Generation and Comprehension of Unambiguous Object Descriptions

Goal

- Image captioning is subjective and ill-posed - many valid ways to describe any given image, making evaluation difficult
- *Referring expression* - An unambiguous text description that applies to exactly one object or region in the image.

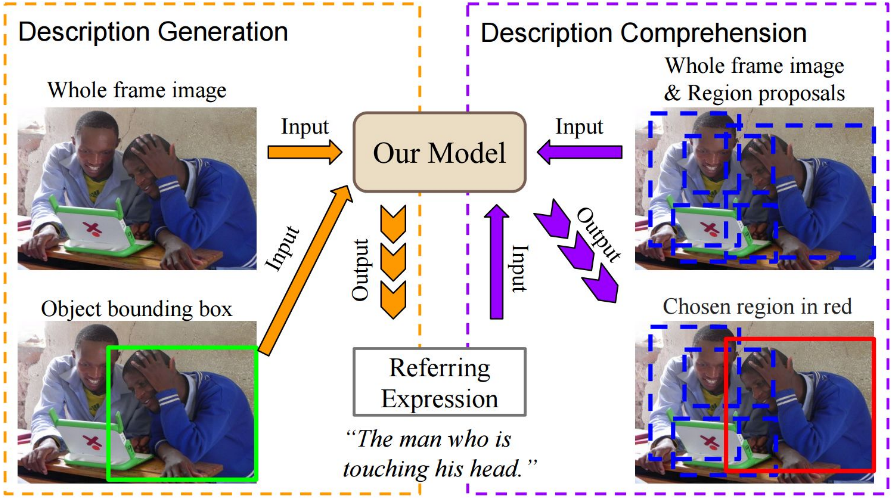
Image caption	**Referring expression**
A man playing soccer	The goalie wearing an orange and black shirt

Goal

Good referring expression -

- Uniquely describes the relevant region or object within its context
- A listener can comprehend and then recover the location of the described object/region

Consider two problems - 1) Description generation 2) Description comprehension



Dataset construction

For each image in MS-COCO dataset, an object is selected if

- There are between 2 and 4 instances of the same object type in the image
- Objects’ bounding boxes occupy at least 5% of image area

Descriptions were generated and verified using MechTurk. Dataset denoted as Google Refexp (G-Ref)

The black and yellow backpack sitting on top of a suitcase.

A yellow and black back pack sitting on top of a blue suitcase.

A girl wearing glasses and a pink shirt.

An Asian girl with a pink shirt eating at the table.

A boy brushing his hair while looking at his reflection.

A young male child in pajamas shaking around hairbrush in the mirror.

An apple desktop computer.

The white IMac computer that is also turned on.

A woman in a flowered shirt.

Woman in red shirt.

The woman in black dress.

A lady in a black dress cuts a wedding cake with her new husband.

Tasks

- Generation - Given image I, a target region R (through bounding box), generate referring expression S* such that S* = $argmax_S$ p(S|R, I) where S is a sentence. Used beam search of size 3

- Comprehension - Generate set *C* of region proposals and select region R* = $argmax_{R \epsilon C}$ p(R|S, I)

$$p(R|S,I) = \frac{p(S|R,I)p(R|I)}{\sum_{R' \in C} p(S|R',I)p(R'|I)}.$$

Assuming uniform prior for p(R|I), R* = $argmax_{R \epsilon C}$ p(S|R, I)

At test time, generate proposals using multibox method, classify each proposal into one of the MS-COCO categories and discard those with low scores to get set *C*.

Baseline

Similar to image captioning models. To train the baseline model, minimize

$$J(\theta) = -\sum_{n=1}^{N} \log p(S_n | R_n, I_n, \theta)$$

Model architecture -

- Use last 1000-d layer of pretrained VGGNet to represent the image and the region.
- Additional 5-d feature $[x_{tl}/W, y_{tl}/H, x_{br}/W, y_{br}/H, s_{bbox}/s_{image}]$ to encode relative size and location of the the region. x_{tl}, y_{tl}, x_{br}, y_{br} - top-left and bottom right coordinates of the bounding box, s - area, H,W - height and width of the image
- This 2005-d vector is given as input at every time step to an LSTM along with a 1024-d word embedding of the word at previous time step.

Proposed method

The baseline method generates expressions based only on the target object (and some context) but does not provide any incentive to generate discriminative sentences.

Discriminative (MMI) training

Minimize,

$$J'(\theta) = -\sum_{n=1}^{N} \log p(R_n | S_n, I_n, \theta),$$

where

$$\log p(R_n | S_n, I_n, \theta) = \log \frac{p(S_n | R_n, I_n, \theta)}{\sum_{R' \in \mathcal{C}(I_n)} p(S_n | R', I_n, \theta)}$$

Equivalent to maximizing mutual information

$$\mathrm{MI}(S, R) = \log \frac{p(S, R)}{p(R)p(S)} = \log \frac{p(S|R)}{p(S)}$$

R_n - ground truth region, R’ - any region. This method is called MMI - SoftMax

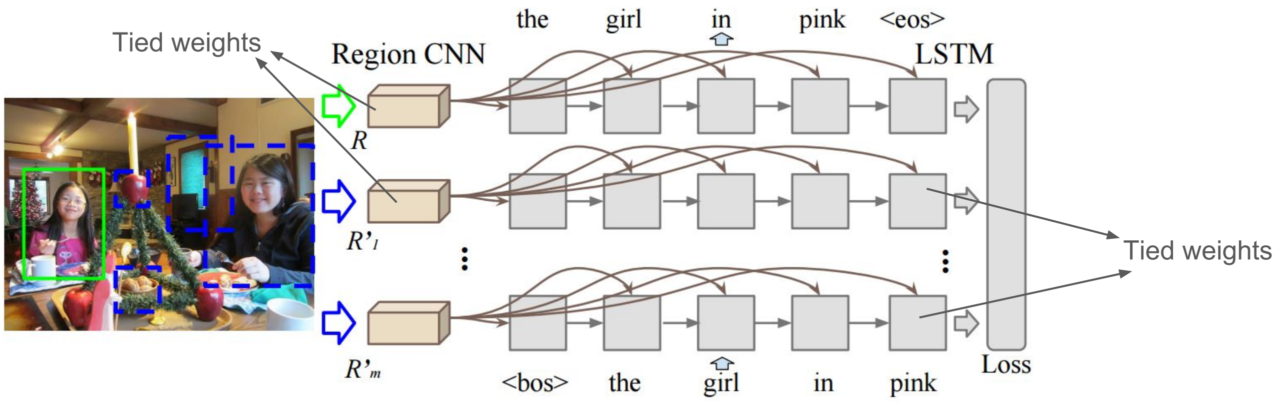
Proposed approach

Intuition - Penalize the model if the generated expression could also be plausible for some other region in the same image

Selecting proposal set *C* during training

- Easy ground truth negatives - All ground truth bounding boxes in the image
- Hard ground truth negatives - Ground truth bounding boxes belonging to the same class as target
- Hard multibox negatives - Multibox proposals with same predicted object labels as target

5 random negatives for each target



Proposed approach

MMI-Max Margin

$$J''(\theta) = -\sum_{n=1}^{N} \{\log p(S_n|R_n, I_n, \theta) -$$
$$\lambda \max(0,\ M - \log p(S_n|R_n, I_n, \theta) + \log p(S_n|R'_n, I_n, \theta))\}$$

- For computational reasons, use the max margin formulation above
- Has similar effect - penalty if difference between log probabilities of ground truth and negative regions is smaller than M
- Requires comparison between only two images (GT + one negative), thereby allowing larger batch sizes and more stable gradients.

Results

Proposals	GT		Multibox	
Descriptions	GEN	GT	GEN	GT
ML (baseline)	0.803	0.654	0.564	0.478
MMI-MM-easy-GT-neg	0.851	0.677	0.590	0.492
MMI-MM-hard-GT-neg	**0.857**	**0.699**	0.591	0.503
MMI-MM-multibox-neg	0.848	0.695	**0.604**	**0.511**
MMI-SoftMax	0.848	0.689	0.591	0.502

Using GT or multibox proposals at test time

Ground truth sentence (comprehension task)

Generated sentence (generation task)

- Proposed approaches perform better
- Maximum margin performs better than SoftMax
- Better to train using multibox negatives when testing on multibox proposals
- Comprehension easier when using generated sentences than ground truth sentences. Intuitively, a model can ‘communicate’ better with itself using its own language than with others

Results

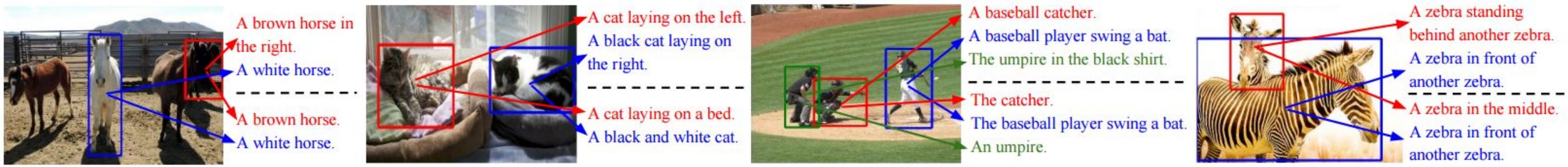
- Previous results were on the UNC-Ref-Val dataset, which was used to select the best hyperparameter settings for all methods.
- Results of MMI-MM-multibox-neg (full model) on other datasets are also better than baseline
- Human evaluation - % descriptions evaluated as better or equal to human captions

  Baseline - 15.9%   Proposed - 20.4%

Proposals	GT		multibox	
Descriptions	GEN	GT	GEN	GT
G-Ref-Val				
Baseline	0.751	0.579	0.468	0.425
Full Model	**0.799**	**0.607**	**0.500**	**0.445**
G-Ref-Test				
Baseline	0.769	0.545	0.485	0.406
Full Model	**0.811**	**0.606**	**0.513**	**0.446**
UNC-Ref-Val				
Baseline	0.803	0.654	0.564	0.478
Full Model	**0.848**	**0.695**	**0.604**	**0.511**
UNC-Ref-Test				
Baseline	0.834	0.643	0.596	0.477
Full Model	**0.851**	**0.700**	**0.603**	**0.518**

Qualitative Results

Generation



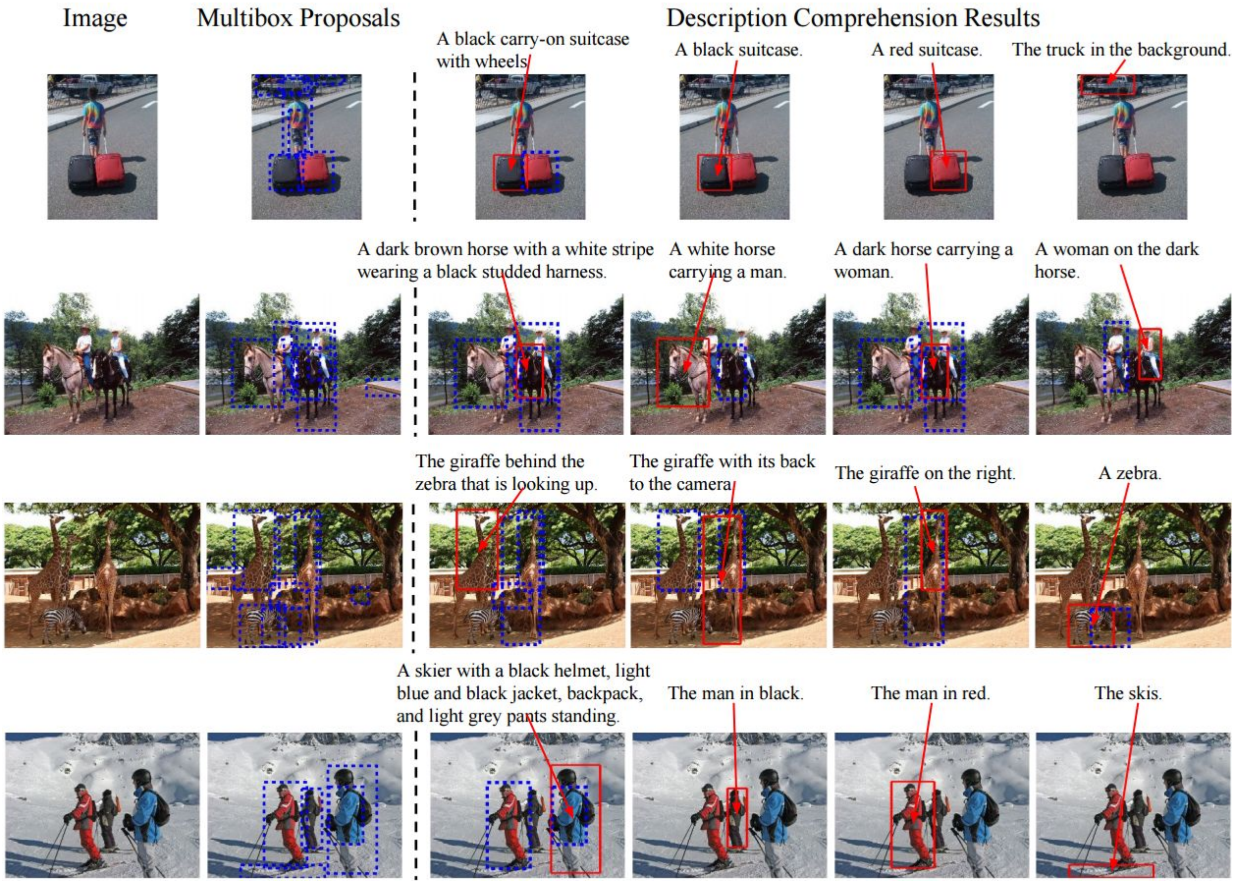
- Descriptions generated by the baseline and the proposed approach are below and above the dashed line respectively
- Proposed approach often removes ambiguity by providing direction/spatial cues such as left, right, behind

Qualitative Results

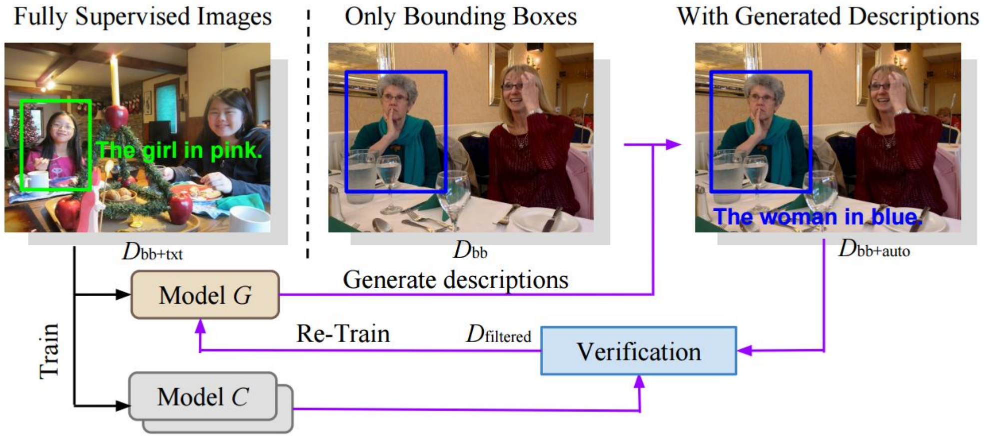
Comprehension

- Col 1: Test image
- Col 2: Multibox proposals
- Col 3: GT description
- Cols 4-6: Probe sentences
- Red bounding box:
- Output bounding box using proposed approach

- Dashed blue bounding boxes (cols 4-6): Other bounding boxes within margin



Semi-supervised training



- D_{bb+txt} - Bounding boxes + text (small set) D_{bb} - Bounding boxes only (large set)
- Learn model G using D_{bb+txt}. Make predictions on D_{bb} to create $D_{bb+auto}$
- Train an ensemble of different models C on D_{bb+txt}
- Use model C to perform comprehension on $D_{bb+auto}$. If each ensemble model maps description to the correct object, keep it, else remove it
- Use $D_{bb+text} \cup D_{bb+auto}$ to retrain model G and repeat

Results

Proposals	GT		multibox	
Descriptions	GEN	GT	GEN	GT
G-Ref				
$D_{\mathrm{bb+txt}}$	0.791	0.561	**0.489**	0.417
$D_{\mathrm{bb+txt}} \cup D_{\mathrm{bb}}$	**0.793**	**0.577**	**0.489**	**0.424**
UNC-Ref				
$D_{\mathrm{bb+txt}}$	0.826	0.655	0.588	0.483
$D_{\mathrm{bb+txt}} \cup D_{\mathrm{bb}}$	**0.833**	**0.660**	**0.591**	**0.486**

Using GT or multibox proposals at test time

Ground truth sentence (comprehension task)

Generated sentence (generation task)