

# Describing Images Using Natural Language

T. Dieltjens    W. Vekemans

Department of Computer Science  
KU Leuven

11 December 2015

# Outline

Motivation

Objectives

Background

Related Work

Datasets

Methodology

Preliminary Results

Future Work

Conclusion

# Outline

Motivation

Objectives

Background

Related Work

Datasets

Methodology

Preliminary Results

Future Work

Conclusion

# Motivation

Why image captioning?

- ▶ connects NLP and CV
- ▶ more than object detection
- ▶ image search
- ▶ visually impaired

# Outline

Motivation

**Objectives**

Background

Related Work

Datasets

Methodology

Preliminary Results

Future Work

Conclusion

# Objectives

- ▶ caption unseen images
- ▶ extend existing implementation
- ▶ semantic information
- ▶ improvement?



a man in a white shirt



construction workers working on a railroad tracks

# Outline

Motivation

Objectives

Background

Related Work

Datasets

Methodology

Preliminary Results

Future Work

Conclusion

# Background

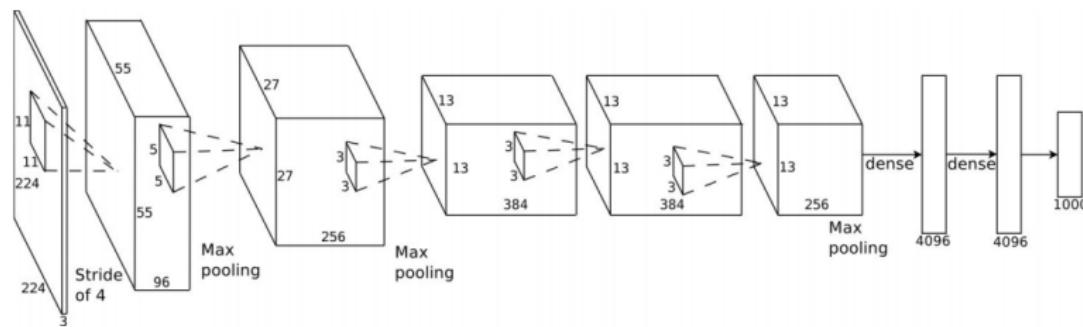
- ▶ CNN
- ▶ RNN
- ▶ LSTM
- ▶ LDA
- ▶ Stacked CCA

# Convolutional Neural Networks (CNN)

- ▶ trainable multistage architectures
- ▶ each stage outputs set of arrays called feature map
- ▶ different types of layers
- ▶ faster to train than feed forward
- ▶ unsupervised learning
- ▶ breakthrough in object recognition

# Convolutional Neural Networks (CNN)

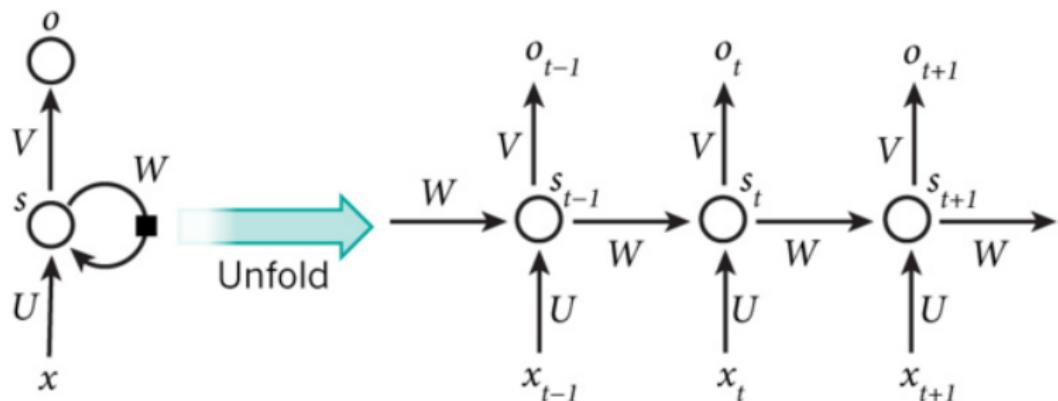
- ▶ ImageNet Classification
- ▶ deeper Models: VGGNet 16 layers  
[Simonyan and Zisserman, 2015]
- ▶ use of layers for other tasks



[Krizhevsky et al., 2012]

# Recurrent Neural Networks (RNN)

- ▶ predicting sequential data
- ▶ encode temporal information
- ▶ use as language model
- ▶ words represented as word vectors

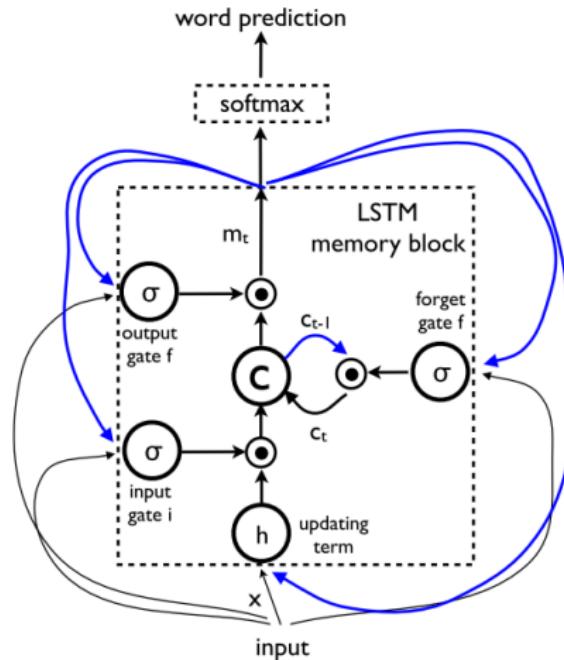


[LeCun et al., 2015]

# Long Short-Term Memory (LSTM)

[Hochreiter and Schmidhuber, 1997]

- ▶ RNN with memory cells
- ▶ capture long-term dependencies



[Vinyals et al., 2015]

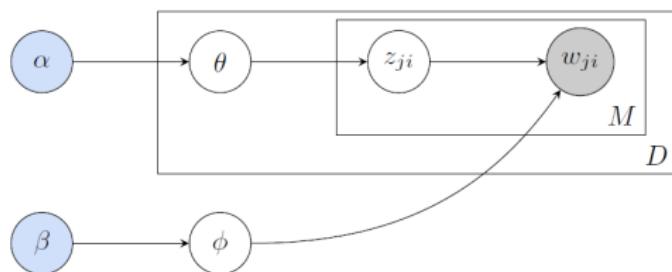
# LDA

[Blei et al., 2012]

- ▶ generative probabilistic model for discrete data
- ▶ documents have topic distribution
- ▶ topics have word distribution
- ▶ train with e.g. Gibbs sampling

# LDA

- ▶  $\theta$ : per-document topic distribution
- ▶  $\phi$ : per-topic word distribution
- ▶  $z$ : topic
- ▶  $w$ : word



$$P(\text{player} | \text{doc}_i) = P(\text{topic}_{\text{sport}} | \text{doc}_i) P(\text{player} | \text{topic}_{\text{sport}})$$

# Stacked CCA

[Gong et al., 2014]

- ▶ CCA
  - ▶ canonical correlation analysis
  - ▶ maps two matrices ( $A, B$ ) into intermediate representation
  - ▶ maximum correlation between projections of  $A$  and  $B$
- ▶ improve embeddings with extra information
  - ▶ CCA on extra dataset:  $A, B$
  - ▶  $\hat{X} = [X, \phi(XA)], \hat{Y} = [Y, \phi(YB)]$
  - ▶ another CCA on top

# Outline

Motivation

Objectives

Background

**Related Work**

Datasets

Methodology

Preliminary Results

Future Work

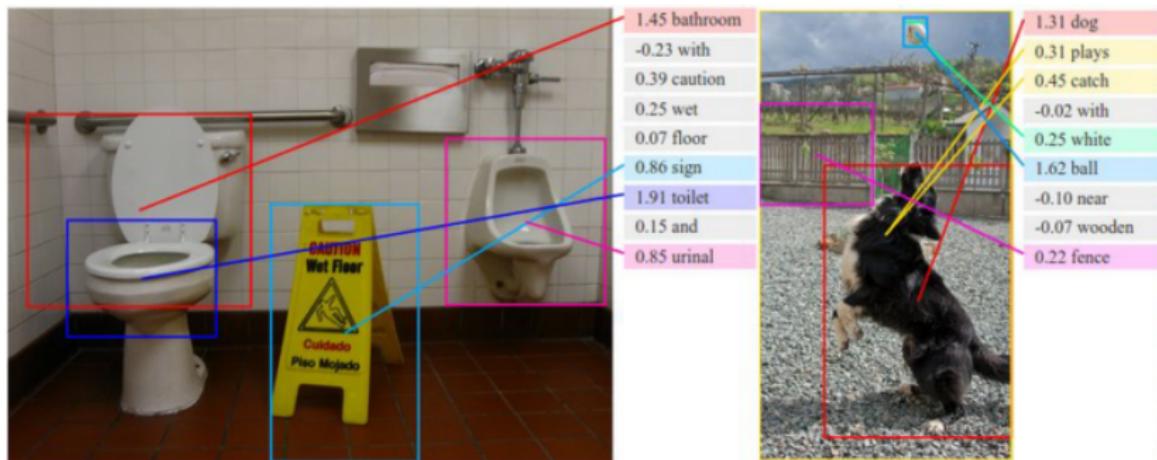
Conclusion

## Related Work

- ▶ Deep Visual-Semantic Alignments for Generating Image Descriptions [Karpathy and Fei-Fei, 2015]
- ▶ Show and Tell: A Neural Image Caption Generator [Vinyals et al., 2015]
- ▶ Guiding Long-Short Term Memory for Image Caption Generation [Jia et al., 2015]

# [Karpathy and Fei-Fei, 2015]

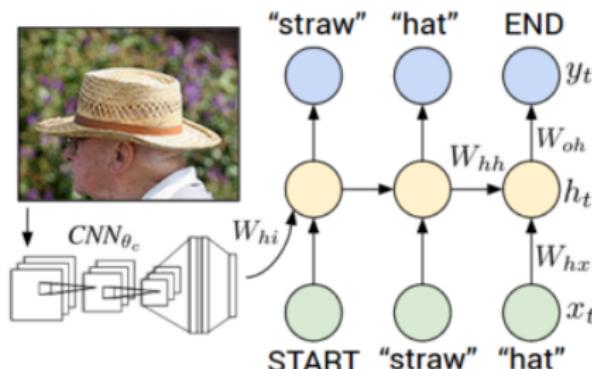
- ▶ two goals
  - ▶ dense image description based on alignments
  - ▶ describing a full image with one sentence
- ▶ alignment model
  - ▶ Region CNN (RCNN) to detect objects [Girshick et al., 2014]
  - ▶ bidirectional RNN to compute word representations
  - ▶ alignment objective



# [Karpathy and Fei-Fei, 2015]

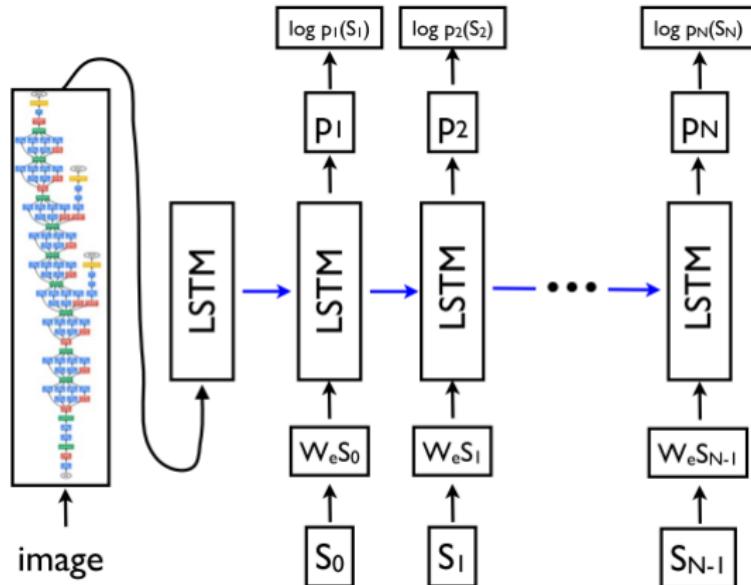
- ▶ Multimodal RNN
  - ▶ CNN
  - ▶ RNN
  - ▶ Softmax classifier
  - ▶ predicting a sentence with beam search

- ▶ Evaluation:
  - ▶ ranking image-sentence retrieval (recall)
  - ▶ sentence predictions (Bleu,METEOR,CIDEr)

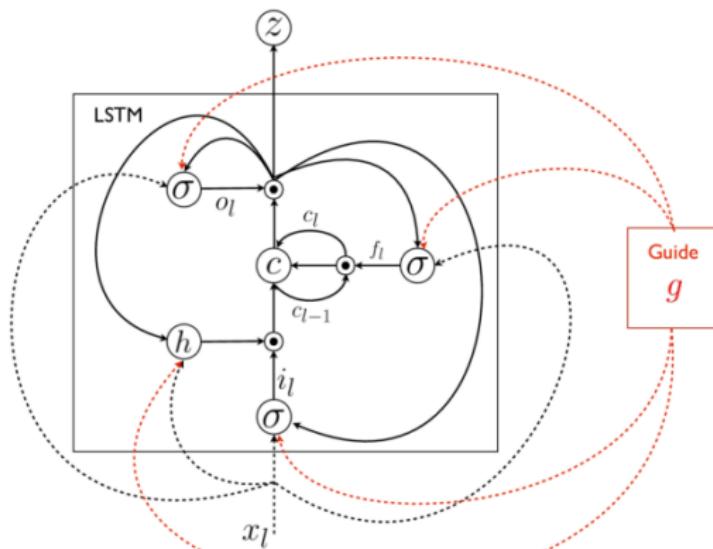


# Neural Image Caption (NIC) [Vinyals et al., 2015]

- ▶ CNN
- ▶ LSTM-based sentence generator
- ▶ train: maximize probability of generating the right caption
- ▶ inference: beam search or sampling



- ▶ extension of NIC implementation by Karpathy
- ▶ add semantic information as extra input to LSTM block (guided LSTM)
  - ▶ close sentences
  - ▶ CCA
  - ▶ image
- ▶ length normalization strategies for beam search



# Outline

Motivation

Objectives

Background

Related Work

**Datasets**

Methodology

Preliminary Results

Future Work

Conclusion

# Datasets

## Manually annotated datasets:

- ▶ Flickr30k [Young et al., 2014]
  - ▶ 31.783 images
  - ▶ 5 captions per image
- ▶ MS COCO [Lin et al., 2014]
  - ▶ Common Objects in COntext
  - ▶ over 330.000 images
  - ▶ 5 captions per image
  - ▶ non-iconic scenes



(a) Iconic object images

(b) Iconic scene images

(c) Non-iconic images

# Datasets

Flickr30k Entities [Plummer et al., 2015]

- ▶ 276.000 annotated bounding boxes
- ▶ corresponding fragments of captions



A man with **pierced ears** is wearing **glasses** and **an orange hat**.

A man with **glasses** is wearing **a beer can crotched hat**.

A man with **gauges** and **glasses** is wearing a **Blitz hat**.

A man in **an orange hat** staring at **something**.

A man wears **an orange hat** and **glasses**.



During a gay pride parade in an Asian city, some people hold up **rainbow flags** to show their support.

A group of youths march down a **street** waving **flags** showing a color spectrum.

Oriental people with **rainbow flags** walking down a **city street**.

A group of people walk down a **street** waving **rainbow flags**.

People are outside waving **flags**.

# Outline

Motivation

Objectives

Background

Related Work

Datasets

**Methodology**

Preliminary Results

Future Work

Conclusion

## Methodology

- ▶ modify NeuralTalk<sup>1</sup> code which implements both Karpathy and Vinyals
- ▶ Feedforward Sequential Memory Neural Networks without Recurrent Feedback [Zhang et al., 2015]
- ▶ LDA
  - ▶ learn LDA on training set
  - ▶ train perceptron to predict topic distribution on given image feature
  - ▶ predict topic distribution on unseen images with network
  - ▶ use topic distribution as additional input for RNN

---

<sup>1</sup><https://github.com/karpathy/neuraltalk>

## Methodology

- ▶ transform Flickr30kEntities to useful data
  - ▶ tf-idf representation for sentence snippets
  - ▶ CNN feature vectors for image snippets
- ▶ add semantic information with guided LSTM
  - ▶ stacked CCA based on Flickr30kEntities data
  - ▶ LDA

# Outline

Motivation

Objectives

Background

Related Work

Datasets

Methodology

Preliminary Results

Future Work

Conclusion

## Results from literature

	B1	B2	B3	B4
<b>Karpathy</b>	57.3	36.9	24.0	15.7
<b>Vinyals</b>	66.3	42.3	27.7	18.3
<b>Jia</b>	64.4	44.6	30.5	20.6
<b>Human</b>	68			

## Experimental Results

	B1	B2	B3	B4
<b>RNN</b>	65.1	43.1	29.0	18.9
<b>RNN + LDA</b>	64.5	43.1	27.9	17.6
<b>LSTM</b>	59.1	37.2	22.8	14.0
<b>Human</b>	68			

# Outline

Motivation

Objectives

Background

Related Work

Datasets

Methodology

Preliminary Results

Future Work

Conclusion

## Future Work

- ▶ finetune current networks
- ▶ finish gLSTM with Flickr30k Entities and LDA
- ▶ investigate dense captioning
  - ▶ Faster RCNN
  - ▶ Flickr30k Entities
- ▶ use Torch to make a faster implementation

# Outline

Motivation

Objectives

Background

Related Work

Datasets

Methodology

Preliminary Results

Future Work

Conclusion

# Conclusion

- ▶ working implementations
- ▶ similar/better results are possible
- ▶ new ideas to further investigate

# Questions?

# References |

- Blei, D. M., Ng, A. Y., and Jordan, M. I. (2012).  
Latent Dirichlet Allocation.  
*Journal of Machine Learning Research*, 3(4-5):993–1022.
- Girshick, R., Donahue, J., Darrell, T., Berkeley, U. C., and Malik, J. (2014).  
Rich feature hierarchies for accurate object detection and semantic segmentation.  
*Cvpr'14*, pages 2–9.
- Gong, Y., Wang, L., Hodosh, M., and Hockenmaier, J. (2014).  
Improving Image-Sentence Embeddings Using Large Weakly Annotated Photo Collections.  
*Computer VisionECCV*, pages 529–545.
- Hochreiter, S. and Schmidhuber, J. (1997).  
Long short-term memory.  
*Neural computation*, 9(8):1735–1780.
- Jia, X., Gavves, E., Fernando, B., and Tuytelaars, T. (2015).  
Guiding Long-Short Term Memory for Image Caption Generation.
- Karpathy, a. and Fei-Fei, L. (2015).  
Deep Visual-Semantic Alignments for Generating Image Des.  
*Cvpr2015*.

## References II

- Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012).  
ImageNet Classification with Deep Convolutional Neural Networks.  
*Advances In Neural Information Processing Systems*, pages 1–9.
- LeCun, Y., Bengio, Y., and Hinton, G. (2015).  
Deep learning.  
*Nature*, 521(7553):436–444.
- Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., and Dollár, P. (2014).  
Microsoft COCO: Common Objects in Context.
- Plummer, B. a., Wang, L., Cervantes, C. M., Caicedo, J. C., Hockenmaier, J., and Lazebnik, S. (2015).  
Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models.
- Simonyan, K. and Zisserman, A. (2015).  
Very Deep Convolutional Networks For Large-Scale Image Recognition.
- Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2015).  
Show and Tell: A Neural Image Caption Generator.

## References III

- Young, P., Lai, A., Hodosh, M., and Hockenmaier, J. (2014).  
From Image Descriptions to Visual Denotations: New Similarity Metrics for Semantic  
Inference over Event Descriptions.  
*Transactions of the Association for Computational Linguistics (TACL)*,  
2(April):67–78.
- Zhang, S., Jiang, H., Wei, S., and Dai, L. (2015).  
Feedforward Sequential Memory Neural Networks without Recurrent Feedback.