From Captions to Visual Concepts and Back

Saurabh Gupta
UC Berkeley

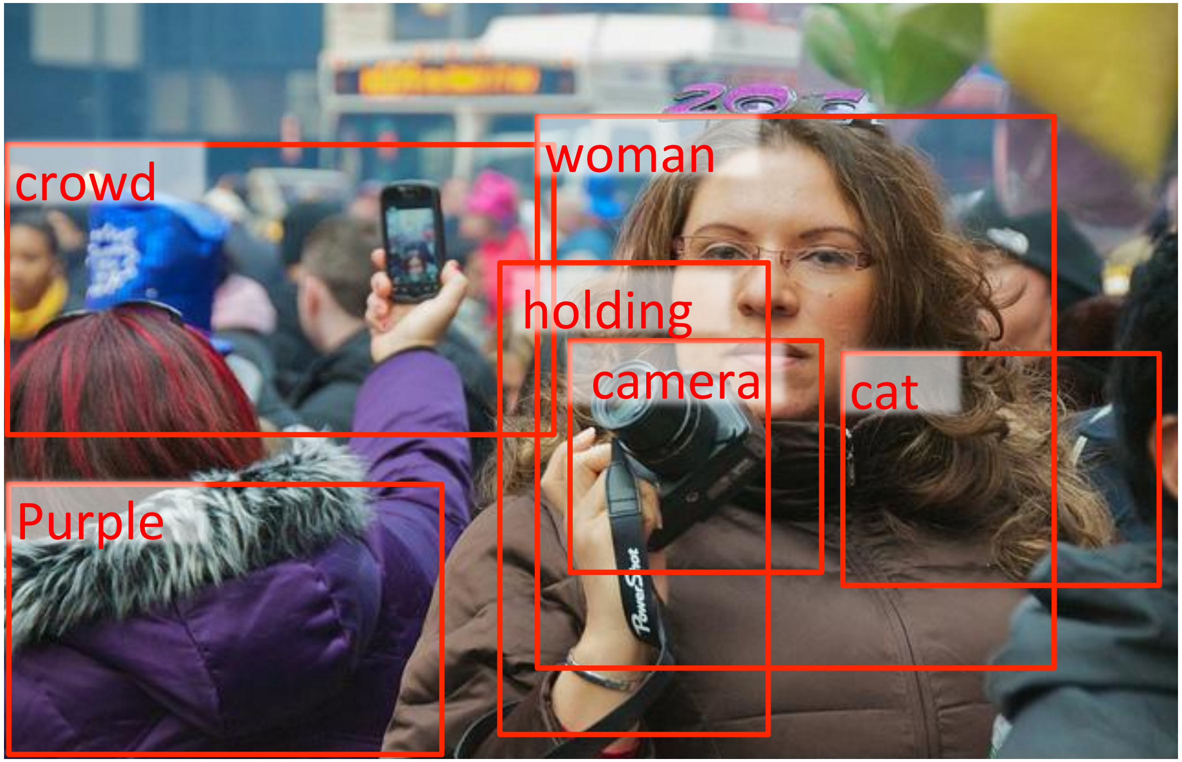
Work done at Microsoft Research

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

- **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, **CVPR 2015**

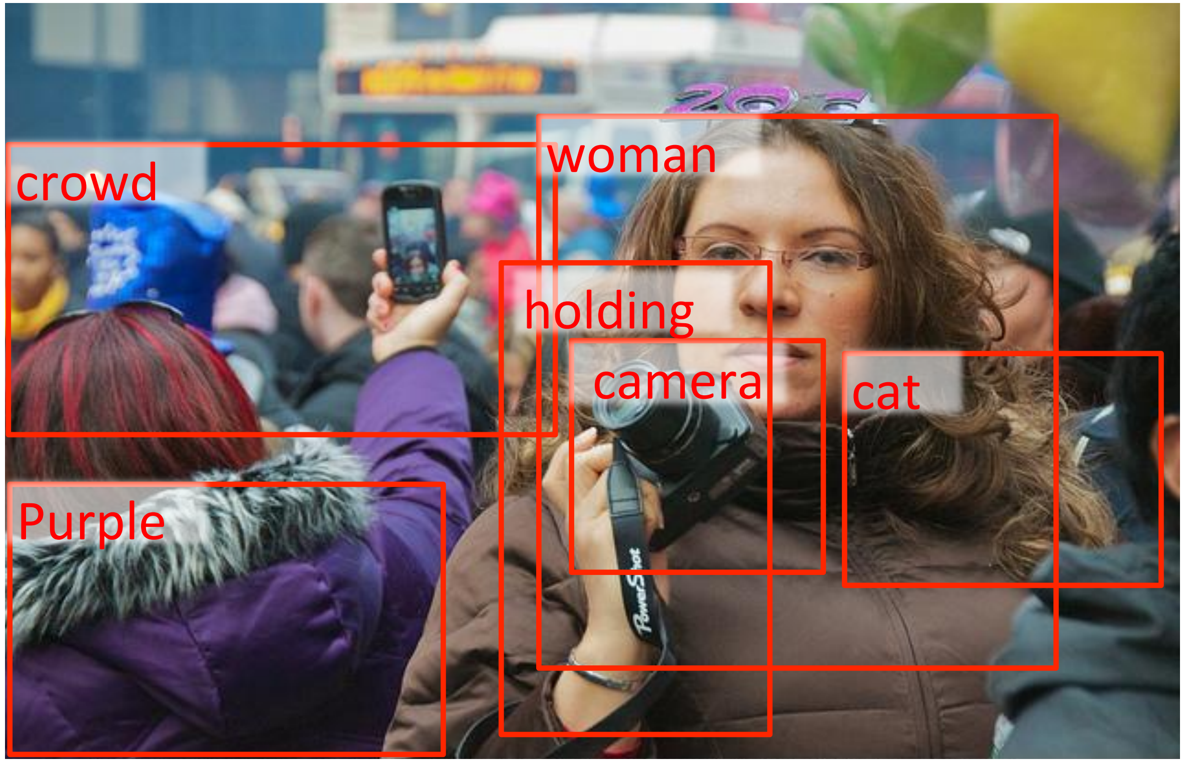
- **Language Models for Image Captioning: The Quirks and What Works,** Jacob Devlin, Hao Cheng, Hao Fang, Saurabh Gupta, Li Deng, Xiaodong He, Geoffrey Zweig, **ACL 2015**

- **Exploring Nearest Neighbor Approaches for Image Captioning** Jacob Devlin, Saurabh Gupta, Ross Girshick, Margaret Mitchell C. Lawrence Zitnick, **arXiv 2015**



1. Word Detection

woman, crowd, cat, camera, holding, purple

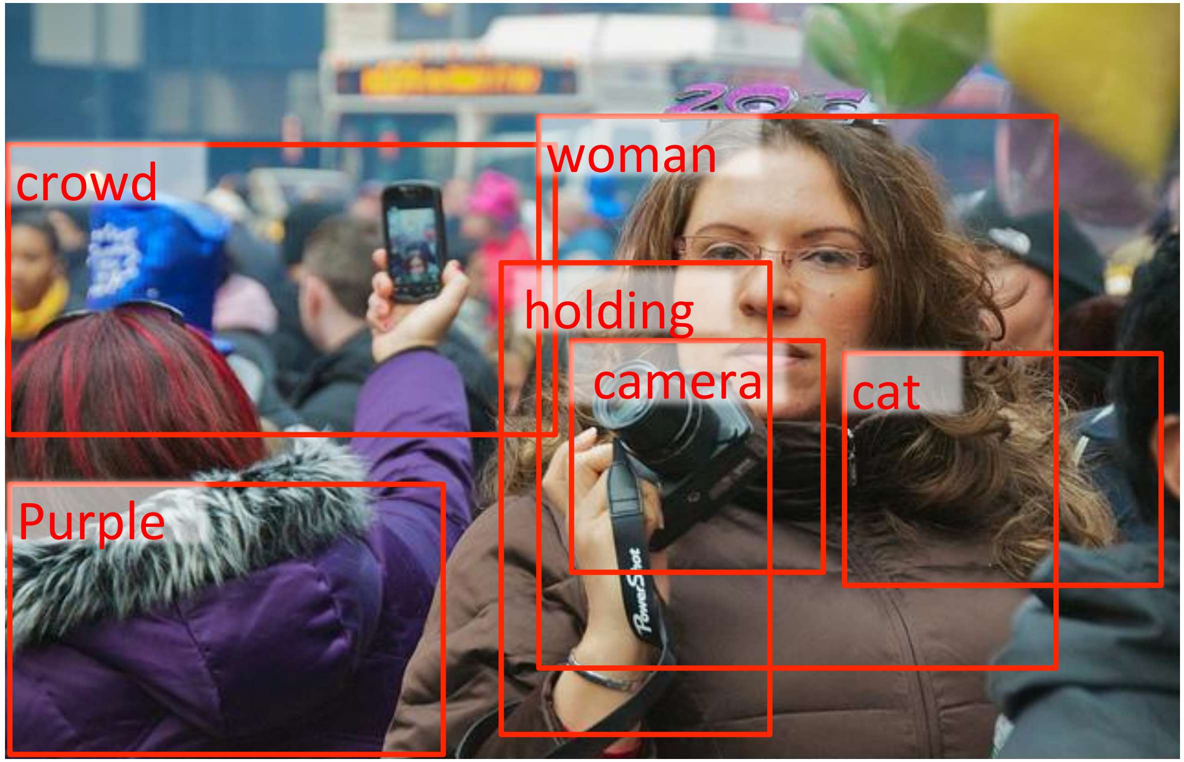


1. Word Detection

2. Sentence Generation

woman, crowd, cat, camera, holding, purple

A purple camera with a woman.
A woman holding a camera in a crowd.
...
A woman holding a cat.

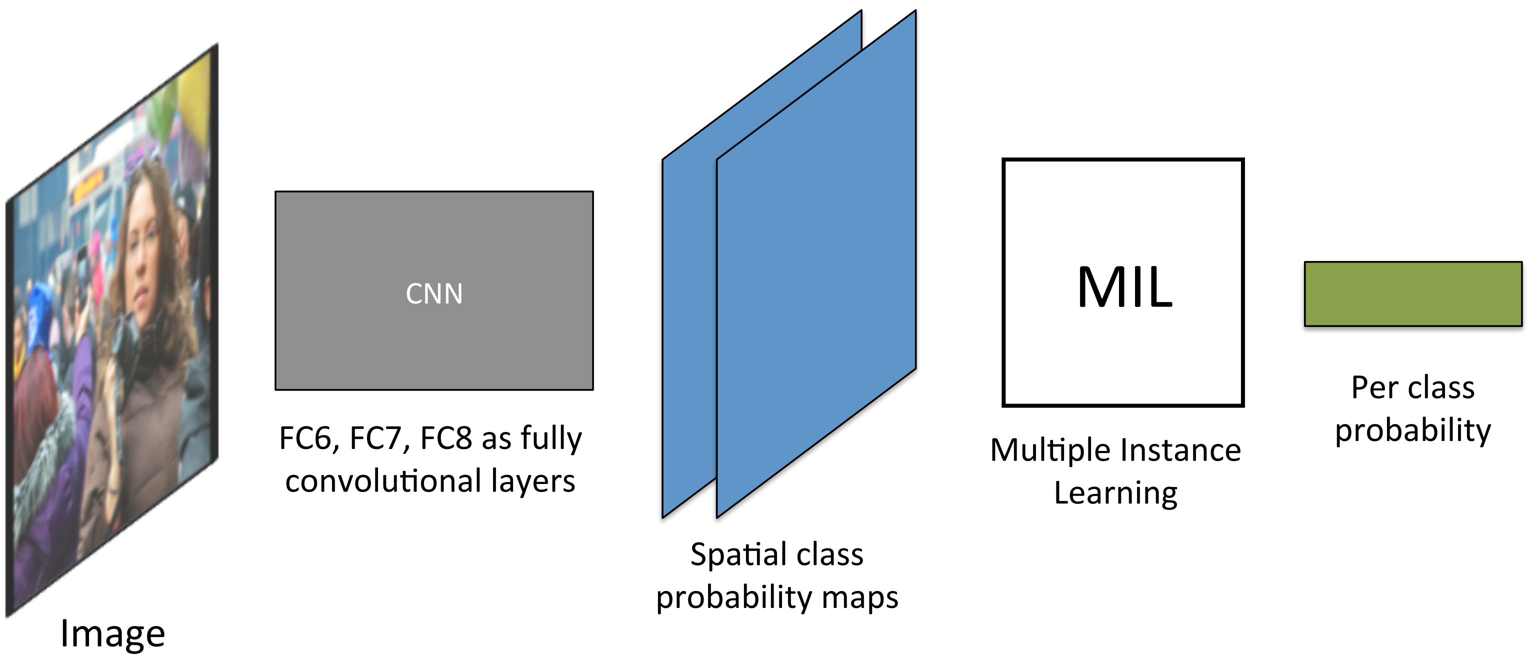


1. Word Detection → 2. Sentence Generation → 3. Sentence Re-Ranking

woman, crowd, cat, camera, holding, purple

A purple camera with a woman.
A woman holding a camera in a crowd.
...
A woman holding a cat.

#1 A woman holding a camera in a crowd.

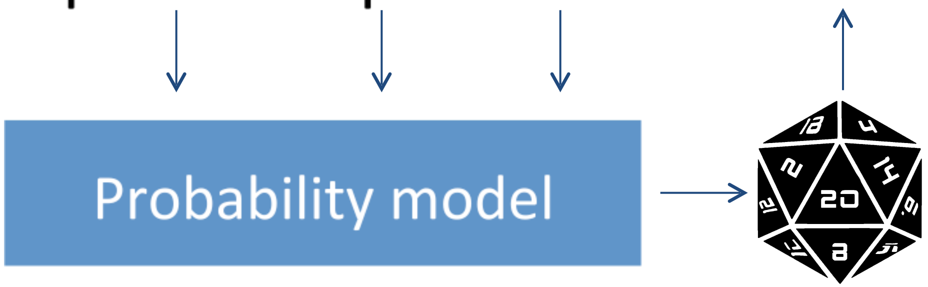

CNN
FC6, FC7, FC8 as fully convolutional layers
Image
Spatial class probability maps
MIL
Multiple Instance Learning
Per class probability

Language models learn to babble

Probability of a word depends on previous

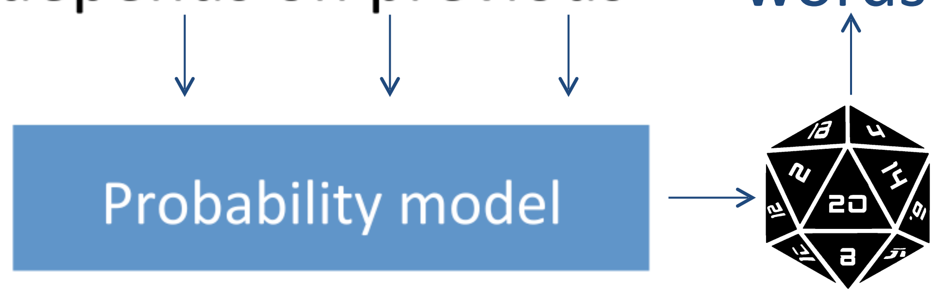
Language models learn to babble

Probability of a word depends on previous



Language models learn to babble

Probability of a word depends on previous words



Language models learn to babble

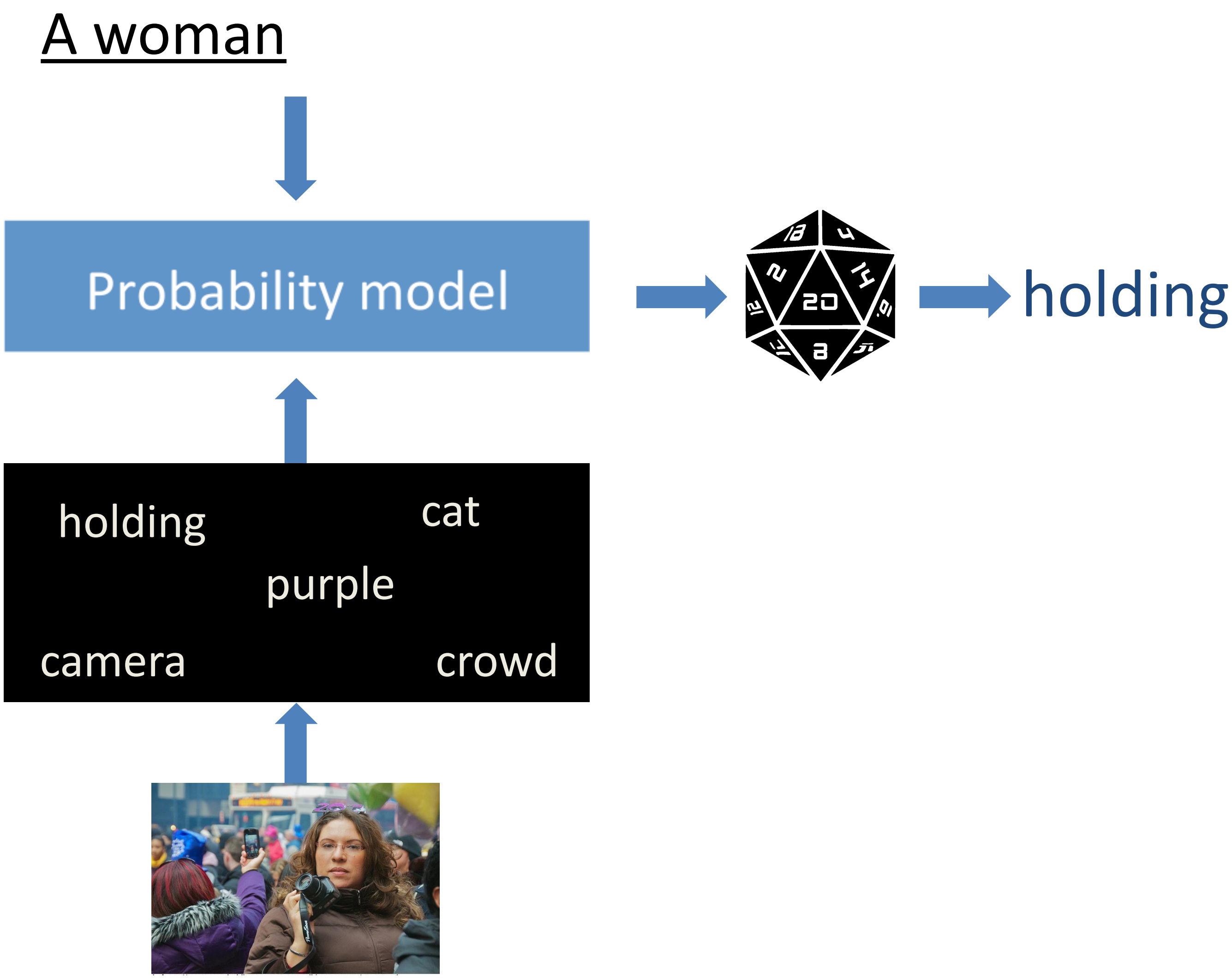
Probability of a word depends on previous words

Probability model

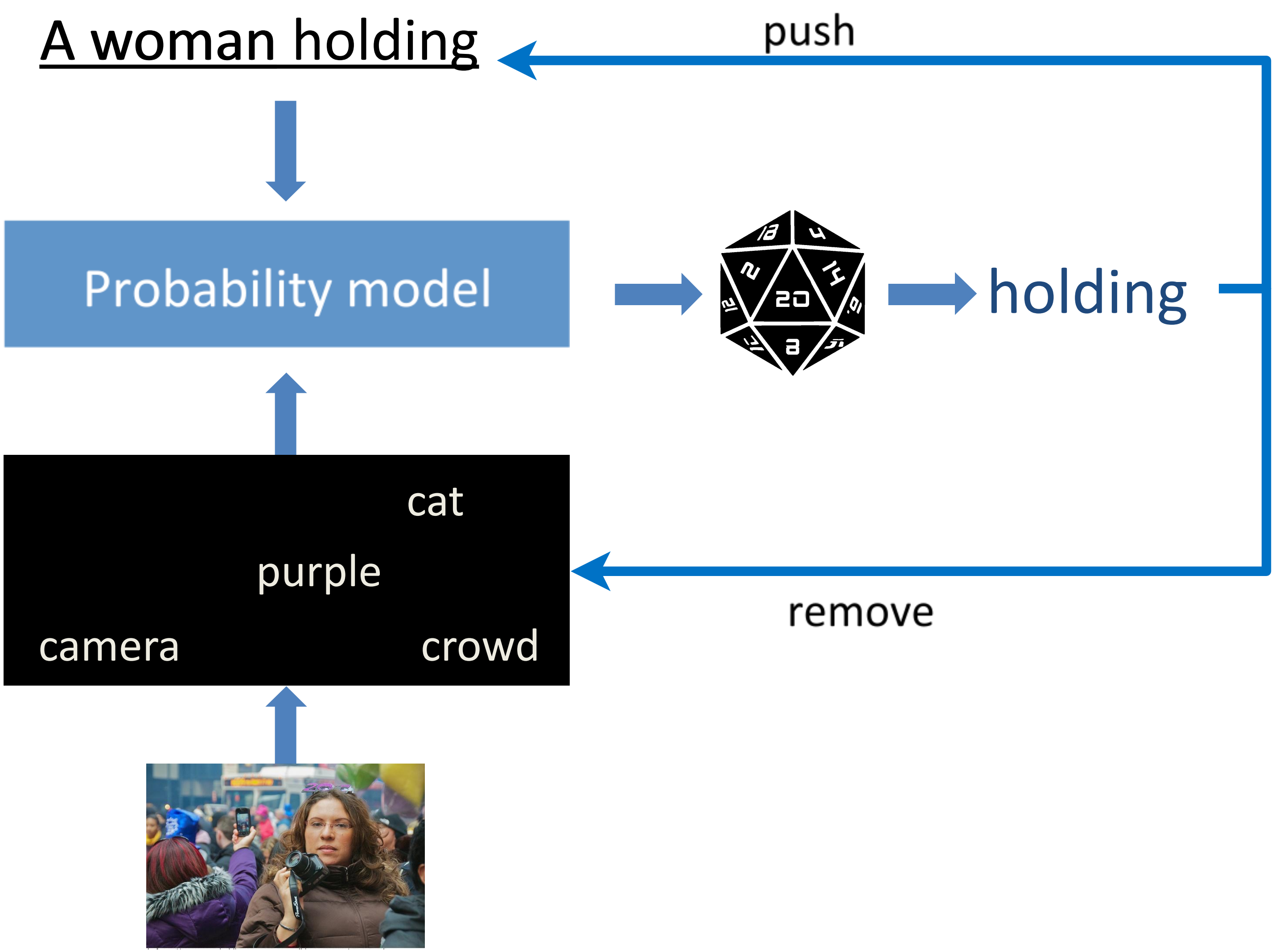
Language model

Nay, I know not:
Is by a sleep to say we end.
The ratifiers and props of every word, They are not the trail of policy so sure As hush as death, anon the dreadful thunder. Doth all the days i' the church.

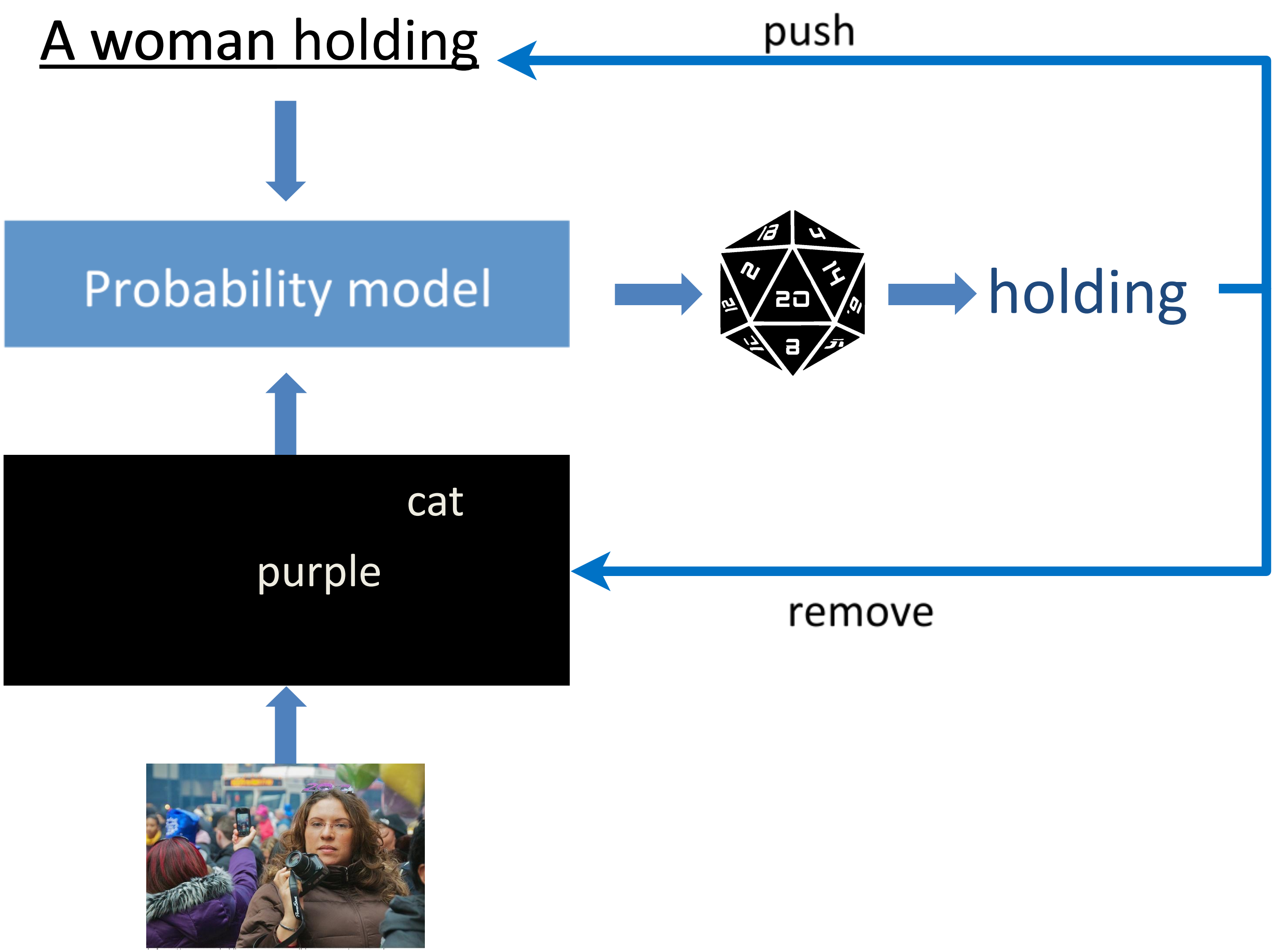
Add a blackboard



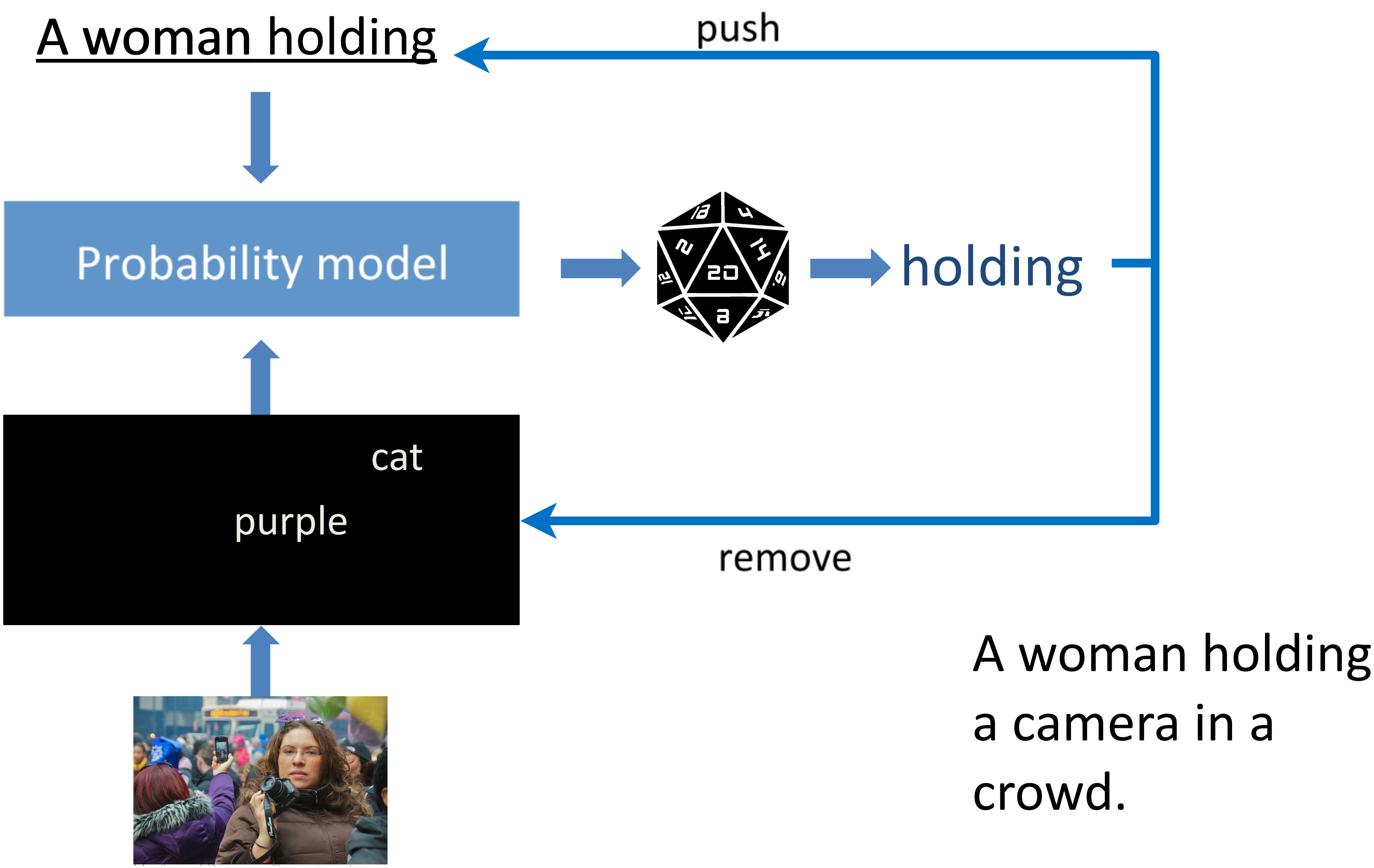
Add a blackboard



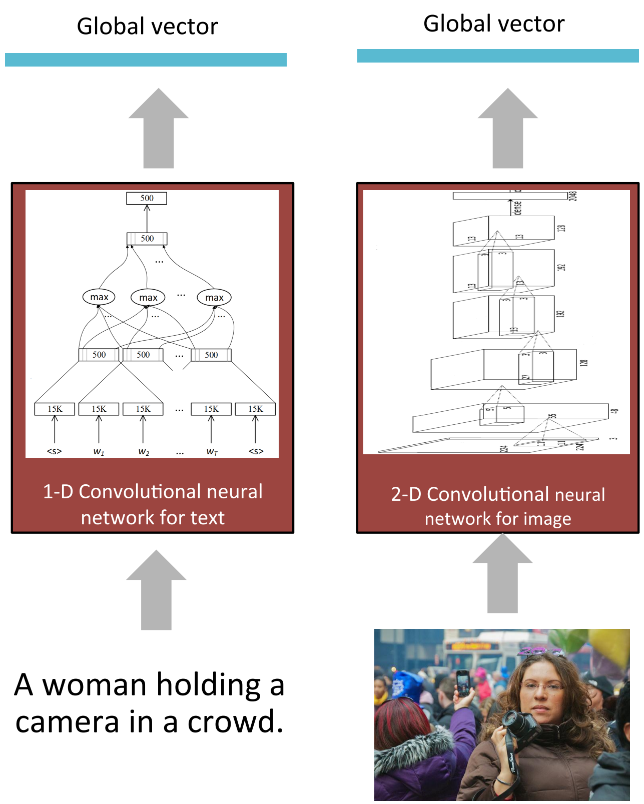
Add a blackboard



Add a blackboard



Re-rank hypotheses *globally*



DMSM - Embedding to maximize similarity between image and its corresponding caption

1. A purple camera with a woman
2. A woman holding a camera in a crowd.
3. A woman holding a cat.
4. ....
5. ....

Sentence and image level features

MERT to optimize for BLEU on val set

Return best hypothesis

Results

	Val c4		Test c40		
System	**BLEU4**	**METEOR**	**BLEU**	**METEOR**	**CIDEr-D**
Our	25.7	**23.6**	56.7	31.8	92.5
G-RNN	25.7	22.6	-	-	-
Our + G-RNN	**27.3**	**23.6**	**60.1**	**33.9**	**93.7**

MSR = Our

MSR Captivator = Our + G-RNN

Results

	Val c4		Test c40		
System	**BLEU4**	**METEOR**	**BLEU**	**METEOR**	**CIDEr-D**
Our	25.7	**23.6**	56.7	31.8	92.5
G-RNN	25.7	22.6	-	-	-
Our + G-RNN	**27.3**	**23.6**	**60.1**	**33.9**	**93.7**

4-5th by automatic metrics, Tied 1st by human evals

MSR = Our

MSR Captivator = Our + G-RNN

Results

	Val c4		Test c40		
System	**BLEU4**	**METEOR**	**BLEU**	**METEOR**	**CIDEr-D**
Our	25.7	**23.6**	56.7	31.8	92.5
G-RNN	25.7	22.6	-	-	-
Our + G-RNN	**27.3**	**23.6**	**60.1**	**33.9**	**93.7**

4-5th by automatic metrics, Tied 1st by human evals

1-2st by automatic metrics

MSR = Our

MSR Captivator = Our + G-RNN

Novelty in Captions?

Novelty in Captions?

	Val c4			
System	**BLEU4**	**METEOR**	**Unique Captions (%)**	**Seen in Training (%)**
Human			99.4	4.8
Our	25.7	23.6	47.0	30.0
G-RNN	25.7	22.6	33.1	60.3
Our + G-RNN	27.3	23.6	28.5	61.3

Novelty in Captions?

	Val c4			
System	**BLEU4**	**METEOR**	**Unique Captions (%)**	**Seen in Training (%)**
Human			99.4	4.8
Our	25.7	23.6	47.0	30.0
G-RNN	25.7	22.6	33.1	60.3
Our + G-RNN	27.3	23.6	28.5	61.3

For a set of 20K images, only 6.6K unique strings were emitted

Novelty in Captions?

		Val c4		
System	**BLEU4**	**METEOR**	**Unique Captions (%)**	**Seen in Training (%)**
Human			99.4	4.8
Our	25.7	23.6	47.0	30.0
G-RNN	25.7	22.6	33.1	60.3
Our + G-RNN	27.3	23.6	28.5	61.3
1-NN	11.2	17.3	-	100

For a set of 20K images, only 6.6K unique strings were emitted

Novelty in Captions?

	Val c4			
System	**BLEU4**	**METEOR**	**Unique Captions (%)**	**Seen in Training (%)**
Human			99.4	4.8
Our	25.7	23.6	47.0	30.0
G-RNN	25.7	22.6	33.1	60.3
Our + G-RNN	27.3	23.6	28.5	61.3
1-NN	11.2	17.3	-	100
k-NN	26.0	22.5	36.6	100

For a set of 20K images, only 6.6K unique strings were emitted

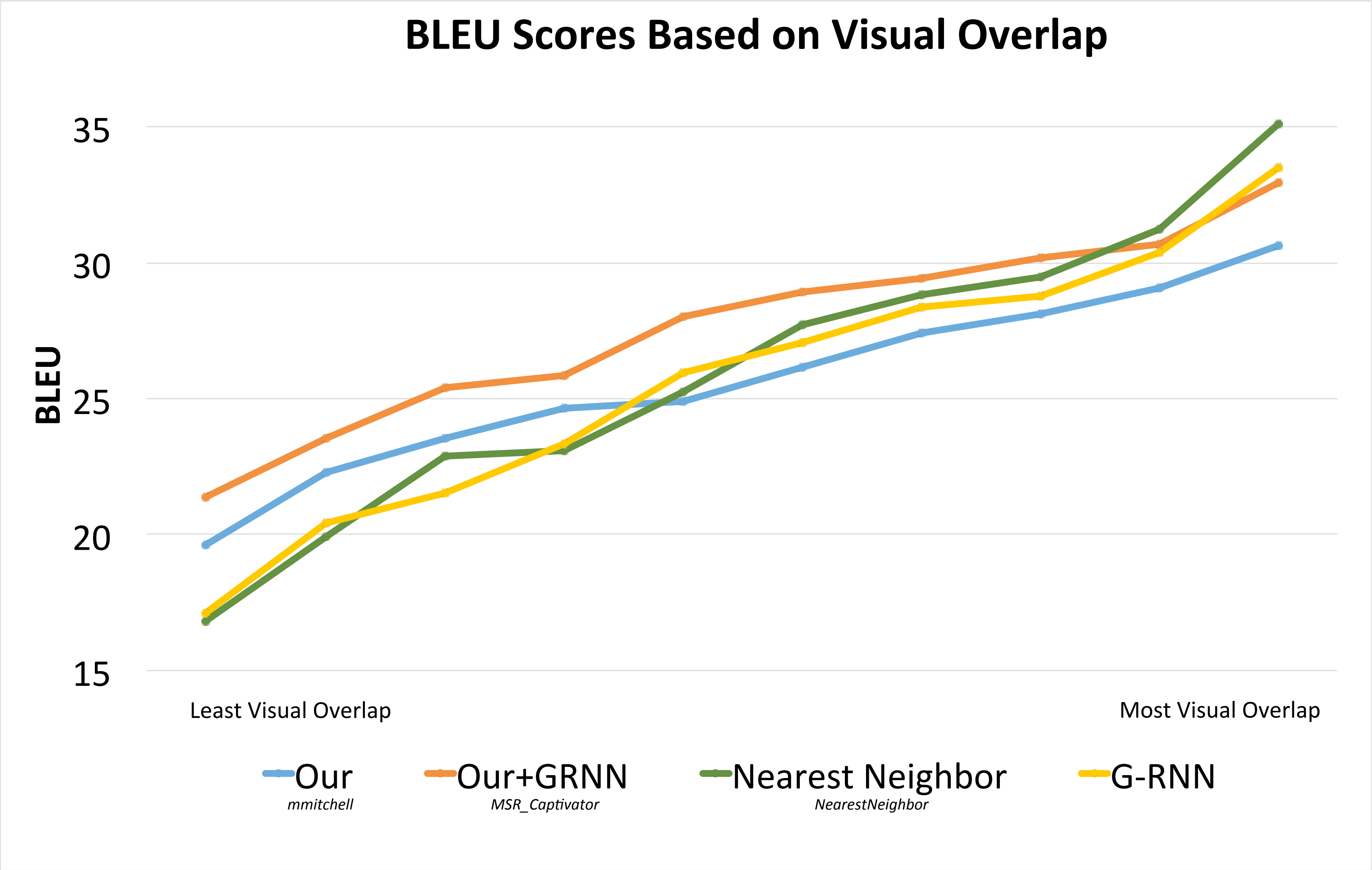
Novelty in Captions?

	Val c4			
System	**BLEU4**	**METEOR**	**Unique Captions (%)**	**Seen in Training (%)**
Human			99.4	4.8
Our	25.7	23.6	47.0	30.0
G-RNN	25.7	22.6	33.1	60.3
Our + G-RNN	27.3	23.6	28.5	61.3
1-NN	11.2	17.3	-	100
k-NN	26.0	22.5	36.6	100

For a set of 20K images, only 6.6K unique strings were emitted

Ranks 7th out of 16 on leaderboard according to automated metrics and human evals

Analysis



BLEU Scores Based on Visual Overlap
BLEU
35
30
25
20
15
Least Visual Overlap
Most Visual Overlap
Our
mmitchell
Our+GRNN
MSR_Captivator
Nearest Neighbor
NearestNeighbor
G-RNN

Interpretability

a man sitting on a couch with a dog

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

baseball (1.00) ball (1.00) player (1.00) throwing (0.86)

a baseball player throwing a ball

a pitcher holds his arm far behind him during a pitch

people (0.89)

standing (0.71)

group (0.68)

doughnuts (0.67)

a group of people standing in front of doughnuts

boxes of donuts orange juice and other snacks are sitting out for empl
oyees

Thank You