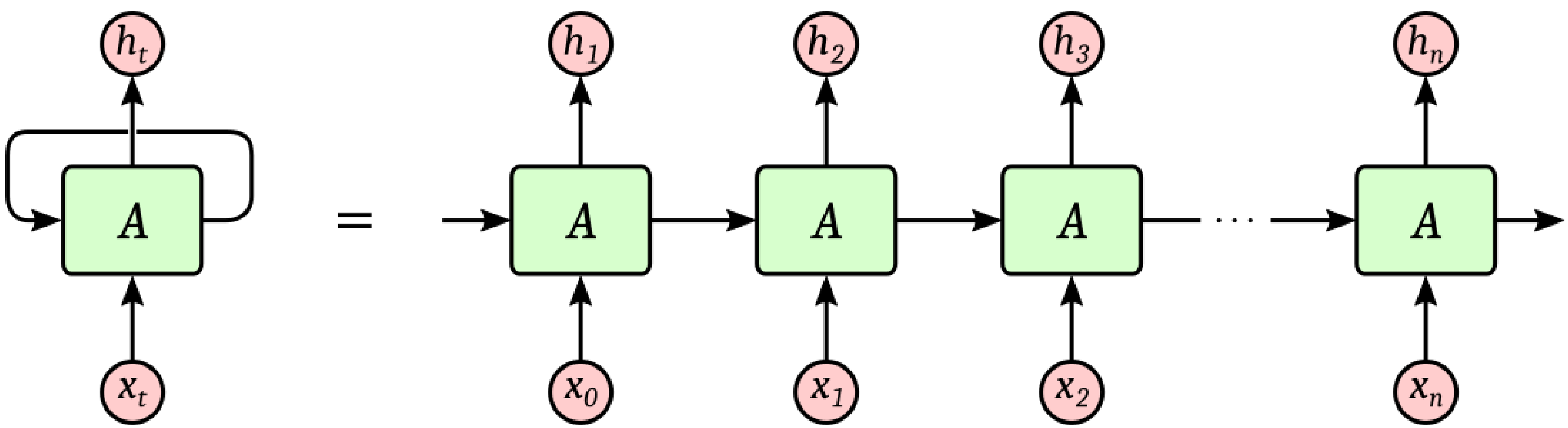


Diagram by Jakub Kvita

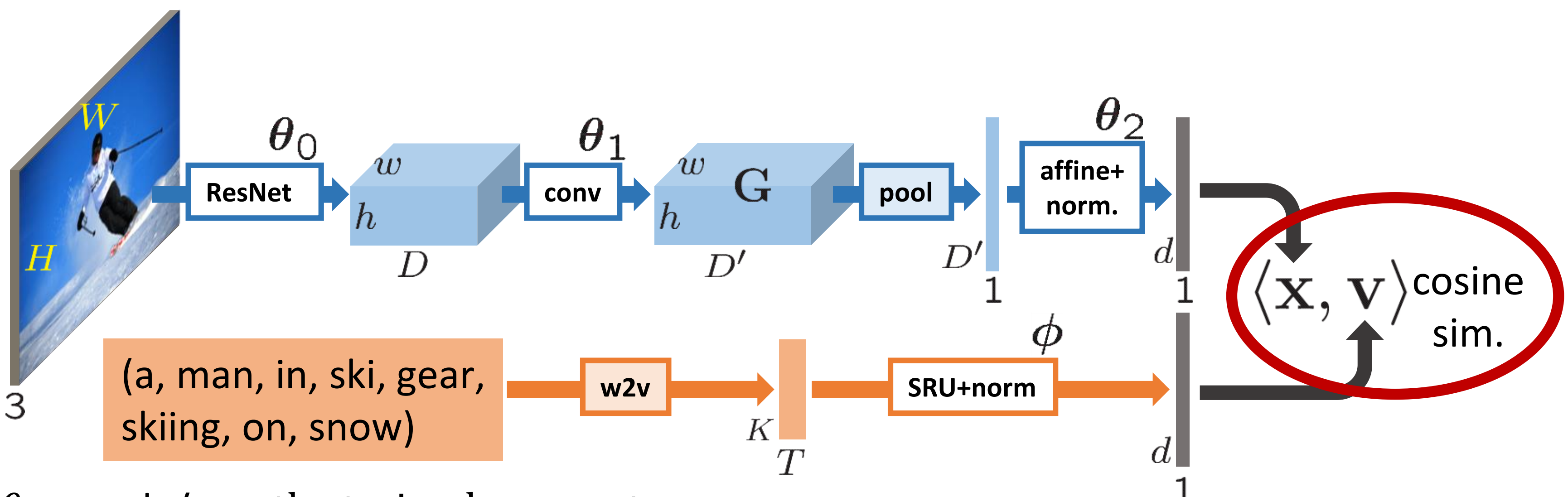
Semantic Embedding Model

Visual pipeline:

- ResNet-152 pretrained.
- Weldon spatial pooling.
- Affine projection
- normalization.

Textual pipeline:

- Pretrained word embedding.
- Simple Recurrent Unit (SRU).
- Normalization.



$\theta_{0:2}$ and ϕ are the trained parameters

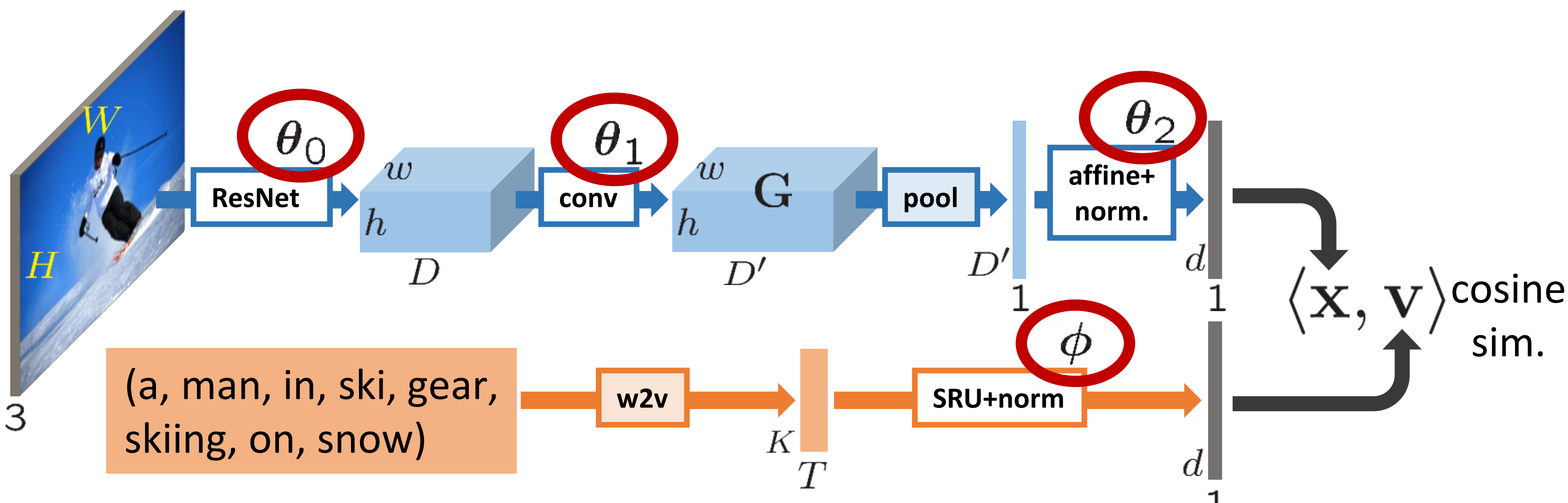
Semantic Embedding Model

Visual pipeline:

- ResNet-152 pretrained.
- Weldon spatial pooling.
- Affine projection
- normalization.

Textual pipeline:

- Pretrained word embedding.
- Simple Recurrent Unit (SRU).
- Normalization.



$\theta_{0:2}$ and ϕ are the trained parameters

Dataset

- MS-CoCo 2014:
  - 110K training images
  - 5 captions per image
  - 2*5k images for validation and test

Dining room table set for a casual meal, with flowers.

Learning strategy: triplet loss

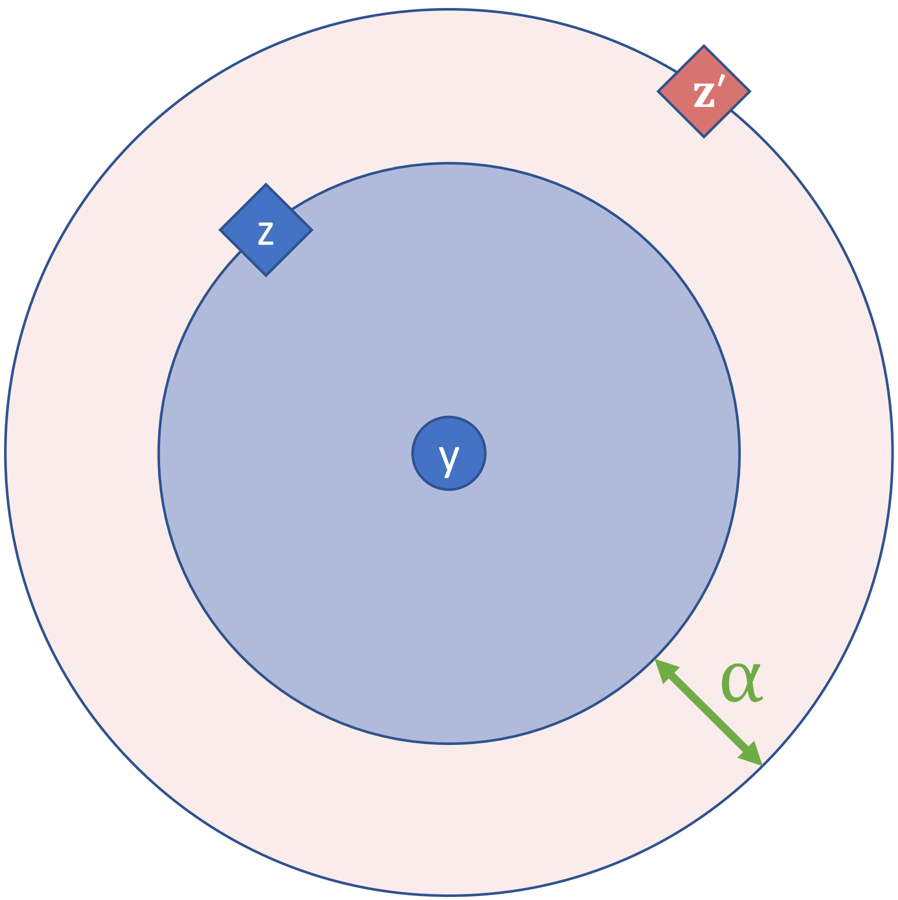
A variant of the standard margin based loss:

- Triplet $(\mathbf{y}, \mathbf{z}, \mathbf{z}')$
- Anchor: $\mathbf{y}$ (E.g image representation)
- Positive: $\mathbf{z}$ (E.g associated caption representation)
- Negative: $\mathbf{z}'$ (E.g contrastive caption representation)
- Margin parameter α

$$\text{loss}(\mathbf{y}, \mathbf{z}, \mathbf{z}') = \max\{0, \alpha - <\mathbf{y}, \mathbf{z}> + <\mathbf{y}, \mathbf{z}'>\}$$

Learning strategy: triplet loss

$$\mathrm{loss}(\mathbf{y},\mathbf{z},\mathbf{z}') = \max\{0, \alpha + \mathrm{d}(\mathbf{y},\mathbf{z}) - \mathrm{d}(\mathbf{y},\mathbf{z}')\}$$



Learning strategy: triplet loss

Hard negative margin based loss:

Loss for a batch $\mathcal{B} = \{(\mathbf{I}_n, \mathbf{S}_n)\}_{n \in B}$ of image sentence pairs:

$$\boxed{\mathcal{L}(\mathbf{\Theta}; \mathcal{B}) = \frac{1}{|B|} \sum_{n \in B} \left(\begin{array}{l} \max\limits_{m \in C_n \cap B} \operatorname{loss}(\mathbf{x}_n, \mathbf{v}_n, \mathbf{v}_m) \\ + \max\limits_{m \in D_n \cap B} \operatorname{loss}(\mathbf{v}_n, \mathbf{x}_n, \mathbf{x}_m) \end{array} \right)}$$

With :

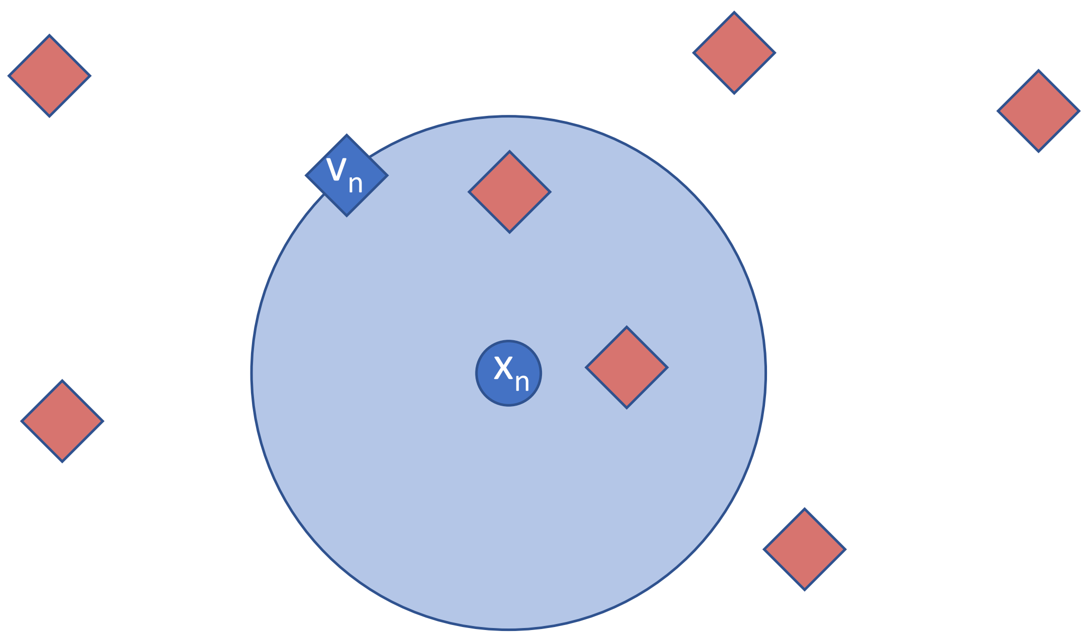
- C_n (resp. D_n) set of indices of caption (resp. image) unrelated to n-th element.

Learning strategy: hard negative triplet loss

SCIENCES SORBONNE UNIVERSITÉ technicolor

Mining hard negative contrastive example:

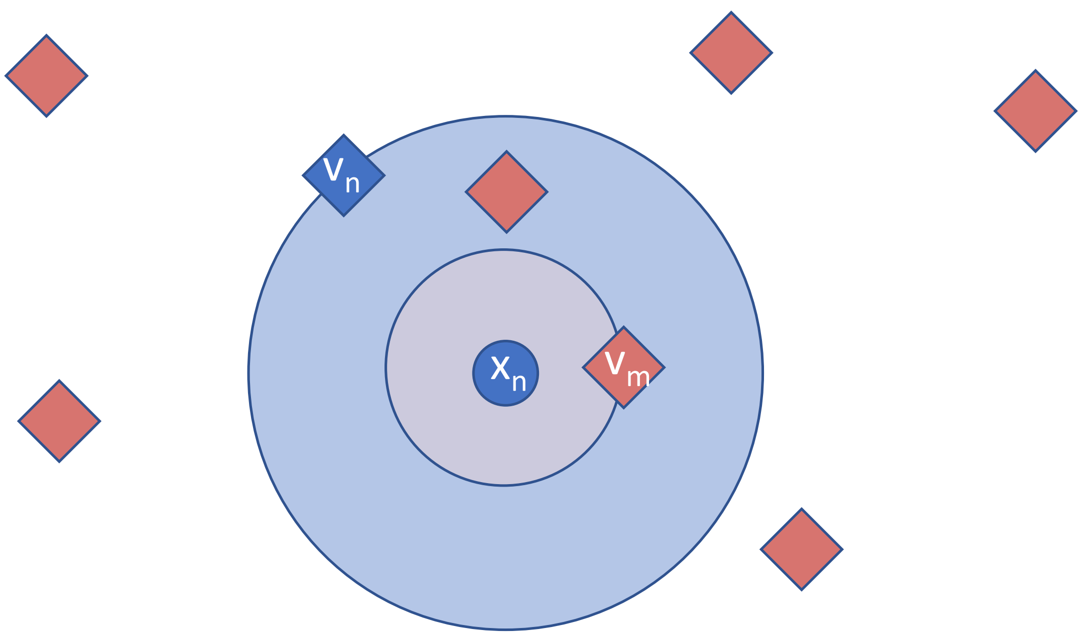
$$\mathcal{L}(\boldsymbol{\Theta};\mathcal{B}) = \frac{1}{|B|}\sum_{n\in B}\left(\max_{m\in C_n\cap B} \operatorname{loss}(\mathbf{x}_n,\mathbf{v}_n,\mathbf{v}_m) + \max_{m\in D_n\cap B} \operatorname{loss}(\mathbf{v}_n,\mathbf{x}_n,\mathbf{x}_m)\right)$$



Learning strategy: hard negative triplet loss

Mining hard negative contrastive example:

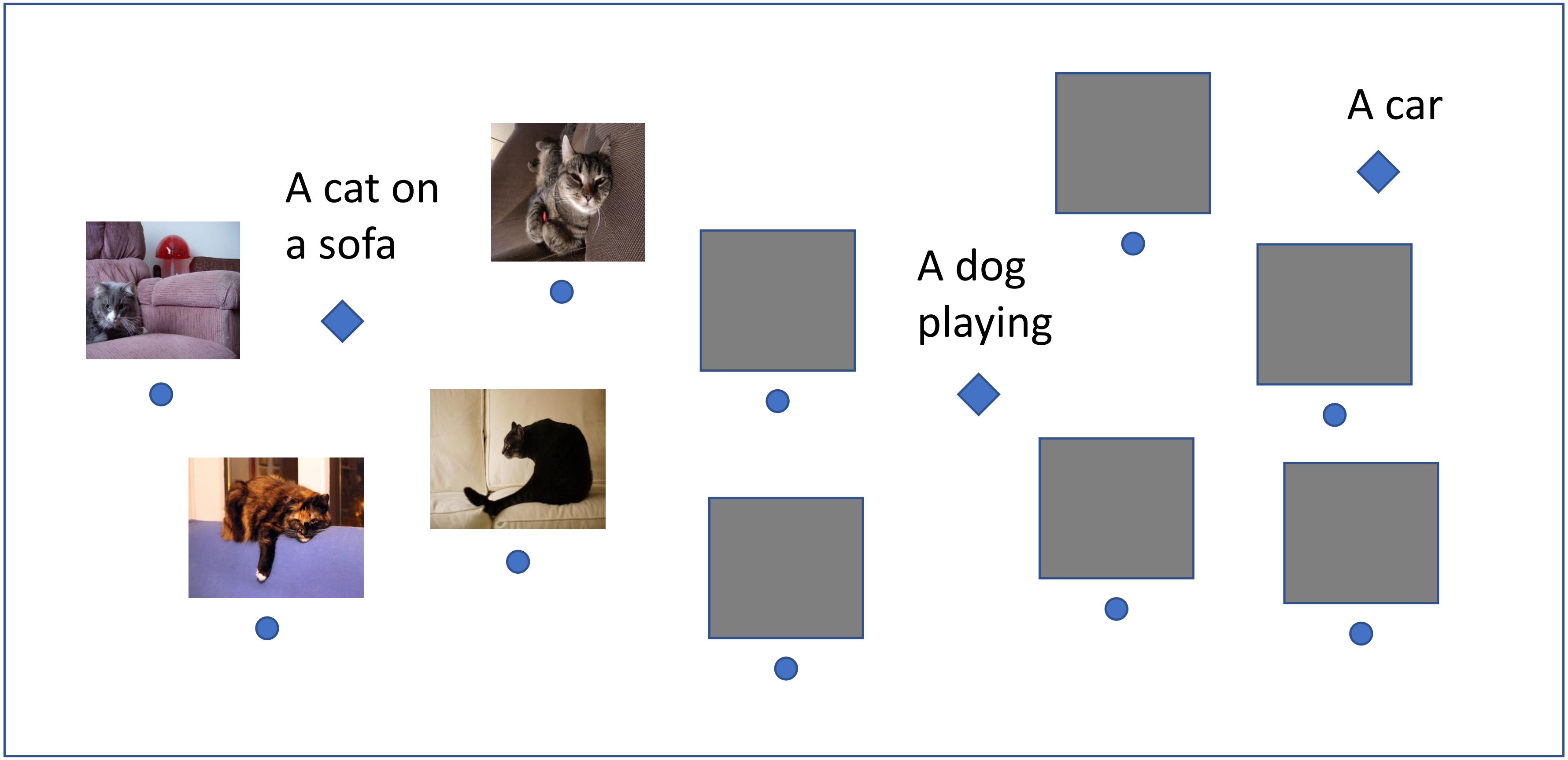
$$\mathcal{L}(\boldsymbol{\Theta};\mathcal{B}) = \frac{1}{|B|}\sum_{n\in B}\left(\max_{m\in C_n\cap B} \text{loss}\,(\mathbf{x}_n,\mathbf{v}_n,\mathbf{v}_m) + \max_{m\in D_n\cap B} \text{loss}\,(\mathbf{v}_n,\mathbf{x}_n,\mathbf{x}_m)\right)$$



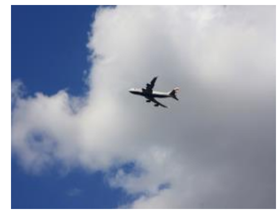
From training to testing

Training finished:

- Visual-semantic space constructed.
- Parameters of the model are fixed.
- Time for testing.



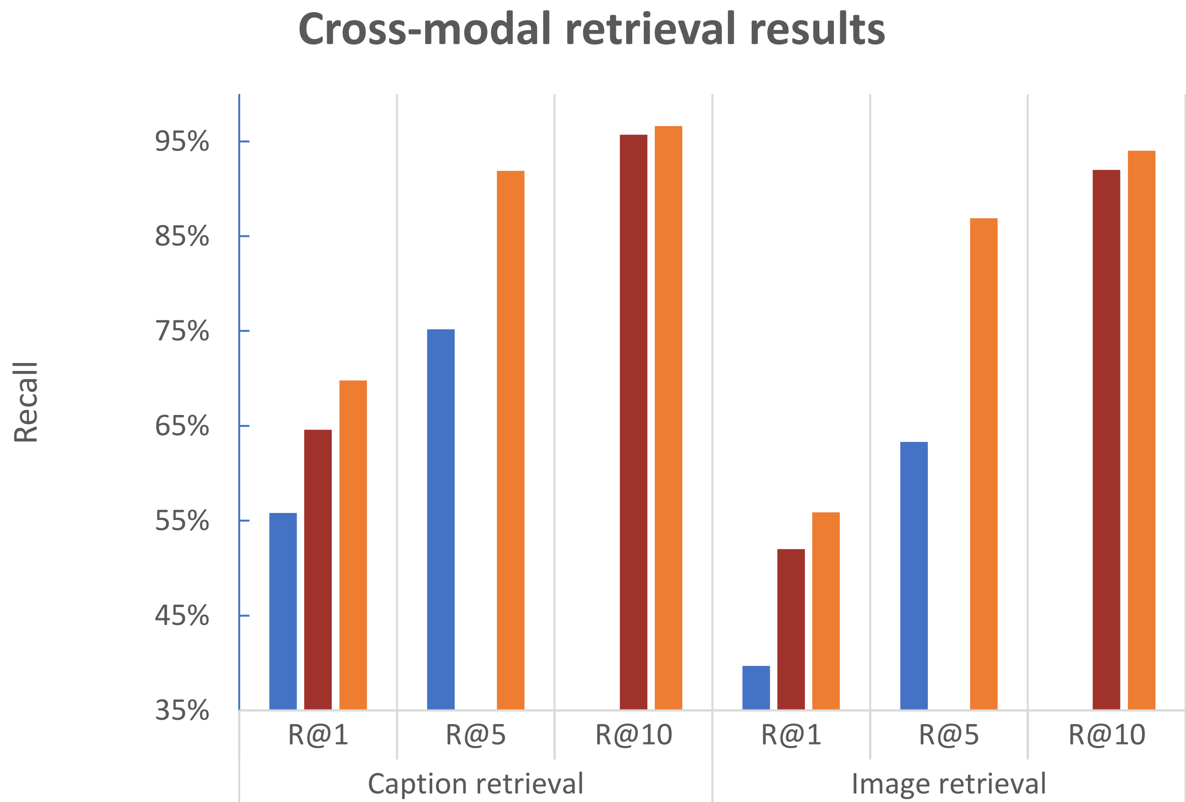
Qualitative evaluation: cross-modal retrieval

Query	Closest elements
A plane in a cloudy sky	    
A dog playing with a frisbee	    
	1. **A herd of sheep standing on top of snow covered field.** 2. There are sheep standing in the grass near a fence. 3. some black and white sheep a fence dirt and grass

Quantitative evaluation: cross-modal retrieval

SCIENCES SORBONNE UNIVERSITÉ technicolor

Cross-modal retrieval: Evaluated on MS-CoCo image/caption pairs.



	Caption retrieval R@1	Caption retrieval R@5	Caption retrieval R@10	Image retrieval R@1	Image retrieval R@5	Image retrieval R@10
2-Way Net [5]	55.80%	75.20%		39.70%	63.30%	
VSE++ [6]	64.60%		95.70%	52%		92%
Ours	69.80%	91.90%	96.60%	55.90%	86.90%	94%

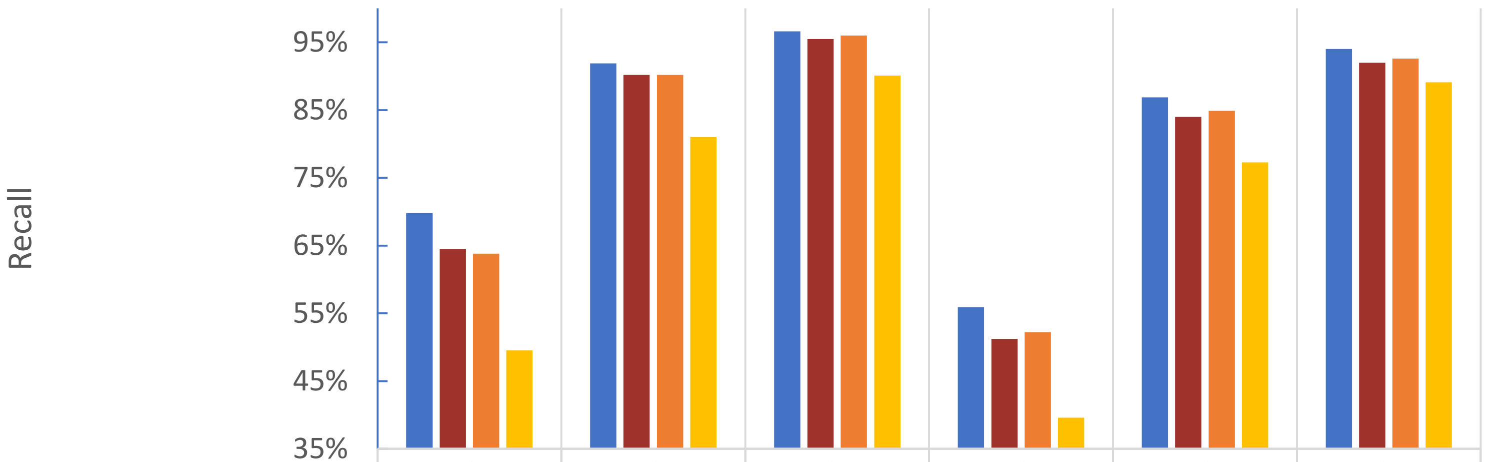
Performance evaluation: ablation study

SCIENCES SORBONNE UNIVERSITÉ technicolor

Performance boost coming from:

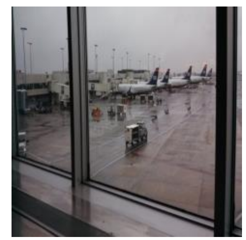
- Architecture choice: SRU and Weldon spatial pooling.
- Efficient learning strategy: hard negative loss.

Ablation study: cross modal retrieval results



	Caption retrieval R@1	Caption retrieval R@5	Caption retrieval R@10	Image retrieval R@1	Image retrieval R@5	Image retrieval R@10
Hard Neg + WLD + SRU 4	69.80%	91.90%	96.60%	55.90%	86.90%	94%
Hard Neg + GAP + SRU 4	64.50%	90.20%	95.50%	51.20%	84.00%	92.00%
Hard Neg + WLD + GRU 1	63.80%	90.20%	96%	52.20%	84.90%	92.60%
Classic + WLD + SRU 4	49.50%	81%	90.10%	39.60%	77.30%	89.10%

Evaluation: cross-modal retrieval and limitations

Query	Closest elements
Multiple wooden spoons are shown on a table top.	    
The plane is parked at the gate at the airport terminal.	    
	1. Two elephants in the eld moving along during the day. 2. Two elephants are standing by the trees in the wild. **3. An elephant and a rhino are grazing in an open wooded area.**
	1. A harbor filled with boats floating on water **2. A small marina with boats docked there** **3. a group of boats sitting together with no one around**

Localization

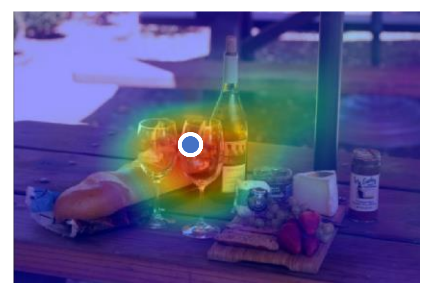
Visual grounding module:

- Weakly supervised, with no additional training.
- Localize a textual query in an image.
- Using the embedding space to select convolutionnal activation maps.

Source image

two glasses

Text query



Visual grounding
Heat map

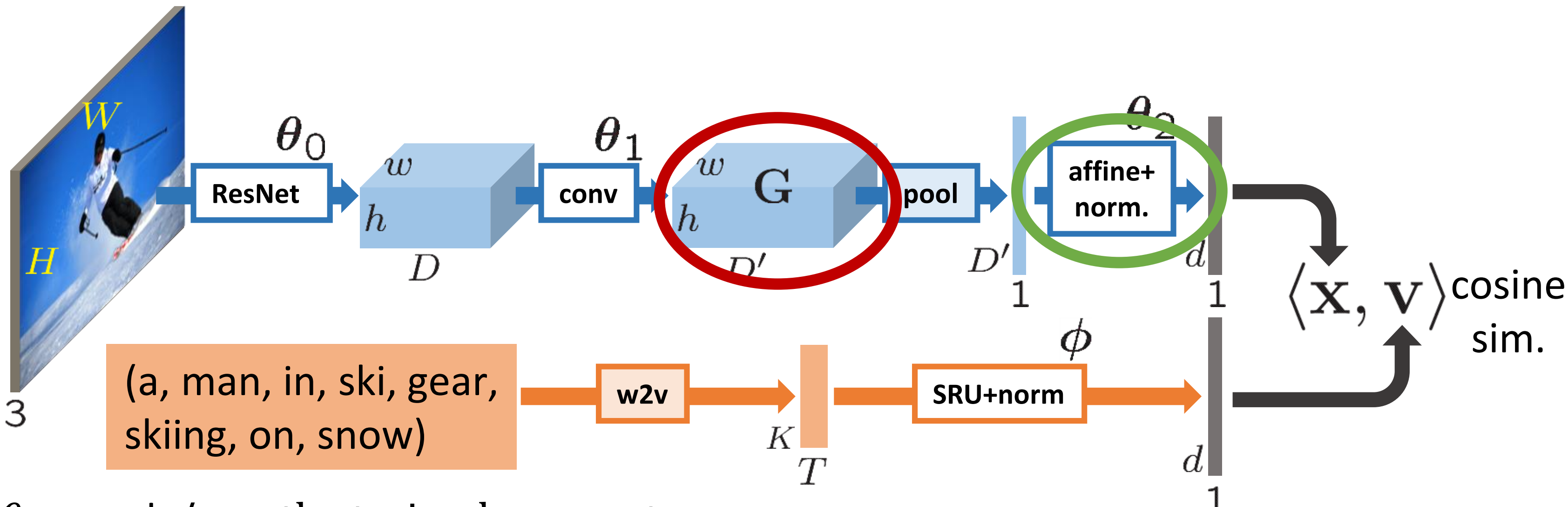
Semantic Embedding Model

Visual pipeline:

- ResNet-152 pretrained.
- Weldon spatial pooling.
- Affine projection
- normalization.

Textual pipeline:

- Pretrained word embedding.
- Simple Recurrent Unit (SRU).
- Normalization.



$\theta_{0:2}$ and ϕ are the trained parameters

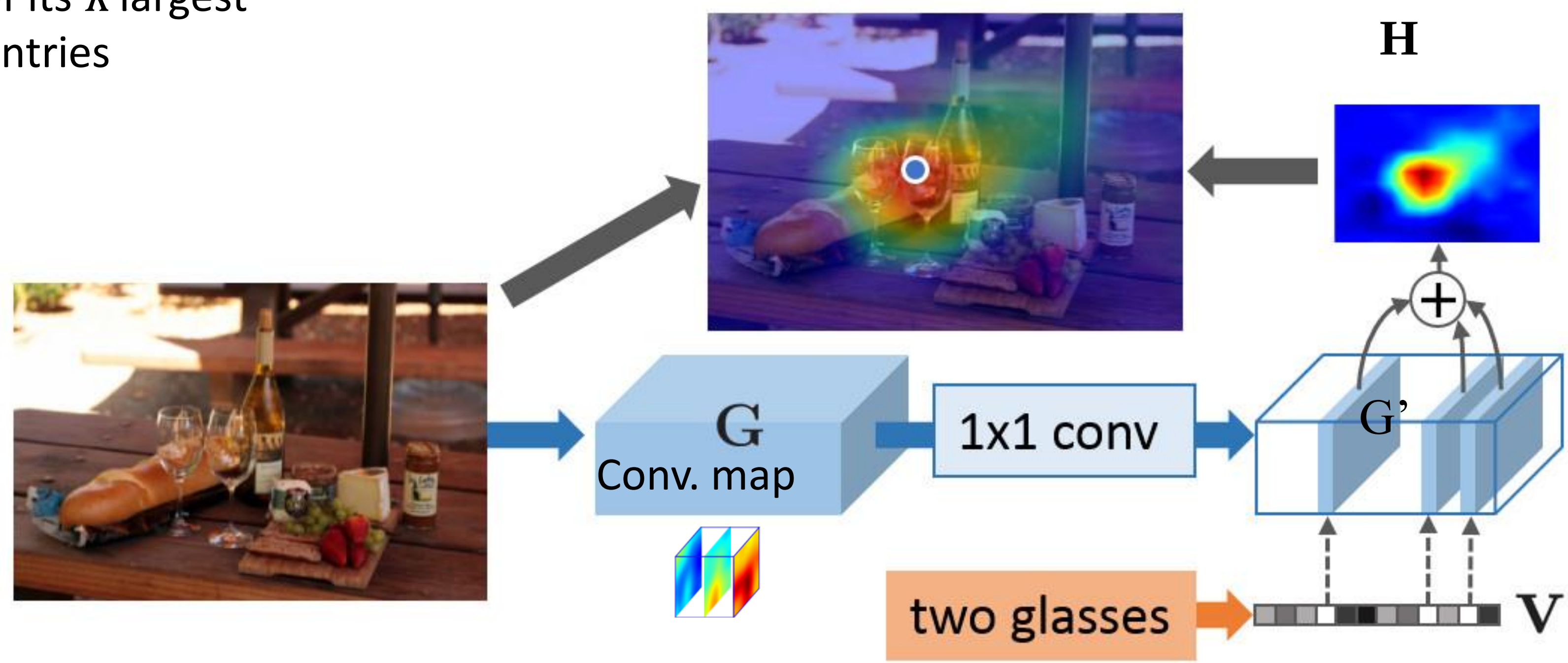
Localization

Generation of heatmap $\mathbf{H}$:

$$\mathbf{G}'[i,j,:] = A\mathbf{G}[i,j,:], \forall(i,j) \in [1,w] \times [1,h]$$

$$\boxed{\mathbf{H} = \sum_{u \in K(\mathbf{v})} |\mathbf{v}[u]| * \mathbf{G}'[:,:,u]}$$

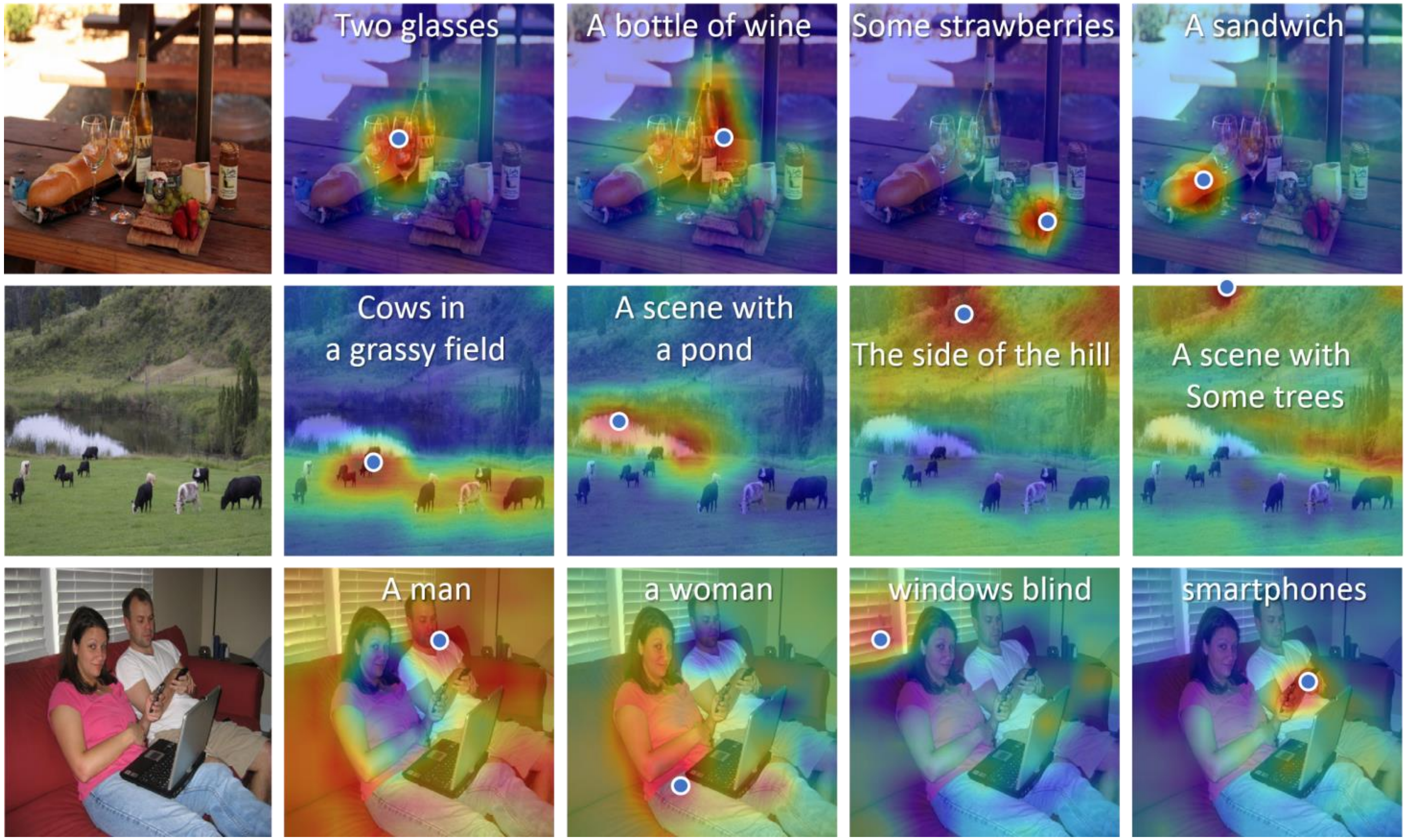
$K(\mathbf{v})$ the set of the indices of its k largest entries



Qualitative evaluation: localization

Visual grounding examples:

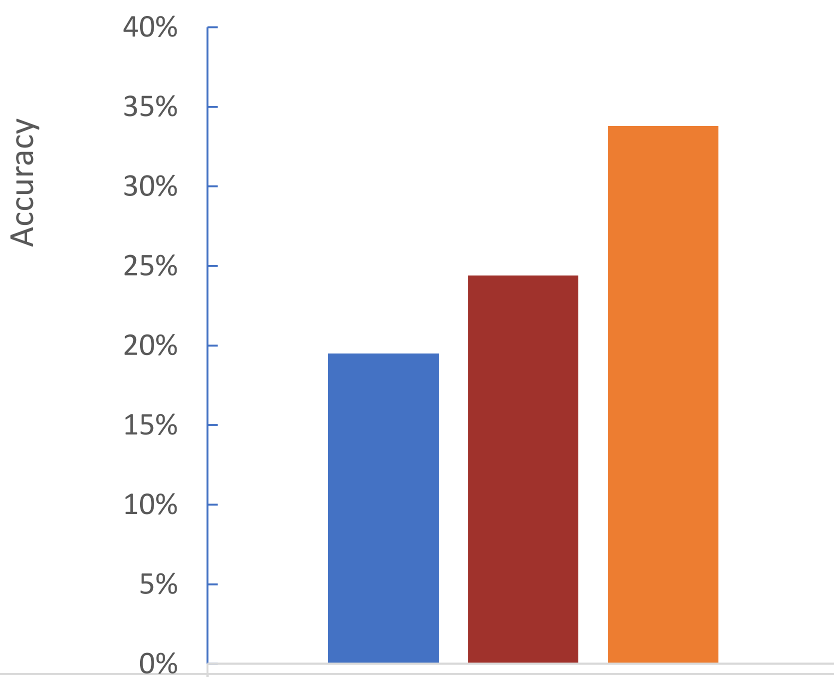
- Generating multiple heat maps with different textual queries.



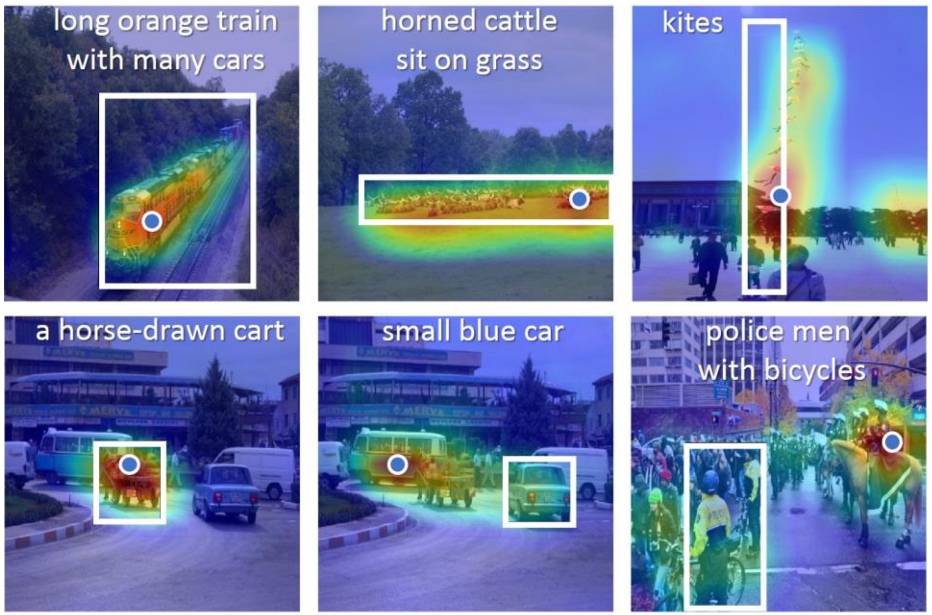
Quantitative evaluation: localization

The pointing game: Localizing phrases corresponding to subregions of the image.

Pointing game results

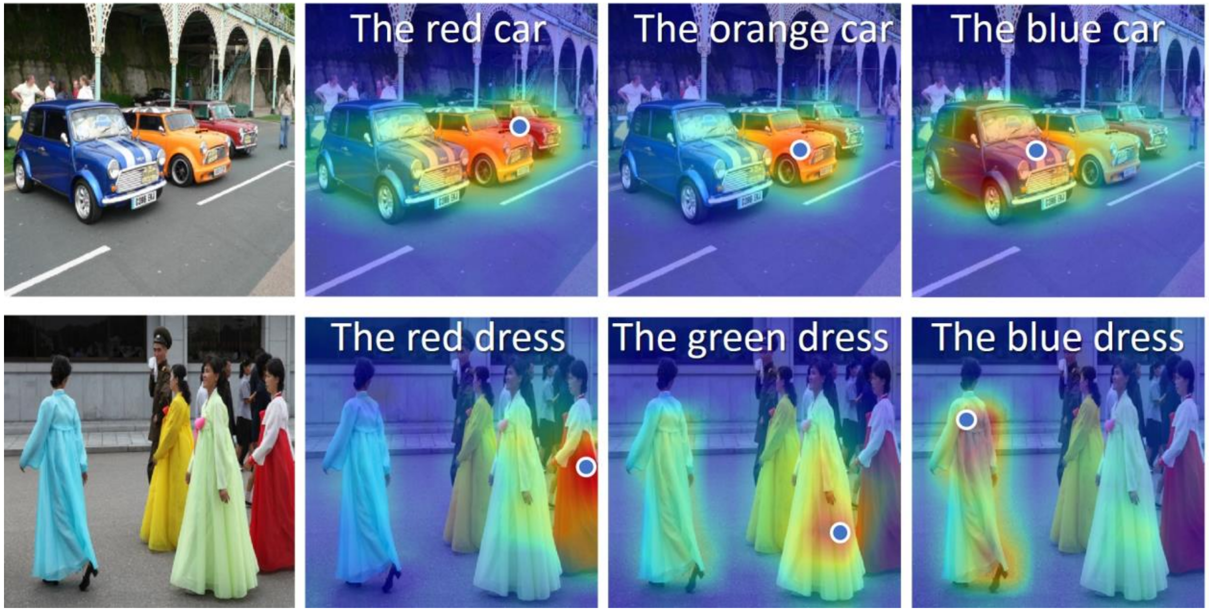


"Center" baseline	19.50%
Linguistic structure [7]	24.40%
Ours	33.80%

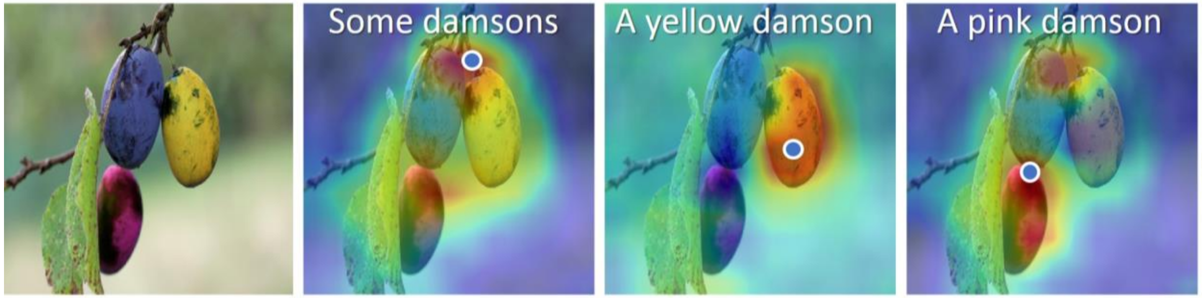


Toward zero-shot localization:

- Emergence of colors understanding:



- Even on artificial images:



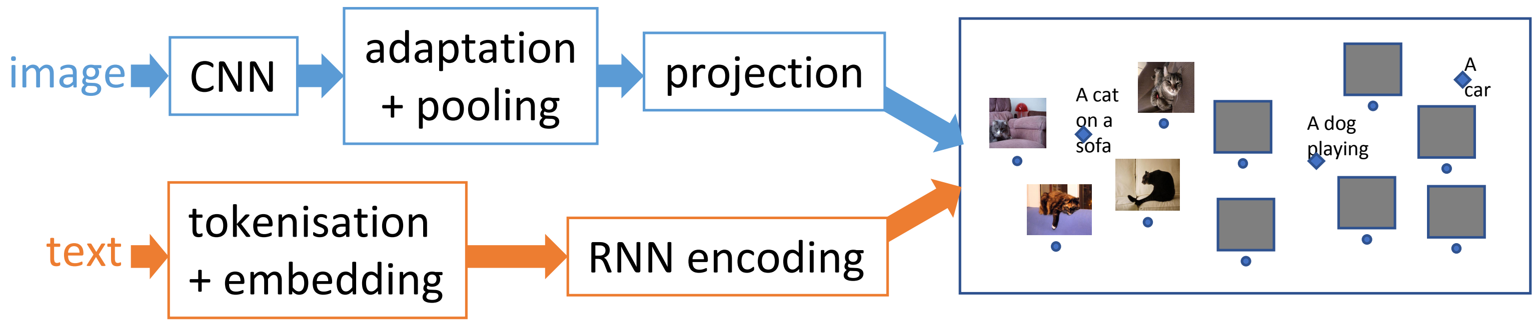
Toward zero-shot localization:

- Generalization to unseen elements:

Conclusion

Summary:

- Semantic-visual embedding model.
- Effective on the cross-modal retrieval task
- Visual grounding of text with no extra supervision.



Localization and retrieval using the embedding space

Thank you!

Paper - *Finding beans in burgers: Deep semantic-visual embedding with localization*