Image Captioning

Anjali Narayan-Chen,
Liaonan Xie, Ruihan Shan

How would you describe this image?

1. A group of children playing on the street , some on bicycles , and one on a scooter , and one in a blue shirt walking .
2. A group of four children cross the street , as an adult looks down the street .
3. Children playing with their bicycles and scooters .
4. Five children playing on a neighborhood street .
5. 3 girls and one boy playing in the street .

✖ My kids and their friends played near our house the other day, while I had to watch for oncoming cars.

P. Young et al., From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions, TACL 2014

What is image captioning?

- **Task:** automatically generate a relevant and accurate caption for an arbitrary image using properly formed English sentences
  - Capture the objects contained in an image
  - Express how objects relate to each other, as well as attributes and activities objects are involved in
  - Visual understanding + language modeling

What is image captioning?

- **Task:** automatically generate a relevant and accurate caption for an arbitrary image using properly formed English sentences
  - Capture the objects contained in an image
  - Express how objects relate to each other, as well as attributes and activities objects are involved in
  - Visual understanding + language modeling
- Challenges
  - Objects don’t come from closed set of labels
  - Objects may or may not be mentioned, depending on visual salience
  - Humans can describe images in a multitude of ways

What is image captioning?

- **Task:** automatically generate a relevant and accurate caption for an arbitrary image using properly formed English sentences
  - Capture the objects contained in an image
  - Express how objects relate to each other, as well as attributes and activities objects are involved in
  - Visual understanding + language modeling
- Challenges
  - Objects don’t come from closed set of labels
  - Objects may or may not be mentioned, depending on visual salience
  - Humans can describe images in a multitude of ways
- Deep neural networks: able to generate image descriptions without relying on hard-coded visual concepts and sentence templates

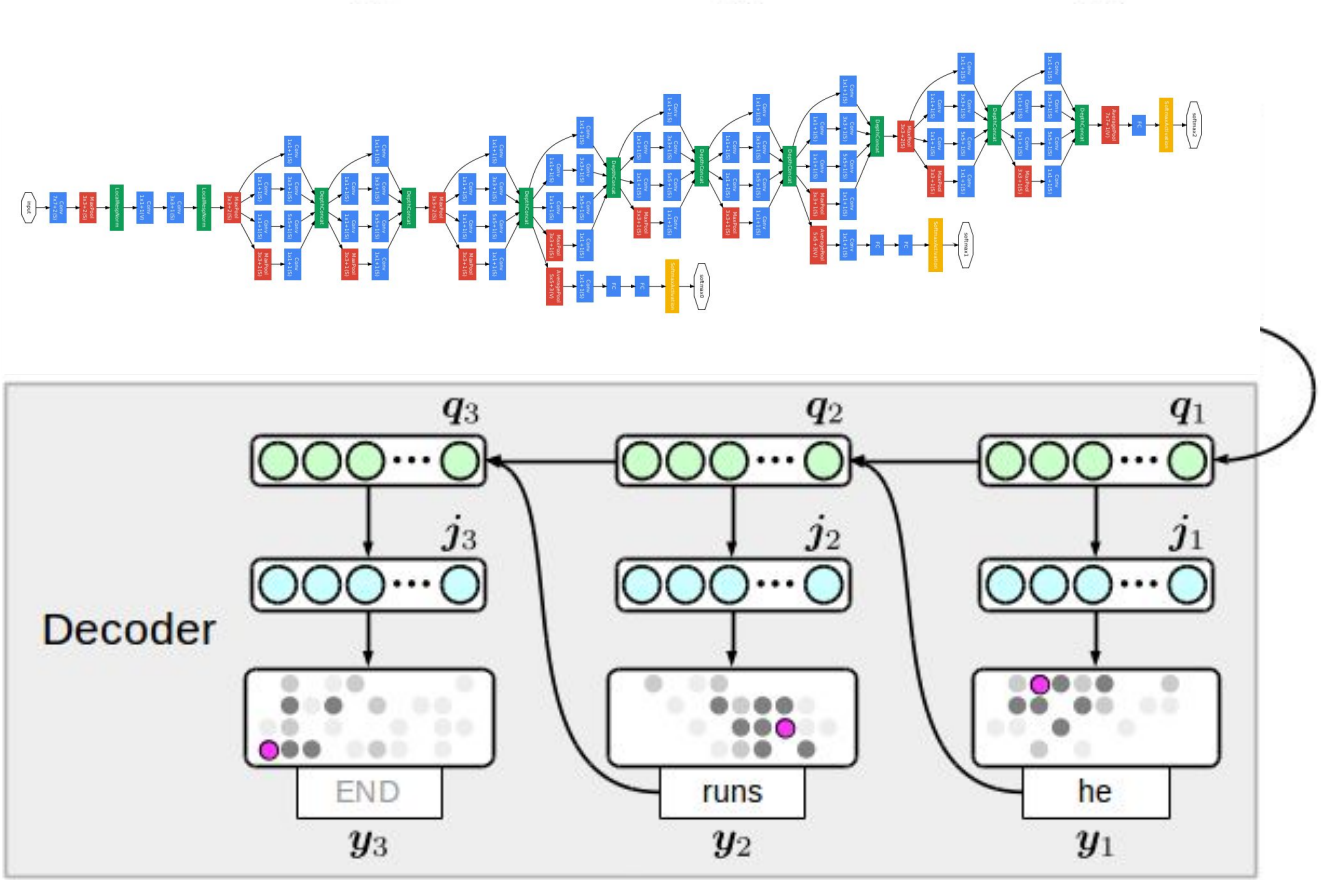
Datasets

- **PASCAL VOC-2008** (Rashtchian et al., 2010)
  - 1K images, each annotated with 5 captions
  - Images randomly selected from PASCAL 2008 object recognition challenge
  - Relatively simple: many pictures do not depict people; many verb-less (or static verb) captions
- **Flickr8K** (Hodosh et al., 2013)
  - 8K images, each annotated with 5 captions
  - Images of people and animals (mostly dogs) doing things
- **Flickr30K** (Young et al., 2014)
  - 30K images, each annotated with 5 captions
  - Extension of Flickr8K
- **MS COCO** (Lin et al., 2014)
  - 300K+ images, 5 captions per image

Outline

1. Basic CNN + RNN architectures: NeuralTalk, Show and Tell
2. Discussion of automated evaluation metrics
3. Alternative approaches and architectures
    a. From Captions to Visual Concepts and Back
    b. Show, Attend, and Tell
    c. Towards Diverse and Natural Image Descriptions via a Conditional GAN
    d. Captioning Images with Diverse Objects

Basic CNN + RNN Architectures

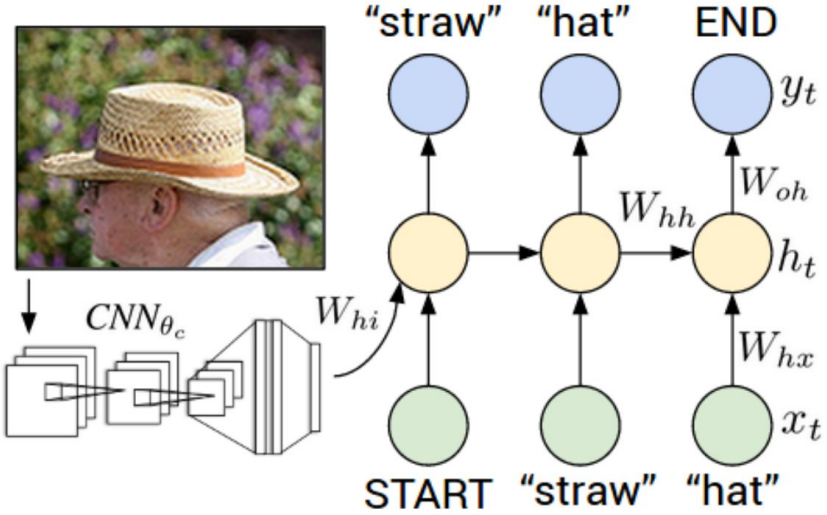


Encoder-decoder: http://qiita.com/odashi_t/items/a1be7c4964fbea6a116e
GoogLeNet: C. Szegedy et al., Going Deeper with Convolutions, CVPR 2015

Deep Visual-Semantic Alignments for Generating Image Descriptions (NeuralTalk)

(Karpathy & Li, 2015)

- CNN features: last layer of VGGNet (Simonyan & Zisserman, 2014)
- Image information fed in as bias term b_v at the first step
- Trained using MLE objective
- Inference
  - *Sampling*: sample words repeatedly until end-of-sentence token is sampled
  - *Beam search*: iteratively consider the set of *k* best sentences up to time *t* as candidates to generate sentences of size *t*+1, and keep only the resulting best *k* of them

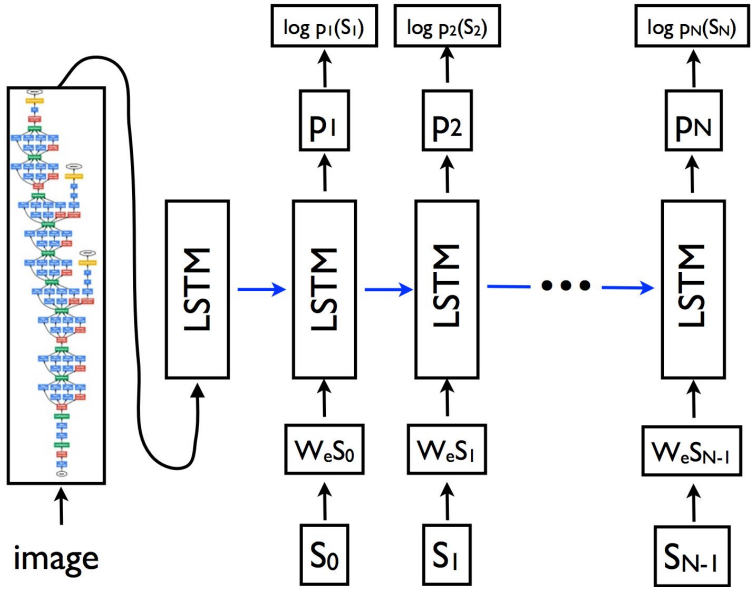


$$b_v = W_{hi}[CNN_{\theta_c}(I)]$$
$$h_t = f(W_{hx}x_t + W_{hh}h_{t-1} + b_h + \mathbb{1}(t=1) \odot b_v)$$
$$y_t = softmax(W_{oh}h_t + b_o).$$

Show and Tell: A Neural Image Caption Generator (NIC)

(Vinyals et al., 2015)

- Vision CNN + language LSTM
  - CNN: last layer of pretrained GoogLeNet (C. Szegedy et al, 2015)
  - Word embeddings W_e map one-hot word vectors to the same space as images
- Image is input once at t = -1 to inform LSTM about image contents
- Trained using MLE objective
- Inference as described previously



$$x_{-1} = \text{CNN}(I)$$
$$x_t = W_e S_t, \quad t \in \{0 \dots N-1\}$$
$$p_{t+1} = \text{LSTM}(x_t), \quad t \in \{0 \dots N-1\}$$

Vinyals et al., Show and Tell: A Neural Image Caption Generator, CVPR 2015

"man in black shirt is playing guitar."

"construction worker in orange safety vest is working on road."

"two young girls are playing with lego toy."

"boy is doing backflip on wakeboard."

"girl in pink dress is jumping in air."

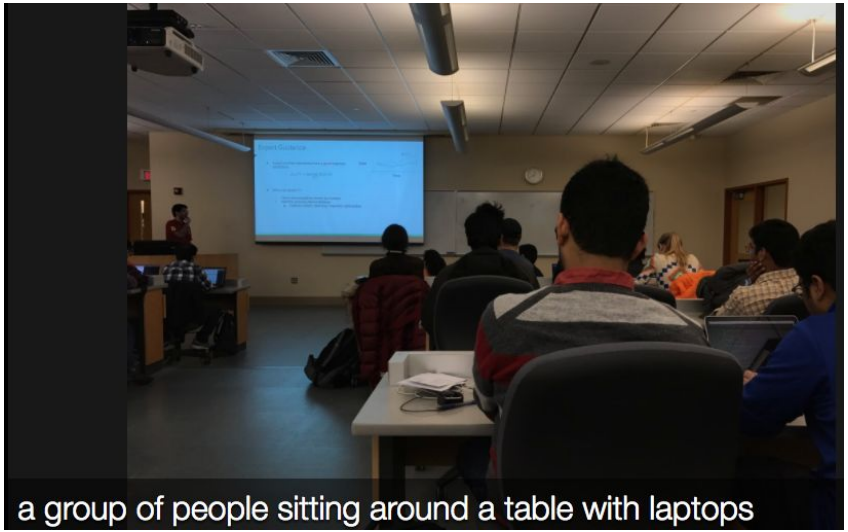
"black and white dog jumps over bar."

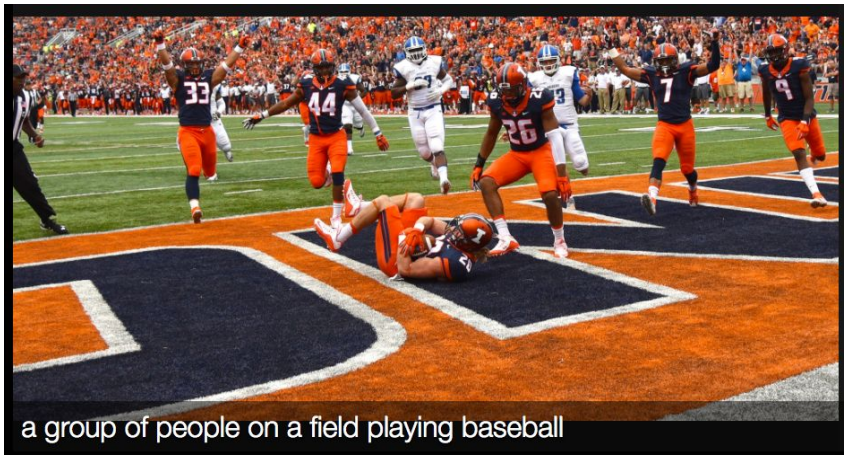
"young girl in pink shirt is swinging on swing."

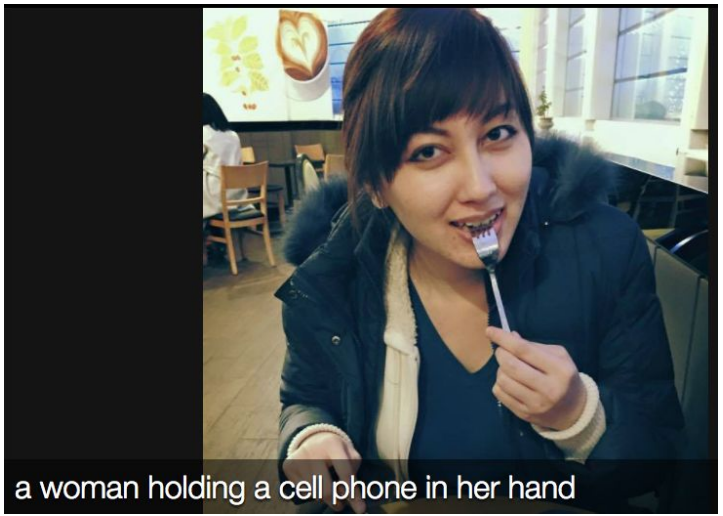
"man in blue wetsuit is surfing on wave."

More examples can be found at http://cs.stanford.edu/people/karpathy/deepimagesent/

NeuralTalk and Walk Demo







Generated using NeuralTalk2

	Flickr8K				Flickr30K				MSCOCO 2014					
Model	B-1	B-2	B-3	B-4	B-1	B-2	B-3	B-4	B-1	B-2	B-3	B-4	METEOR	CIDEr
Nearest Neighbor	—	—	—	—	—	—	—	—	48.0	28.1	16.6	10.0	15.7	38.3
Mao et al. [38]	58	28	23	—	55	24	20	—	—	—	—	—	—	—
Google NIC [54]	63	41	27	—	66.3	42.3	27.7	18.3	66.6	46.1	32.9	24.6	—	—
LRCN [8]	—	—	—	—	58.8	39.1	25.1	16.5	62.8	44.2	30.4	—	—	—
MS Research [12]	—	—	—	—	—	—	—	—	—	—	—	21.1	20.7	—
Chen and Zitnick [5]	—	—	—	14.1	—	—	—	12.6	—	—	—	19.0	20.4	—
Our model	57.9	38.3	24.5	16.0	57.3	36.9	24.0	15.7	62.5	45.0	32.1	23.0	19.5	66.0

Karpathy & Li, Deep Visual-Semantic Alignments for Generating Image Descriptions, CVPR 2015

Technique	BLEU-4 Improvement
Better Image Model [24]	2
Beam Size Reduction	2
Fine-tuning Image Model	1
Scheduled Sampling [48]	1.5
Ensembles	1.5

+ batch norm

indicating overfitting

after training LSTM

Vinyals et al., Show and Tell: Lessons learned from the 2015 MSCOCO Image Captioning Challenge, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2016

Evaluation Metrics for Generated Image Descriptions

A person riding a motorcycle on a dirt road.

Two dogs play in the grass.

A skateboarder does a trick on a ramp.

A dog is jumping to catch a frisbee.

A group of young people playing a game of frisbee.

Two hockey players are fighting over the puck.

A little girl in a pink hat is blowing bubbles.

A refrigerator filled with lots of food and drinks.

A herd of elephants walking across a dry grass field.

A close up of a cat laying on a couch.

A red motorcycle parked on the side of the road.

A yellow school bus parked in a parking lot.

Describes without errors | Describes with minor errors | Somewhat related to the image | Unrelated to the image

Vinyals et al., Show and Tell: A Neural Image Caption Generator, CVPR 2015

1. A woman in a green shirt is getting food ready with a child , while sitting on rocks .
2. A mother and child having a picnic on a big rock with blue utensils .
3. A woman serving food for a little boy outside on a large rock .
4. A woman and a baby eating (having a picnic) .
5. A mother and child picnic on some rocks .

P. Young et al., From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions, TACL 2014

BLEU (BiLingual Evaluation Understudy)

(Papineni et al., 2002)

- *"The closer a machine translation is to a professional human translation, the better it is."*

Papineni et al., BLEU: a Method for Automatic Evaluation of Machine Translation, ACL 2002

BLEU (BiLingual Evaluation Understudy)

(Papineni et al., 2002)

- "*The closer a machine translation is to a professional human translation, the better it is.*"
- Analyzes co-occurrences of *n*-grams between candidate and reference sentences
  - Modified (clipped) *n*-gram precision
  - Brevity penalty to penalize short candidate sentences

Papineni et al., BLEU: a Method for Automatic Evaluation of Machine Translation, ACL 2002

BLEU (BiLingual Evaluation Understudy)

(Papineni et al., 2002)

- "*The closer a machine translation is to a professional human translation, the better it is.*"
- Analyzes co-occurrences of *n*-grams between candidate and reference sentences
  - Modified (clipped) *n*-gram precision
  - Brevity penalty to penalize short candidate sentences
- Has been shown in MT literature to be an insufficient metric (Callison-Burch et al., 2006)
  - Many large variations of a generated sentence can score identically
  - Higher BLEU score is not necessarily indicative of higher human-judged quality

Papineni et al., BLEU: a Method for Automatic Evaluation of Machine Translation, ACL 2002

BLEU (BiLingual Evaluation Understudy)

(Papineni et al., 2002)

- *"The closer a machine translation is to a professional human translation, the better it is."*
- Analyzes co-occurrences of *n*-grams between candidate and reference sentences
  - Modified (clipped) *n*-gram precision
  - Brevity penalty to penalize short candidate sentences
- Has been shown in MT literature to be an insufficient metric (Callison-Burch et al., 2006)
  - Many large variations of a generated sentence can score identically
  - Higher BLEU score is not necessarily indicative of higher human-judged quality

Candidate: the the the the the the the.

Reference 1: The cat is on the mat.

Reference 2: There is a cat on the mat.

Modified Unigram Precision $= 2/7$.

Papineni et al., BLEU: a Method for Automatic Evaluation of Machine Translation, ACL 2002

Reference captions:

1. Latino man holding sign on the sidewalk outside promoting Quiznos-Subs .
2. A man is holding an advertisement for Quiznos Subs .
3. A man is holding a Quiznos sign next to a street .
4. A man is holding a Quiznos Sub sign .

Candidate caption:

❓ Quiznos worker wearing sign .

BLEU-4 = 0.106

P. Young et al., From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions, TACL 2014

METEOR

(Banerjee & Lavie, 2005)

More flexible MT metric that calculates sentence-level similarity scores as a harmonic mean of unigram precision & recall, based on:

- Exact token matching
- Stemmed tokens
- WordNet synonyms
- Paraphrases

SYSTEM	Jim went home	SYSTEM	Jim walks home
REFERENCE	Joe goes home	REFERENCE	Joe goes home

Examples from Statistical Machine Translation slides
Banerjee & Lavie, METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments, ACL 2005

Comparing Automatic Evaluation Measures for Image Description
(Elliott & Keller, 2014)

On Flickr8K, all measures are either *weakly* or *moderately* correlated with human judgments

	Flickr 8K co-efficient $n = 17,466$
METEOR	0.524
ROUGE SU-4	0.435
Smoothed BLEU	0.429
Unigram BLEU	0.345
TER	-0.279

Elliott & Keller, Comparing Automatic Evaluation Measures for Image Description, ACL 2014

CIDEr: Consensus-based Image Description Evaluation

(Vedantam et al., 2015)

- *"Does a caption describe an image as most people tend to describe it?"*

Vedantam et al., CIDEr: Consensus-based Image Description Evaluation, CVPR 2015

CIDEr: Consensus-based Image Description Evaluation

(Vedantam et al., 2015)

- "*Does a caption describe an image as most people tend to describe it?*"
- Automatically evaluate for image I_i how well a candidate sentence c_i matches the **consensus** of a set of image descriptions $S_i = \{s_{i1}, \ldots, s_{im}\}$

Vedantam et al., CIDEr: Consensus-based Image Description Evaluation, CVPR 2015

CIDEr: Consensus-based Image Description Evaluation

(Vedantam et al., 2015)

- "*Does a caption describe an image as most people tend to describe it?*"
- Automatically evaluate for image I_i how well a candidate sentence c_i matches the **consensus** of a set of image descriptions $S_i = \{s_{i1}, \ldots, s_{im}\}$
- Intuitively, a measure of consensus should:
  - Encode how often *n*-grams in the candidate sentence are present in the reference sentences
  - *n*-grams not present in the reference sentences should not be in the candidate sentence
  - *n*-grams that commonly occur across all images in the dataset should be given lower weight, since they are likely to be less informative

Vedantam et al., CIDEr: Consensus-based Image Description Evaluation, CVPR 2015

CIDEr: Consensus-based Image Description Evaluation

(Vedantam et al., 2015)

- "*Does a caption describe an image as most people tend to describe it?*"
- Automatically evaluate for image I_i how well a candidate sentence c_i matches the **consensus** of a set of image descriptions $S_i = \{s_{i1}, \ldots, s_{im}\}$
- Intuitively, a measure of consensus should:
  - Encode how often *n*-grams in the candidate sentence are present in the reference sentences
  - *n*-grams not present in the reference sentences should not be in the candidate sentence
  - *n*-grams that commonly occur across all images in the dataset should be given lower weight, since they are likely to be less informative
- In practice: perform a **Term Frequency Inverse Document Frequency (TF-IDF)** (Robertson, 2004) weighting for each *n*-gram

Vedantam et al., CIDEr: Consensus-based Image Description Evaluation, CVPR 2015

	CIDEr-D	Meteor	ROUGE-L	BLEU-1	BLEU-2	BLEU-3	BLEU-4	date
Watson Multimodal[46]	1.123	0.268	0.559	0.773	0.609	0.461	0.344	2016-11-16
MSM@MSRA[29]	1.049	0.266	0.552	0.751	0.588	0.449	0.343	2016-10-25
G-RMI(PG-SPIDEr-TAG)[17]	1.042	0.255	0.551	0.751	0.591	0.445	0.331	2016-11-11
MetaMind/VT_GT[25]	1.042	0.264	0.55	0.748	0.584	0.444	0.336	2016-12-01
ATT-IMG (MSM@MSRA)[5]	1.023	0.262	0.551	0.752	0.59	0.449	0.34	2016-06-13
G-RMI (PG-BCMR)[16]	1.013	0.257	0.55	0.754	0.591	0.445	0.332	2016-10-30
DONOT_FAIL_AGAIN[13]	1.01	0.262	0.542	0.734	0.564	0.425	0.32	2016-11-22
DLTC@MSR[12]	1.003	0.257	0.543	0.74	0.575	0.436	0.331	2016-09-04
Postech_CV[38]	0.987	0.255	0.539	0.743	0.575	0.431	0.321	2016-06-13
feng[15]	0.986	0.255	0.54	0.743	0.578	0.434	0.323	2016-11-06
			...					
Human[21]	0.854	0.252	0.484	0.663	0.469	0.321	0.217	2015-03-23

According to CIDEr, humans are in 38th place!! 😱

MSCOCO challenge leaderboard

SPICE: Semantic Propositional Image Caption Evaluation

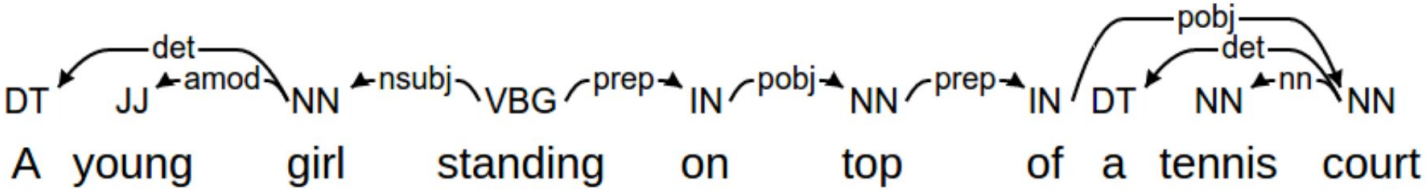
(Anderson et al., 2016)

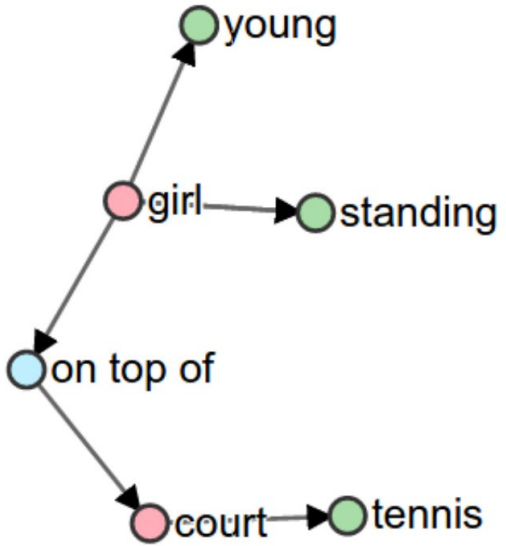
Previous metrics are primarily sensitive to *n*-gram overlap:

(a) A young girl *standing on top of a* tennis court.
(b) A giraffe *standing on top of a* green field.

(c) A shiny metal pot filled with some diced veggies.
(d) The pan on the stove has chopped vegetables in it.

Instead, **semantic propositional content** should be an important component of evaluation





{ (girl), (court), (girl, young), (girl, standing)
(court, tennis), (girl, on-top-of, court) }

Anderson et al., SPICE: Semantic Propositional Image Caption Evaluation, ECCV 2016

	M1	M2	M3	M4	M5
Metric	ρ	ρ	ρ	ρ	ρ
BLEU-1	0.24	0.29	0.72	-0.54	0.44
BLEU-4	0.05	0.1	0.58	-0.63	0.3
METEOR	0.53	0.57	0.86	-0.1	0.74
CIDEr	0.53	0.47	0.81	-0.21	0.65
SPICE	**0.88**	**0.89**	**0.89**	**0.46**	**0.97**

M1	Percentage of captions 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 detail of the captions from 1–5 (lacking details - very detailed).
M5	Percentage of captions that are similar to human description.

Anderson et al., SPICE: Semantic Propositional Image Caption Evaluation, ECCV 2016

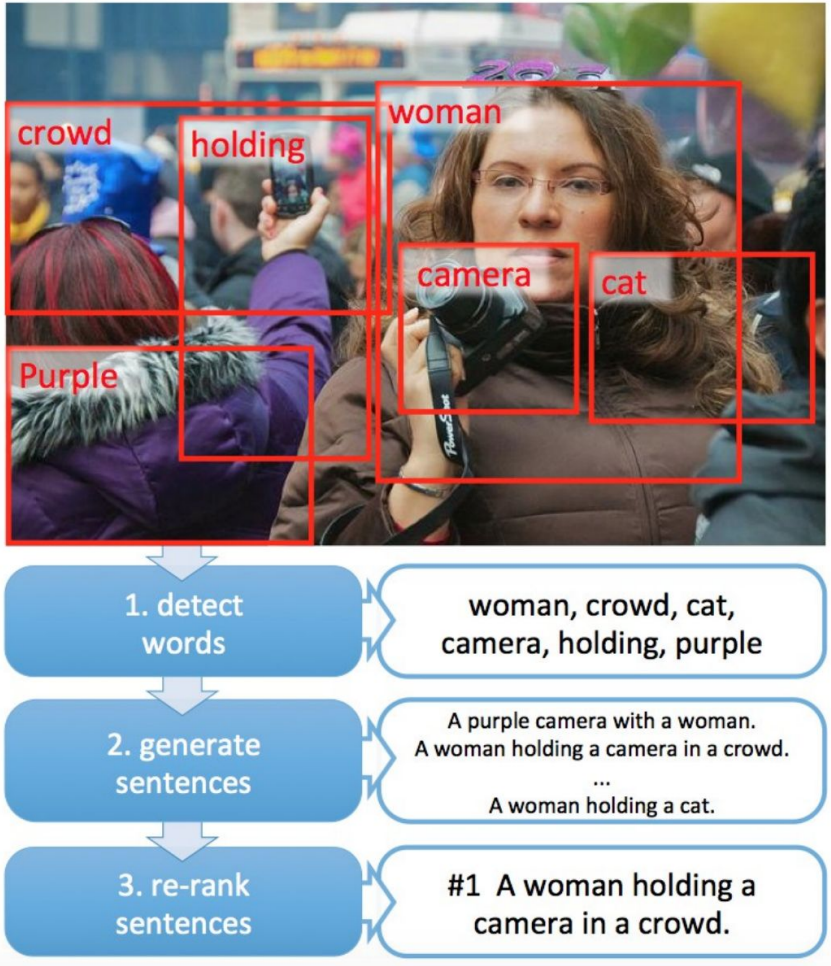
Takeaways

- How to automatically evaluate generated image descriptions such that quantitative results agree with human judgments is still an open problem
- Most popular approaches today are reliant on *n*-grams
- Success “by the numbers” of automatic captioning systems should be taken with a grain of salt
- That said, we can move on to...

Alternative Approaches and Architectures

From Captions to Visual Concepts and Back

(Fang et al., 2015)



Fang et al., From Captions to Visual Concepts and Back, CVPR 2015

Word Detection

Advantage

- Some abstract concepts (e.g., “beautiful”) may be highly correlated to specific visual patterns

Method

- Image sub-regions rather than the full image
- Featurize each region using rich convolutional neural net (CNN)
- Map the features of each region to words in the caption

✖ How do we learn the mappings without ground truth labels?

Fang et al., From Captions to Visual Concepts and Back, CVPR 2015

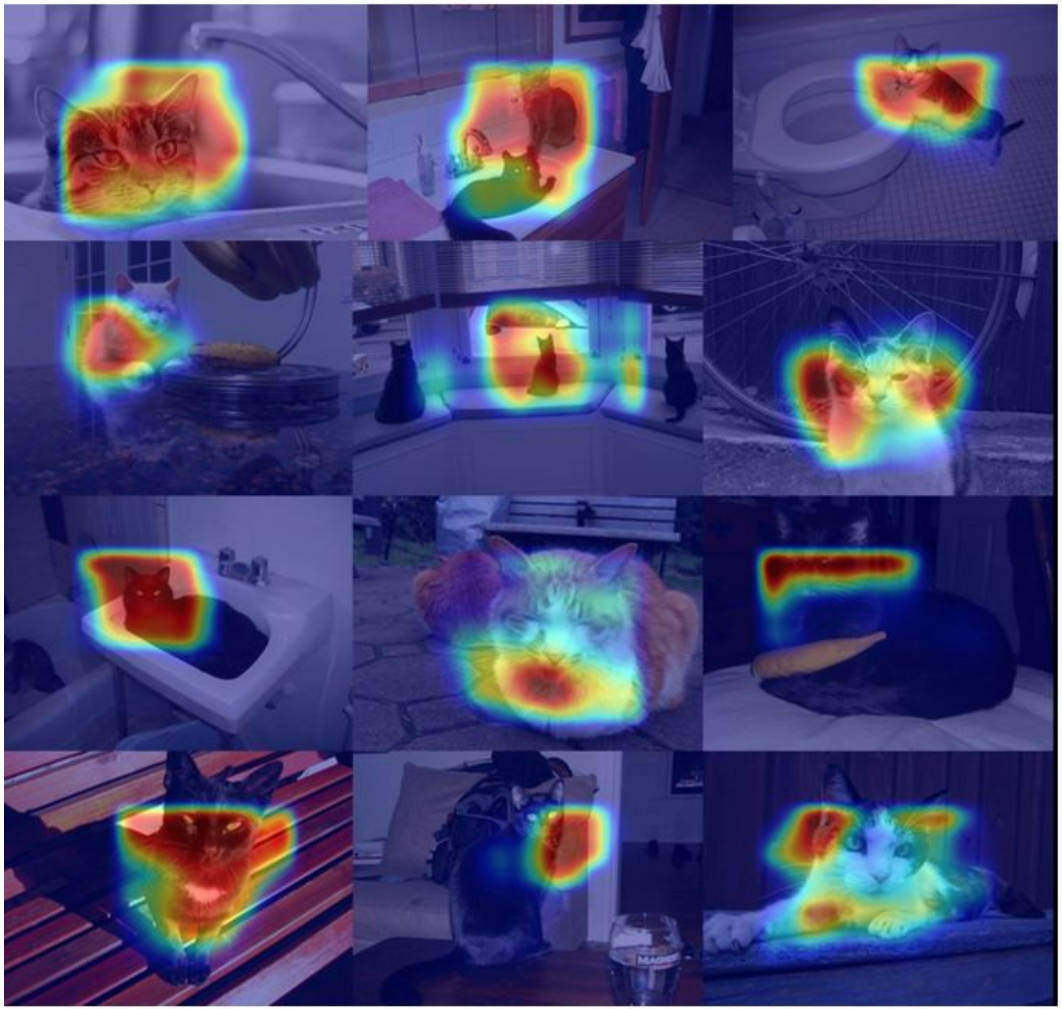
Multiple Instance Learning (MIL)

Multiple Instance Learning is a weakly supervised approach, where:

- Learner receives a set of labeled bags, each containing many instances
- Each image is described as a bag $X = \{X_1, \ldots, X_N\}$
  where each X_i is a bounding box instance of a sub-region in the image
- For each word $w \in \mathcal{V}$, MIL takes a set of "positive" and "negative" bags
  - A bag X is positive if word w is in image X's description

Iteratively select instances within positive bags
Retrain the detector using the updated positive label
Used to obtain mapping of regions to most likely words

Fang et al., From Captions to Visual Concepts and Back, CVPR 2015



Fang et al., From Captions to Visual Concepts and Back, CVPR 2015

Sentence Generation

Given an image with visually detected words, now we want to generate sentences

Statistical model:

Maximum Entropy Language Model conditioned on the set of visually detected words

$$L(\Lambda) = \sum_{s=1}^{S} \sum_{l=1}^{\#(s)} \log \Pr(\bar{w}_l^{(s)} | \bar{w}_{l-1}^{(s)}, \cdots, \bar{w}_1^{(s)}, <s>, \tilde{\mathcal{V}}_{l-1}^{(s)})$$

Sentences in the dataset

Sentence length

Set of words with high likelihood detections that have yet to be mentioned in the sentence

Fang et al., From Captions to Visual Concepts and Back, CVPR 2015

Sentence Generation

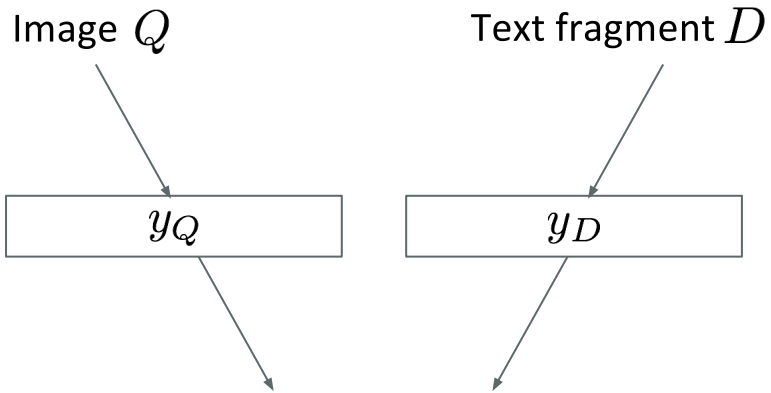
Generation process: left-to-right **beam search**

- Maintain a stack of *L* partial hypotheses
- Extend every path with a set of likely words; resulting length *L*+1 paths are stored
- Retain the top *k* paths of length *L*+1; others are pruned away

Then, form an *M*-best list of complete sentences ranked by log likelihood that cover at least *T* image objects

Fang et al., From Captions to Visual Concepts and Back, CVPR 2015

Multimodal Similarity



Deep multimodal similarity model
maps images and text
to the same semantic space

Relevance *R*(*Q*,*D*) = cosine similarity

Fang et al., From Captions to Visual Concepts and Back, CVPR 2015

Sentence Reranking

Given the *M*-best list of complete sentences, we want to re-rank them based on global similarities between image and text

The posterior probability of the text being relevant to the image is defined as:

$$P(D|Q) = \frac{\exp(\gamma R(Q, D))}{\Sigma_{D' \in \mathbb{D}} \exp(\gamma R(Q, D'))}$$

All possible candidate captions

Smoothing factor

Relevance between image and text

Fang et al., From Captions to Visual Concepts and Back, CVPR 2015

a pot of broccoli on a stove

a wok with a cooked broccoli meal in it

a herd of cattle standing on top of a grass covered field

several cows are gathered together in a grassy field

a yellow sign on a dirt road

a floodway sign sitting on the side of a road in a field

a group of people posing for a picture on a ski lift

three people wearing ski gear sitting on a ski lift

a man sitting on a couch with a dog

a man sitting on a chair with a dog in his lap

a baseball player throwing a ball

a pitcher holds his arm far behind him during a pitch

Fang et al., From Captions to Visual Concepts and Back, CVPR 2015

	Flickr8K				Flickr30K				MSCOCO 2014					
Model	B-1	B-2	B-3	B-4	B-1	B-2	B-3	B-4	B-1	B-2	B-3	B-4	METEOR	CIDEr
Nearest Neighbor	—	—	—	—	—	—	—	—	48.0	28.1	16.6	10.0	15.7	38.3
Mao et al. [38]	58	28	23	—	55	24	20	—	—	—	—	—	—	—
Google NIC [54]	63	41	27	—	66.3	42.3	27.7	18.3	66.6	46.1	32.9	24.6	—	—
LRCN [8]	—	—	—	—	58.8	39.1	25.1	16.5	62.8	44.2	30.4	—	—	—
MS Research [12]	—	—	—	—	—	—	—	—	—	—	—	21.1	20.7	—
Chen and Zitnick [5]	—	—	—	14.1	—	—	—	12.6	—	—	—	19.0	20.4	—
Our model	57.9	38.3	24.5	16.0	57.3	36.9	24.0	15.7	62.5	45.0	32.1	23.0	19.5	66.0

Karpathy & Li, Deep Visual-Semantic Alignments for Generating Image Descriptions, CVPR 2015

Show, Attend, and Tell

(Xu et al., 2015)

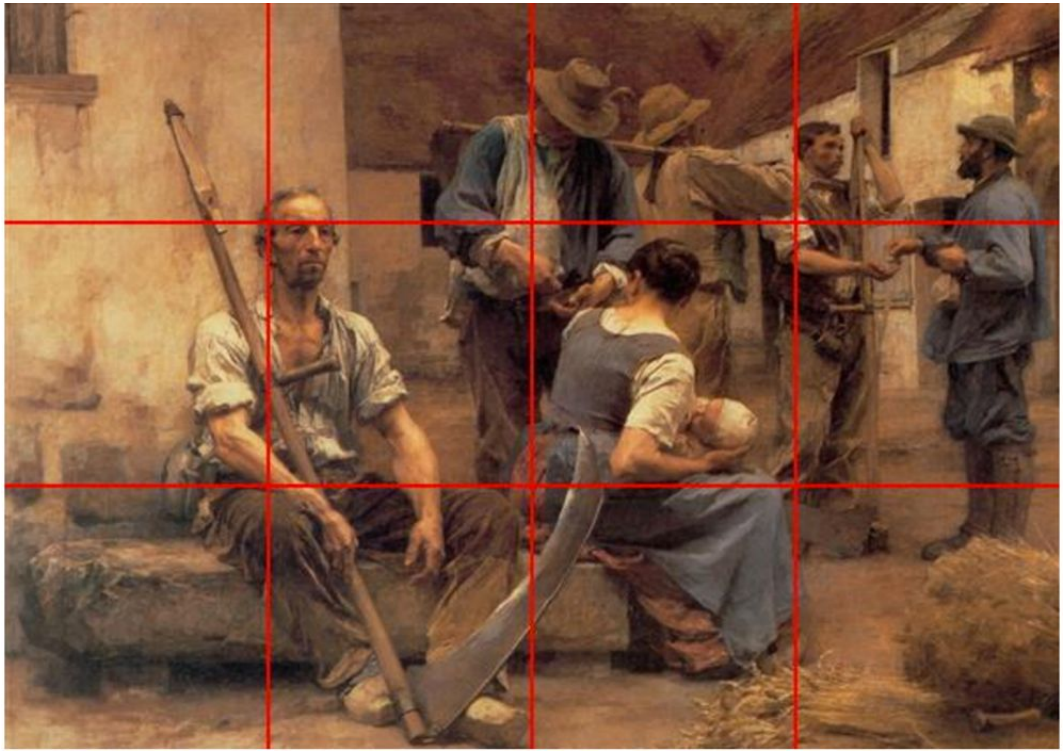
Instead of learning word detectors over image regions, consider learning an **attention model** instead

- What is visual attention?
- How to augment Show and Tell with visual attention
- Soft vs. hard attention

Xu et al., Show, Attend, and Tell: Neural Image Caption Generation with Visual Attention, ICML 2015

La Paye des moissonneurs (1882), Paris, musée d'Orsay

Intuition: how do humans analyze images?



Focus on certain salient aspect of the current image

Attend to different parts of image sequentially over time

La Paye des moissonneurs (1882), Paris, musée d'Orsay

Intuition: how do humans analyze images?

Focus on certain salient aspect of the current image

Attend to different parts of image sequentially over time

La Paye des moissonneurs (1882), Paris, musée d'Orsay

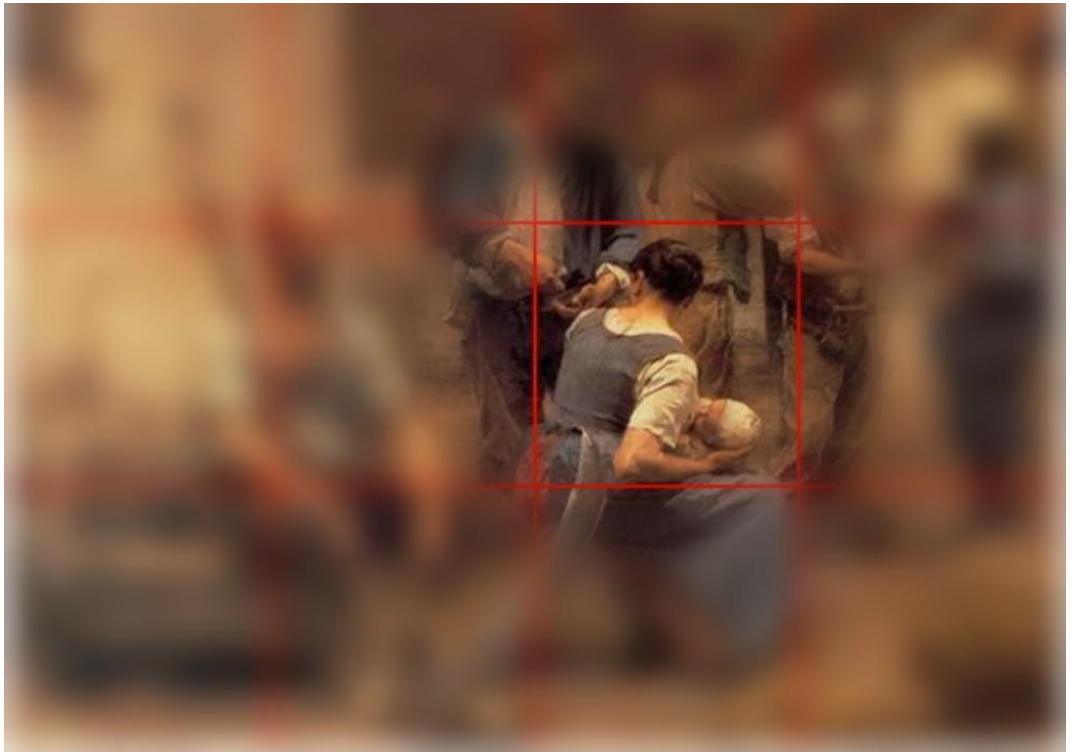
Intuition: how do humans analyze images?

Focus on certain salient aspect of the current image

Attend to different parts of image sequentially over time

La Paye des moissonneurs (1882), Paris, musée d'Orsay

Intuition: how do humans analyze images?

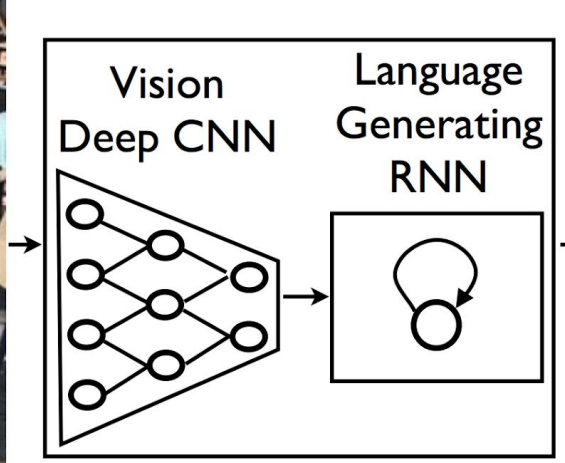


Focus on certain salient aspect of the current image

Attend to different parts of image sequentially over time

La Paye des moissonneurs (1882), Paris, musée d'Orsay

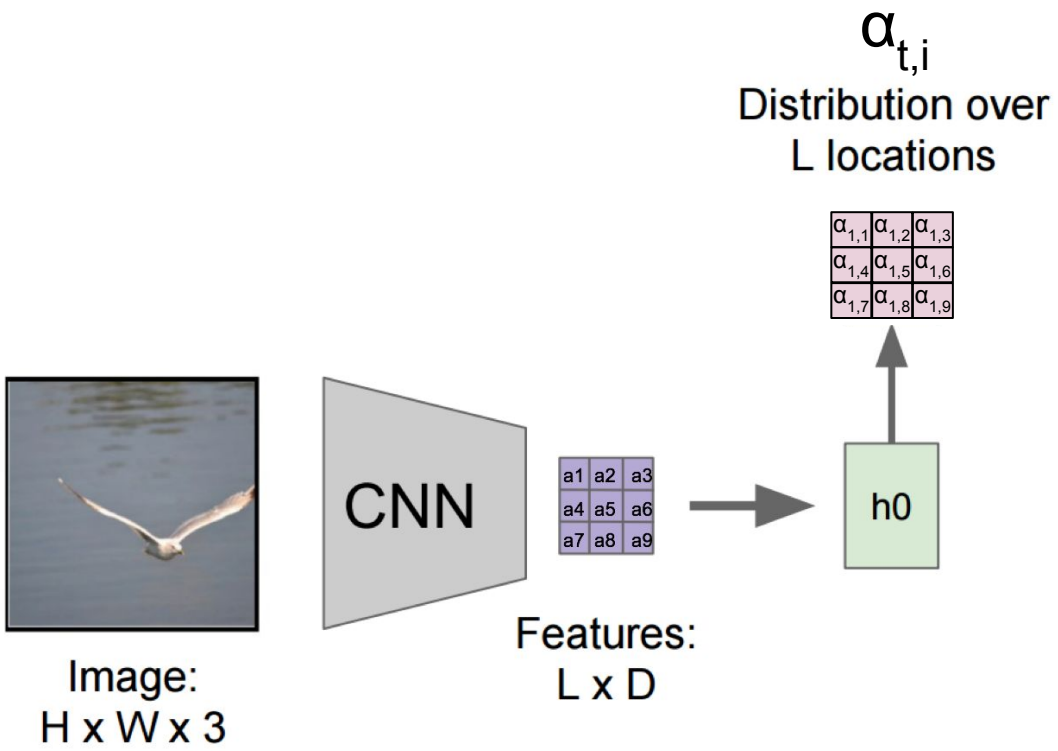
Limitations of Vanilla NIC



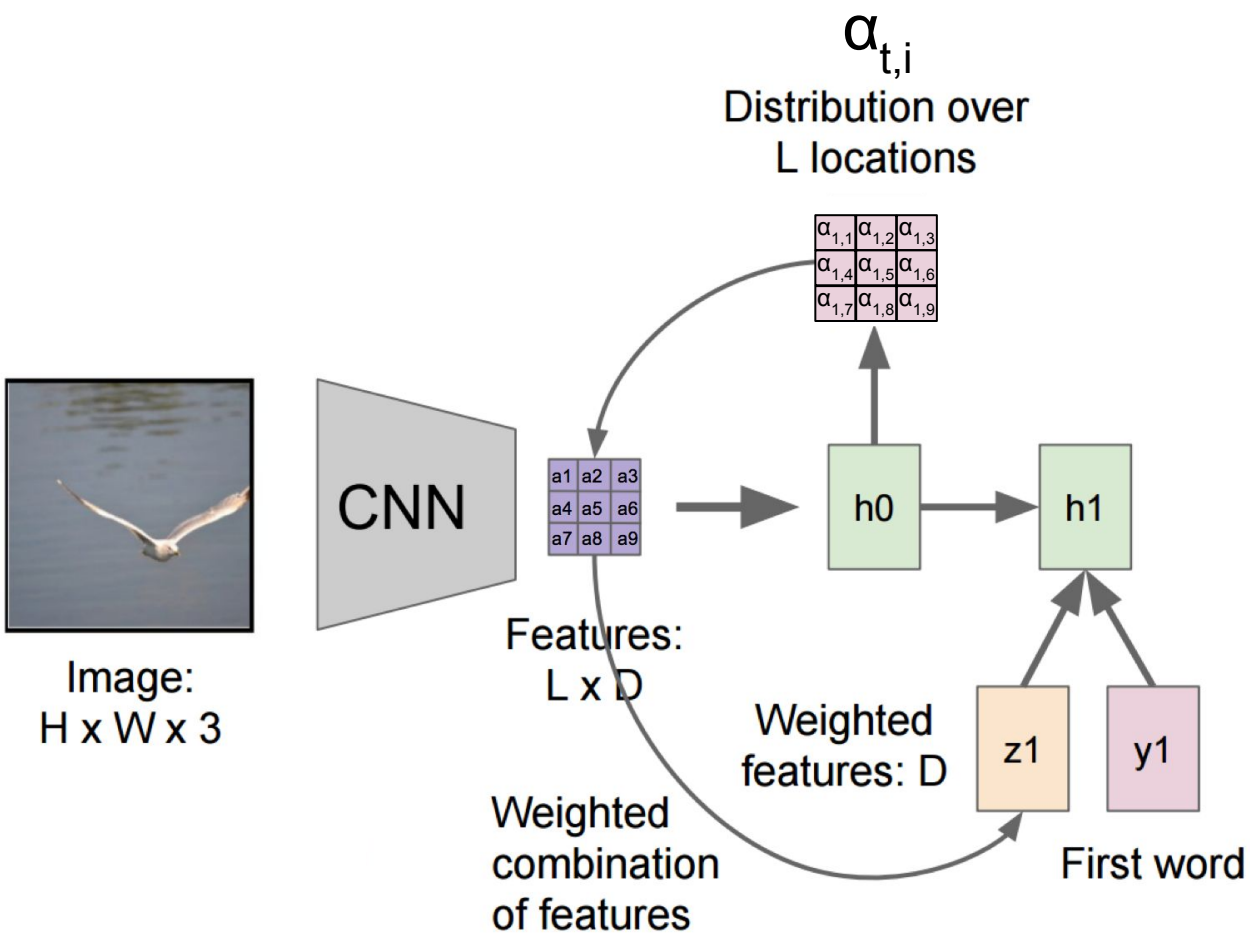
A group of people shopping at an outdoor market.

There are many vegetables at the fruit stand.

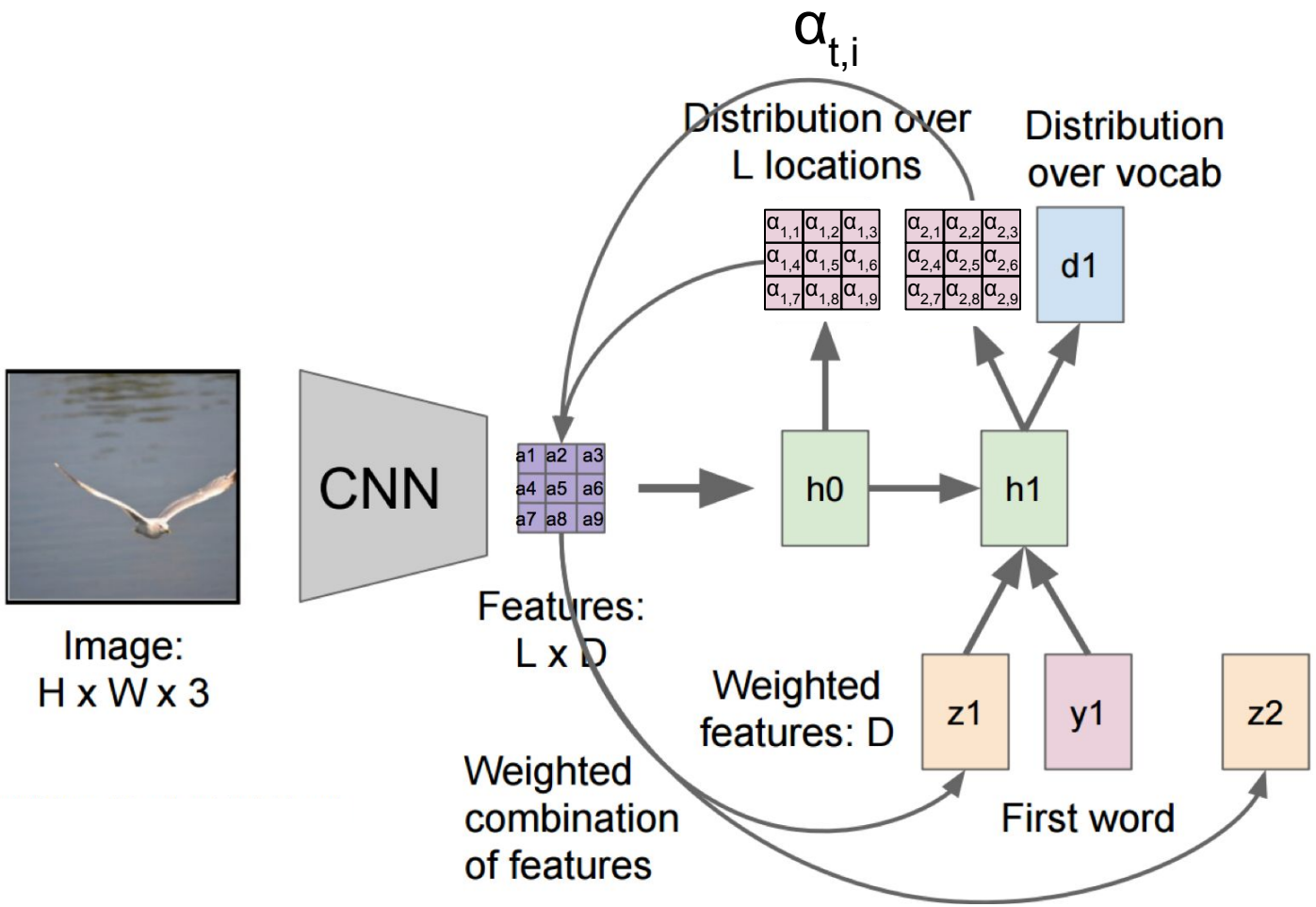
Vinyals et al., Show and Tell: A Neural Image Caption Generator, CVPR 2015



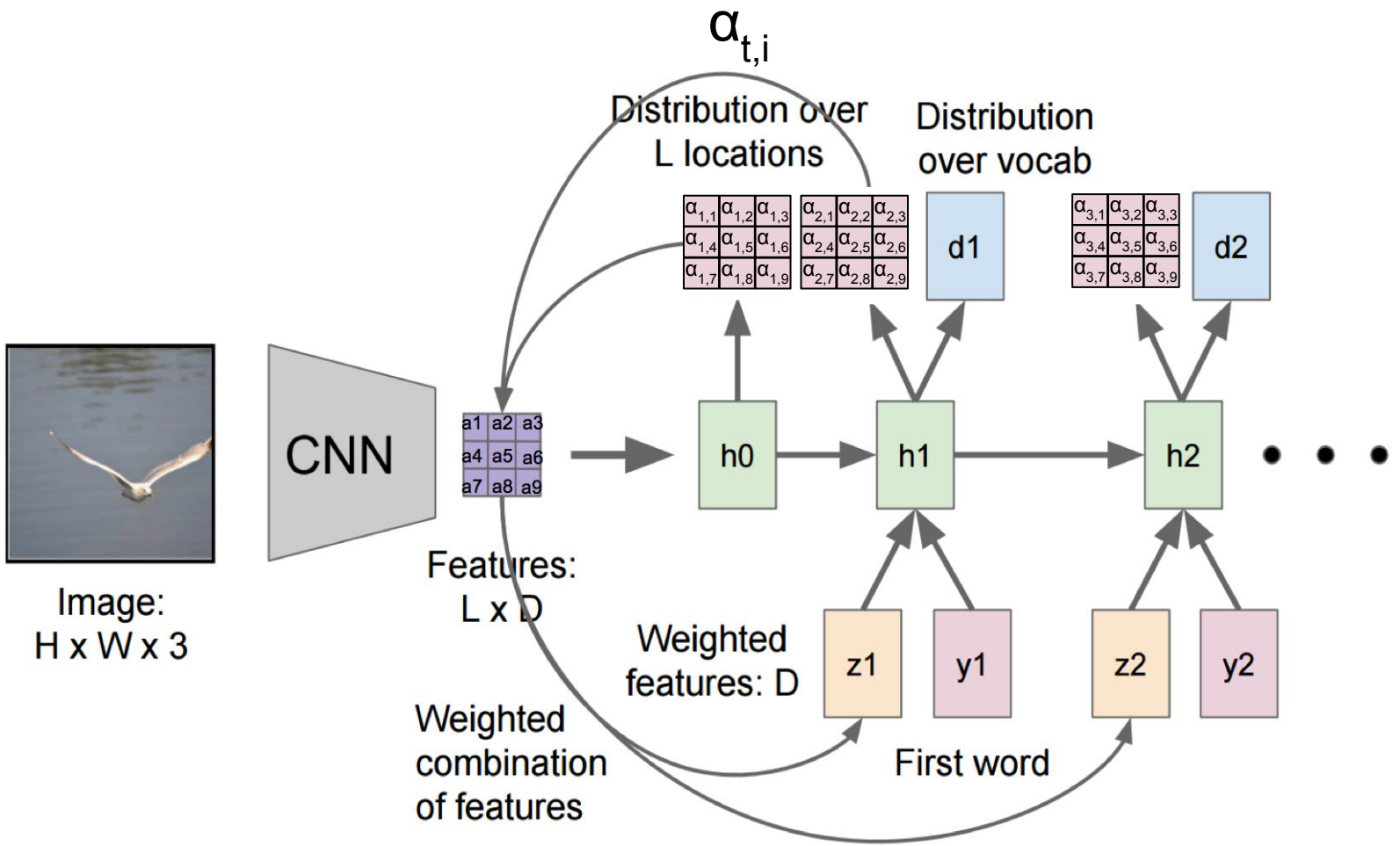
Slides from Stanford 231n
Xu et al., Show, Attend, and Tell: Neural Image Caption Generation with Visual Attention, ICML 2015



Slides from Stanford 231n
Xu et al., Show, Attend, and Tell: Neural Image Caption Generation with Visual Attention, ICML 2015



Slides from Stanford 231n
Xu et al., Show, Attend, and Tell: Neural Image Caption Generation with Visual Attention, ICML 2015



Slides from Stanford 231n
Xu et al., Show, Attend, and Tell: Neural Image Caption Generation with Visual Attention, ICML 2015

Show, Attend, and Tell

(Xu et al., 2015)

- Compute weights α_i of each annotation vector a_i by an attention model f_{att}
  - f_{att}: multilayer perceptron conditioned on h_{t-1}
  - α_i are used in different ways:
    - Deterministic (**soft** attention)
    - Stochastic (**hard** attention)
- Context vector $\mathbf{z}_t$: dynamic representation of the relevant part of the image input at time t
- Φ returns a single vector given the set of annotation vectors a_i and their corresponding weights α_i

$$z = \phi(\{a_i\}, \{\alpha_i\})$$

Xu et al., Show, Attend, and Tell: Neural Image Caption Generation with Visual Attention, ICML 2015

Soft Attention

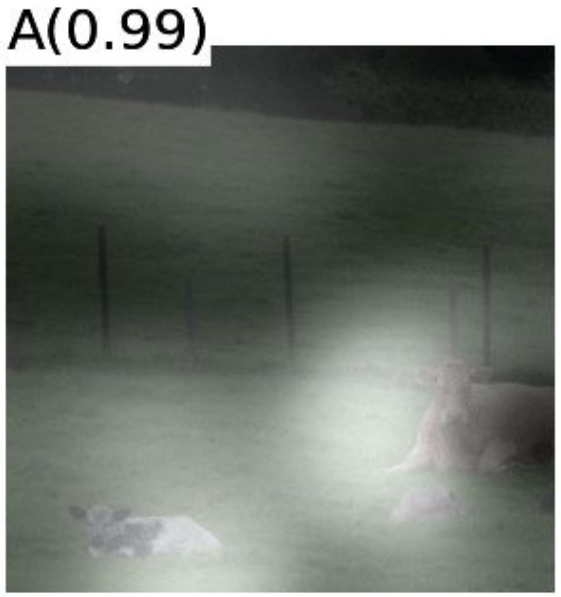
$\boldsymbol{z}_t$ is calculated by taking the weighted sum of all feature vectors a

$$z_t = \sum_{i=1}^{L} \alpha_t[i] \cdot a_i$$

- Differentiable
- Deterministic: α_i's assign relative importance to give to location i in blending the a_i's together
- Learned using standard backpropagation

Xu et al., Show, Attend, and Tell: Neural Image Caption Generation with Visual Attention, ICML 2015

Soft Attention: Examples



More examples at project website
Xu et al., Show, Attend, and Tell: Neural Image Caption Generation with Visual Attention, ICML 2015

Hard Attention

At time step *t*, the index into the feature vectors is sampled from the current location distribution vector α_t

$$k = sample(\alpha_t)$$

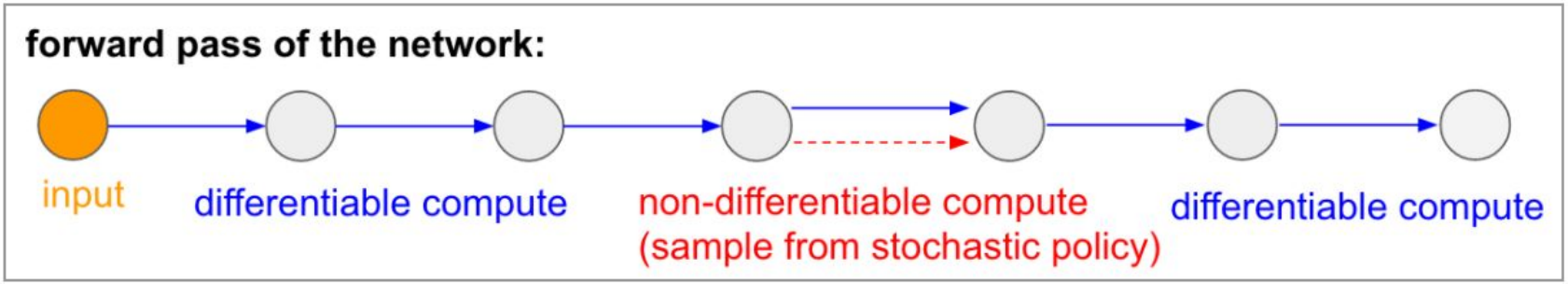
$$z_t = a_k$$

- Stochastic: α_i's assign probability that location *i* is the right place to focus for producing the next word
- Focuses on one image region at a time
- Non-differentiable due to sampling
  - How to train?

Xu et al., Show, Attend, and Tell: Neural Image Caption Generation with Visual Attention, ICML 2015

Hard Attention

- Non-differentiable due to sampling
  - How to train?



Picture credit to Andrej Karpathy's blog
Xu et al., Show, Attend, and Tell: Neural Image Caption Generation with Visual Attention, ICML 2015

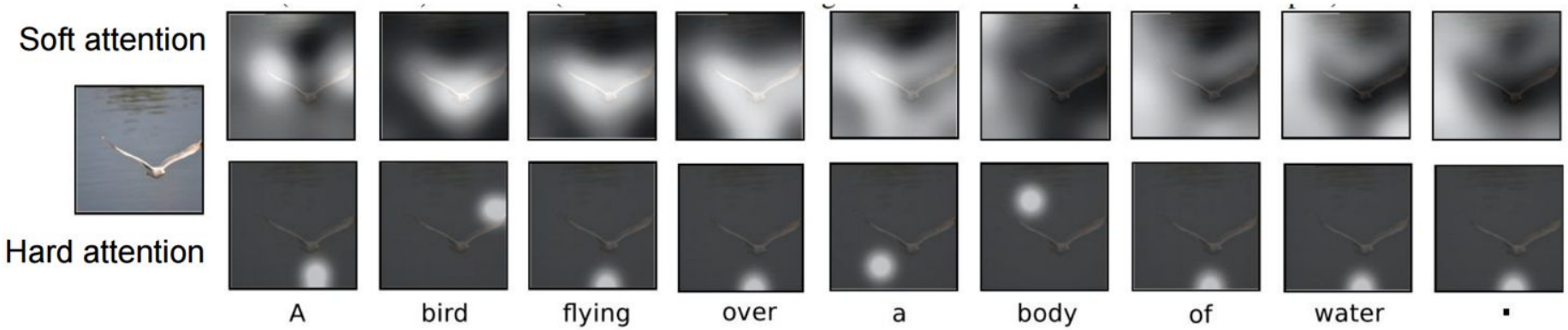
Hard Attention: Training

Set up as a **reinforcement learning problem**

- Action: choosing which area to attend to next
- Reward: caption quality proportional to the log likelihood of the target sentence

Xu et al., Show, Attend, and Tell: Neural Image Caption Generation with Visual Attention, ICML 2015

Soft vs. Hard Attention



Xu et al., Show, Attend, and Tell: Neural Image Caption Generation with Visual Attention, ICML 2015

Comparison

Dataset	Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
Flickr8k	Google NIC	63	41	27	-	-
	Soft-Attention	**67**	44.8	29.9	19.5	18.93
	Hard-Attention	**67**	**45.7**	**31.4**	**21.3**	**20.30**
Flickr30k	Google NIC	66.3	42.3	27.7	18.3	-
	Soft-Attention	66.7	43.4	28.8	19.1	**18.49**
	Hard-Attention	**66.9**	**43.9**	**29.6**	**19.9**	18.46
COCO	Google NIC	66.6	46.1	32.9	24.6	-
	Soft-Attention	70.7	49.2	34.4	24.3	**23.90**
	Hard-Attention	**71.8**	**50.4**	**35.7**	**25.0**	23.04

Xu et al., Show, Attend, and Tell: Neural Image Caption Generation with Visual Attention, ICML 2015

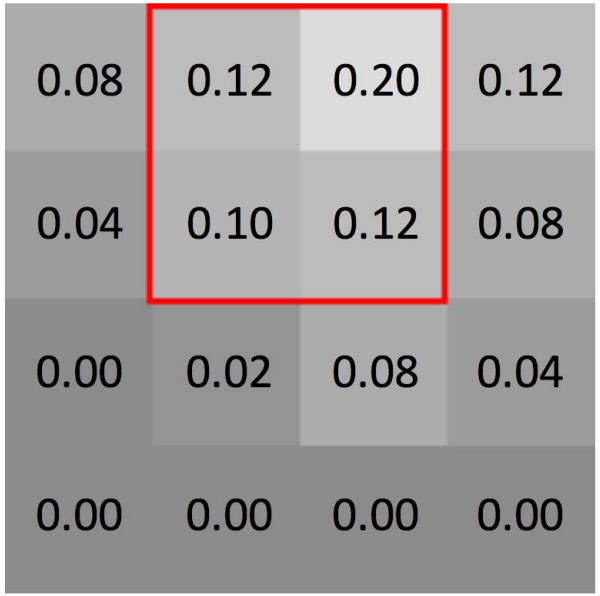
Attention Correctness in Neural Image Captioning

(Liu et al., 2016)

- To what extent are attention maps consistent with human perceptions?
- Will more human-like attention maps result in better captioning performance?

Liu, Chenxi, et al. Attention correctness in neural image captioning, submitted 2016.

Attention Correctness



Attention correctness:
the sum of the weights α
within the ground truth region

Liu, Chenxi, et al. Attention correctness in neural image captioning, submitted 2016.

Where to get ground truth for attention maps?

Flickr30k Entities (Plummer et al., 2015)

IMAGE 2433178831
The entity refers to people

SENTENCES

1. A smiling man with long hair and a beard is standing outside , wearing a teal jumpsuit , a hat , and protective ear-gear .
2. A man wearing a heave green work jacket and orange ear protectors is smiling at the camera .
3. Here is a picture of a man with long hair in a ponytail who works at this state park .
4. A long-haired man wearing a jumpsuit with headphones over his green cap .
5. A man wearing a green jacket and ear protection stands on the street .

ENTITIES

1 2 3 4 5 6 7 8 9 10 11 12

Show All Clear

More examples at Flickr30K Entities website
Plummer et al., Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models, ICCV 2015

Supervised Attention Model

Previous attention model was implicitly trained with respect to ground truth attention maps

- Negative log probability of the ground truth words in the sentence

Now, modify the loss function to take advantage of prior knowledge about the attention map

- Supervision from ground truth bounding boxes

Liu, Chenxi, et al. Attention correctness in neural image captioning, submitted 2016.

Do more human-like attention maps result in better captioning performance?

Girl rock climbing on the rock wall.

A young smiling child hold his toy alligator up to the camera.

Two male friends in swimming trunks jump on the beach while people in the background lay in the sand.

A black dog swims in water with a colorful ball in his mouth.

Caption	Model	Baseline	Correctness
Ground Truth	Implicit	0.3214	0.3836
	Supervised	0.3214	**0.4329**

Liu, Chenxi, et al. Attention correctness in neural image captioning, submitted 2016.

Do more human-like attention maps result in better captioning performance?

Dataset	Model	BLEU-3	BLEU-4	METEOR
Flickr30k	Implicit	28.8	19.1	18.49
	Strong Sup	**30.2**	**21.0**	**19.21**

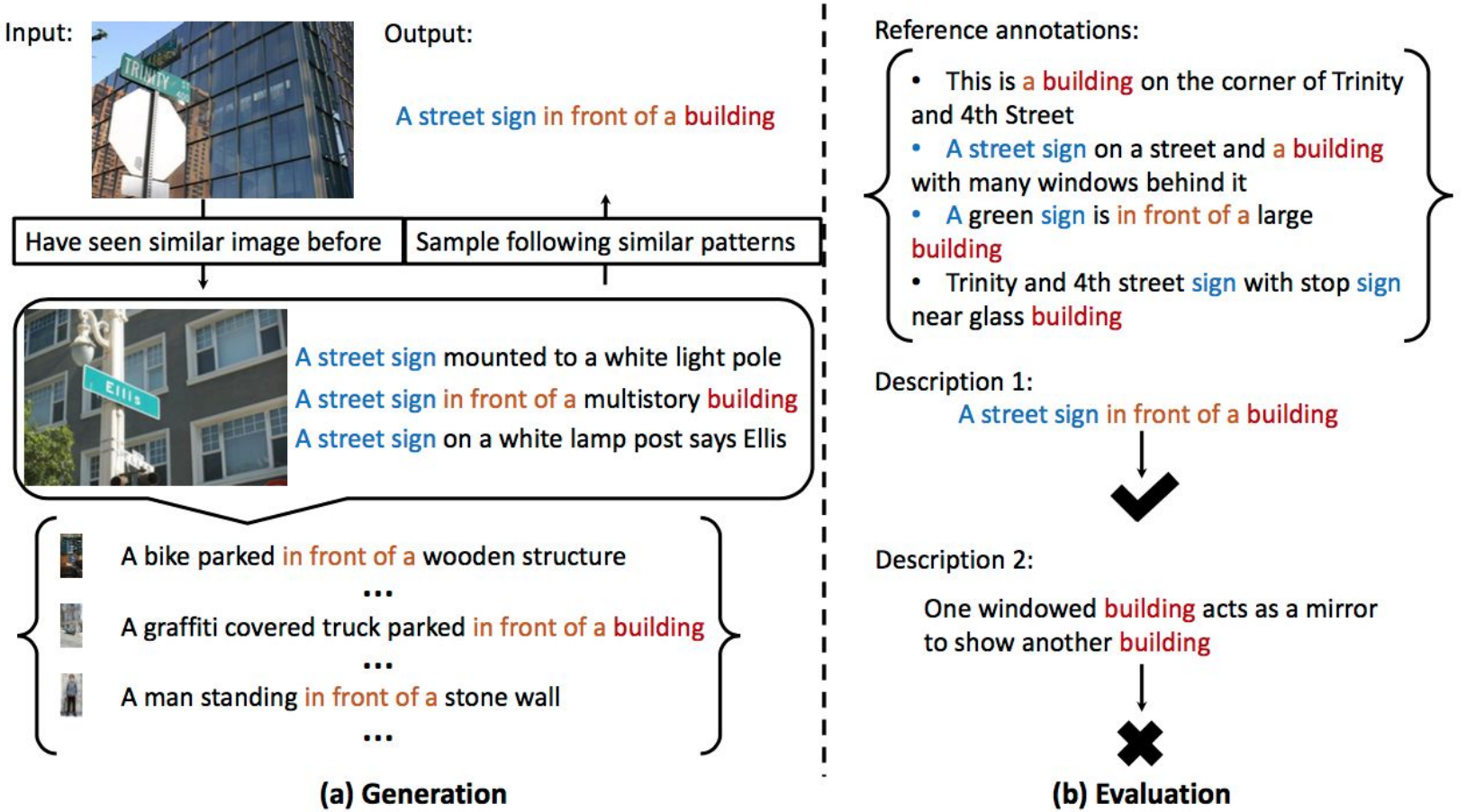
Liu, Chenxi, et al. Attention correctness in neural image captioning, submitted 2016.

What's wrong with MLE-based approaches?

A cow standing in a field next to houses
A cow standing in a field with houses
A cow standing in a field of grass

A train that is pulling into a station
A train that is going into a train station
A train that is parked in a train station

Dai et al., Towards Diverse and Natural Image Descriptions via a Conditional GAN, submitted 2017



Dai et al., Towards Diverse and Natural Image Descriptions via a Conditional GAN, submitted 2017

Towards Diverse and Natural Image Descriptions via a Conditional GAN
(Dai et al., 2017)

MLE approaches tend to generate similar expressions

- Objective encourages use of *n*-grams that appeared in training sentences
- Conventional evaluation metrics tend to favor sentences containing matched *n*-grams

What can be improved?

- Captions should have fidelity, naturalness, and diversity
- Existing efforts primarily focus on fidelity

Approach: encourage diversity and naturalness via conditional GAN

Dai et al., Towards Diverse and Natural Image Descriptions via a Conditional GAN, submitted 2017

How to define naturalness?

“When nine hundred years old you reach, look as good you will not.”

-- Return of the Jedi

“Do. Or do not. There is no try.”

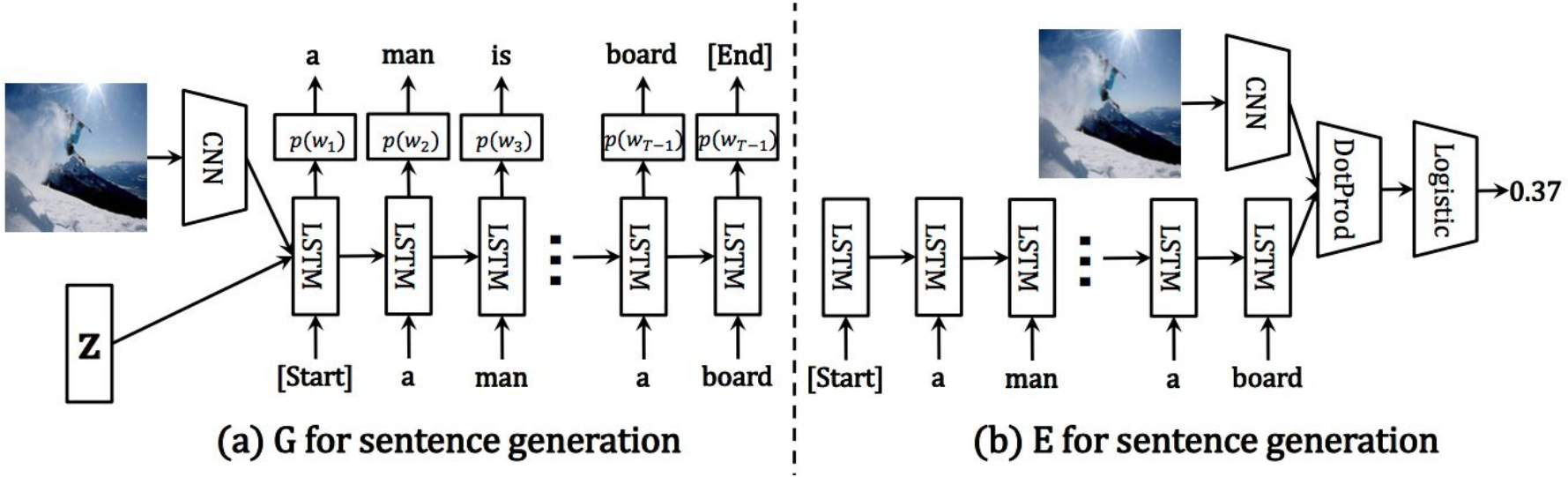
-- The Empire Strikes Back

“Not if anything to say about it I have”

-- Revenge of the Sith

Dai et al., Towards Diverse and Natural Image Descriptions via a Conditional GAN, submitted 2017

Towards Diverse and Natural Image Descriptions via a Conditional GAN
(Dai et al., 2017)



G produces different descriptions given an image

E measures how well a description fits an image

Dai et al., Towards Diverse and Natural Image Descriptions via a Conditional GAN, submitted 2017

Training *G*

Problem: production of sentences is a discrete sampling process

- This is nondifferentiable!
- Solution: **policy gradient**
  - Each word is an action
  - A sentence is considered a sequence of actions

Dai et al., Towards Diverse and Natural Image Descriptions via a Conditional GAN, submitted 2017

Training *G*

Problem: a sentence can only be evaluated when it is completely generated

- Leads to practical difficulties
  - Vanishing gradients
  - Overly slow convergence
- Solution: **early feedback** by computing expected future reward
  - When we have a partial sentence, continue to sample the remaining words using an LSTM
  - Generate *N* such sentences and compute average score to approximate expected future reward

Dai et al., Towards Diverse and Natural Image Descriptions via a Conditional GAN, submitted 2017

Training *E*

Two variants of *E* are trained

1. **E-GAN:** distinguishes between artificial descriptions and real descriptions in training set
2. **E-NGAN:** combination of
   a. E-GAN objective
   b. Explicitly suppresses mismatched descriptions to ensure semantic relevance

Dai et al., Towards Diverse and Natural Image Descriptions via a Conditional GAN, submitted 2017

Evaluation

Metrics: compare conventional metrics with those determined by **E-NGAN** and **E-GAN**

Models: comparison between human, Show and Tell (G-MLE) and conditional GAN (G-GAN)

		BLEU-3	BLEU-4	METEOR	ROUGE_L	CIDEr	SPICE	E-NGAN	E-GAN
COCO	human	0.290	0.192	0.240	0.465	0.849	**0.211**	0.527	**0.626**
	G-MLE	**0.393**	**0.299**	**0.248**	**0.527**	**1.020**	0.199	0.464	0.427
	G-GAN	0.305	0.207	0.224	0.475	0.795	0.182	**0.528**	0.602
Flickr	human	0.269	0.185	0.194	0.423	0.627	0.159	0.482	**0.464**
	G-MLE	**0.372**	**0.305**	**0.215**	**0.479**	**0.767**	**0.168**	0.465	0.439
	G-GAN	0.153	0.088	0.132	0.330	0.202	0.087	**0.582**	0.456

Dai et al., Towards Diverse and Natural Image Descriptions via a Conditional GAN, submitted 2017

				
z_1	a baseball player holds a bat up to hit the ball	a man riding a snowboard down a slope	a group of people sitting around a table having a meal in a restaurant	a group of men dressed in suits posing for a photo
z_2	a baseball player holding white bat and wear blue baseball uniform	a person standing on a snowboard sliding down a hill	a young man sitting at a table with coffee and a lot of food	a couple of men standing next to each other wearing glasses
z_3	a professional baseball player holds up his bat as he watches	a man is jumping over a snow covered hill	a pretty young man sitting next to two men in lots of people	some people dressed in costume and cups
z_4	a baseball player is getting ready to hit a baseball in the outfield	a man riding a snowboard through a wave	group of people sitting around wooden tables with wine glasses	a man is in a dress making a funny face
z_5	a baseball player who is about to hit the ball	a skier is in mid air over a jump in the snow	a group of people sitting at picnic tables in a room	a group of people standing and posing for a photo

Dai et al., Towards Diverse and Natural Image Descriptions via a Conditional GAN, submitted 2017

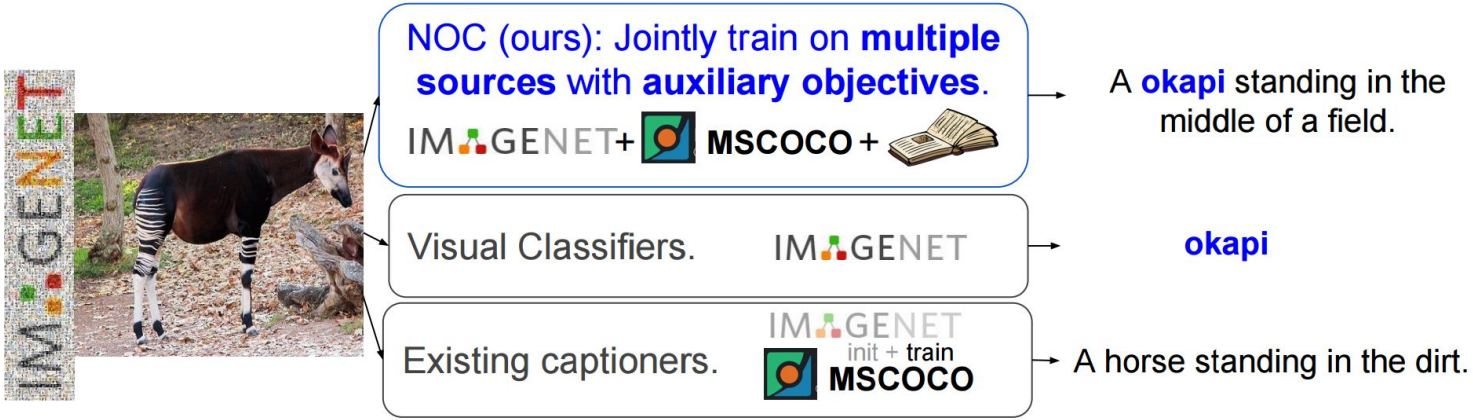
				
G-MLE	a man flying through the air while riding a snowboard	a man flying through the air while riding a snowboard	a man flying through the air while riding a snowboard	a man flying through the air while riding a skateboard
G-GAN	a man on a skateboard in a snowy park	a man skiing down the slope near a mountain	a man performing a grind trick on a skateboard ramp	a man with stunts on his skis in the snow
				
G-MLE	a group of people standing around a boat	a group of people sitting around a table	a group of people sitting at a table	a group of people sitting around a living room
G-GAN	the bench is sitting on the ground by the water	a group of people watching each other	a table with a lot of stuff on it	furnished living room with furniture and built area

Dai et al., Towards Diverse and Natural Image Descriptions via a Conditional GAN, submitted 2017

Captioning Images with Diverse Objects

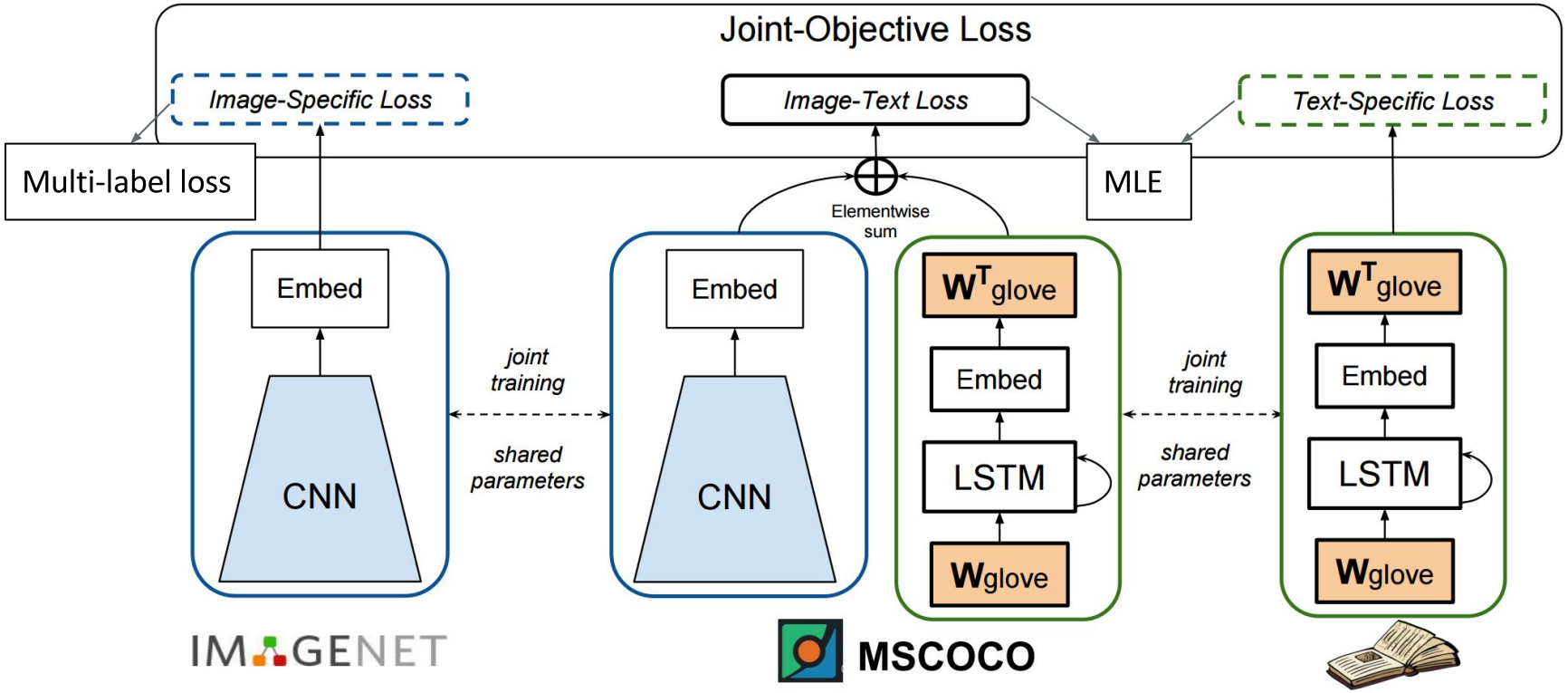
(Venugopalan et al., 2016)

- How can a model describe object categories not present in existing image-caption datasets?
- Take advantage of **external sources**
  - Labeled images from object recognition datasets
  - Semantic knowledge extracted from unannotated text



Venugopalan et al., Captioning Images with Diverse Objects, submitted 2016

Novel Object Captioner (NOC)



Venugopalan et al., Captioning Images with Diverse Objects, submitted 2016

Joint Training with Auxiliary Losses

Most approaches: pre-train image & text models, tune caption model alone

- But models tend to “forget” weights for objects seen only in external data sources

Instead, train image, text, and caption models simultaneously on different data sources

- Final objective minimizes weighted combination of losses

Weights of the caption model are shared with the image network and language network

- Model can be trained simultaneously on independent image-only data, unannotated text data, and paired image-caption data

Venugopalan et al., Captioning Images with Diverse Objects, submitted 2016

Racket

DCC: A man playing a **racket** on a court.
NOC (Ours): A tennis player swinging a **racket** at a ball.

Bus

DCC: A group of people on a snowy road next to trees.
NOC (Ours): **Bus** driving down a snowy road next to trees.

Bottle

DCC: A glass of wine sitting on a table with a glass of wine.
NOC (Ours): A table with a **bottle** of wine and a glass of wine.

Suitcase

DCC: A close up of a person sitting on a wooden bench.
NOC (Ours): A bunch of **suitcases** stacked on top of each other.

Model	bottle	bus	couch	microwave	pizza	racket	suitcase	zebra	Avg. F1	Avg. METEOR
DCC	4.63	29.79	**45.87**	**28.09**	64.59	52.24	13.16	79.88	39.78	21.00
NOC (ours)	**17.78**	**68.79**	25.55	24.72	**69.33**	**68.06**	**39.86**	**89.02**	**49.14**	**21.38**

Venugopalan et al., Captioning Images with Diverse Objects, submitted 2016

Novel Objects (COCO)

Tennis player preparing to hit the ball with a **racket**.

A **bus** driving down a busy street with people standing around.

A cat sitting on a **suitcase** next to a bag.

Rare Words

A man in a red and white shirt and a red and white **octopus**.

A red **trolley train** sits on the tracks near a building

A close up of a plate of food with a **spatula**.

Novel Objects (ImageNet Images)

A white and red **cockatoo** standing in a field.

A woman is holding a large **megaphone** in her hand.

A table with a plate of **sashimi** and vegetables.

A **woodpecker** sitting on a tree branch in the woods.

A **orca** is riding a small wave in the water.

A large **flounder** is resting on a rock

A **otter** is sitting on a rock in the sun.

A **saucepan** full of soup and a pot on a stove.

A man is standing on a field with a **caddie**.

Errors (ImageNet)

A man holding a baseball bat standing in front of a building

A cat is laying inside of a small white **aardvark**.

A **barracuda** on a blue ocean with a **barracuda**.

Venugopalan et al., Captioning Images with Diverse Objects, submitted 2016

Gladiator (n10131815) Error: Semantics

NOC: A man wearing a **gladiator** wearing a **gladiator** hat.

Taper (n13902793) Error: Counting

NOC: A group of three **taper** sitting on a table.

Trifle (n07613480) Error: Repetition

NOC: A **trifle** cake with **trifle** cake on top of a **trifle** cake.

Lory (n01820348) Error: Recognition

NOC: A bird sitting on a branch with a colorful bird sitting on it.

Venugopalan et al., Captioning Images with Diverse Objects, submitted 2016

	Word Incorporation		Image Description	
Objects subset →	Union	Intersection	Union	Intersection
NOC is better	**43.78**	34.61	**59.84**	51.04
DCC is better	25.74	34.12	40.16	48.96
Both equally good	6.10	9.35		-
Neither is good	24.37	21.91		-

Venugopalan et al., Captioning Images with Diverse Objects, submitted 2016

Conclusion

- Image captioning is an interesting task combining vision and language that poses unique challenges
  - How do we generate human-like captions?
  - How can we automatically evaluate captions?
- Various architectures have been proposed
  - CNN + RNN
  - Compositional pipeline
  - Attention models (in vogue)
- Combating inherent problems with generated captions
  - Diversity
  - Rare words

Reading list

- 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.
- Vinyals, Oriol, et al. "Show and tell: A neural image caption generator." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015.
- Vinyals, Oriol, et al., "Show and tell: Lessons learned from the 2015 MSCOCO Image Captioning Challenge." *IEEE Transactions on Pattern Analysis and Machine Intelligence.* 2016.
- Fang, Hao, et al., "From captions to visual concepts and back."*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015.
- Xu, Kelvin et al., "Show, Attend, and Tell: Neural Image Caption Generation with Visual Attention." ICML. Vol. 14. 2015.
- Liu, Chenxi, et al. "Attention correctness in neural image captioning." arXiv preprint arXiv:1605.09553 (2016).
- Dai, Bo, et al. "Towards Diverse and Natural Image Descriptions via a Conditional GAN." arXiv preprint arXiv:1703.06029. 2017.
- Venugopalan, Subhashini, et al. "Captioning images with diverse objects." arXiv preprint arXiv:1606.07770 (2016).