Redefine the Diversity of Image Captioning

Chuan Cen, Yuan Liang, Ruoyao Wang, Xuanyu Wang, Fei Yi

UNIVERSITY OF MICHIGAN

Introduction

Mapping from image to text is essentially “one-to-many”. The captions, in our opinion, can diverse in three dimensions:

1. **Content**: objects or regions to describe
2. **The level of Detail**: richness of captions to describe the content, usually reflected by length of generated sentences
3. **Form**: variety of sentence structure or word choice, under constraint of the same content and level of detail

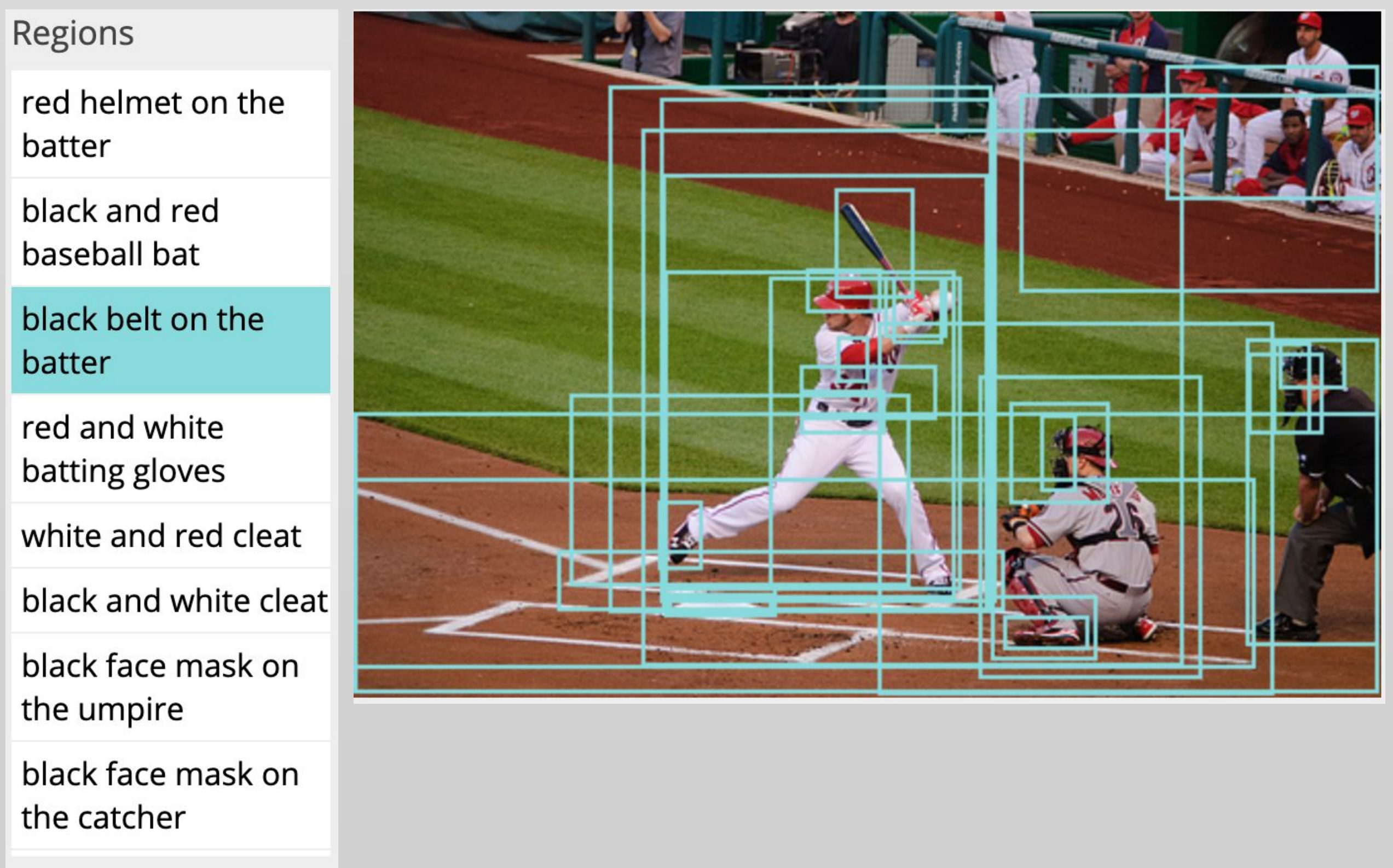
In this project we propose a model that is able to generate various captions for an image diverse in the three aspects.

Contribution

1. **New model for diverse caption generation**
   We introduce a new model architecture that is able to generate diverse captions in terms of content, detail and form.
2. **Novel Multimodal Similarity Loss**
   We introduce a novel multimodal similarity loss to measure difference between image feature and text with the following advantages.
   a) It can deal with similarity between multiple subregions of image with text.
   b) It achieved better performance than previous metrics.
3. **Spot problems when using Language Model as criterion**
   We figured out why the LM failed when used as a criterion training decoder, and give future directions for solving this issue.

Dataset

1. VISUALGENOME v1.2 with 108,249 images and around 4,000,000 regional captions and around 20000 full captions
2. MSCOCO 2017 with 600,000 image annotations



Regions

red helmet on the batter

black and red baseball bat

black belt on the batter

red and white batting gloves

white and red cleat

black and white cleat

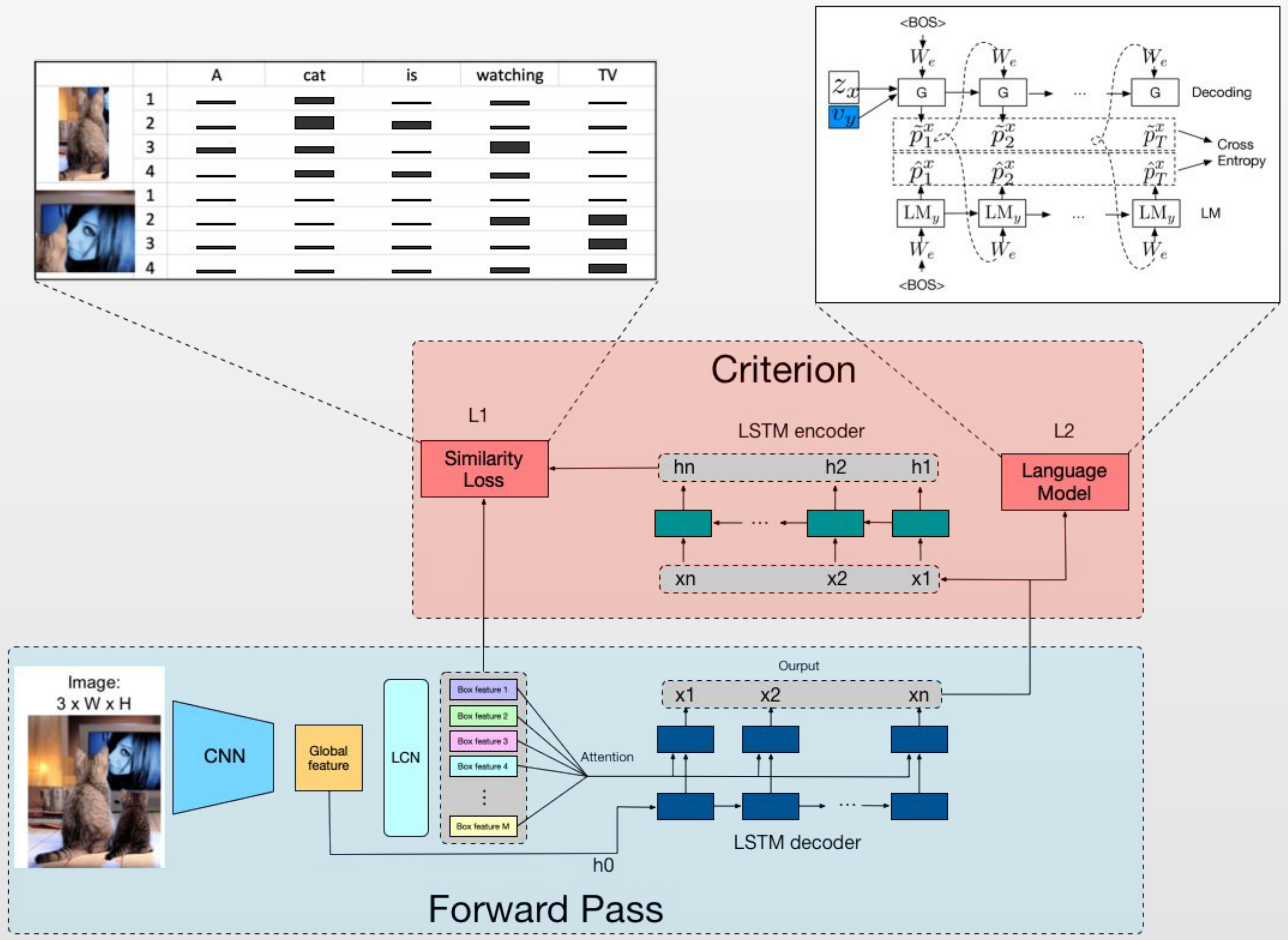
black face mask on the umpire

black face mask on the catcher

VISUALGENOME

COCO Common Objects in Context

Architecture



Methodology

1. **Overall**
   a) **Training phase**
      i. We first pre-train an LSTM Encoder using similarity loss between boxes features and regional captions.
      ii. After that we train our Language Model using 20,000 full captions from VisualGenome and 600,000 image annotations from MSCOCO.
      iii. Finally we train our LSTM Decoder whose outputs were supervised by the pre-trained LSTM Encoder and Language Model.
   b) **Testing phase**
      We draw various number of boxes from images to generate captions of different contents and details with LSTM Decoder.
2. **<u>For diversity 1&2:</u> *“Bidirectional Multi-region Multi-modal Similarity Loss”***

$$s = e \cdot \tilde{v}^T \in \mathbb{R}^{T \times (M \cdot H_r \cdot W_r)}$$

   a) **Attention on Sub-regions**

$$v_i' = \sum_{j=0}^{M \cdot H_r \cdot W_r} \alpha_{i,j} \tilde{v}_j \quad \alpha_{i,j} = \frac{\exp(s_{i,j})}{\sum_{k=0}^{M \cdot H_r \cdot W_r} \exp(s_{i,k})}$$

   b) **Attention on Words**

$$e_m' = \sum_{i=0}^{T} \beta_{m,i} e_i \qquad \beta_{m,i} = \sum_{u \in Reg^{(m)}} \frac{\exp(s_{i,u})}{\sum_{l \in Reg^{(m)}} \sum_{k=0}^{T} \exp(s_{k,l})}$$

$$R(Q,D) = \log\left(\sum_{i=0}^{T-1} \exp(R_e(\tilde{v}_i', e_i))\right) + \log\left(\sum_{m=0}^{M-1} \exp(R_v(v_m, e_m))\right)$$

$$P(D_i|Q_i) = \frac{\exp(R(Q_i, D_i))}{\sum_{j=1}^{N} \exp(R(Q_i, D_j))}$$

   c) **Similarity Loss**

$$\mathcal{L}_1' = -\sum_{i=1}^{N} P(D_i|Q_i) - \sum_{i=1}^{N} P(Q_i|D_i)$$

   d) **Regularization Term**

$$\mathcal{L}_{reg} = \sum_{i=1}^{N} \|\beta^{(i)}\beta^{(i)T} - diag(\beta^{(i)}\beta^{(i)T})\|_F$$

3. **<u>For diversity 3:</u> *Gumbel-softmax & beam search***
   a) Gumbel-softmax is a technique we used to solve gradient backpropagation issue, but we also found it useful when we want to sample diverse captions.
   b) Beam search iteratively generate multiple text candidates from decoder, hence increase diversity.
4. **<u>For fluency:</u> *Language Model Loss***

$$\mathcal{L}_{LM} = \mathbb{E}_{y \sim Y}[-\log p_{LM}(y)] \approx \frac{1}{N} \sum_{i=1}^{N} -\log p_{LM}(y^{(i)})$$

Results & Analysis

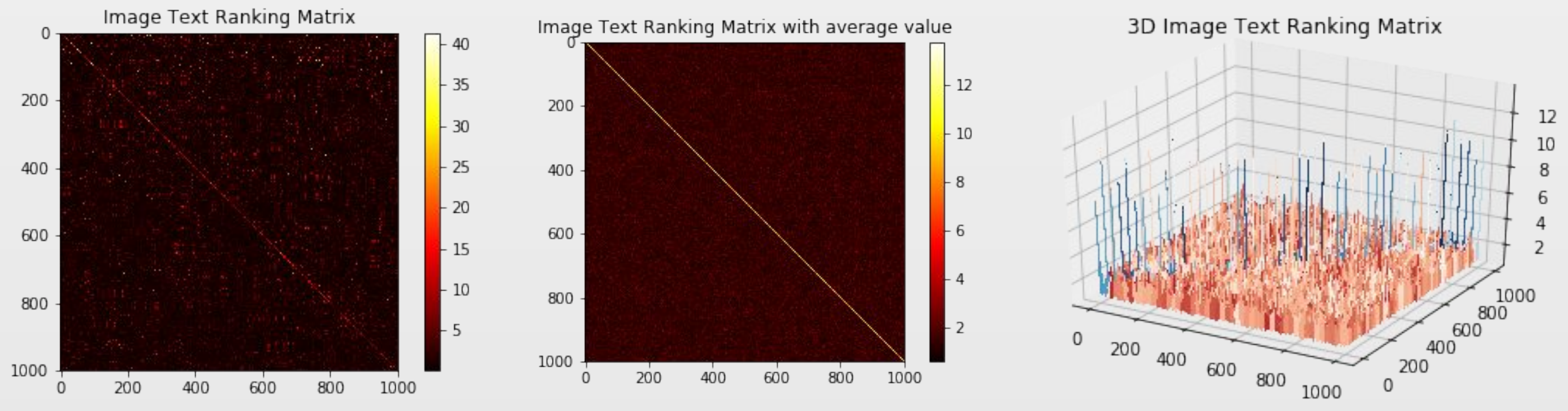
1. Similarity loss performance

1). Image sentence ranking experiment result

data	model	Image annotation R@1	R@5	R@10	Med r	Image search R@1	R@5	R@10	Med r
train	DAMSM	0.041	0.162	0.249	68.3	0.037	0.140	0.239	69.5
	Our Model	**0.098**	**0.316**	**0.477**	**29.1**	**0.088**	**0.308**	**0.452**	**30.206**
test	DAMSM	0.032	0.153	0.213	74.9	0.027	0.132	0.225	76.3
	Our Model	**0.081**	**0.27**	**0.415**	**38.615**	**0.069**	**0.263**	**0.397**	**39.824**

Table 1: Image-Sentence ranking experiment results. R@K is Recall@K (high is good). Med r is the median rank (low is good). In the table, The size of sample training and testing data set are both 1000. And the statistics are obtained from 5 group of sample data

2). Image sentence ranking matrix visualization

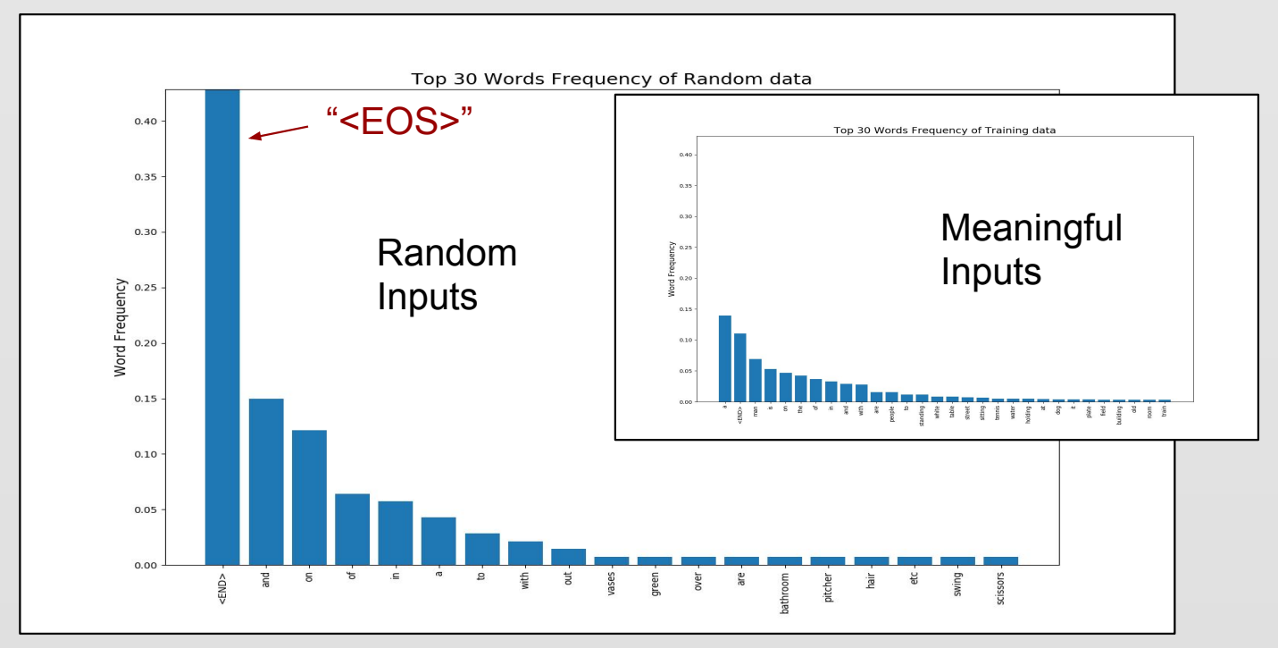


2. Why language model collapse?

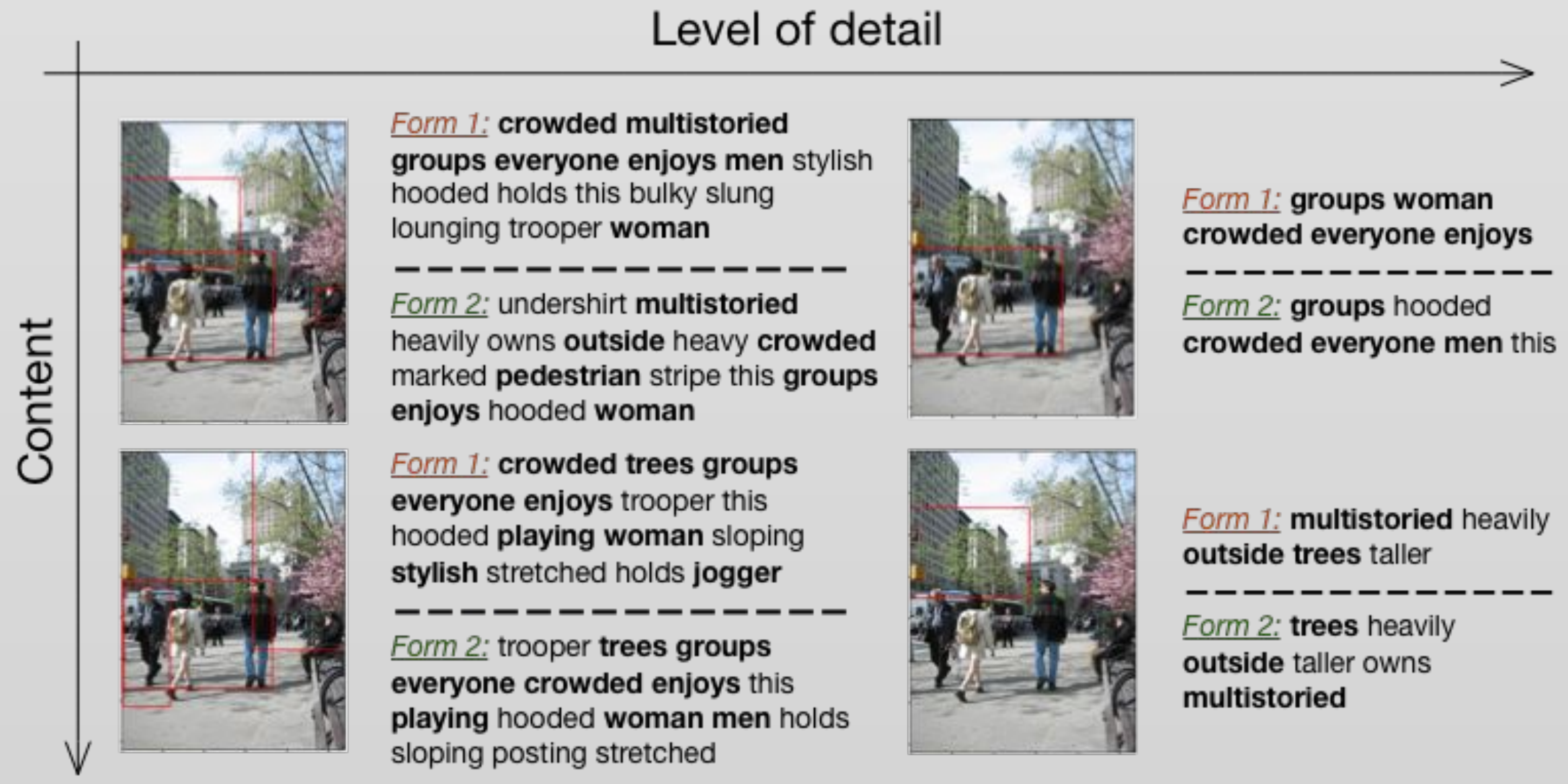
Language model trained with real sentences tends to output “<EOS>” when fed with random word sequences, hence gives wrong supervision signals when training the LSTM decoder and causes failures...

Input Sequence:
apricots conifer plaques beautiful nowhere against button pierced seasoning bandage backround males

LM suggestions of next word:
and and and and in a and <END> <END> <END> <END>



3. Image caption sample results *(trained without LM)*



Form 1: **crowded multistoried groups everyone enjoys men** stylish hooded holds this bulky slung lounging trooper **woman**

Form 2: undershirt **multistoried** heavily owns **outside** heavy **crowded** marked **pedestrian** stripe this **groups enjoys** hooded **woman**

Form 1: **groups woman crowded everyone enjoys**

Form 2: **groups** hooded **crowded everyone men** this

Form 1: **crowded trees groups everyone enjoys** trooper this hooded **playing woman** sloping **stylish** stretched holds **jogger**

Form 2: trooper **trees groups everyone crowded enjoys** this **playing** hooded **woman men** holds sloping posting stretched

Form 1: **multistoried** heavily **outside trees** taller

Form 2: **trees** heavily **outside** taller owns **multistoried**

4.Image caption evaluation

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	CIDER	ROUGE-L	METEOR
Train	1.20E-03	9.52E-11	2.09E-13	1.08E-14	1.46E-02	1.18E-02	6.39E-03
Test	6.35E-03	6.77E-11	1.64E-13	8.93E-15	8.17E-03	6.73E-03	4.44E-03

Table 2: Evaluation of sample results using metric BLEU(1-4), CIDER, ROUGE-L and METEOR from training set and validation set

Conclusion

1. Our similarity loss achieved ***significantly better performance*** in image-text ranking task compared with referred previous work (DAMSM).
2. The ***LM loss will collapse*** when fed with random or noisy sentences, hence not a ideal criterion for natural language generation.
3. Our trained model was able to ***generate relevant words*** to the boxes selected, but not logical sentences due to LM failure.