VALSE 视觉与学习青年学者研讨会 Vision And Learning SEminar

Microsoft Research

idea INTERNATIONAL DIGITAL ECONOMY ACADEMY 粤港澳大湾区数字经济研究院

VALSE Tutorial

A Tutorial On Vision Language Intelligence

Pengchuan Zhang

Microsoft Research (MSR)

https://www.microsoft.com/en-us/research/

Lei Zhang

International Digital Economy Academy (IDEA)

https://idea.edu.cn/

A Tutorial On Vision Language Intelligence

Lecture 1: Representation and Attention

Pengchuan Zhang and Lei Zhang
Microsoft Research (MSR)
International Digital Economy Academy (IDEA)

Multimodal Intelligence

The **moose** (in North America) or **elk** (in Eurasia) (*Alces alces*), is a member of the New World deer subfamily and is the largest and heaviest extant species in the deer family. Most adult male moose have distinctive broad, palmate ("open-hand shaped") antlers; most other members of the deer family have antlers with a dendritic ("twig-like") configuration. Moose typically inhabit boreal forests and temperate broadleaf and mixed forests of the Northern Hemisphere in temperate to subarctic climates. (wikipedia.org)

Male (bull)

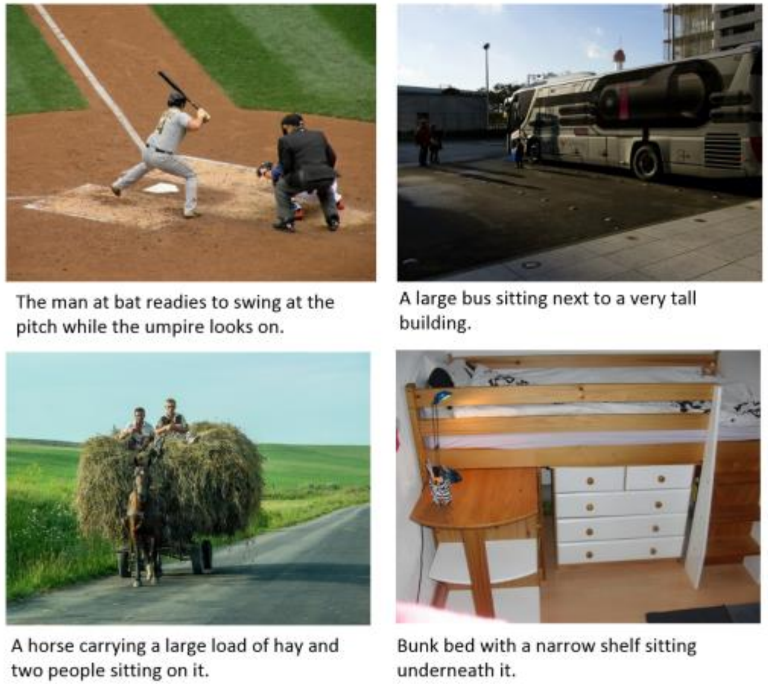
Female (cow)

Vision-Language Tasks

	Text-to-Image Retrieval	Image-to-Text Retrieval	VQA	Image Captioning	Text-to-Image Generation
Input	Query: A couple of zebra walking across a dirt road. A pool of images.	Query:  A pool of texts.	Image:  Q: why did the zebra cross the road?	Image: 	Text: A couple of zebra walking across a dirt road.
Output		A couple of zebra walking across a dirt road.	A: to get to the other side (Selected from a pool of 3,129 answers in VQAv2)	A couple of zebra walking across a dirt road.	
	Understanding	**Understanding**	**Understanding**	**Generation**	**Generation**

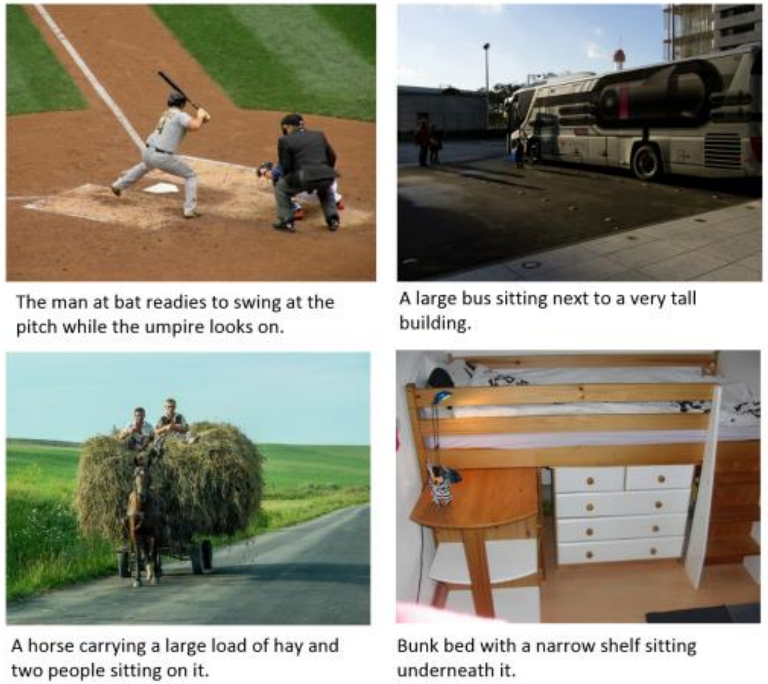
Vision-Language Research Problems

- Vision-language tasks require understanding concepts in both vision and language, and fusing information from both modalities



Vision-Language Research Problems

- Vision-language tasks require understanding concepts in both vision and language, and fusing information from both modalities
- Understanding concepts in both vision and language (single modality representation):
  - vision representation (CNN -> Faster-RCNN -> Transformer)
  - language representation (RNN -> Transformer)
- Vision Language information fusion (attention)



Roadmap

- **Start with task-specific problem in vision-language (image captioning)**
  - **Single modality representation and cross-modality fusion**
- Vision-language joint representation learning empowered by pre-training
  - How to train VL-aligned representation?
- Vision-Language for improving Vision tasks and Language tasks
  - More data, Larger models for pre-training, zero/fewer-shot for fine-tuning

Representation and attention

- Baby talk [CVPR 2011], "Every Picture Tells a Story"[ECCV2010]
  - object, attribute, and relation detectors learned from separate hand-labeled training data
  - Hard-coded templates for caption generation
- Show and Tell [CVPR 2015] (S2S)
  - Deep Visual-Semantic Alignments [CVPR 2015], From Captions to Visual Concepts and Back [CVPR 2015], m-RNN [ICLR 2015], Long-term Recurrent Convolutional Networks [CVPR 2015]
  - One global image feature from a CNN encoder
  - Language generating RNN
- Show, Attend and Tell [ICML 2015] (S2S with attention)
  - A 2D grid of image features from a CNN encoder
  - Visual attention was introduced
- Bottom-up and Top-down Attention [CVPR 2018] (Object-centric attention)
  - Object-centric image features from an object detector
  - BUTD attention was introduced

Representation and attention

- Baby talk [CVPR 2011], "Every Picture Tells a Story"[ECCV2010]
  - object, attribute, and relation detectors learned from separate hand-labeled training data
  - Hard-coded templates for caption generation
- Show and Tell [CVPR 2015] (S2S)
  - Deep Visual-Semantic Alignments [CVPR 2015], From Captions to Visual Concepts and Back [CVPR 2015], m-RNN [ICLR 2015], Long-term Recurrent Convolutional Networks [CVPR 2015]
  - One global image feature from a CNN encoder
  - Language generating RNN
- Show, Attend and Tell [ICML 2015] (S2S with attention)
  - A 2D grid of image features from a CNN encoder
  - Visual attention was introduced
- Bottom-up and Top-down Attention [CVPR 2018] (Object-centric attention)
  - Object-centric image features from an object detector
  - BUTD attention was introduced

Background in Image Captioning
– Early Research before Deep Learning

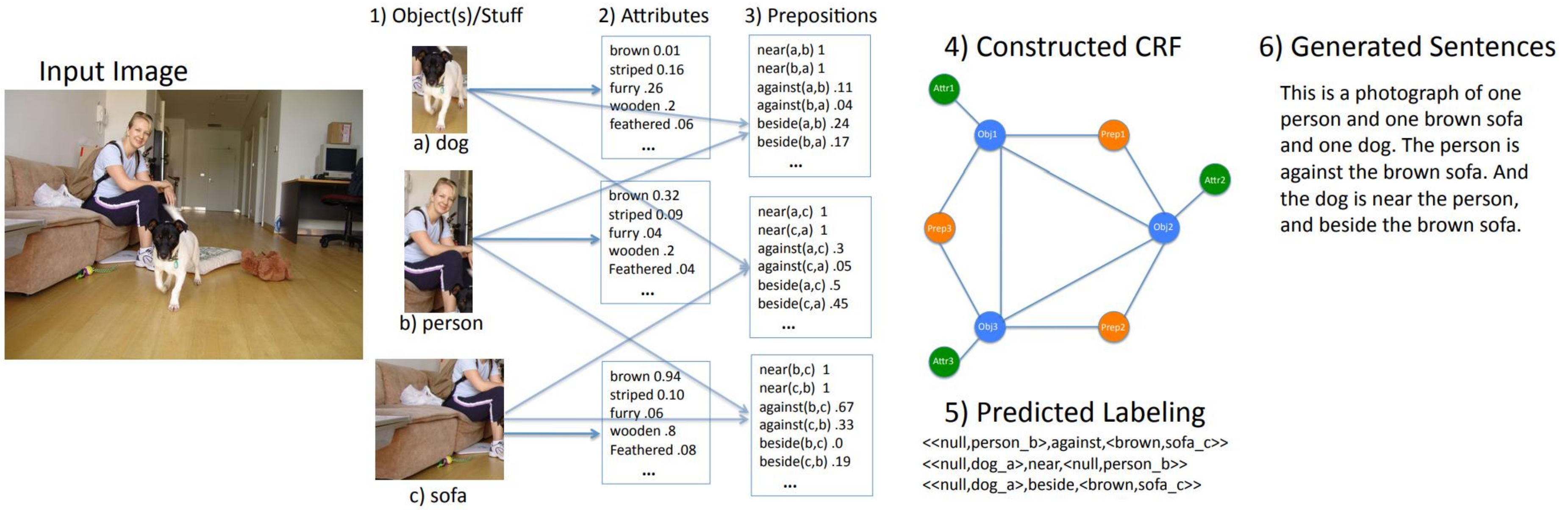


Figure 2. System flow for an example image: 1) object and stuff detectors find candidate objects, 2) each candidate region is processed by a set of attribute classifiers, 3) each pair of candidate regions is processed by prepositional relationship functions, 4) A CRF is constructed that incorporates the unary image potentials computed by 1-3, and higher order text based potentials computed from large document corpora, 5) A labeling of the graph is predicted, 6) Sentences are generated based on the labeling.

Pros: one of the first works in image captioning

Cons: poor diversity in generated captions
- visual concept recognizers with limited vocabulary
- template-based caption generation system

G. Kulkarni et al., "Baby talk: Understanding and generating simple image descriptions," CVPR 2011
Farhadi A. et al., “Every Picture Tells a Story: Generating Sentences from Images,” ECCV 2010

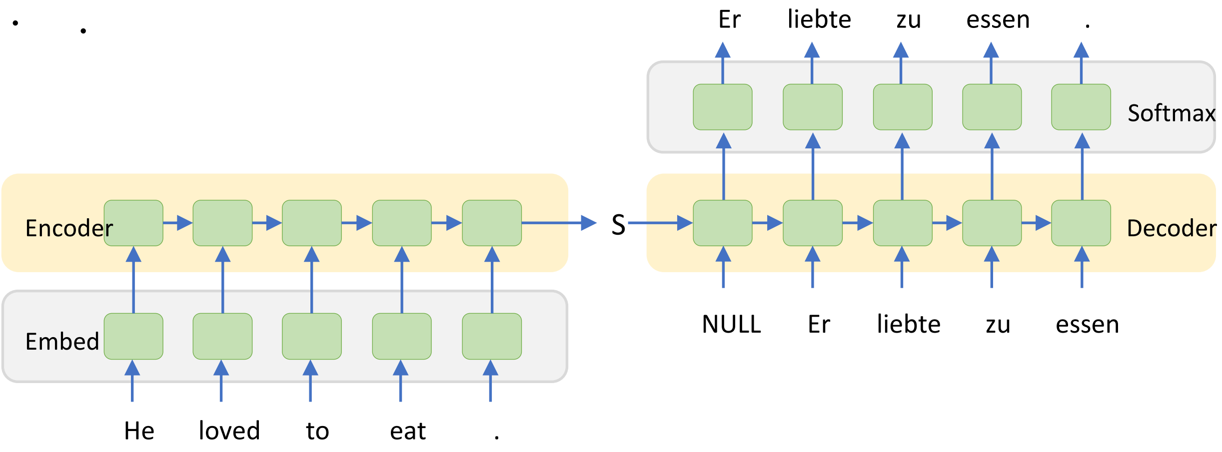
Representation and attention

- Baby talk [CVPR 2011], "Every Picture Tells a Story"[ECCV2010]
  - object, attribute, and relation detectors learned from separate hand-labeled training data
  - Hard-coded templates for caption generation
- Show and Tell [CVPR 2015] (S2S)
  - Deep Visual-Semantic Alignments [CVPR 2015], From Captions to Visual Concepts and Back [CVPR 2015], m-RNN [ICLR 2015], Long-term Recurrent Convolutional Networks [CVPR 2015]
  - One global image feature from a CNN encoder
  - Language generating RNN
- Show, Attend and Tell [ICML 2015] (S2S with attention)
  - A 2D grid of image features from a CNN encoder
  - Visual attention was introduced
- Bottom-up and Top-down Attention [CVPR 2018] (Object-centric attention)
  - Object-centric image features from an object detector
  - BUTD attention was introduced

Deep Learning in Image Captioning

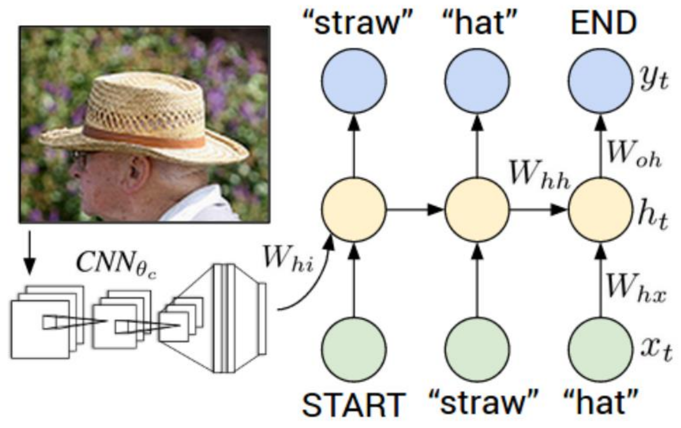
– Early Research in 2014/2015

Machine Translation: Sequence to sequence



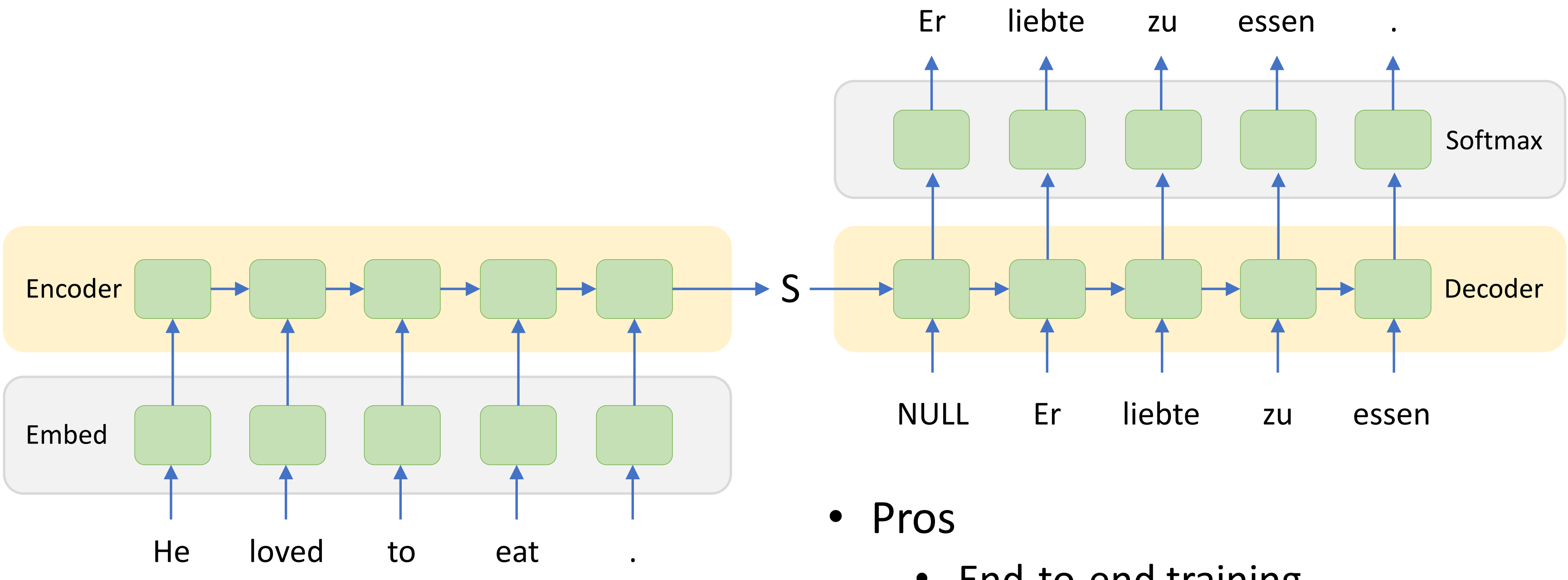
Sutskever, et al. “Sequence to sequence learning with neural networks.” NIPS 2014

Image Captioning: Image to sequence



Karpathy, Andrej, and Li Fei-Fei. "Deep visual-semantic alignments for generating image descriptions." [CVPR 2015]
Show and Tell [CVPR 2015]
m-RNN [ICLR 2015]
Long-term Recurrent Convolutional Networks [CVPR 2015]

RNN/LSTM-based Seq2Seq Learning



- Pros
  - End-to-end training
  - Potential to learn one latent embedding for multiple tasks -> multi-task learning
- Cons
  - Hard to learn long-range dependencies

Sutskever, et al. "Sequence to sequence learning with neural networks." NIPS 2014

A compositional approach

Understand the image stage by stage:

Image word detection

Deep-learned features, applied to likely items in the image, trained to produce words in captions

Language generation

Maxent language model (MELM), trained on caption, conditional on words detected from the image

Global semantic re-ranking

Hypothetical captions re-ranked by deep multimodal similarity model (DMSM) looking at the entire image

"From Captions to Visual Concepts and Back" [CVPR 2015]

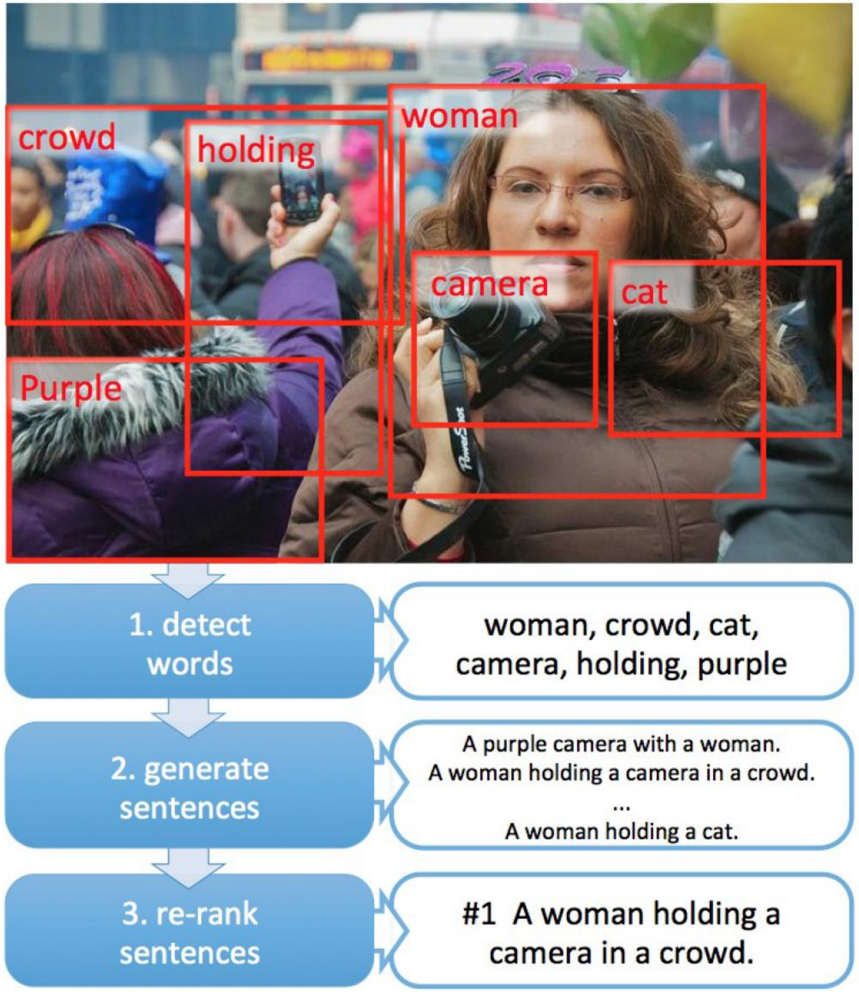


Figure 1. An illustrative example of our pipeline.

MS COCO Captioning Challenge 2015

TABLE 9
Automatic scores of the top five competition submissions.

	CIDER	METEOR	ROUGE	BLEU-4	Rank
Google [46]	0.943	0.254	0.53	0.309	1st
MSR Captivator [34]	0.931	0.248	0.526	0.308	2nd
m-RNN [28]	0.917	0.242	0.521	0.299	3rd
MSR [23]	0.912	0.247	0.519	0.291	4th
m-RNN (2) [28]	0.886	0.238	0.524	0.302	5th
Human	0.854	0.252	0.484	0.217	8th

TABLE 10
Human generated scores of the top five competition submissions.

	M1	M2	M3	M4	M5	Rank
Google [46]	0.273	0.317	4.107	2.742	0.233	1st
MSR [23]	0.268	0.322	4.137	2.662	0.234	1st
MSR Captivator [34]	0.250	0.301	4.149	2.565	0.233	3rd
Montreal/Toronto [31]	0.262	0.272	3.932	2.832	0.197	3rd
Berkeley LRCN [30]	0.246	0.268	3.924	2.786	0.204	5th
Human	0.638	0.675	4.836	3.428	0.352	1st

Big gap at that time!

Human evaluation metrics:
M1: Percentage of captions that are evaluated as better or equal to human caption
M2: Percentage of captions that pass the Turing Test.
M3: Average correctness of the captions on a scale 1-5 (incorrect - correct).
M4: Average amount of detail of the captions on a scale 1-5 (lack of details - very detailed).
M5: Percentage of captions that are similar to human description.

Show and Tell: Lessons learned from the 2015 MSCOCO Image Captioning Challenge https://arxiv.org/pdf/1609.06647.pdf

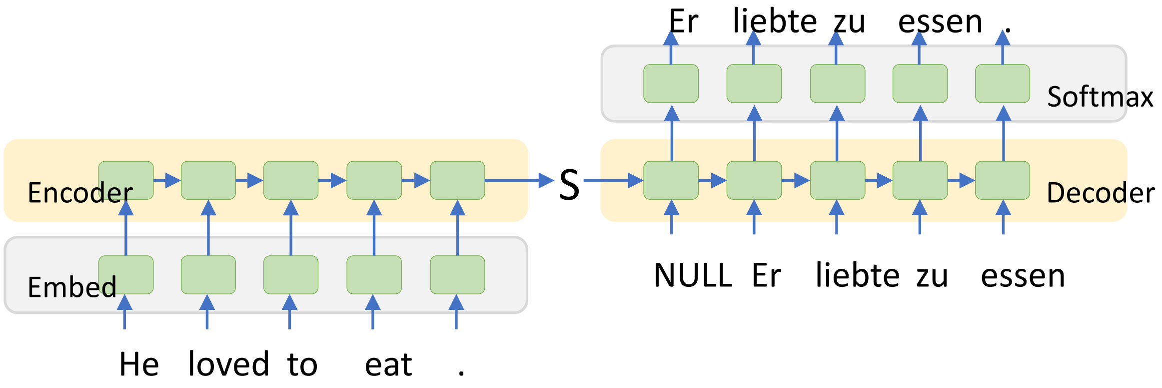
Representation and attention

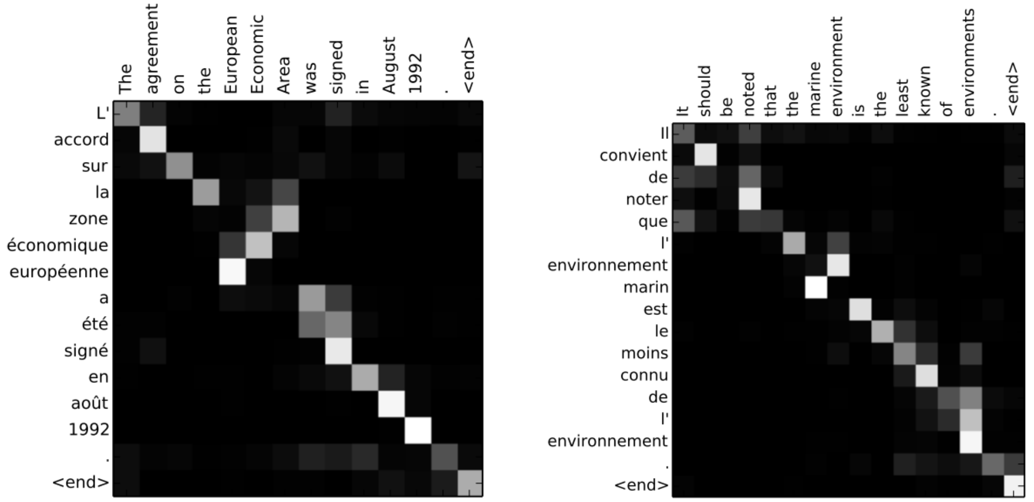
- Baby talk [CVPR 2011], "Every Picture Tells a Story"[ECCV2010]
  - object, attribute, and relation detectors learned from separate hand-labeled training data
  - Hard-coded templates for caption generation
- Show and Tell [CVPR 2015] (S2S)
  - Deep Visual-Semantic Alignments [CVPR 2015], From Captions to Visual Concepts and Back [CVPR 2015], m-RNN [ICLR 2015], Long-term Recurrent Convolutional Networks [CVPR 2015]
  - One global image feature from a CNN encoder
  - Language generating RNN
- Show, Attend and Tell [ICML 2015] (S2S with attention)
  - A 2D grid of image features from a CNN encoder
  - Visual attention was introduced
- Bottom-up and Top-down Attention [CVPR 2018] (Object-centric attention)
  - Object-centric image features from an object detector
  - BUTD attention was introduced

Attention-based Decoder

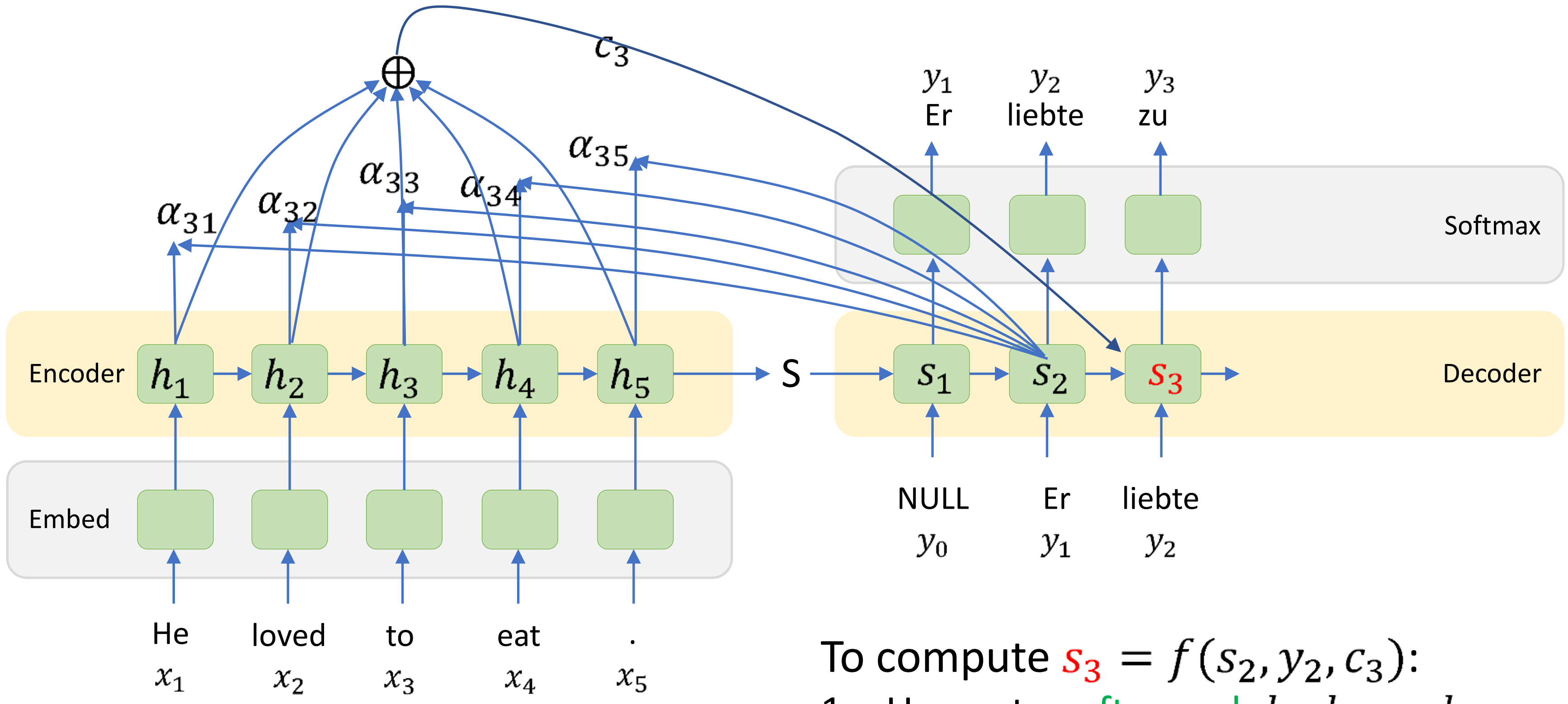
- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. "**Neural machine translation by jointly learning to align and translate.**" ICLR 2015.

  Abstract
  … In this paper, we conjecture that the use of a fixed-length vector is a bottleneck in improving the performance of this basic encoder–decoder architecture, and propose to extend this by allowing a model to automatically (soft-)search for parts of a source sentence that are relevant to predicting a target word, without having to form these parts as a hard segment explicitly. …





(a) (b)



$$s_i = f(s_{i-1}, y_{i-1}, c_i)$$

$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j.$$

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})},$$

$$e_{ij} = a(s_{i-1}, h_j)$$

To compute $s_3 = f(s_2, y_2, c_3)$:

1. Use s_2 to soft-search $h_1, h_2, \ldots, h_5$

   $e_{31} = a(s_2, h_1), e_{32} = a(s_2, h_2), \ldots, e_{35} = a(s_2, h_5)$

2. Apply softmax to $e_{31}, \ldots, e_{35}$ and get

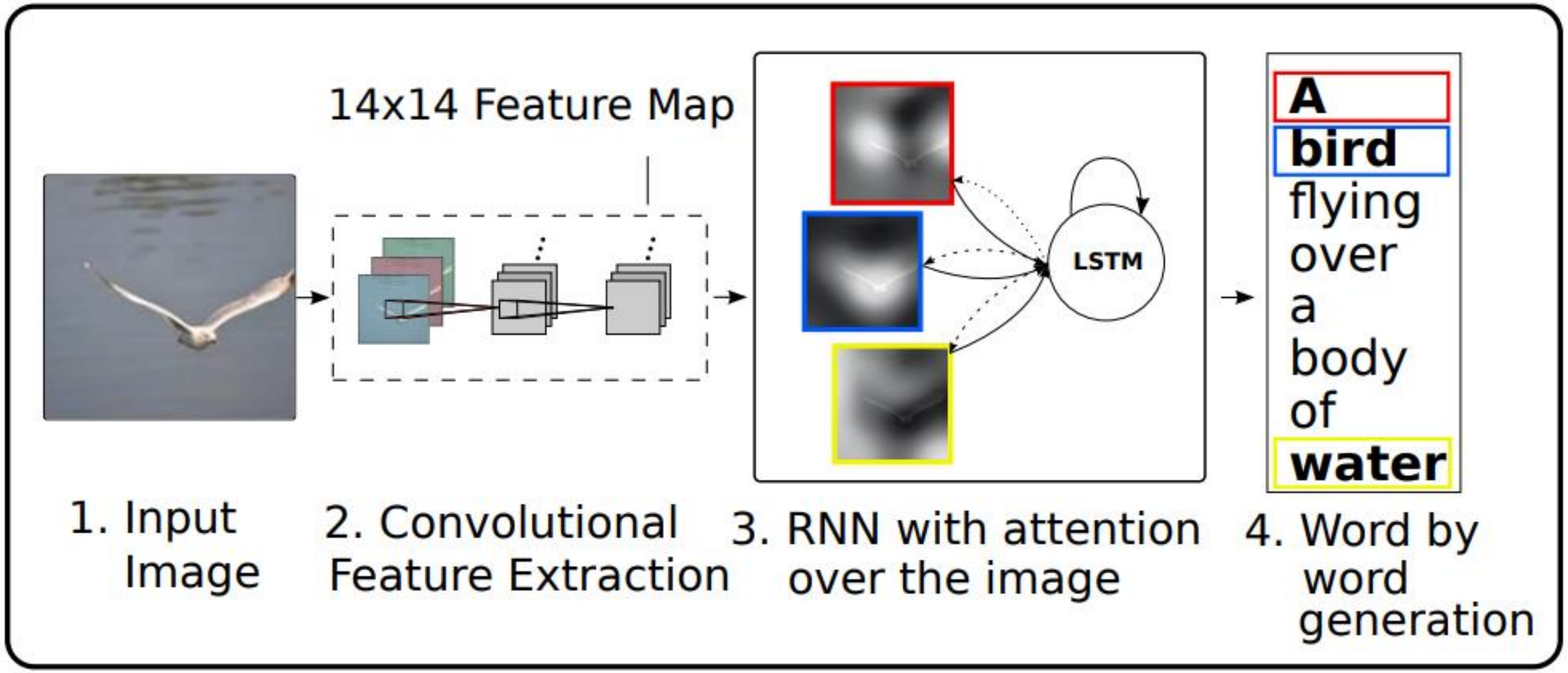
   $\alpha_{31}, \alpha_{32}, \ldots, \alpha_{35}$

3. Weighted sum over $h_1, h_2, \ldots, h_5$

$$c_3 = \sum_{j=1}^{5} \alpha_{3j} h_j = \alpha_{31} h_1 + \alpha_{32} h_2 + \cdots + \alpha_{35} h_5$$

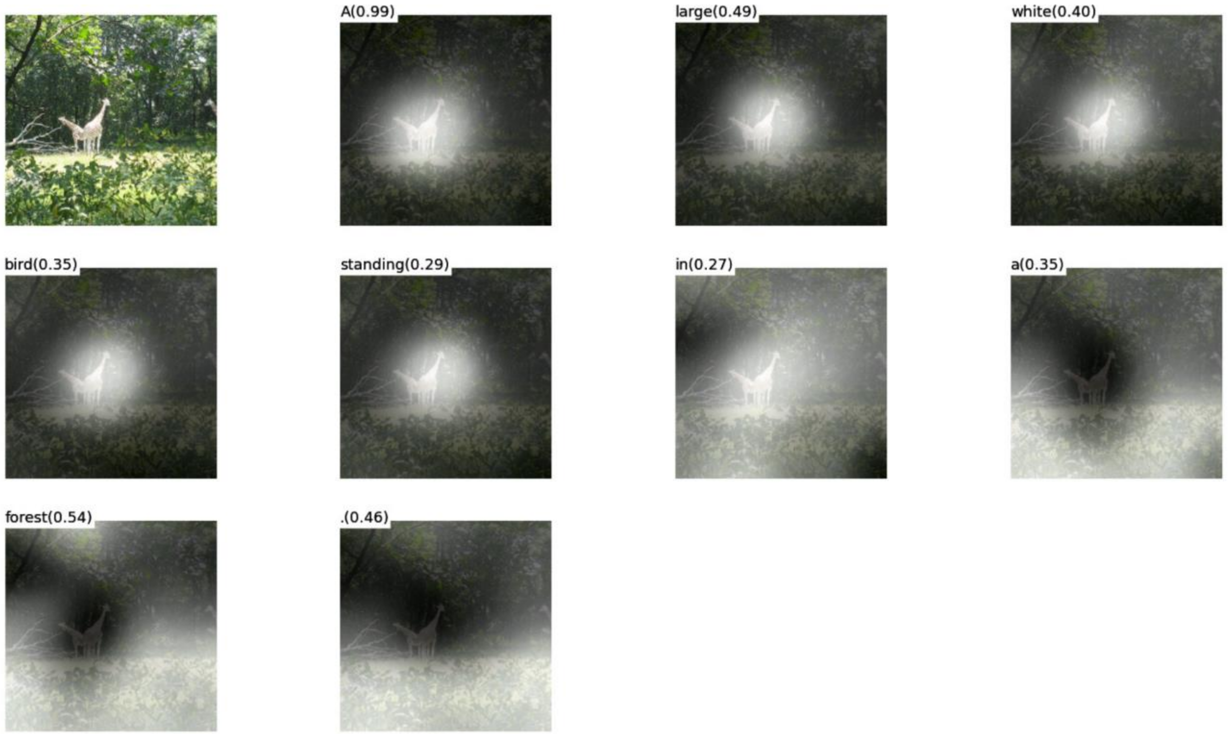
Attention in Image Captioning

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. "**Show, attend and tell**: Neural image caption generation with visual attention." ICML 2015.



A woman is throwing a frisbee in a park.

A large white bird standing in a forest.



- The model learns alignments that correspond very strongly with human intuition.
- It is possible to exploit such visualizations to get an intuition as to why those mistakes were made.

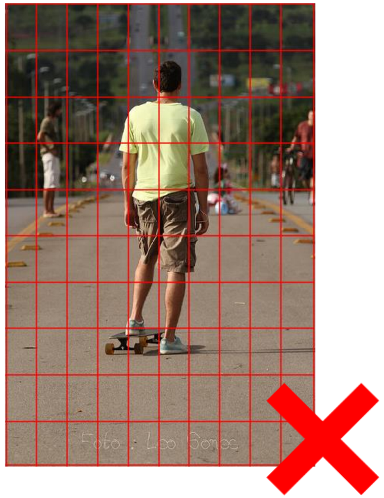
Representation and attention

- Baby talk [CVPR 2011], "Every Picture Tells a Story"[ECCV2010]
  - object, attribute, and relation detectors learned from separate hand-labeled training data
  - Hard-coded templates for caption generation
- Show and Tell [CVPR 2015] (S2S)
  - Deep Visual-Semantic Alignments [CVPR 2015], From Captions to Visual Concepts and Back [CVPR 2015], m-RNN [ICLR 2015], Long-term Recurrent Convolutional Networks [CVPR 2015]
  - One global image feature from a CNN encoder
  - Language generating RNN
- Show, Attend and Tell [ICML 2015] (S2S with attention)
  - A 2D grid of image features from a CNN encoder
  - Visual attention was introduced
- Bottom-up and Top-down Attention [CVPR 2018] (Object-centric attention)
  - Object-centric image features from an object detector
  - BUTD attention was introduced

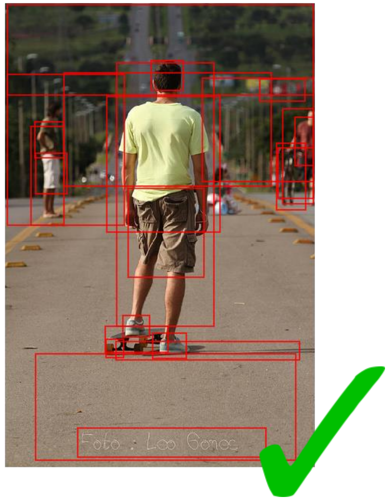
Attention Empowered by Object Detection

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. "Bottom-up and top-down attention for image captioning and visual question answering." CVPR 2018.

Key Idea – Introduce object detection to improve visual representation.

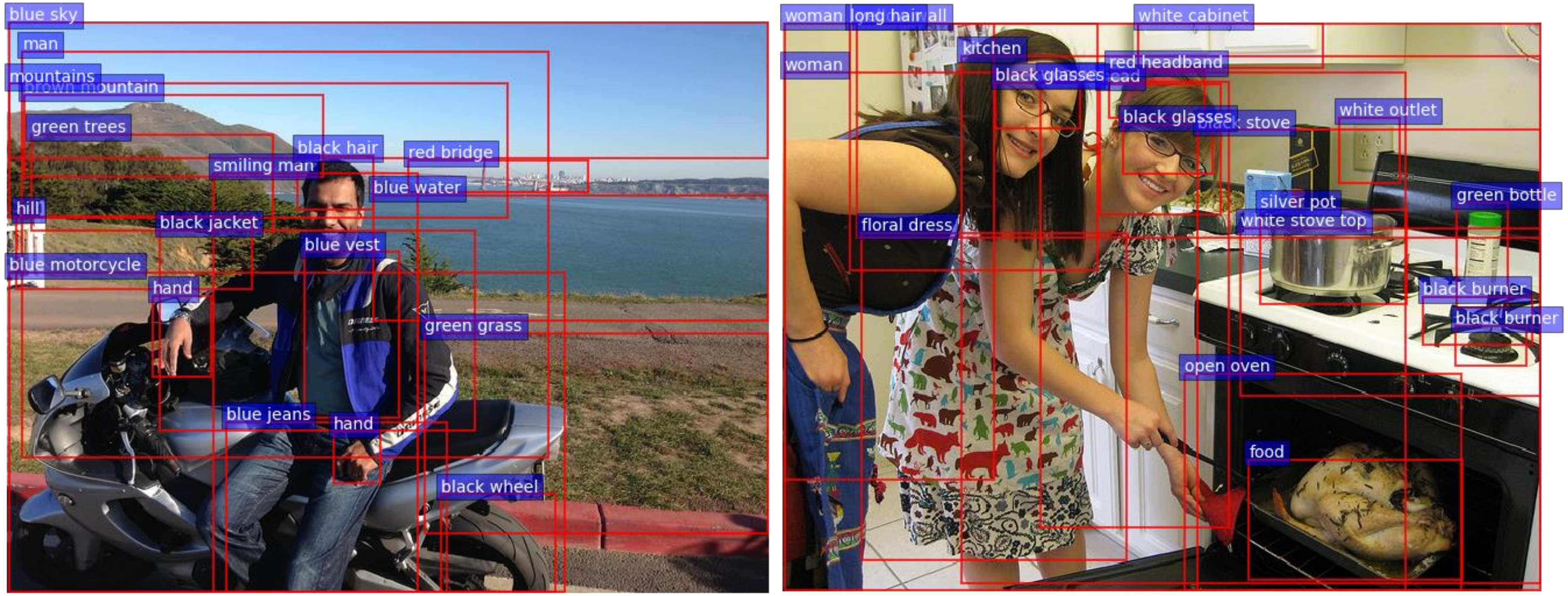


Typical approach: spatial output of CNN



Our approach: object-centric attention

Pre-training



- Pre-train Faster R-CNN on Visual Genome[1] (1600 objects / 400 attributes)
- Bottom-up selection of salient regions based on object confidence scores
- Take the mean-pooled ResNet-101[2] feature from each region

[1]http://visualgenome.org, [2]CVPR 2016

Qualitative differences - captioning

Grid feature: A man sitting on a ~~toilet~~ in a bathroom.

Object-centric feature: A man sitting on a couch in a bathroom.

Results

VQA v2 – validation set:

	Yes/No	Number	Other	Overall
Ours: ResNet (1×1)	76.0	36.5	46.8	56.3
Ours: ResNet (14×14)	76.6	36.2	49.5	57.9
Ours: ResNet (7×7)	77.6	37.7	51.5	59.4
Ours: Up-Down	**80.3**	**42.8**	**55.8**	**63.2**
Relative Improvement	3%	14%	8%	6%

- Significant improvements over baseline
- 1st VQA challenge 2017
- 1st MSCOCO test leaderboard (July 2017)

MS COCO captions – Karpathy test set:

	Cross-Entropy Loss						CIDEr Optimization					
	BLEU-1	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE	BLEU-1	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
SCST:Att2in [37]	-	31.3	26.0	54.3	101.3	-	-	33.3	26.3	55.3	111.4	-
SCST:Att2all [37]	-	30.0	25.9	53.4	99.4	-	-	34.2	26.7	55.7	114.0	-
Ours: ResNet	74.5	33.4	26.1	54.4	105.4	19.2	76.6	34.0	26.5	54.9	111.1	20.2
Ours: Up-Down	**77.2**	**36.2**	**27.0**	**56.4**	**113.5**	**20.3**	**79.8**	**36.3**	**27.7**	**56.9**	**120.1**	**21.4**
Relative Improvement	4%	8%	3%	4%	8%	6%	4%	7%	5%	4%	8%	6%

Similar Trend for VQA, Image/Text Retrieval, and Text-to-image generation

- Global vector representation and simple fusion
  - VQA: Visual Question Answering [ICCV 2015]
  - Unifying visual-semantic embeddings with multimodal neural language models [NeurIPS 2014 DL workshop]
  - Generative adversarial text-to-image synthesis [ICML 2016], StackGAN [ICCV2017]
- Grid feature representation and cross-modal attention
  - Stacked Attention Networks [CVPR 2016]
  - Instance-aware Image and Sentence Matching with Selective Multimodal LSTM [CVPR 2017]
  - AttnGAN [CVPR 2018]
- Object-centric feature representation and BUTD attention:
  - Bottom-up and Top-down Attention [CVPR 2018]
  - Stacked Cross Attention for Image-Text Matching [ECCV 2018]
  - ObjGAN [CVPR 2019]

Zhang, Chao, et al. "Multimodal intelligence: Representation learning, information fusion, and applications." *IEEE Journal of Selected Topics in Signal Processing* 14.3 (2020): 478-493.