Joint Representation for Image and text

Dexiong Chen and Xi Shen

Objective : multimodal learning for image and text

- Image search: retrieving images for a given text description
- Image annotation : searching descriptions for a given image
- Image description : generating descriptions for a given image

Difficult Task

- Integrating vision, language understanding and learning

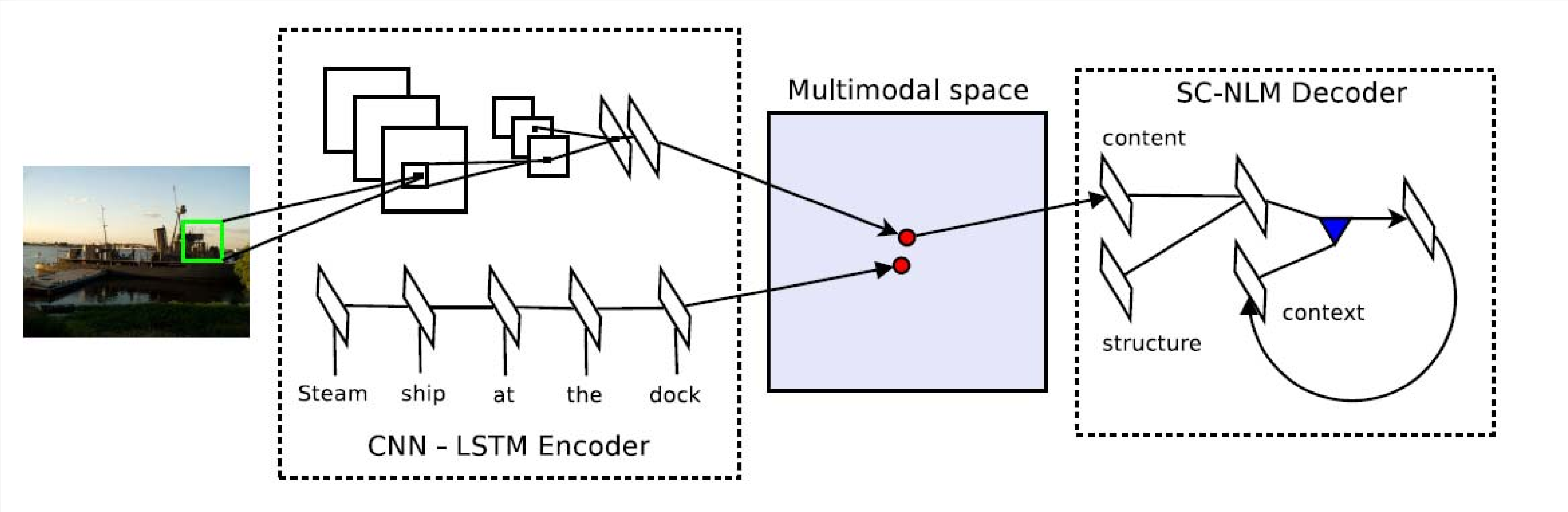
Database : Flickr [M. Hodosh and al. 2013]

- Each image contains five descriptions

- A football player pauses during a game.
- A football player wears a green jersey with the number "4" on it.
- Greenbay football player is being handed a towel on the field.
- Green Bay Packer player cooling off
- Someone takes a cloth off of a Green Bay Packers football player.

- Flickr8K : 6000, 1000, 1000 for train, validation and test.
  Flickr30K : 28000, 1000, 1000 for train, validation and test

Previous work : MNLM model [R. Kiros and al. 2014]



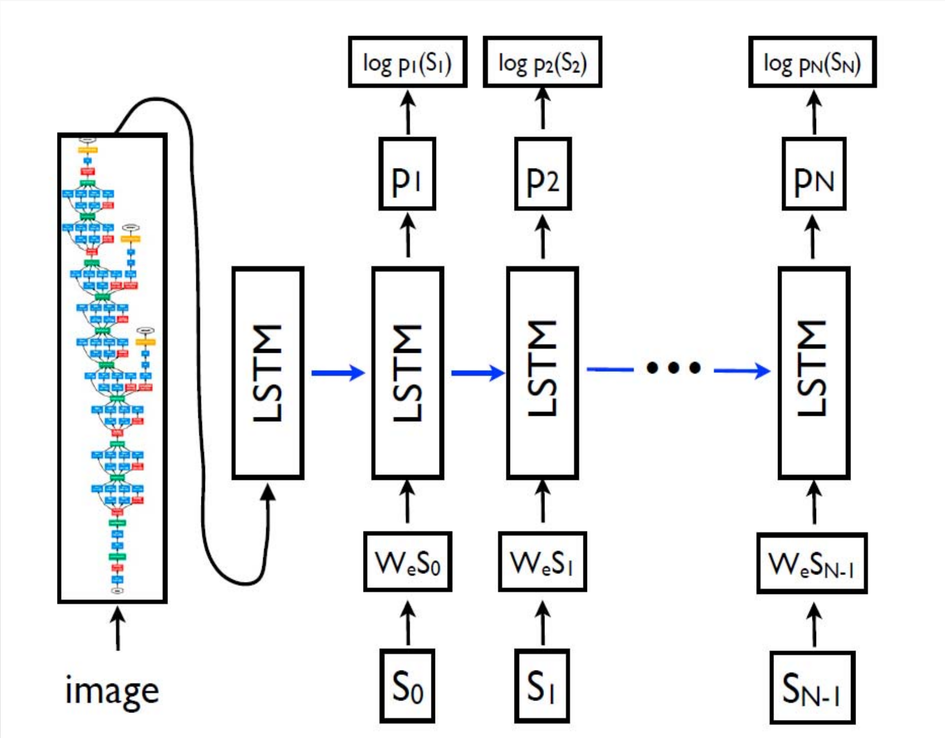
Encoder loss (kiros loss) :

$$\min_{\boldsymbol{\theta}} \sum_{\mathbf{x}} \sum_{k} \max\{0, \alpha - s(\mathbf{x}, \mathbf{v}) + s(\mathbf{x}, \mathbf{v}_k)\} + \sum_{\mathbf{v}} \sum_{k} \max\{0, \alpha - s(\mathbf{v}, \mathbf{x}) + s(\mathbf{v}, \mathbf{x}_k)\}$$

Performance of MNLM: image search and image annotation results

Model	Image Annotation				Image Search			
	R@1	**R@5**	**R@10**	**Med r**	**R@1**	**R@5**	**R@10**	**Med r**
Flickr8K								
DeFrag	12.6	32.9	44.0	14	9.7	29.6	42.5	15
m-RNN-AlexNet	14.5	37.2	48.5	11	11.5	31.0	42.4	15
NIC	20	-	61	6	19	-	64	5
MNLM	18.0	40.9	55.0	8	12.5	37.0	51.5	10
Our reimplemented MNLM	17.8	46.9	59.4	6.5	15.2	38.8	52.3	9
Flickr30K								
DeFrag	16.4	40.2	54.7	8	10.3	31.4	44.5	13
m-RNN-AlexNet	18.4	40.2	50.9	10	12.6	31.2	41.5	16
m-RNN-VggNet	35.4	63.8	73.7	3	22.8	50.7	63.1	5
NIC	17	-	56	7	17	-	57	7
DeepVS	22.2	48.2	61.4	4.8	15.2	37.7	50.5	9.2
MNLM	23.0	50.7	62.9	5	16.8	42.0	56.5	8
Our reimplemented MNLM	24.1	51.2	62.9	5	19.0	44.4	57.7	7

Previous work : NIC model [O. Vinyals and al. 2015]



Log-perplexity (log-ppl) loss:

$$L(I, S) = -\sum_{t=1}^{N} \log p_t(S_t)$$

Advantage: Enable text generation

Performance of NIC: Bleu scores for image description [K. Papineni and al. 2002]

$$\mathrm{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases}$$

$$\mathrm{BLEU} = \mathrm{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

c : length of candidate sentence

R : length of reference sentence

N = 4

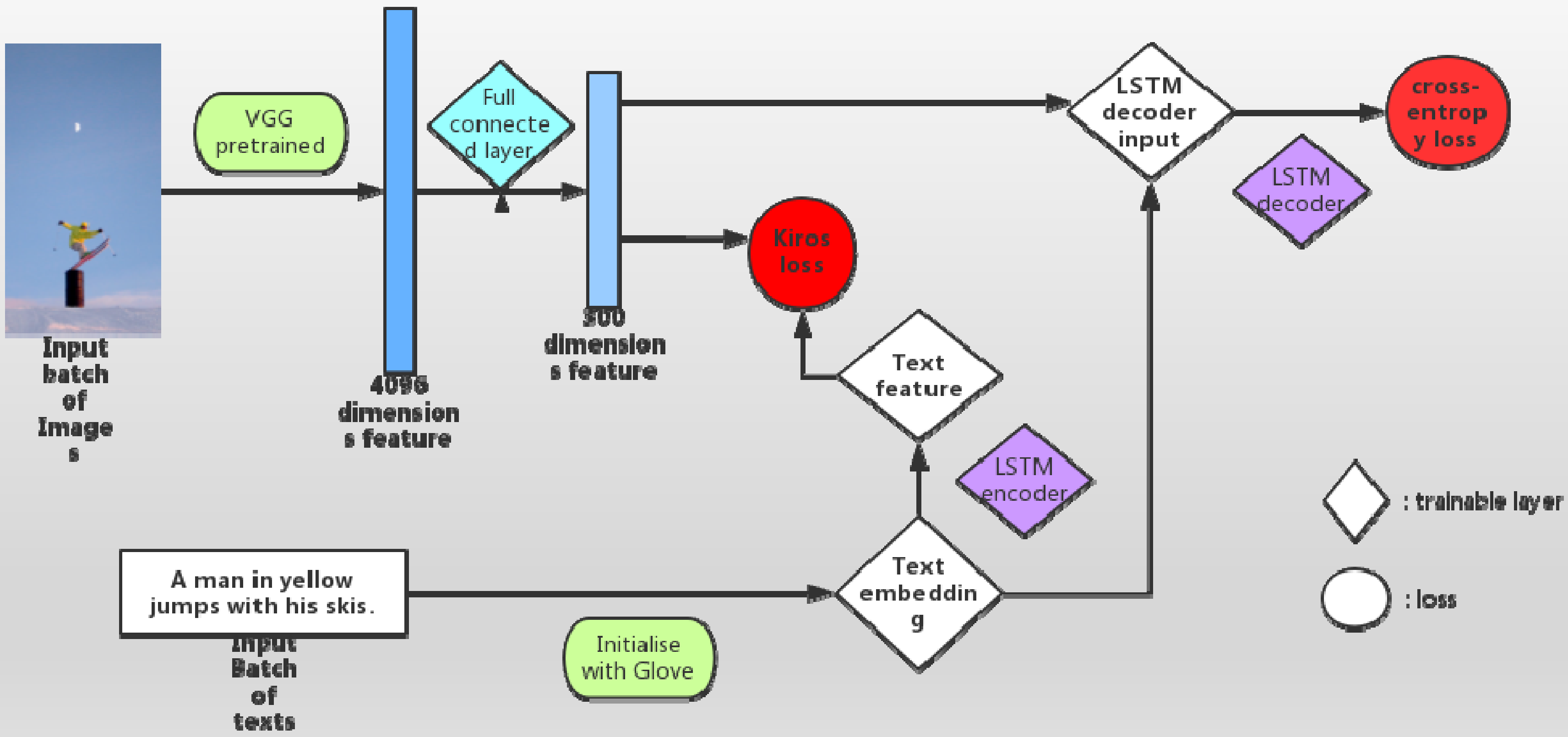
wn = ¼

pn : n gram precision

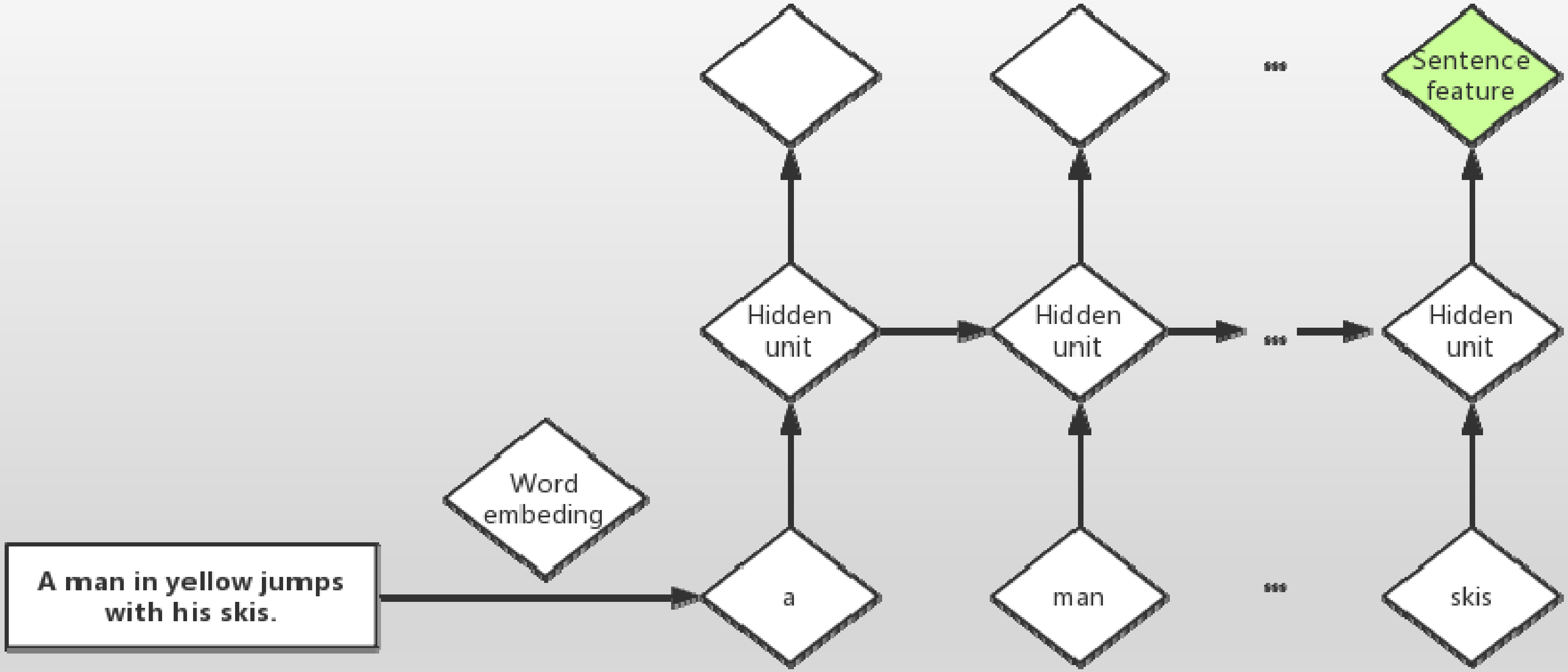
Motivation: can image features learned by MNLM be used to generate sentence ?

- Use latent space features to train NIC -> meaningful descriptions
- High probability to get same descriptions for similar images -> overfitting
- Why not combine two models ? -> Our encoder-decoder model

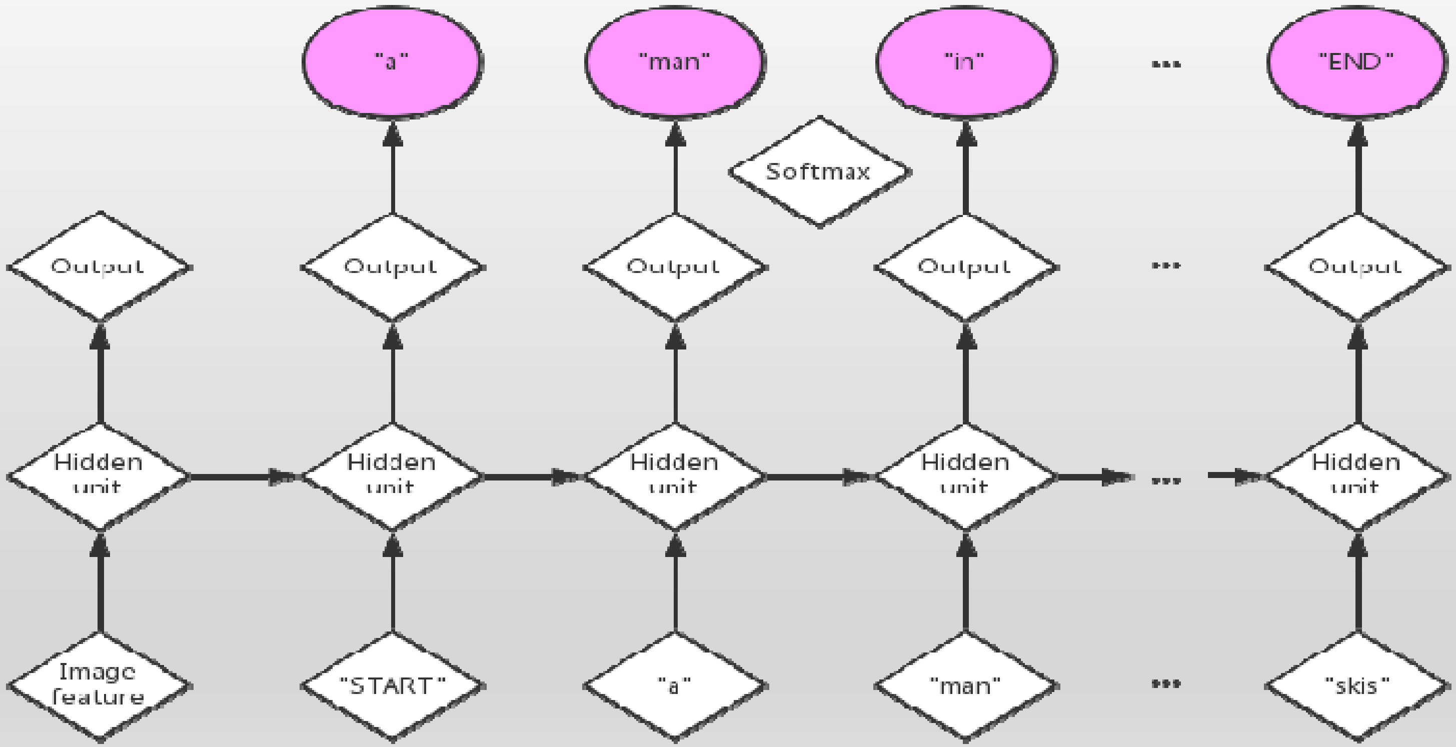
Our model : combination of the both models



Detail of the model (1) : LSTM [S. Hochreiter and J. Schmidhuber, 1997] for encoder



Detail of the model (2) : LSTM [S. Hochreiter and J. Schmidhuber, 1997] for decoder



Detail of the model (3)

- Total loss = *lambda* * kiros_loss + log-ppl_loss, *lambda* is determined by experiments

- Decoder LSTM enables to generate image describtion

- Image retrieval and image annotation can be determined by finding the smallest loss of :

  Match_Error(Image, Text) = 1 - cosine_similarity(Image_feature, Text_feature)

Implementation details

- Toolbox : Theano and Keras [P.W.D. Charles 2013]
- Optimization method : rmsprop [T Tieleman and al. 2012]
- Implemented model : MNLM, NIC and the combined model
- Early stop, dropout [N Srivastava and al. 2014] etc.
- Besides the usual hyper-parameters, our model contains extra hyper-parameters such as lambda etc.
- Run more than 80 models.
- Train and validate on the same loss.
- Beam search for sentence generation, match_error for image annotation and image retrieval

Image search and annotation results

Flickr 8k										
Model	Image search					Image annotation				
	R@1	R@5	R@10	Median	Mean	R@1	R@5	R@10	Median	Mean
MNLM [R. Kiros and al. 2014]	12.5	37.0	51.5	10	-	18.0	40.9	55.0	8	
NIC [O. Vinyals and al. 2015]	19	-	64	5	-	20	-	61	6	
Our reimplemented MNLM	15.2	38.8	52.3	9	34.7	17.8	46.9	59.4	6.5	36.8
Our model (Only Kiros loss)	17.54	44.06	58.63	7	27.96	22.70	50.80	65.20	5	31.575

Image search and annotation results

Flickr 30k								
Model	Image search				Image annotation			
	R@1	R@5	R@10	Median	R@1	R@5	R@10	Median
MNLM [R. Kiros and al. 2014]	16.8	42.0	56.5	8	23.0	50.7	62.9	5
NIC [O. Vinyals and al. 2015]	17	-	57	7	17	-	56	7
BRNN [A. Karpathy and al. 2015]	15.2	37.7	50.5	9.2	22.2	48.2	61.4	4.8
Our reimplemented MNLM	19.0	44.4	57.7	7	24.1	51.2	62.9	5
Our model (Only Kiros loss)	17.56	45.42	57.4	7	24.8	51.1	63.4	5

Image description results

Flickr 8k				
Model	BLEU			
	BLEU-1	BLEU-2	BLEU-3	BLEU-4
NIC [O. Vinyals and al. 2015]	63	41	27	-
BRNN [A. Karpathy and al. 2015]	57.9	38.3	24.5	16.0
Our model	60.0	40.9	27.1	17.6

Flickr 30k				
Model	BLEU			
	BLEU-1	BLEU-2	BLEU-3	BLEU-4
NIC [O. Vinyals and al. 2015]	66.3	42.3	27.7	18.3
BRNN [A. Karpathy and al. 2015]	57.9	36.9	24.0	15.7
Our model	62.7	41.4	27.4	17.8

Quantitative examples: retrieval examples

A group of dogs runs beside a pond through a field.

a boy wearing a teal shirt is riding a skateboard on a sidewalk


two brown dogs playfully fight in the snow 0.50041947
two brown dogs wrestle in the snow 0.52399729
a brown dog holding a huge stick in its mouth running in the snow 0.5615539
a black dog and a brown dog in snow 0.58192966
the brown dog running in the snow is carrying a large stick also covered with snow 0.58854175

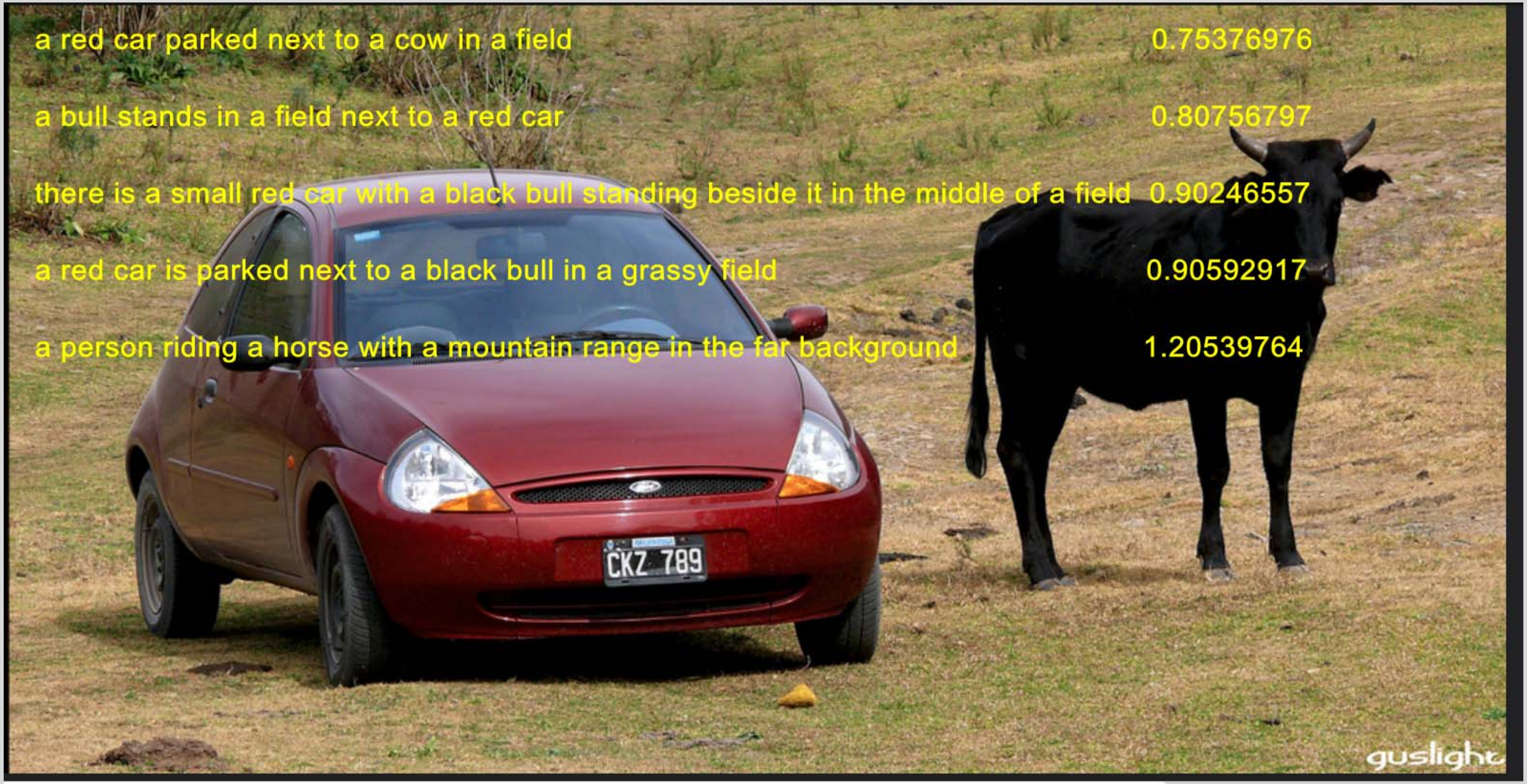
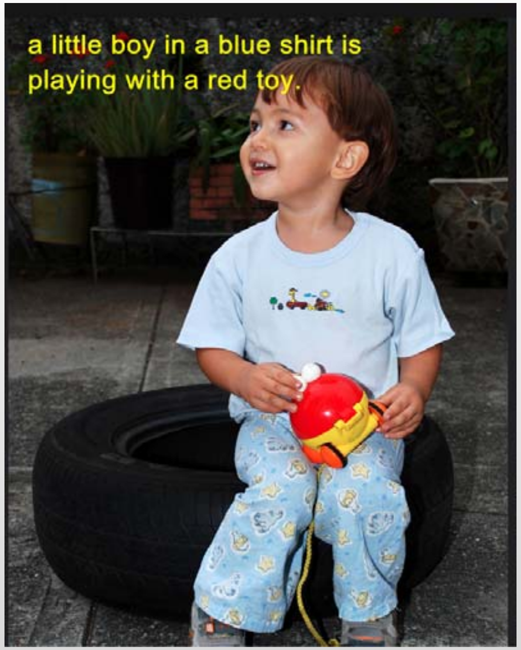

a red car parked next to a cow in a field 0.75376976
a bull stands in a field next to a red car 0.80756797
there is a small red car with a black bull standing beside it in the middle of a field 0.90246557
a red car is parked next to a black bull in a grassy field 0.90592917
a person riding a horse with a mountain range in the far background 1.20539764
CKZ 789
guslight

Image Description Examples




Two men stand outside of a building

a young boy jumping on a trampoline

a girl in a pink bikini is walking on the beach

Discussion

- NIC model has fine-tuned CNN, whereas our model does not.
- MNLM performs better with more data, since more difference between the unrelated image and sentence will be learnt
- The two parts (MNLM and NIC) in our model, share the same image feature.
- In our model, two losses can be viewed as a regularized term for another

- The cosine similarity enables to get good result, experiment shows the consideration of log-ppl loss can improve the image search result.

- *lambda* is not difficult to tune, in praticice, making the kiros loss and log-ppl loss converge to the same scale. (*lambda* * kiros loss ≈ log-ppl loss)

- Flickr8k allows us to choose and fix the best hyperparameters for other larger datasets.

- Dropout (in LSTM) is helpful to avoid overfitting.

- Initializing the word embedding with Glove or not does not impact the results

- Breakdown in correlation between the validation set log-likelihood and BLEU score (same observation as [K. Xu et al. 2015])

- Take about 2 hours to train flickr8k and 10 hours to train flickr30k on a GPU NVIDIA GeForce GT 750M

Future work

- Analysis and illustration of learned embedding space for words, using techniques like PCA.
- Instead of only considering kiros loss for ranking, we can take into account of log-ppl loss.
- Using other metrics to evaluate generated descriptions, such as METEOR.
- Use a larger dataset: MS COCO, study the transfer learning and scalability of our model.
- Check if fine-tuning CNN can improve results.

Conclusion

- An encoder-decoder model which learns in the meantime multimodal representation and text reconstruction.
- Combination of two models does improve performance.
- A simple but powerful model which accomplishes several tasks at the same time.