From Captions to Visual Concepts and Back

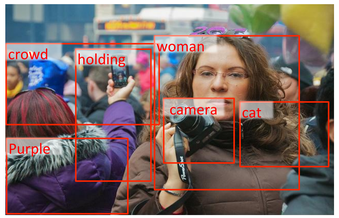
Hao Fang*, Saurabh Gupta*, Forrest Iandola*, Rupesh Srivastava*, Li Deng, Piotr Dollár,
Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C. Platt, C. Lawrence Zitnick, Geoffrey Zweig

Overview

We present a novel approach for automatically generating image descriptions using:

- Multiple Instance Learning (MIL) for visually detecting words
- A maximum entropy language model
- Sentence ranking using MERT and a Deep Multimodal Similarity Model (DMSM)

Pipeline



1. Word Detection → woman, crowd, cat, camera, holding, purple

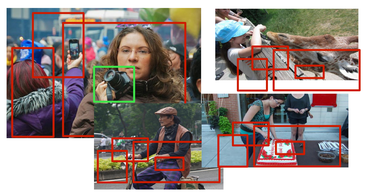
2. Sentence Generation → A purple camera with a woman. A woman holding a camera in a crowd. ... A woman holding a cat.

3. Sentence Re-Ranking → #1 A woman holding a camera in a crowd.

1. Word Detection

Multiple Instance Learning (MIL)

Camera



SoftMax:

$$p_{ij}^w = \frac{1}{1+\exp\left(-v_w^t \phi(b_{ij}) - u_w\right)}$$

Noisy Or:

$$p_i^w = 1 - \prod_{j \in b_i} (1 - p_{ij}^w)$$

Image

CNN

FC6, FC7, FC8 as fully convolutional layers

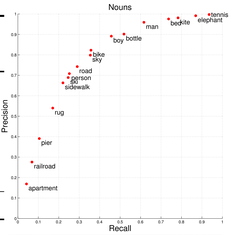
Spatial class probability maps

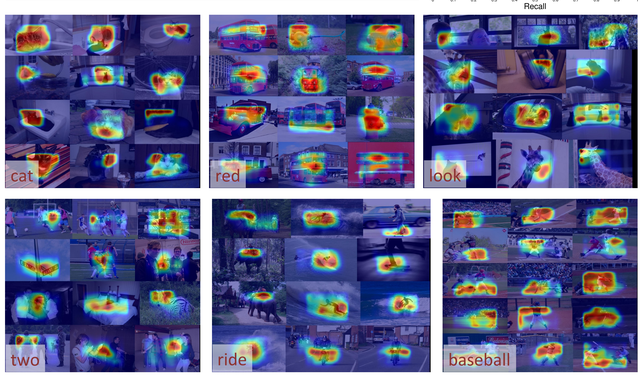
MIL

Per class probability

Results

	PHR								AP
	NN	VB	JJ	DT	PRP	IN	Oth	All	All
Classification (AlexNet)	39	28	37	37	26	32	25	36	27
Classification (VGG)	45	31	37	**40**	**30**	**34**	**26**	41	31
MIL (AlexNet)	46	29	40	38	26	32	22	41	30
MIL (VGG)	**52**	**33**	**44**	39	29	**34**	24	**46**	**34**
Human Agreement	64	35	36	43	32	34	32	53	-





2. Sentence Generation

Language Model

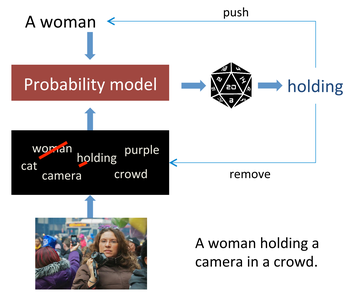
Probability of a word depends on previous words

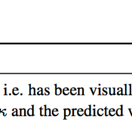
Probability model

Word Probability

$$\Pr(w_l = \bar{w}_l | \bar{w}_{l-1}, \cdots, \bar{w}_1, <s>, \tilde{\mathcal{V}}_{l-1}) = \frac{\exp\left[\sum_{k=1}^K \lambda_k f_k(\bar{w}_l, \bar{w}_{l-1}, \cdots, \bar{w}_1, <s>, \tilde{\mathcal{V}}_{l-1})\right]}{\sum_{v \in \mathcal{V} \cup </s>} \exp\left[\sum_{k=1}^K \lambda_k f_k(v, \bar{w}_{l-1}, \cdots, \bar{w}_1, <s>, \tilde{\mathcal{V}}_{l-1})\right]}$$

Attribute Conditioning





A woman holding a camera in a crowd.

	Definition	Description
Attribute	0/1	$\bar{w}_l \in \tilde{\mathcal{V}}_{l-1}$	Predicted word is in the attribute set, i.e. has been visually detected and not yet used.
N-gram+	0/1	$\bar{w}_{l-N+1}, \cdots, \bar{w}_l = \kappa$ and $\bar{w}_l \in \tilde{\mathcal{V}}_{l-1}$	N-gram ending in predicted word is κ and the predicted word is in the attribute set.
N-gram-	0/1	$\bar{w}_{l-N+1}, \cdots, \bar{w}_l = \kappa$ and $\bar{w}_l \notin \tilde{\mathcal{V}}_{l-1}$	N-gram ending in predicted word is κ and the predicted word is not in the attribute set.
End	0/1	$\bar{w}_l = \kappa$ and $\tilde{\mathcal{V}}_{l-1} = \emptyset$	The predicted word is κ and all attributes have been mentioned.
Score	$\mathbb{R}$	score($\bar{w}_l$) when $\bar{w}_l \in \tilde{\mathcal{V}}_{l-1}$	The log-probability of the predicted word when it is in the attribute set.

3. Sentence Re-Ranking

Re-rank the *m*-best sentences using Minimum Error Rate Training (MERT). Ranking is based on the following features:

1. The log-likelihood of the sequence.
2. The length of the sequence.
3. The log-probability per word of the sequence.
4. The logarithm of the sequence's rank in the log-likelihood.
5. 11 binary features indicating whether the number of mentioned objects is x ($x = 0, \ldots, 10$).
6. The DMSM score between the sequence and the image.

Q = image, D = caption, R = relevance

Relevance:

$$R(Q,D) = \text{cosine}(y_Q, y_D) = \frac{{y_Q}^T y_D}{\|y_Q\| \|y_D\|}$$

Caption probability:

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

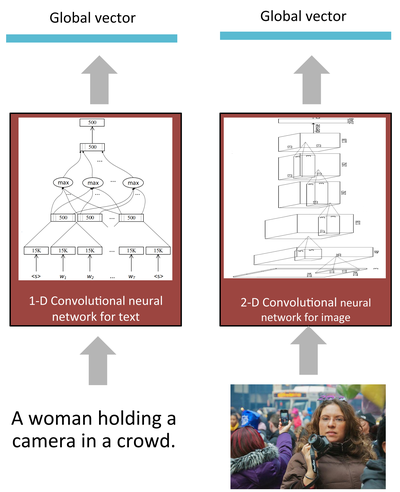
Candidate captions — Smoothing factor

Objective:

$$L(\Lambda) = -\log \prod_{(Q,D^+)} P(D^+|Q)$$

Correct caption

DMSM



Results

Ablation Study — Human Study

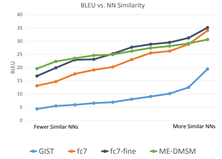
System	PPLX	BLEU	METEOR	= human	> human	>= human
Unconditioned	24.1	1.2	6.8			
Shuffled Human	-	1.7	7.3			
Baseline	20.9	16.9	18.9	9.9	2.4	12.3
Baseline + score	20.2	20.1	20.5	16.9	3.9	20.8
Baseline + score +DMSM	20.2	21.1	20.7	18.7	4.6	23.3
Baseline + score + DMSM [ft]	19.2	23.3	22.2			
VGG + score [ft]	18.1	23.6	22.8			
VGG + score + DMSM [ft]	18.1	25.7	23.6	26.2	7.8	34.0
Human written caption	-	19.3	24.1			

MS COCO Caption Test Server

	CIDEr-D	Meteor	ROUGE-L	BLEU-1	BLEU-2	BLEU-3	BLEU-4
MSR[5]	0.925	0.331	0.662	0.88	0.789	0.678	0.567
Human[3]	0.91	0.335	0.626	0.88	0.744	0.603	0.471
Berkeley LRCN[1]	0.891	0.322	0.656	0.871	0.772	0.653	0.534
Google[2]	0.842	0.327	0.649	0.872	0.766	0.648	0.538
m-RNN (Baidu/ UCLA)[8]	0.828	0.312	0.647	0.872	0.771	0.654	0.543
MLBL[4]	0.752	0.294	0.635	0.848	0.747	0.633	0.517
NeuralTalk[6]	0.692	0.28	0.603	0.828	0.701	0.566	0.446
Tsinghua Bigeye[7]	0.682	0.273	0.616	0.866	0.756	0.628	0.493

Analysis

	Unique Captions	Seen in Training
Human	99.4	4.8
k-Nearest Neighbor	36.6	100
LSTM / RNN Style	33.1	60.3
Our	47.0	30.0



	= human	> human	>= human
k-Nearest Neighbor	22.1	5.5	27.6
Our	26.2	7.8	34.0