Image-Text Representation and Image-Text Applications

Yang Liu, Qing Ye
University of Illinois at Urbana-Champaign
CS598 LAZ

April 6, 2017

Outline

- Part I: Computer Vision Tasks Introduction
  - Image Detection
  - Image Text Tasks: Image Captioning, Phrase Localization, Image-Sentence Retrieval
- Part II: Foundation: How to represent image and text? $\Longrightarrow$ image-text representation.
- Part III: Three Image-Text Applications:
  - Learning to ground by reconstruction
  - Description generation and comprehension
  - Dense captioning

Tasks

Object Detection

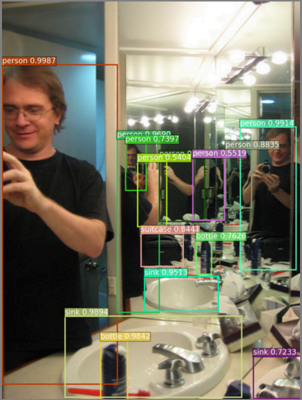


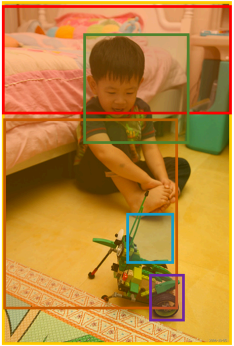
Image Captioning

A graying man in a suit is perplexed at a business meeting.
A businessman in a yellow tie gives a frustrated look.
A man in a yellow tie is rubbing the back of his neck.
A man with a yellow tie looks concerned.

Bell, Sean, et al. "Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks." Young, Peter, et al. "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions."

Tasks

Phrase Localization



A small Asian boy [0.45] is sitting on the floor [0.82] of a bedroom [0.87] being entertained and smiling at a lego toy [0.77] that looks like a bug [0.87] on wheels [0.81] .

Retrieval

(a) Results for the query on a popular image search engine.

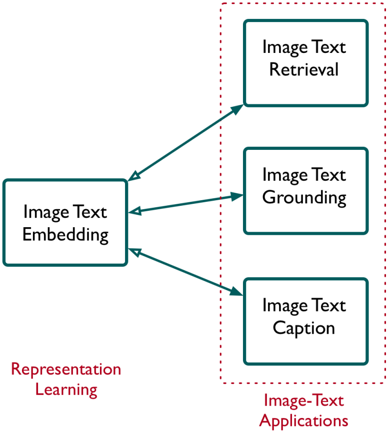
(b) Expected results for the query.

Figure 1: Image search using a complex query like "man holding fish and wearing hat on white boat" returns unsatisfactory results in (a). Ideal results (b) include correct *objects* ("man", "boat"), *attributes* ("boat is white") and *relationships* ("man on boat").

Plummer, Bryan A., et al. "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models." Johnson, Justin, et al. "Image retrieval using scene graphs."

Learning Image-Text Representation

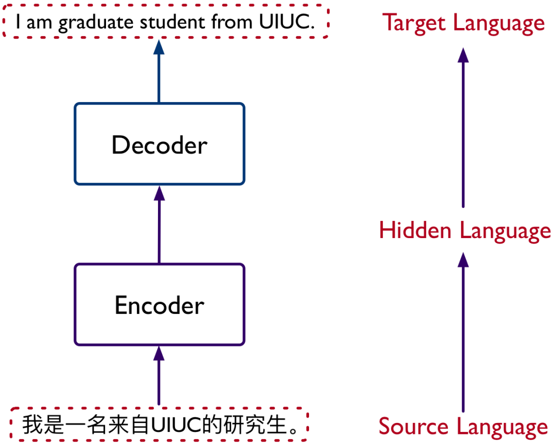
- Represents Image and Text $\Longrightarrow \mathbf{v} \in \mathbb{R}^n$.
- Similar words/images $\Longrightarrow$ similar vectors.
- Challenges: Multi-modal Learning. (Semantic sparsity in image and text)



Overview: Image-Text Embedding

- Task: Similar Semantic Unit $\Longrightarrow$ Similar Vectors
- Define similarity:
  - Symmetric Similarity :
    - Cosine Similarity
  - Asymmetric Similarity : Order-Embedding
- Task to train on:
  - Image-Sentence Matching: Ranking Loss
  - Caption Generation etc.

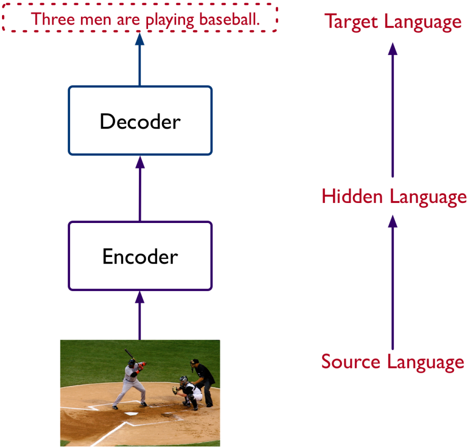
Motivation: Machine Translation



Unifying Visual-Semantic Embeddings

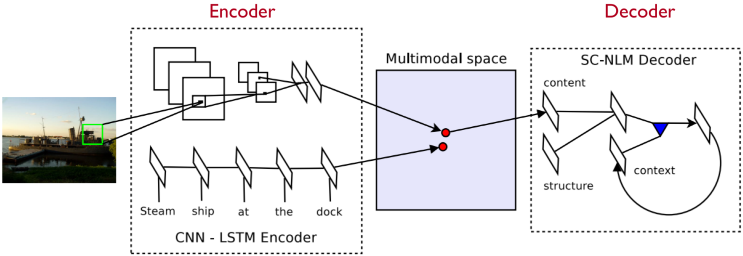
Motivation

Image as Source Language!



Unifying Visual-Semantic Embeddings

Method Framework



Kiros, Ryan, Ruslan Salakhutdinov, and Richard S. Zemel. "Unifying visual-semantic embeddings with multimodal neural language models."

Unifying Visual-Semantic Embeddings

Encoder

- Input:
  - Image feature: $\mathbf{q}$
  - Sentence: $w_1, w_2, \ldots, w_N$

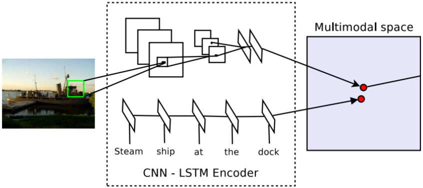


Image Representation: $\mathbf{W_I} \cdot q$

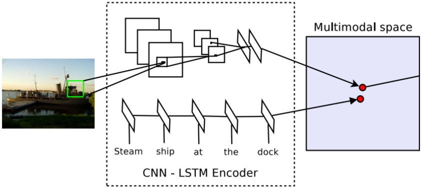
$$\mathbf{score}(\mathbf{x}, \mathbf{v}) = \frac{\mathbf{x^T} \cdot \mathbf{v}}{||\mathbf{x}||_2 \cdot ||\mathbf{v}||_2}$$

$v = LSTM(w_1, w_2, w_3 \ldots w_N)$

Kiros, Ryan, Ruslan Salakhutdinov, and Richard S. Zemel. "Unifying visual-semantic embeddings with multimodal neural language models."

Unifying Visual-Semantic Embeddings

Encoder Loss



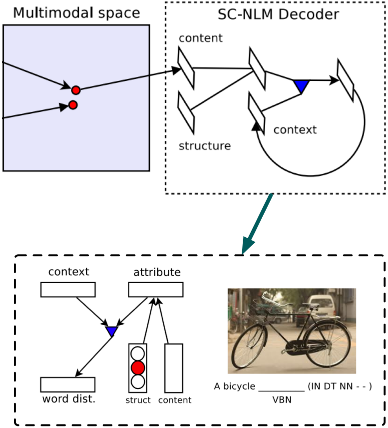
- Loss Function: Ranking Loss

$$\min_{\theta} \underbrace{\sum_{\mathbf{x}} \sum_{k} \max(0, \alpha - s(x, v) + s(x, v_k)}_{\text{rank sentences}} + \underbrace{\sum_{\mathbf{v}} \sum_{k} \max(0, \alpha - s(v, x) + s(v, x_k))}_{\text{rank images}}$$

Jamie Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. ”Unifying visual-semantic embeddings with multimodal neural language models.”

Unifying Visual-Semantic Embeddings

Method Framework



- $\mathbf{u}$: content embedding
- $w_1, w_2, ..., w_N$: word sequence
- $t_1, t_2, ..., t_N$: POS-tagging

$$Pr(w_n = i | w_{1:n-1}, \mathbf{u})$$

$$\max \log Pr(w_n = i | w_{1:n-1}, \mathbf{u})$$

Jamie Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. "Unifying visual-semantic embeddings with multimodal neural language models."

Unifying Visual-Semantic Embeddings

Evaluation Result

- Dataset: Flickr30K
- Evaluation: Recall and Median Ranking
- Baselines:
  - Random Ranking
  - SDT-RNN: Single Image, Recursive NN
  - DeFrag: Image Fragments

Table: Performance on Image-Sentence Retrieval(AlexNet)

Model	Image Annotation				Image Search			
	R@1	R@5	R@10	Med r	R@1	R@5	R@10	Med r
Random Ranking	0.1	0.6	1.1	631	0.1	0.5	1.1	500
SDT-RNN	9.6	29.8	41.1	16	8.9	29.8	41.1	16
DeFrag	19.2	44.5	58.0	6.0	12.9	35.4	47.5	10.8
MNLM(Kiros et al.)	14.8	39.2	50.9	10	11.8	34.0	46.3	13

Jamie Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. "Unifying visual-semantic embeddings with multimodal neural language models."

Unifying Visual-Semantic Embeddings

Multimodal linguistic regularities

Word Analogy

- "man" as "king" is "woman" to ?

$$v(\text{king}) - v(\text{man}) + v(\text{woman}) = ?$$

Unifying Visual-Semantic Embeddings

Multimodal linguistic regularities

Word Analogy

- ”man” as ”king” is ”woman” to ?

$$v(\text{king}) - v(\text{man}) + v(\text{woman}) = v(\text{queen})\text{ !}$$

Unifying Visual-Semantic Embeddings

Multimodal linguistic regularities

Word-Image Analogy

- "red" as  is "blue" as to ?

$$v(\;) - v(\text{red}) + v(\text{blue}) = ?$$

- Note that here they used a linear encoder rather than LSTM.

Jamie Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. "Unifying visual-semantic embeddings with multimodal neural language models."

Unifying Visual-Semantic Embeddings

Multimodal linguistic regularities

Word-Image Analogy

- ”red” as  is ”blue” as to ?

$v($  $) - v(\text{red}) + v(\text{blue}) = v($  $)!$

- Note that here they used a linear encoder rather than LSTM.

Jamie Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. ”Unifying visual-semantic embeddings with multimodal neural language models.”

Unifying Visual-Semantic Embeddings

Multimodal linguistic regularities

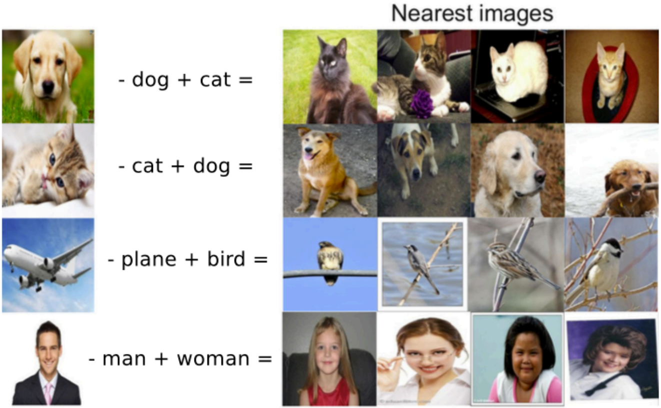


Figure: Object Transferring

Unifying Visual-Semantic Embeddings

Multimodal linguistic regularities

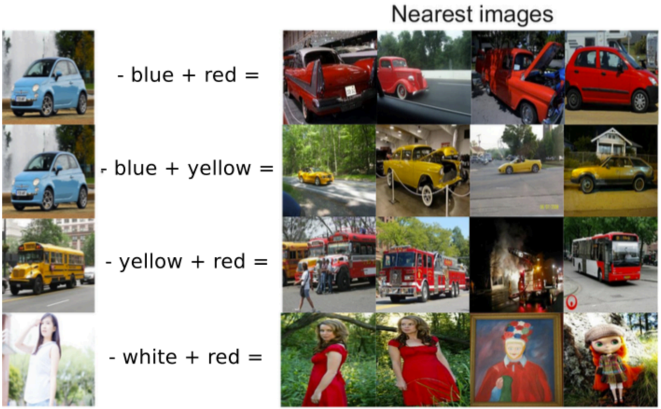


Figure: Color Transferring

Jamie Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. "Unifying visual-semantic embeddings with multimodal neural language models."

Unifying Visual-Semantic Embeddings

Multimodal linguistic regularities

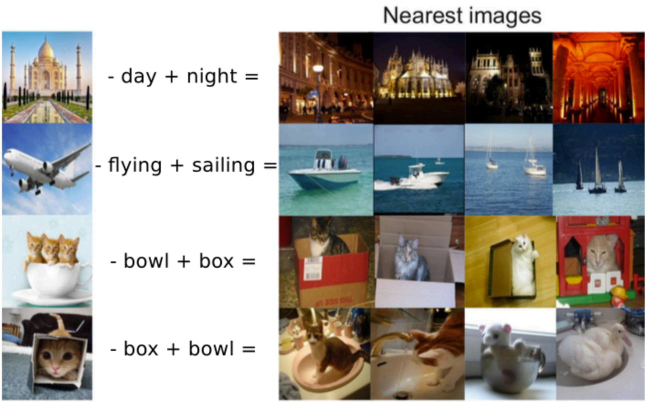


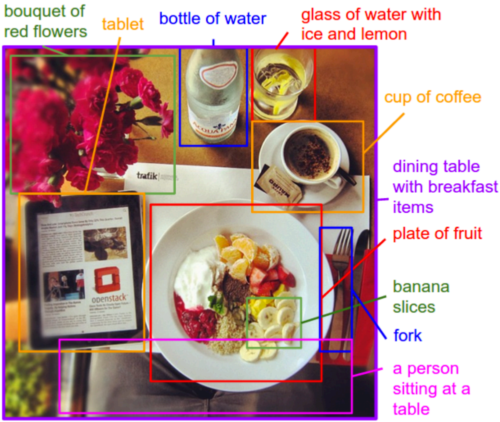
Figure: Structure Transferring

Jamie Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. "Unifying visual-semantic embeddings with multimodal neural language models."

Beyond one object

It is hard to describe an image with one caption!

Deep Visual-Semantic Alignment Model

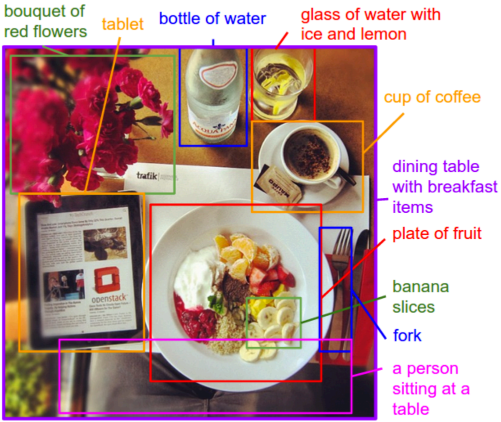


- Hard to describe an image with natural language.
- Caption may include multiple entities.

Figure: Caption Includes Multiple Entities!

Karpathy, Andrej, and Li Fei-Fei. ”Deep visual-semantic alignments for generating image descriptions.”

Deep Visual-Semantic Alignment Model



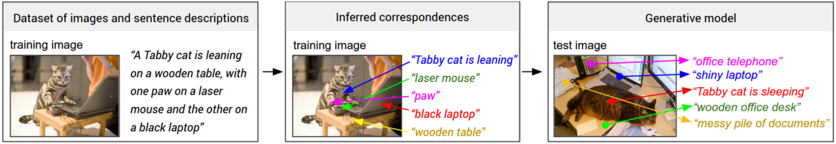
- Hard to describe an image with natural language.
- Caption may include multiple entities.

Figure: Caption Includes Multiple Entities!

Karpathy, Andrej, and Li Fei-Fei. "Deep visual-semantic alignments for generating image descriptions."

DVSA Model

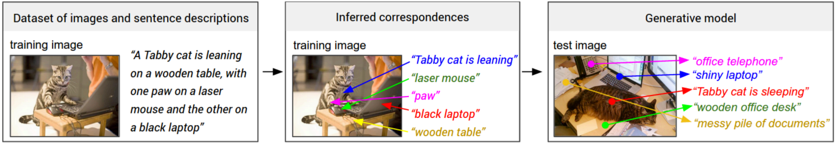
- Objective: Predict the descriptions for image regions



Karpathy, Andrej, and Li Fei-Fei. "Deep visual-semantic alignments for generating image descriptions."

DVSA Model

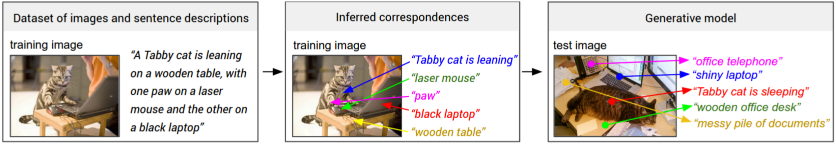
- Objective: Predict the descriptions for image regions
- Framework:



Karpathy, Andrej, and Li Fei-Fei. "Deep visual-semantic alignments for generating image descriptions."

DVSA Model

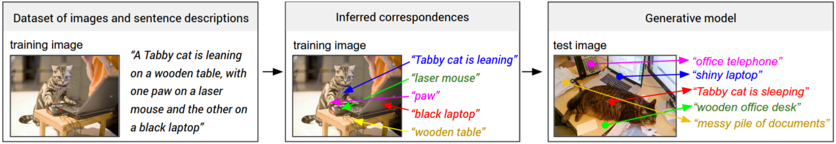
- Objective: Predict the descriptions for image regions
- Framework:
  - Learning Correspondences: Align sentence snippets to visual regions.



Karpathy, Andrej, and Li Fei-Fei. "Deep visual-semantic alignments for generating image descriptions."

DVSA Model

- Objective: Predict the descriptions for image regions
- Framework:
  - Learning Correspondences: Align sentence snippets to visual regions.
  - Generate Description: Generate description for bounding boxes.



Karpathy, Andrej, and Li Fei-Fei. "Deep visual-semantic alignments for generating image descriptions."

DVSA Model: Learning MultiModal Embedding

Image Representation

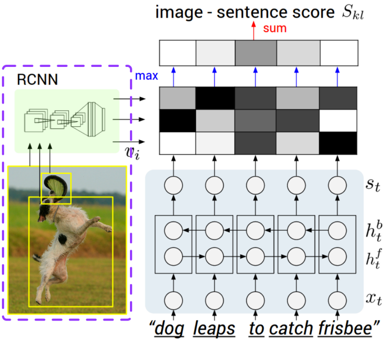


Figure: Aligning Image with Text

Transformed image embedding

$$\mathbf{v} = \mathbf{W_m} \cdot \left[CNN_{\theta_c} \left(\mathbf{I_b} \right) \right] + \mathbf{b_m}$$

Pre-trained CNN Model with θ_c.

Input : image + top 19 bounding boxes

Karpathy, Andrej, and Li Fei-Fei. "Deep visual-semantic alignments for generating image descriptions."

DVSA Model: Learning MultiModal Embedding

Sentence Representation

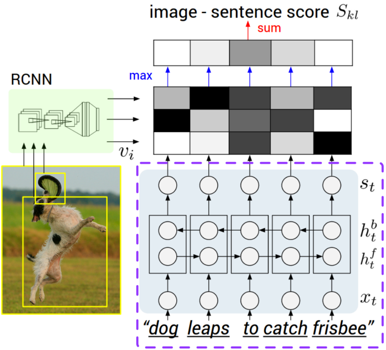


Figure: Aligning Image with Text

Semantic Word Representation

$$\mathbf{h}_t^f = \mathbf{LSTM}\left(\mathbf{h}_{t-1}^f, \mathbf{x}_t\right)$$

$$\mathbf{h}_t^b = \mathbf{LSTM}\left(\mathbf{h}_{t+1}^b, \mathbf{x}_t\right)$$

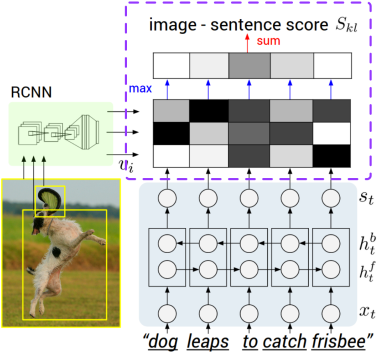
$$\mathbf{s}_t = \mathbf{f}(\mathbf{W}_d(\mathbf{h}_t^f + \mathbf{h}_t^b)) + \mathbf{b}_d$$

sentence representation at word t

Karpathy, Andrej, and Li Fei-Fei. "Deep visual-semantic alignments for generating image descriptions."

DVSA Model: Learning MultiModal Embedding

Aligning Image with Text



$$S_{kl} = \sum_{t \in g_l} \sum_{i \in g_k} \max(0, v_i^T \cdot s_t)$$

terms in one sentence

image regions

Figure: Aligning Image with Text

Karpathy, Andrej, and Li Fei-Fei. "Deep visual-semantic alignments for generating image descriptions."

DVSA Model: Learning MultiModal Embedding

Aligning Image with Text

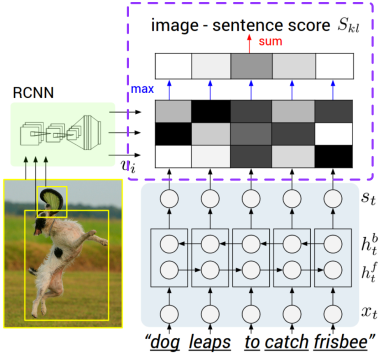


Figure: Aligning Image with Text

$$S_{kl} = \sum_{t \in g_l} \sum_{i \in g_k} \max(0, v_i^T \cdot s_t)$$

terms in one sentence

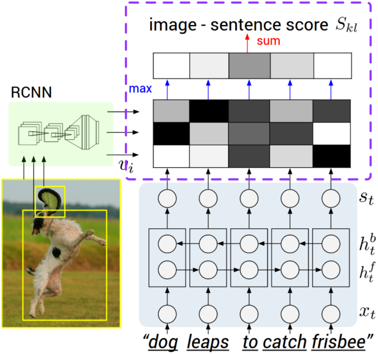
image regions

Every word just aligns to single best image region!

Karpathy, Andrej, and Li Fei-Fei. "Deep visual-semantic alignments for generating image descriptions."

DVSA Model: Learning MultiModal Embedding

Aligning Image with Text



$$S_{kl} = \sum_{t \in g_l} \max_{i \in g_k} \left(v_i^T \cdot s_t \right)$$

Figure: Aligning Image with Text

Karpathy, Andrej, and Li Fei-Fei. "Deep visual-semantic alignments for generating image descriptions."

DVSA Model: Learning MultiModal Embedding

Aligning Image with Text

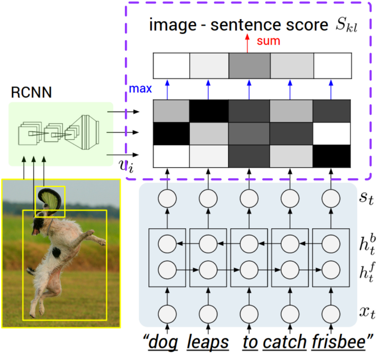


Figure: Aligning Image with Text

- Total Loss Function:

$$\mathcal{C}(\theta) = \sum_k [\underbrace{\sum_l \max(0, S_{kl} - S_{kk} + 1)}_{\text{rank images}} + \underbrace{\sum_l \max(0, S_{lk} - S_{kk} + 1)}_{\text{rank sentences}}]$$

Karpathy, Andrej, and Li Fei-Fei. "Deep visual-semantic alignments for generating image descriptions."

DVSA Model: Description Generation

Simple multi-modal RNN

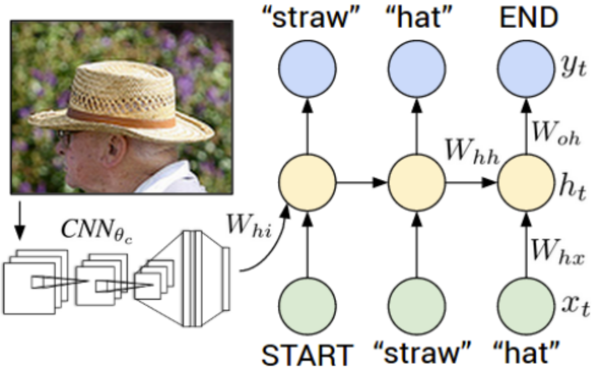


Figure: Multimodal Recurrent Neural Network

Karpathy, Andrej, and Li Fei-Fei. "Deep visual-semantic alignments for generating image descriptions."

DVSA Model: Learning MultiModal Embedding

Experiments: Image-Sentence Alignment

- Dataset: Flickr30K
- Evaluation: Recall and Median Ranking
- Baselines:
  - SDT-RNN: Single Image, Recursive NN
  - Kiros et al. : Single Image, LSTM
  - DeFrag: Image Fragments, Dependency Embedding

Table: Performance on Image-Sentence Alignment(AlexNet)

Model	Image Annotation				Image Search			
	R@1	R@5	R@10	Med r	R@1	R@5	R@10	Med r
SDT-RNN	9.6	29.8	31.1	16	8.9	29.8	41.1	16
MNLM(Kiros et al.)	14.8	39.2	50.9	10	11.8	34.0	46.3	13
DeFrag	19.2	44.5	58.0	6.0	12.9	35.4	47.5	10.8
DVSA(BRNN)	22.2	48.2	61.4	4.8	15.2	37.7	50.5	9.2

Karpathy, Andrej, and Li Fei-Fei. "Deep visual-semantic alignments for generating image descriptions."

DVSA Model: Learning MultiModal Embedding

Experiments: Image-Sentence Alignment

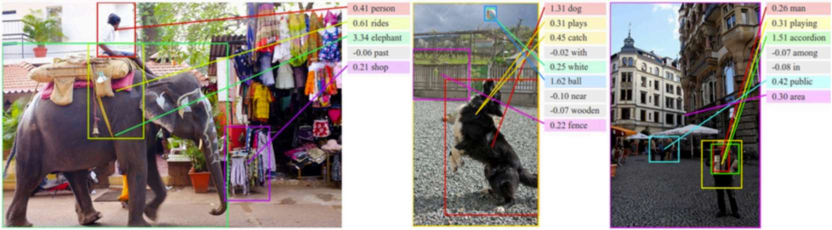


Figure: Example alignments

Karpathy, Andrej, and Li Fei-Fei. "Deep visual-semantic alignments for generating image descriptions."

DVSA Model: Learning MultiModal Embedding

Experiments: Description Generation

Figure: Result for Description Generation

Karpathy, Andrej, and Li Fei-Fei. ”Deep visual-semantic alignments for generating image descriptions.”

Fancy Models do not work well

Table: Performance on Image-Sentence Alignment(Flickr30K)

Model	Image Annotation			Image Search		
	R@1	R@5	R@10	R@1	R@5	R@10
DVSA(BRNN)(AlexNet)	22.2	48.2	61.4	15.2	37.7	50.5
MNLM(AlexNet)	23.0	50.7	62.9	16.8	42.0	56.5
CCA(Whole Image)(VGGNet)	36.5	62.2	73.3	24.7	53.4	66.8

- Methods:

Plummer, Bryan A., et al. "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models."

Fancy Models do not work well

Table: Performance on Image-Sentence Alignment(Flickr30K)

Model	Image Annotation			Image Search		
	R@1	R@5	R@10	R@1	R@5	R@10
DVSA(BRNN)(AlexNet)	22.2	48.2	61.4	15.2	37.7	50.5
MNLM(AlexNet)	23.0	50.7	62.9	16.8	42.0	56.5
CCA(Whole Image)(VGGNet)	36.5	62.2	73.3	24.7	53.4	66.8

- **Methods:**
  - DVSA(BRNN): DVSA Model Mentioned before

Plummer, Bryan A., et al. "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models."

Fancy Models do not work well

Table: Performance on Image-Sentence Alignment(Flickr30K)

Model	Image Annotation R@1	R@5	R@10	Image Search R@1	R@5	R@10
DVSA(BRNN)(AlexNet)	22.2	48.2	61.4	15.2	37.7	50.5
MNLM(AlexNet)	23.0	50.7	62.9	16.8	42.0	56.5
CCA(Whole Image)(VGGNet)	36.5	62.2	73.3	24.7	53.4	66.8

- Methods:
  - DVSA(BRNN): DVSA Model Mentioned before
  - MNLM: Multimodal Neural Language Models

Plummer, Bryan A., et al. "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models."

Fancy Models do not work well

Table: Performance on Image-Sentence Alignment(Flickr30K)

Model	Image Annotation R@1	R@5	R@10	Image Search R@1	R@5	R@10
DVSA(BRNN)(AlexNet)	22.2	48.2	61.4	15.2	37.7	50.5
MNLM(AlexNet)	23.0	50.7	62.9	16.8	42.0	56.5
CCA(Whole Image)(VGGNet)	36.5	62.2	73.3	24.7	53.4	66.8

- Methods:
  - DVSA(BRNN): DVSA Model Mentioned before
  - MNLM: Multimodal Neural Language Models
  - CCA: A classical linear method even in textbook

Plummer, Bryan A., et al. "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models."

Fancy Models do not work well

Table: Performance on Image-Sentence Alignment(Flickr30K)

Model	Image Annotation R@1	Image Annotation R@5	Image Annotation R@10	Image Search R@1	Image Search R@5	Image Search R@10
DVSA(BRNN)(AlexNet)	22.2	48.2	61.4	15.2	37.7	50.5
MNLM(AlexNet)	23.0	50.7	62.9	16.8	42.0	56.5
CCA(Whole Image)(VGGNet)	36.5	62.2	73.3	24.7	53.4	66.8

- Methods:
  - DVSA(BRNN): DVSA Model Mentioned before
  - MNLM: Multimodal Neural Language Models
  - CCA: A classical linear method even in textbook
- How to go beyond naive baseline?

Plummer, Bryan A., et al. "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models."

Structure-Preserving Image-Text Embedding

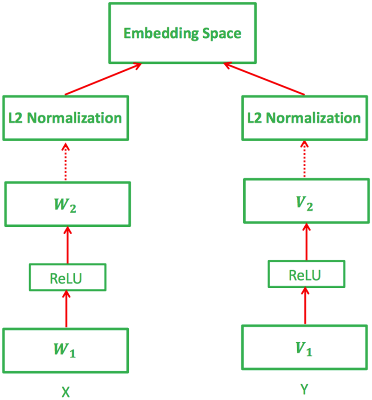


Figure: Network Structure

- Two Branch Network Embeddings
- Minimize the ranking loss

Wang, Liwei, Yin Li, and Svetlana Lazebnik. "Learning deep structure-preserving image-text embeddings."

Structure-Preserving Image-Text Embedding Bi-direcitonal Ranking Constraints

Image Embedding

$$d\left(\mathbf{x_i}, \mathbf{y_j}\right) + m < d(\mathbf{x_i}, \mathbf{y_k}) \quad \forall \mathbf{y_j} \in \mathbf{Y_i^+}, \quad \forall \mathbf{y_k} \in \mathbf{Y_i^-}$$

matching sentences of image i

non-matching sentences of image i

$$d\left(\mathbf{x_i}, \mathbf{y_j}\right) + m < d(\mathbf{x_k}, \mathbf{y_j}) \quad \forall \mathbf{x_i} \in \mathbf{X_j^+}, \quad \forall \mathbf{x_k} \in \mathbf{X_j^-}$$

Wang, Liwei, Yin Li, and Svetlana Lazebnik. "Learning deep structure-preserving image-text embeddings."

Structure-Preserving Image-Text Embedding
Structure Preserving Constraints

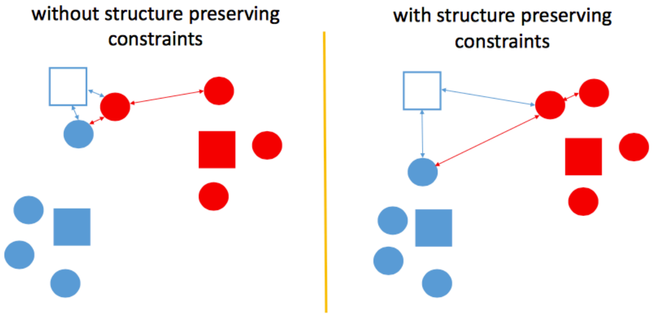


Figure: Illustration of Structure-Preserving

Wang, Liwei, Yin Li, and Svetlana Lazebnik. "Learning deep structure-preserving image-text embeddings."

Structure-Preserving Image-Text Embedding

Structure Preserving Constraints

Image Embedding

$$d(\mathbf{x_i}, \mathbf{x_j}) + m < d(\mathbf{x_i}, \mathbf{x_k}) \forall \mathbf{x_j} \in \mathbf{N}(\mathbf{x_i}), \forall \mathbf{x_k} \notin \mathbf{N}(\mathbf{x_i})$$

neighborhood of x_i: images sharing same meaning

Structure-Preserving Image-Text Embedding

Loss Function

$$
\begin{aligned}
L(X,Y) = &\sum_{i,j,k} \max[0, m + d(x_i, y_j) - d(x_i, y_k)] \\
&+ \lambda_1 \sum_{i',j',k'} \max[0, m + d(x_{j'}, y_{i'}) - d(x_{k'}, y_{i'})] \\
&+ \lambda_2 \sum_{i,j,k} \max[0, m + d(x_i, x_j) - d(x_i, x_k)] \\
&+ \lambda_3 \sum_{i',j',k'} \max[0, m + d(y_{i'}, y_{j'}) - d(y_{i'}, y_{k'})]
\end{aligned}
$$

Figure: Total Loss Function

Wang, Liwei, Yin Li, and Svetlana Lazebnik. ”Learning deep structure-preserving image-text embeddings.”

Structure-Preserving Image-Text Embedding

Experiment

- Dataset: Flicker30K
- Task: Image-Sentecne Retrieval
- Features:
  - Image: VGG Features
  - Sentence: Fisher Vector
- Training: SGD with momentum

Table: Performance on Image-Sentence Alignment

Model	Image Annotation			Image Search		
	R@1	R@5	R@10	R@1	R@5	R@10
BRNN(AlexNet)	22.2	48.2	61.4	15.2	37.7	50.5
MNLM(AlexNet)	22.2	48.2	61.4	15.2	37.7	50.5
CCA(VGGNet,FV)	36.5	62.2	73.3	24.7	53.4	66.8
Wang et al.(VGGNet, FV)	40.3	68.9	79.9	29.7	60.1	72.1

Wang, Liwei, Yin Li, and Svetlana Lazebnik. "Learning deep structure-preserving image-text embeddings."

Order-Embedding: Motivation

- Previous Methods project semantic similar units $\Longrightarrow$ similar vectors.

Vendrov, Ivan, et al. "Order-embeddings of images and language."

Order-Embedding: Motivation

- Previous Methods project semantic similar units $\Longrightarrow$ similar vectors.
- Challenges: Hard to define image/text similarity.

Vendrov, Ivan, et al. "Order-embeddings of images and language."

Order-Embedding: Motivation

- Previous Methods project semantic similar units $\Longrightarrow$ similar vectors.
- Challenges: Hard to define image/text similarity.
- Is it necessary to have symmetric similarity score?

Vendrov, Ivan, et al. "Order-embeddings of images and language."

Order-Embedding: Motivation

- Previous Methods project semantic similar units $\Longrightarrow$ similar vectors.
- Challenges: Hard to define image/text similarity.
- Is it necessary to have symmetric similarity score?
- Use Asymmetric Score!

Vendrov, Ivan, et al. "Order-embeddings of images and language."

Order-Embedding: Motivation

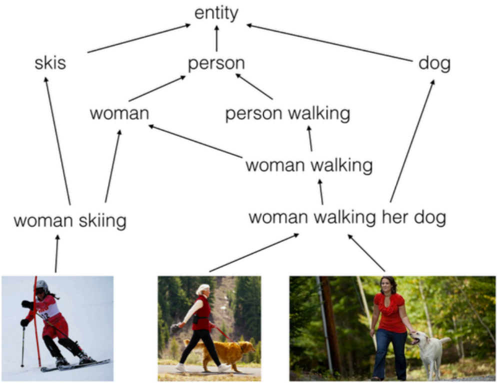


Figure: Order-Embedding: Motivation

Vendrov, Ivan, et al. "Order-embeddings of images and language."

Order-Embedding: Definition

- Order embedding function $f : \mathbf{X} \rightarrow \mathbf{Y}$:

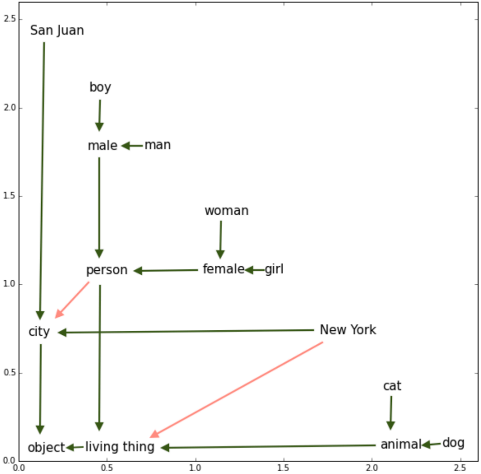
$$f(u) \preceq f(v) \Longleftrightarrow u \preceq v, \forall u, v \in \mathbf{X}$$

- How to define $\preceq$?

$$x \preceq y \Longleftrightarrow \bigwedge_{i=1}^{N} x_i \geq y_i$$
$$x, y \in \mathbb{R}_{+}^{N}$$

Vendrov, Ivan, et al. "Order-embeddings of images and language."

Order-Embedding: Example Embedding on WordNet



Vendrov, Ivan, et al. "Order-embeddings of images and language."

Order-Embedding: Define Loss Function

- Penalty given ordered pair (x, y):

$$E(x, y) = ||max(0, (y - x))||^2$$

- Loss Function:

$$\sum_{(u,v)\in P} E(f(u), f(v)) + \sum_{(u',v')\in N} \max\{0, \alpha - E(f(u'), f(v'))\}$$

Vendrov, Ivan, et al. "Order-embeddings of images and language."

Order-Embedding: Performance on WordNet

- Positive Set: All $(u, v) \in$ WordNet.
- Negative Set: Corrupted version of (u, v).

Table: Performance on WordNet Prediction

Algorithm	Accuracy
transitive closure	88.2
word2gauss	86.6
order-embeddings(symmetric)	84.2
order-embeddings(bilinear)	86.3
order-embeddings	90.6

Vendrov, Ivan, et al. "Order-embeddings of images and language."

Order-Embedding: For Image Caption

- Caption-image pairs are two-level partial order
- Similarity Score: $\mathbf{s}(\mathbf{c}, \mathbf{i}) = -\mathbf{E}(\mathbf{f_i}(\mathbf{i}), \mathbf{f_c}(\mathbf{c}))$
- Performance on MS-COCO (1k test)

Table: Performance on Image-Sentence Retrieval

Model	Image Annotation			Image Search		
	R@1	R@10	Med r	R@1	R@10	Med r
MNLM(AlexNet)	43.4	85.8	2	31.0	79.9	3
DVSA(AlexNet)	38.4	80.5	1	27.4	74.8	3
order-embeddings symm. (VGGNet)	45.4	88.7	2.0	36.3	85.8	2.0
order-embeddings(VGGNet)	46.7	88.9	2.0	37.9	85.9	2.0

Vendrov, Ivan, et al. "Order-embeddings of images and language."

Order-Embedding: For Image Caption

Captions	Image rank	
	cosine	order-emb
a sitting area with furniture and flowers makes a back-drop for a boy with headphones sitting in the fore-ground at one of the chairs at a dining table that holds glasses and a handbag working at a laptop	4	8
a kid is wearing headphone while on a laptop	286	24
view of top of a white building with tan speckled area an uncovered awning with a pigeon in fight below and a red umbrella behind balcony wall	3	5
a pigeon flying near white beams of a building	91	6

Vendrov, Ivan, et al. ”Order-embeddings of images and language.”

Image-Text Representation:Summary

Algorithms	Similarity Score	Task(Objective)	Contribution
MultiModal Language Model	cosine	Ranking	Multi-Modal LM
DVSA	cosine	Ranking	Each term only associated with one region
Structure-Preserving Embeddings	cosine	Ranking	structure-preserving constraints
Order-Embedding	Asymmetric	Ranking	Innovative similarity score

Table: Summaries of Multiple Image-Text Embedding Methods

Image-Text Representation:Summary

Algorithms	Similarity Score	Task(Objective)	Contribution
MultiModal Language Model	cosine	Ranking	Multi-Modal LM
DVSA	cosine	Ranking	Each term only associated with one region
Structure-Preserving Embeddings	cosine	Ranking	structure-preserving constraints
Order-Embedding	Asymmetric	Ranking	Innovative similarity score

Table: Summaries of Multiple Image-Text Embedding Methods

Takeaway Questions

How to develop new image-text embedding algorithms?

- New Similarity Score?
- Work on new Task like VQA? New Loss?
- Optimizing ranking loss more effectively?

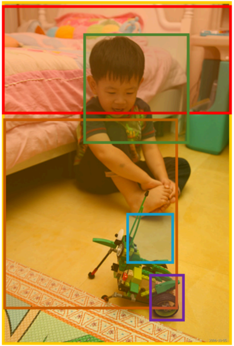
Go beyond Image-Text Representation...

Go beyond Image-Text Representation...

Applications

Tasks

Phrase Localization



A small Asian boy [0.45] is sitting on the floor [0.82] of a bedroom [0.87] being entertained and smiling at a lego toy [0.77] that looks like a bug [0.87] on wheels [0.81] .

Retrieval

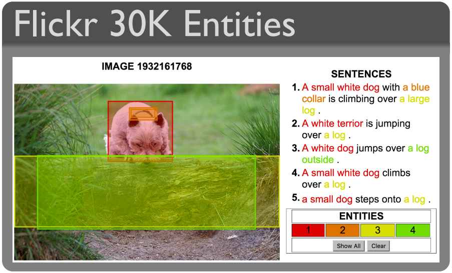
(a) Results for the query on a popular image search engine.

(b) Expected results for the query.

Figure 1: Image search using a complex query like "man holding fish and wearing hat on white boat" returns unsatisfactory results in (a). Ideal results (b) include correct *objects* ("man", "boat"), *attributes* ("boat is white") and *relationships* ("man on boat").

Plummer, Bryan A., et al. "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models." Johnson, Justin, et al. "Image retrieval using scene graphs."

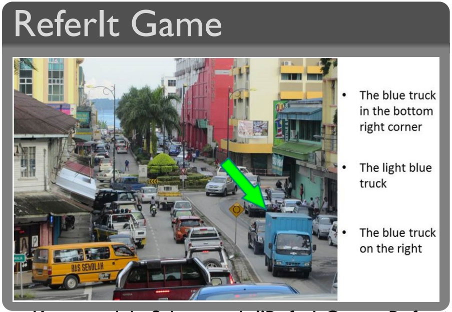
Dataset



- Augments the 158k captions from Flickr30k with 244k co-reference chains
- Links mentions of the same entities across different captions
- Associates the entities with 276k manually annotated bounding boxes

Plummer, Bryan A., et al. ”Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models.”

Datasets



- 130k expressions
- 96k objects
- 19k photos

Kazemzadeh, Sahar, et al. "ReferItGame: Referring to Objects in Photographs of Natural Scenes."

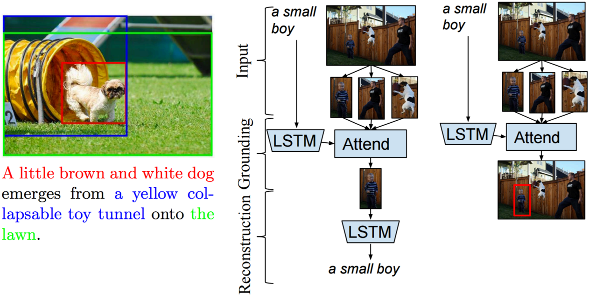
Grounding by reconstruction

- Localization annotation is costly
  - Flickr30k Entities only has 31k images with 158k captions
  - ReferIt Game has only 19k images with 130k expressions
- Needs to develop an unsupervised/semi-supervised method
  - GroundR: Learn to localize phrases relying only on sentence/visual data without localization annotations

Rohrbach, Anna, et al. "Grounding of textual phrases in images by reconstruction."

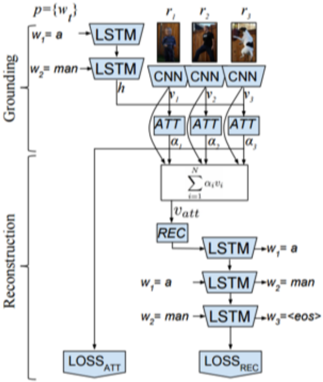
Grounding by reconstruction

- Method:
  - Learning to ground: Selecting a bounding box from region proposals
  - Learning to reconstruct: Reconstructing the phrase only from the attended boxes



(a) Predicted grounding. (b) Training time. (c) Test time.

Grounding by reconstruction



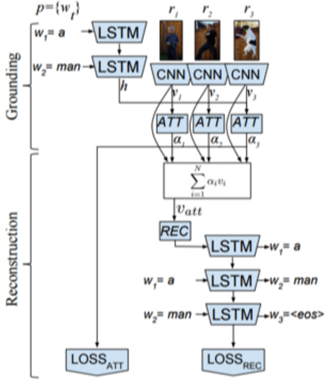
(b) Semi-supervised

$$\bar{\alpha}_i = f_{ATT}(p, r_i)$$

$$L_{att} = -\frac{1}{B}\sum_{b=1}^{B}\log(P(\hat{j}|\bar{\alpha}))$$

- Needs an objective function to attend the correct region
- Use two layer perceptron to compute the attention on the phrase and region
- Use softmax to obtain normalized attention weights

Grounding by reconstruction



(b) Semi-supervised

$$L_{rec} = -\frac{1}{B}\sum_{b=1}^{B}\log(P(p|v'_{att}))$$

Grounding by reconstruction

- Qualitative study on Flickr30K Entities
- Top is unsupervised and bottom is unsupervised. Note that the top one is much more accurate

A little girl in a pink shirt is looking at a toy doll.

A woman is riding a bicycle on the pavement.

A girl with a red cap, hair tied up and a gray shirt is fishing in a calm lake.

Grounding by reconstruction

- Qualitative study on ReferItGame
- Red box is the predicted box, green is the ground truth

two people on right

picture of a bird flying above sand

dat alpaca up in front, total coffeelate swag

palm tree coming out of the top of the building

guy with blue shirt and yellow shorts

hut to the nearest left of the person on the right

Grounding by reconstruction

‘

Table: Accuracy on the Flickr30k Entities Dataset

Approach	Accuracy VGG-CLS	VGG-DET
Unsupervised training		
GroundR	24.66	28.94
Supervised training		
CCA	27.42	-
GroundR	41.56	47.81
Proposal upperbound	77.90	77.90

Table: Accuracy on ReferItGame

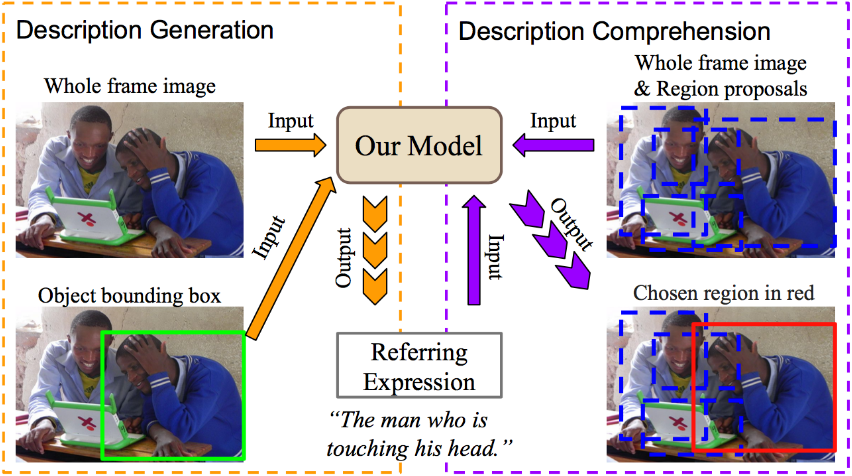
Approach	Accuracy VGG-CLS	VGG-DET
Unsupervised training		
GroundR	10.69	10.70
Supervised training		
SCRC	-	17.93
GroundR	23.44	26.93
Proposal upperbound	59.38	59.38

Description Generation and Comprehension

- Despite the recent interest in tasks such image caption, it is difficult to evaluate
- We will formulate the problem into two problems that can be objectively evaluated
  - Description generation: Generate a text expression that uniquely pinpoints a highlighted object
  - Description comprehension: Select an object given a text expression that refers to the object
- Hopefully, by modelling a listener, we can achieve better performance in both tasks by discriminate the object of interest from other objects in the image

Mao, Junhua, et al. ”Generation and comprehension of unambiguous object descriptions.”

Description Generation and Comprehension



Description Generation and Comprehension

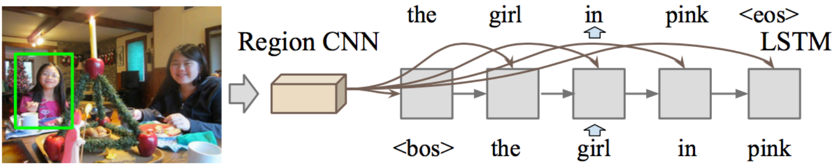
- Description Generation:
  - Compute $argmax_S P(S|R, I)$
  - S is a sentence, R is a region and I is an image
- Comprehension:
  - Needs to compute $R^* = argmax_{R' \in \mathcal{C}} p(R|S, I)$
  - $P(R|S, I) = \frac{P(S|R,I)p(R|I)}{\sum_{R' \in \mathcal{C}} P(S|R',I)P(R'|I)}$
- We need a method to model $p(S|R, I)$.

Description Generation and Comprehension

- Model:
  - Uses VGGNet and a 5 dimensional vector encoding the bounding box to generate the image features of 2005 dimensions
  - Feed the feature vector into an LSTM sequence model to parameterized the distribution $p(S|R, I)$

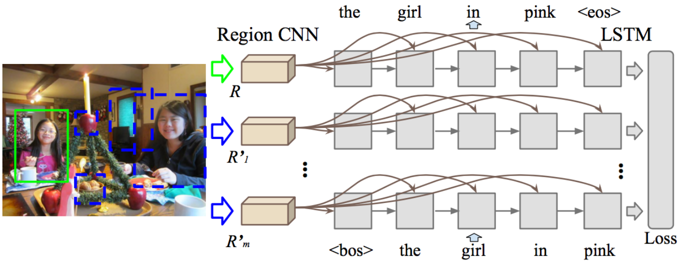
Description Generation and Comprehension

- Training: Maximum Likelihood Training
  - Minimize the negative log probability over the entire dataset
  - Objective: $J(\theta) = -\sum_{n=1}^{N} \log p(S_n | R_n, I_n, \theta)$



Description Generation and Comprehension

- Training: Maximum Likelihood Training
  - Drawback: only generate description on the target object
  - We need to introduce negative examples
- Instead we will use a “softmax” loss to discriminate the target object against other objects
- Discriminative Training (softmax loss)
  - $J'(\theta) = -\sum_{n=1}^{N} \log p(R_n|S_n, I_n, \theta)$
  - $\log p(R_n|S_n, I_n, \theta) = \log \frac{p(S_n|R_n, I_n, \theta)}{\sum_{R' \in \mathcal{C}(I_n)} p(S_n|R', I_n, \theta)}$

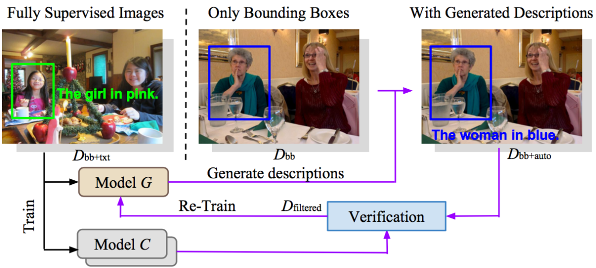


Description Generation and Comprehension

- Discriminative Training (softmax loss)
  - The softmax loss is computational expensive to calculate so we use a max-margin instead
  - $\max(0, M - \log p(S_n|R_n, I_n, \theta) + \log p(S_n|R'_n, I_n, \theta))$
  - R' is a negative example

Description Generation and Comprehension

- Semi-supervised learning
  - For training with a small dataset D_{bb+txt} with bounding box and description and a large dataset D_{bb} of images with bounding boxes but no descriptions
  - First train a model G on D_{bb+txt} and generate description on D_{bb} to get a new dataset $D_{bb+auto}$
  - Retrain G on $D_{bb+txt} \bigcup D_{bb+auto}$
- Also, train an ensemble of different models C on the dataset D_{bb+txt} for verification



Description Generation and Comprehension

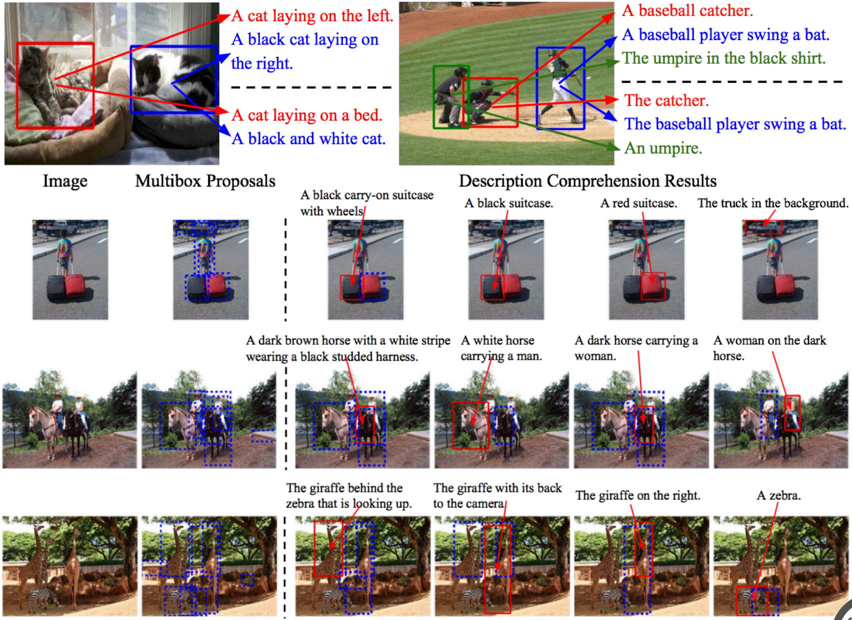
- Performance of full model vs base model
- GT(Comprehension) ground truth is based on Intersection over Union
- GEN(Generation) is manually labeled but Amazon Mechanical Turk workers

Proposals	GT		multibox	
Descriptions	GEN	GT	GEN	GT
Google Refexp-Val				
Baseline	0.751	0.579	0.468	0.425
Full Model	**0.799**	**0.607**	**0.500**	**0.445**
Google Refexp-Test				
Baseline	0.769	0.545	0.485	0.406
Full Model	**0.811**	**0.606**	**0.513**	**0.446**

- Performance of the full model on a small labeled dataset vs wuth automatically labeled data

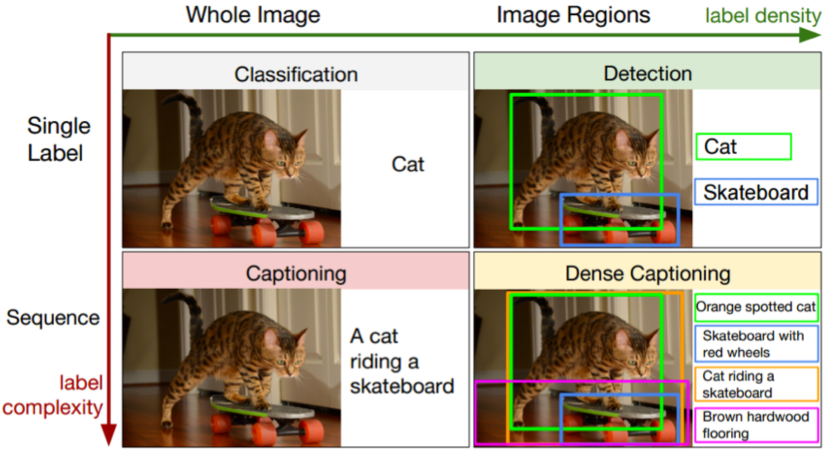
Proposals	GT		multibox	
Descriptions	GEN	GT	GEN	GT
Google Refexp				
$D_{\text{bb+txt}}$	0.791	0.561	**0.489**	0.417
$D_{\text{bb+txt}} \cup D_{\text{bb}}$	**0.793**	**0.577**	**0.489**	**0.424**

Description Generation and Comprehension



Dense Captioning

- The Task



Johnson, Justin, Andrej Karpathy, and Li Fei-Fei. "Densecap: Fully convolutional localization networks for dense captioning."

Dense Captioning

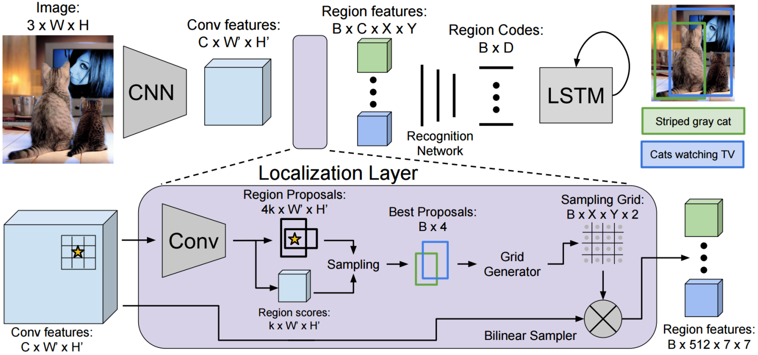
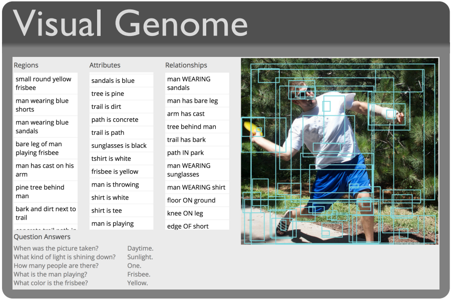


Figure 2. Model overview. An input image is first processed a CNN. The Localization Layer proposes regions and smoothly extracts a batch of corresponding activations using bilinear interpolation. These regions are processed with a fully-connected recognition network and described with an RNN language model. The model is trained end-to-end with gradient descent.

Johnson, Justin, Andrej Karpathy, and Li Fei-Fei. "Densecap: Fully convolutional localization networks for dense captioning."

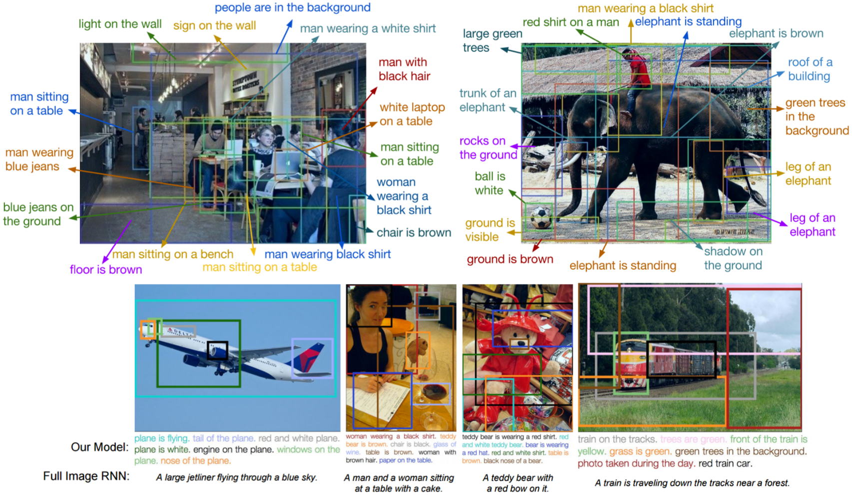
Datasets



- A dataset with 108k images
- 5.4M Regional description
- 1.7M Visual question Answering
- 3.8M Object instances
- 2.8M attributes
- 2.3M relationships

Krishna, Ranjay, et al. "Visual genome: Connecting language and vision using crowdsourced dense image annotations."

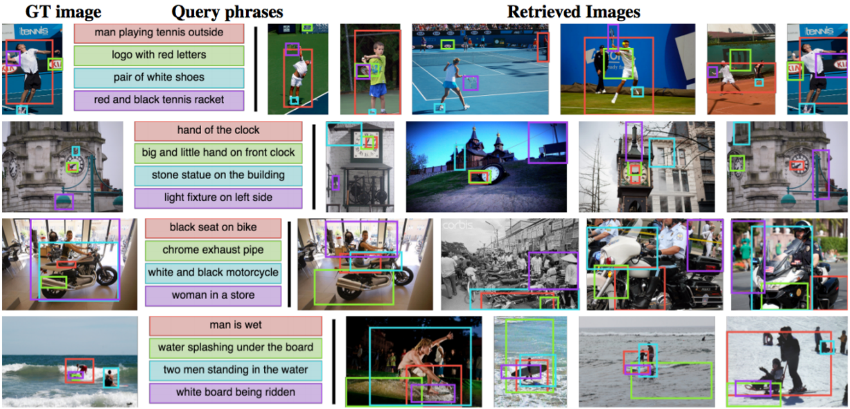
Dense Captioning



Dense Captioning

Region source	Language (METEOR) EB	RPN	GT	Dense captioning (AP) EB	RPN	GT
Full image RNN [22]	0.173	0.197	0.209	2.42	4.27	*14.11*
Region RNN [22]	0.221	0.244	0.272	1.07	4.26	*21.90*
FCLN on EB [14]	**0.264**	**0.296**	0.293	4.88	3.21	*26.84*
Our model (FCLN)	**0.264**	0.273	**0.305**	**5.24**	**5.39**	***27.03***

Dense Captioning



Dense Captioning

	Ranking			
	R@1	R@5	R@10	Med. rank
Full Image RNN [22]	0.10	0.30	0.43	13
EB + Full Image RNN [22]	0.11	0.40	0.55	9
Region RNN [14]	0.18	0.43	0.59	7
Our model (FCLN)	**0.27**	**0.53**	**0.67**	**5**

Summary

- Beyond embedding
  - GroundR: Learn to attend bounding box and reconstruct phrases
    - Can learn to ground semi-supervised or unsupervised
    - Suspicious evaluation
  - Mao et.al: Learn to train a "listener" to discriminate non-target regions against target regions
    - Can learn to generate descriptions in semi-supervised way
    - Can only select bounding boxes (as oppose to proposing bounding boxes).
    - No comparison against other models
  - DenseCap: Use a localization network to perform end-to-end training
    - Does not need external bounding box proposal
    - The localization network can be inserted into any neural network to enable localized predictions

Today's Summary

- Part I: Computer Vision Tasks Introduction
- Part II: Foundation: How to represent image and text? $\Longrightarrow$ image-text representation.
  - Similarity Definition: Both symmetric and asymmetric
  - Task: Image-Sentence Ranking (Retrieval, Grounding, Captioning, Question Answering...)
  - Objective Function: Ranking Loss, Reconstruction Loss
- Part III: Three Image-Text Applications:
  - GroundR: Grounding by reconstruction
  - Referring Expression: Description generation and comprehension
  - DenseCap: Generating captioning and perform localization

Readinglist I

- Plummer, Bryan A., et al. ”Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models.” Proceedings of the IEEE International Conference on Computer Vision. 2015.
- Krishna, Ranjay, et al. ”Visual genome: Connecting language and vision using crowdsourced dense image annotations.” arXiv preprint arXiv:1602.07332 (2016).
- Kiros, Ryan, Ruslan Salakhutdinov, and Richard S. Zemel. ”Unifying visual-semantic embeddings with multimodal neural language models.” arXiv preprint arXiv:1411.2539 (2014).
- Karpathy, Andrej, and Li Fei-Fei. ”Deep visual-semantic alignments for generating image descriptions.” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015.

Readinglist II

- Wang, Liwei, Yin Li, and Svetlana Lazebnik. ”Learning deep structure-preserving image-text embeddings.” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.
- Vendrov, Ivan, et al. ”Order-embeddings of images and language.” arXiv preprint arXiv:1511.06361 (2015).
- Rohrbach, Anna, et al. ”Grounding of textual phrases in images by reconstruction.” European Conference on Computer Vision. Springer International Publishing, 2016.
- Mao, Junhua, et al. ”Generation and comprehension of unambiguous object descriptions.” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.

Readinglist III

- Johnson, Justin, Andrej Karpathy, and Li Fei-Fei. ”Densecap: Fully convolutional localization networks for dense captioning.” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.

Reference I

[1] S. Bell, C. Lawrence Zitnick, K. Bala, and R. Girshick. Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2874–2883, 2016.

[2] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. arXiv preprint arXiv:1606.01847, 2016.

[3] J. Johnson, A. Karpathy, and L. Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4565–4574, 2016.

Reference II

[4] J. Johnson, R. Krishna, M. Stark, L.-J. Li, D. Shamma, M. Bernstein, and L. Fei-Fei. Image retrieval using scene graphs. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3668–3678, 2015.

[5] R. Kiros, R. Salakhutdinov, and R. S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. arXiv preprint arXiv:1411.2539, 2014.

[6] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. arXiv preprint arXiv:1602.07332, 2016.

Reference III

[7] J. Mao, J. Huang, A. Toshev, O. Camburu, A. L. Yuille, and K. Murphy. Generation and comprehension of unambiguous object descriptions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 11–20, 2016.

[8] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In Proceedings of the IEEE International Conference on Computer Vision, pages 2641–2649, 2015.

[9] A. Rohrbach, M. Rohrbach, R. Hu, T. Darrell, and B. Schiele. Grounding of textual phrases in images by reconstruction. In European Conference on Computer Vision, pages 817–834. Springer, 2016.

Reference IV

[10] I. Vendrov, R. Kiros, S. Fidler, and R. Urtasun. Order-embeddings of images and language. arXiv preprint arXiv:1511.06361, 2015.

[11] L. Wang, Y. Li, and S. Lazebnik. Learning deep structure-preserving image-text embeddings. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 5005–5013, 2016.

[12] M. Wang, M. Azab, N. Kojima, R. Mihalcea, and J. Deng. Structured matching for phrase localization. In European Conference on Computer Vision, pages 696–711. Springer, 2016.

Reference V

[13] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions of the Association for Computational Linguistics, 2:67–78, 2014.