Towards Human Quality Image Captioning: Deep Semantic Learning of Text and Images

Xiaodong He

DLTC, Microsoft Research, Redmond, WA, USA

Invited talk at Deep Learning Workshop, 8 August 2015, San Francisco

Collaborators:

Hao Fang
Saurabh Gupta
Forrest Iandola
Rupesh Srivastava
Li Deng
Piotr Dollár
Jianfeng Gao
Margaret Mitchell
John Platt
Lawrence Zitnick
Geoffrey Zweig
Jacob Devlin
…

Why should vision people ever care about language?

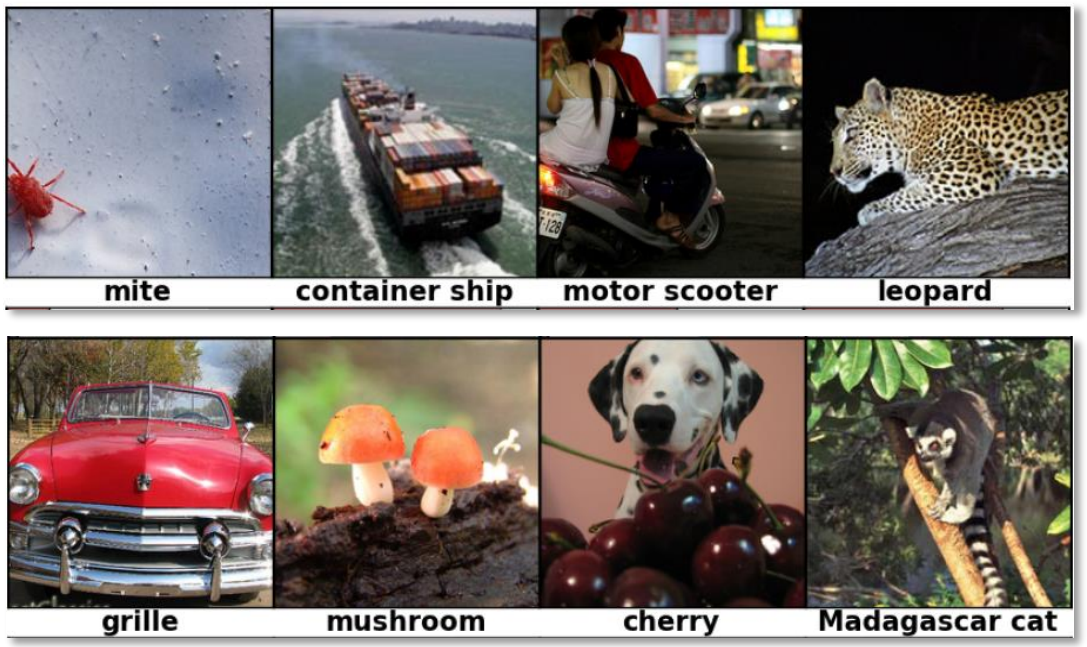
1. How to *teach* machines to understand images?

2. How to *test* if a machine understands an image or not?

How to teach machines to understand images?

For image classification, we can label each image by a category and train the machine to predict

E.g., ImageNet provides hundreds to thousands of images for each category, aka **synset**, in the WordNet.



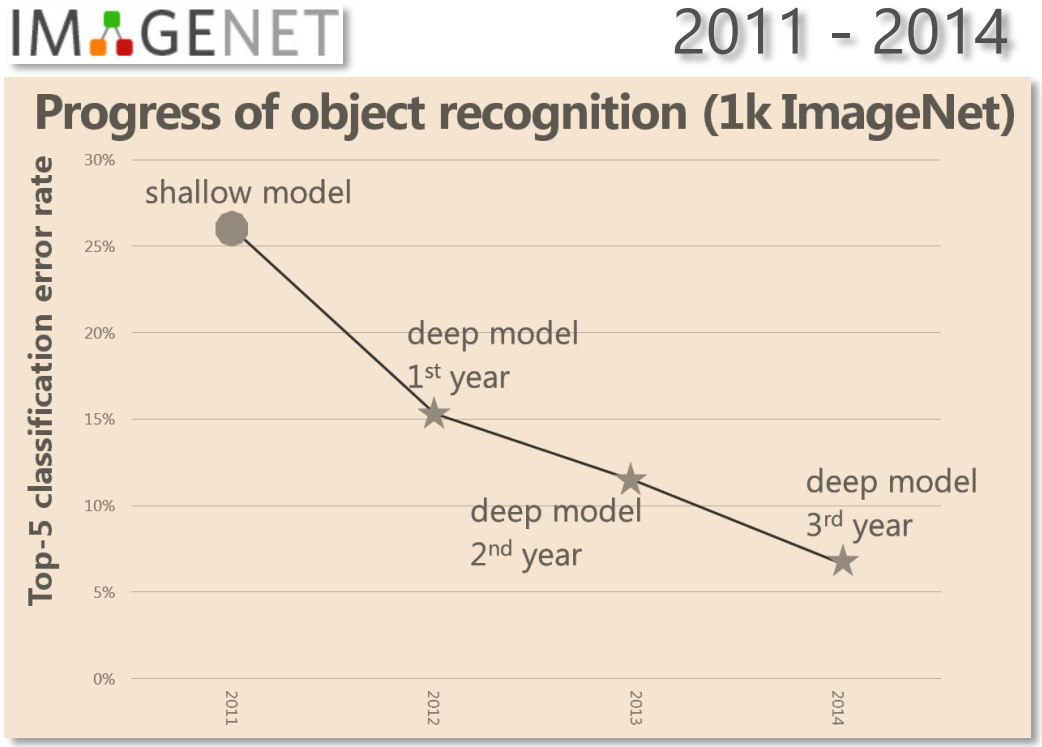
[Russakovsky, Deng, et al., 2014]

How to test if a machine understand a image?

For image classification, just check the prediction error rate

Dramatic progress in recent years thanks to deep CNN [LeCun, Bottou, Bengio, Haffner, 1998, Krizhevsky, Sutskever, Hinton, 2012].

First time surpassed human-level performance (top5 err < 5%) on ImageNet classification in 2015 [He, Zhang, Ren, Sun, 2015]



But for complex scenes with a rich context, not possible to define all fine-grained subtle differences by categorization.

The best supervision is a description in natural language

e.g., MS COCO dataset provides 5 descriptions for each image that has a rich content.

Each description is:

- a coherent story with clear semantic meaning.
- focused on salient info.
- reflecting the common sense.

Could be a big variety, though.

- a woman is playing a frisbee with a dog.
- a woman is playing frisbee with her large dog.
- a girl holding a frisbee with a dog coming at her.
- a woman kneeling down holding a frisbee in front of a white dog.
- a young lady is playing frisbee with her dog.

[Lin, et al., 2014]

How to know if a machine understands a complex scene?

-- let's do a Turing Test!

ask the machine to describe the image in human language and see whether it reads like generated by a human

a man riding a skateboard down a street.

Therefore, language understanding is a key part for building strong vision intelligence

- Serves as the natural supervision to teach the machine to understand images *as human do*
- Provide the natural *communication* between the human and the machine with vision intelligence

The MSR system

[Fang, Gupta, Iandola, Srivastava, Deng, Dollar, Gao, He, Mitchell, Platt, Zitnick, Zweig, "From Captions to Visual Concepts and Back," CVPR, June 2015]

Understand the image stage by stage:

Image word detection

Deep-learned features, applied to likely items in the image, trained to produce words in captions

Language generation

Maxent language model (MELM), trained on caption, conditional on words detected from the image

Global semantic re-ranking

Hypothetical captions re-ranked by deep-learned multimodal similarity model (DMSM) looking at the entire image

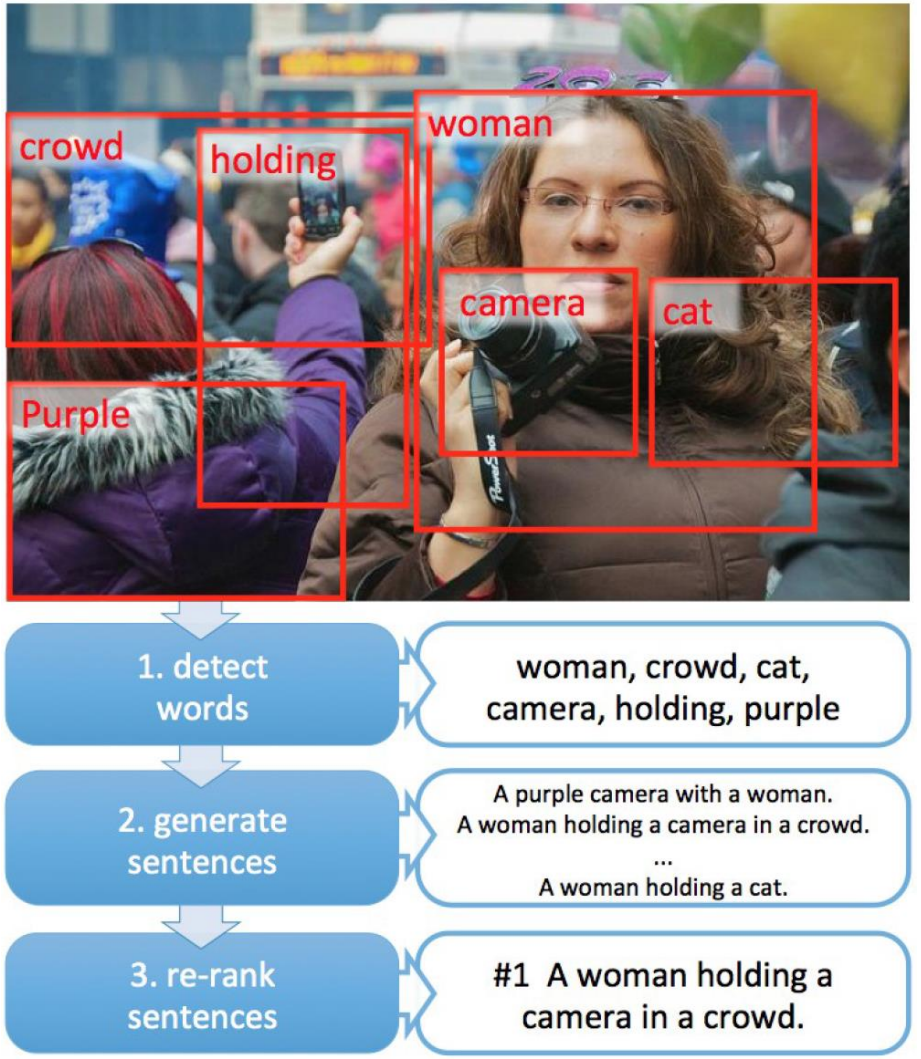


Figure 1. An illustrative example of our pipeline.

Stage 1: Multiple Instance Learning (MIL)

- Treat training caption as bag of image labels
- Train one binary classifier per label on all images
- "Noisy-Or" classifier
  - Image divided into 12x12 overlapping regions
  - fc7 vector used for image features

$$p_i^w = 1 - \prod_{j \in r_i} (1 - \sigma(f_{ij} \cdot v_w))$$

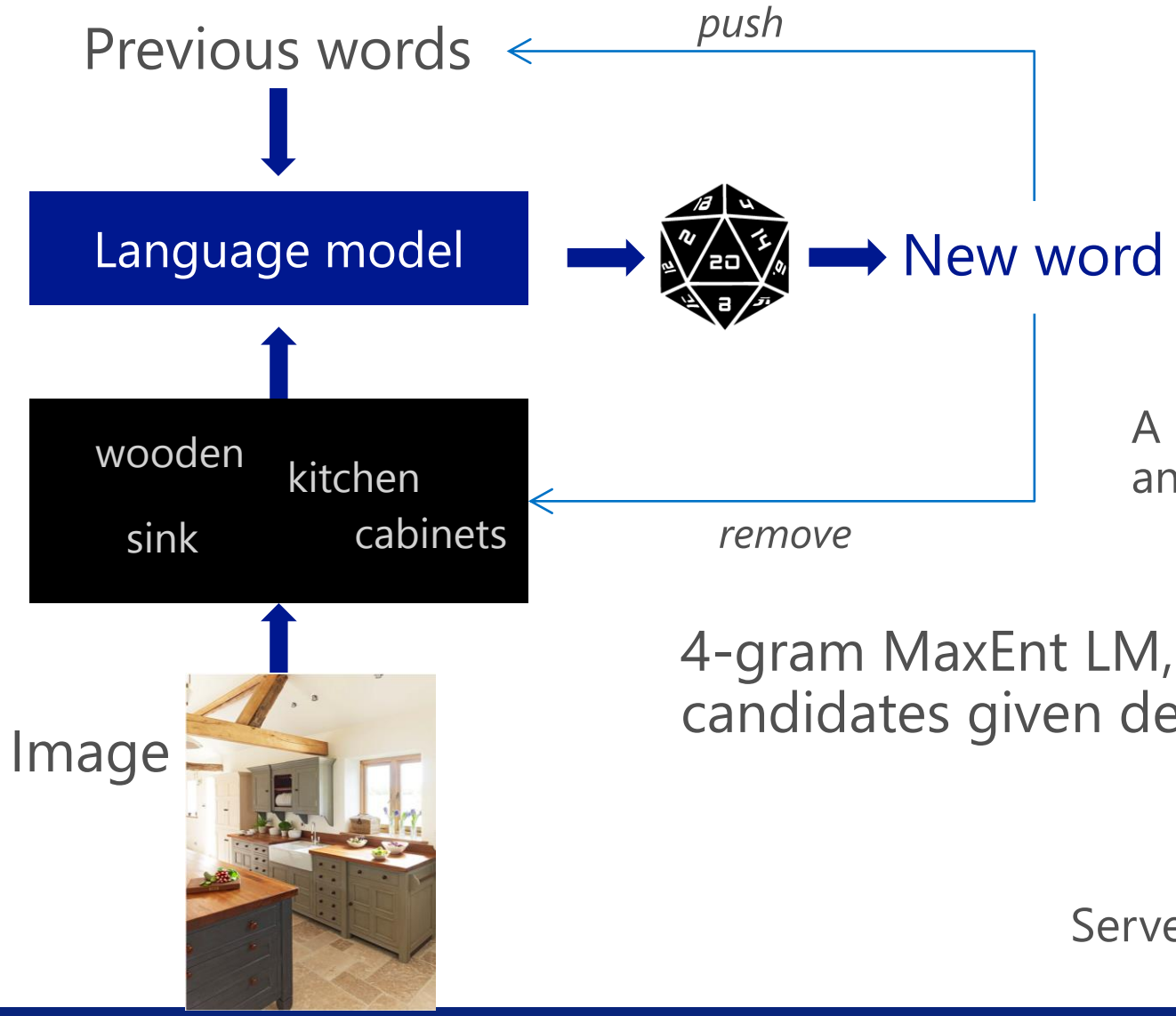
i = image id
r_i = regions
f_{ij} = fc7 vector
v_w = learned classifier weights
$\sigma(x)$ = sigmoid

sitting

sitting (0.83)

Models the visual "attention".

Stage 2: Language models with a blackboard



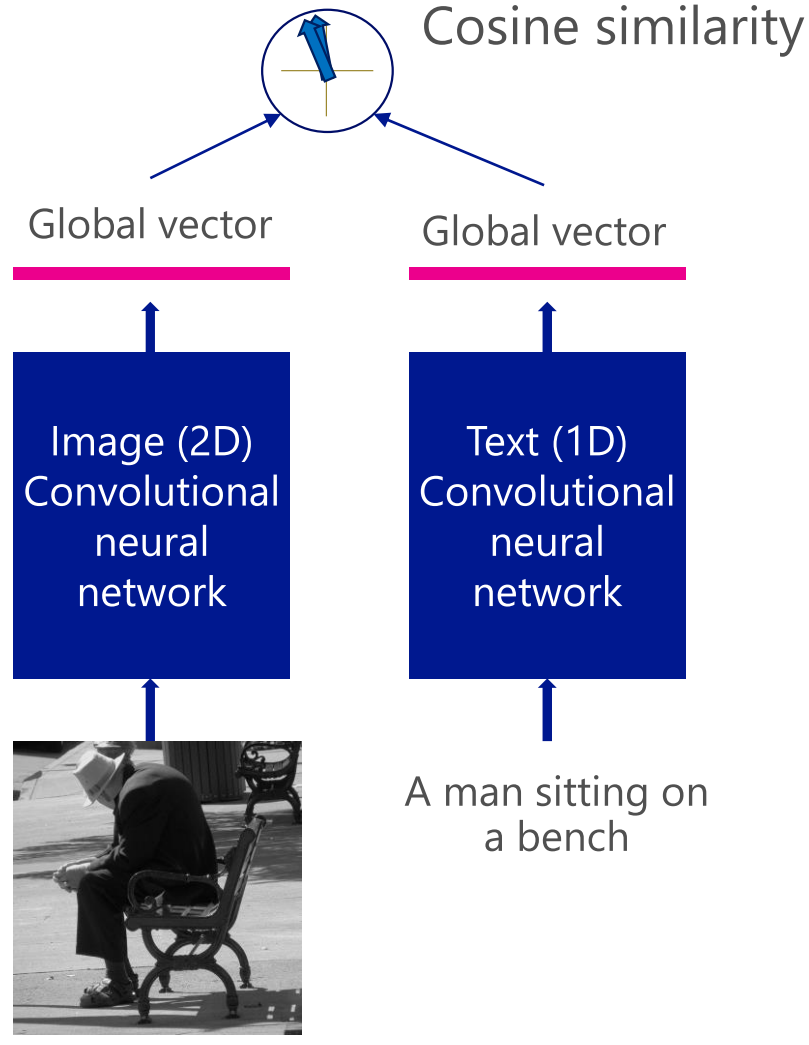
A kitchen with wooden cabinets and a sink

4-gram MaxEnt LM, generates 500 caption candidates given detected words

Serves as a grammar checker.

Stage 3: Deep Multimodal Similarity Model

- Project sentence and image into a comparable semantic vector space
- Whole sentence language model
- DMSM + basic features → re-ranked caption list



Q = image, D = caption, R = relevance

Relevance: $R(Q,D) = \text{cosine}(y_Q, y_D) = \frac{{y_Q}^T y_D}{\|y_Q\| \|y_D\|}$

Caption probability: $P(D|Q) = \frac{\exp(\gamma R(Q,D))}{\Sigma_{D' \in \mathbb{D}} \exp(\gamma R(Q,D'))}$

Candidate captions ↗ ↖ Smoothing factor

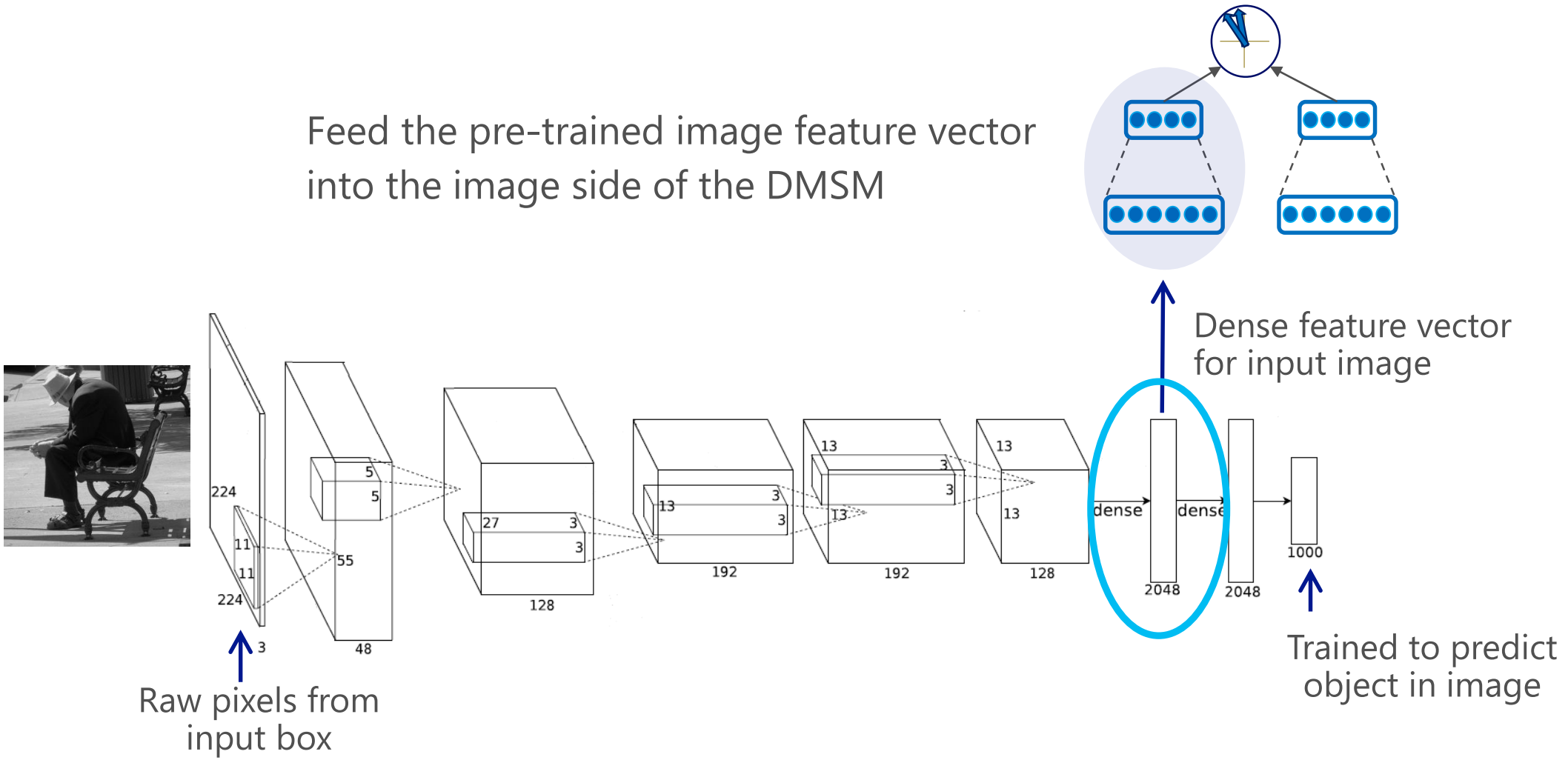
Objective: $L(\Lambda) = -\log \prod_{(Q,D^+)} P(D^+|Q)$

↖ Correct caption

Serves as a semantic matching checker.

The convolutional network at the image side

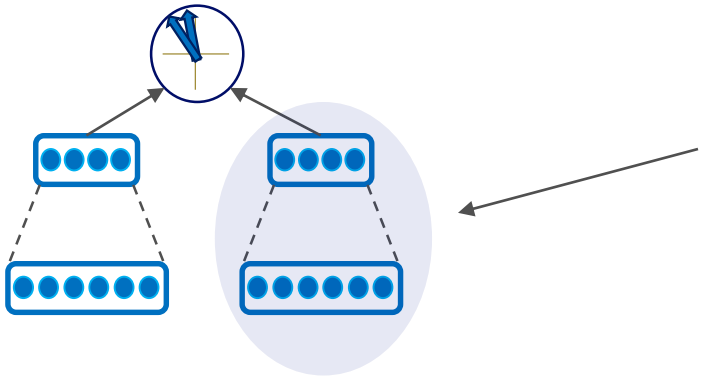
Feed the pre-trained image feature vector into the image side of the DMSM



Tuned image features from AlexNet (Krizhevsky et al., 2012) or VGG (Simonyan and Zisserman, 2014).

The convolutional network at the caption side

Models fine-grained structural language information in the caption



Using a convolutional neural network for the text caption side

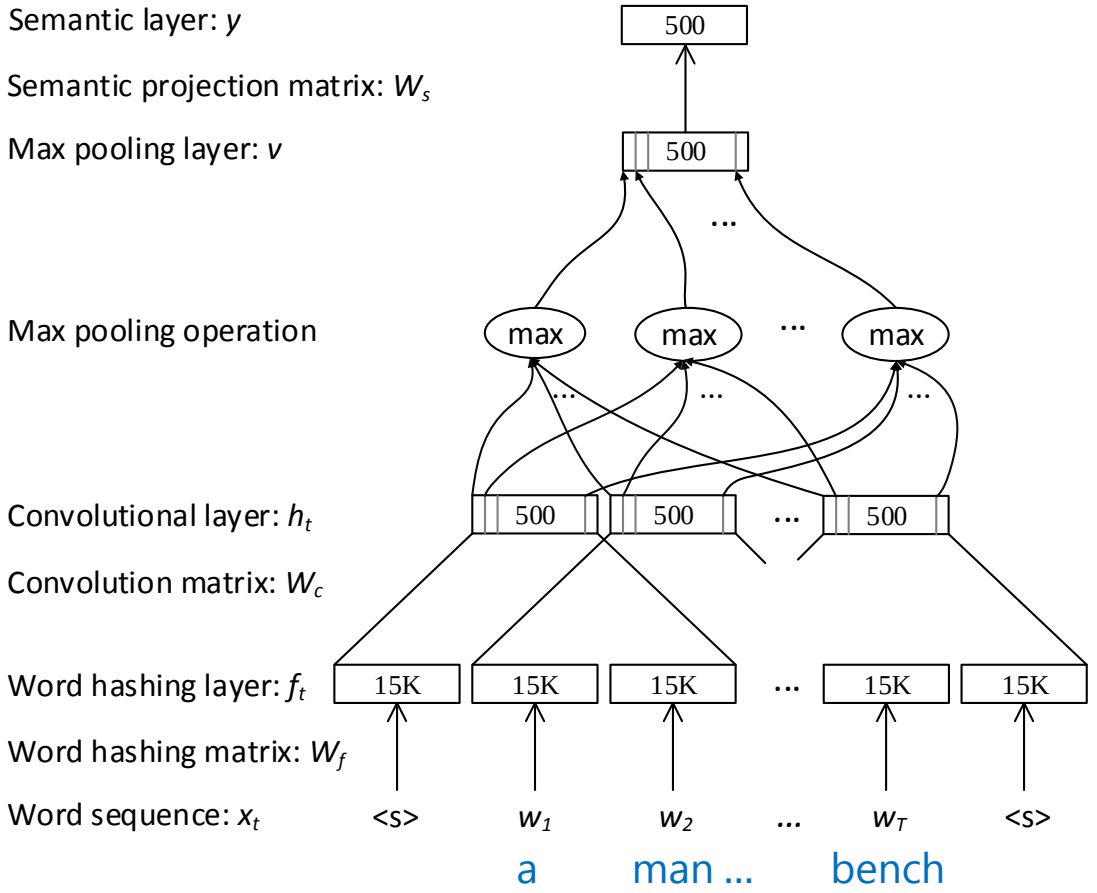


Figure Credit: [Shen, He, Gao, Deng, Mesnil, WWW, April 2014]

The MS COCO Benchmark

Microsoft COCO
Common Objects in Context

http://mscoco.org/

What is Microsoft COCO?

Microsoft COCO is a new image recognition, segmentation, and captioning dataset. Microsoft COCO has several features:

- **Object segmentation**
- **Recognition in Context**
- **Multiple objects per image**
- **More than 300,000 images**
- **More than 2 Million instances**
- **80 object categories**
- **5 captions per image**

Collaborators

Tsung-Yi Lin Cornell Tech

Michael Maire TTI Chicago

Serge Belongie Cornell Tech

Lubomir Bourdev Facebook AI

Ross Girshick Microsoft Research

James Hays Brown University

Pietro Perona Caltech

Deva Ramanan UC Irvine

Larry Zitnick Microsoft Research

Piotr Dollár Facebook AI

CORNELL NYCTECH

Caltech

facebook

Brown University

UCIrvine
University of California, Irvine

Microsoft Research

The man at bat readies to swing at the pitch while the umpire looks on.

A large bus sitting next to a very tall building.

MS COCO Challenge: generate descriptive captions for images

How much can machines understand complex scenes nowadays?

Measure the quality of the captions by human judge (e.g., *Turing Test*).

Great progress, but still a ***big gap*** vs. *Human*.
(huge room for improvement)

The state-of-the-art at the MS COCO Captioning Challenge 2015

		% of captions that pass the Turing Test	Official Rank
Human		67.5%	--
MSR	[Fang+ 15]	32.2%	1st(tie)
Google	[Vinyals+ 15]	31.7%	1st(tie)
MSR Captivator	[Devlin+ 15]	30.1%	3rd(tie)
Montreal/Toronto	[Xu+ 15]	27.2%	3rd(tie)
Berkeley LRCN	[Donahue+ 15]	26.8%	5th

Examples:

MSR: a clock tower in the middle of the street

MSR: a stop light on a city street

MSR: a living room filled with furniture and a flat screen tv sitting on top of a brick building

MSR: a display in a grocery store filled with lots of food on a table

More examples:

MSR: a man riding a skateboard down a street

MSR: a group of people posing for a picture

MSR: a cat sitting on top of a bed

MSR: a baby elephant standing next to a fence

Interpretability

baseball (1.00)

a **baseball**

Provided grounded evidence for the generated caption.

Thanks to the MIL based image word detection

Interpretability

player (1.00)

a baseball **player**

Provided grounded evidence for the generated caption.

Thanks to the MIL based image word detection

Interpretability

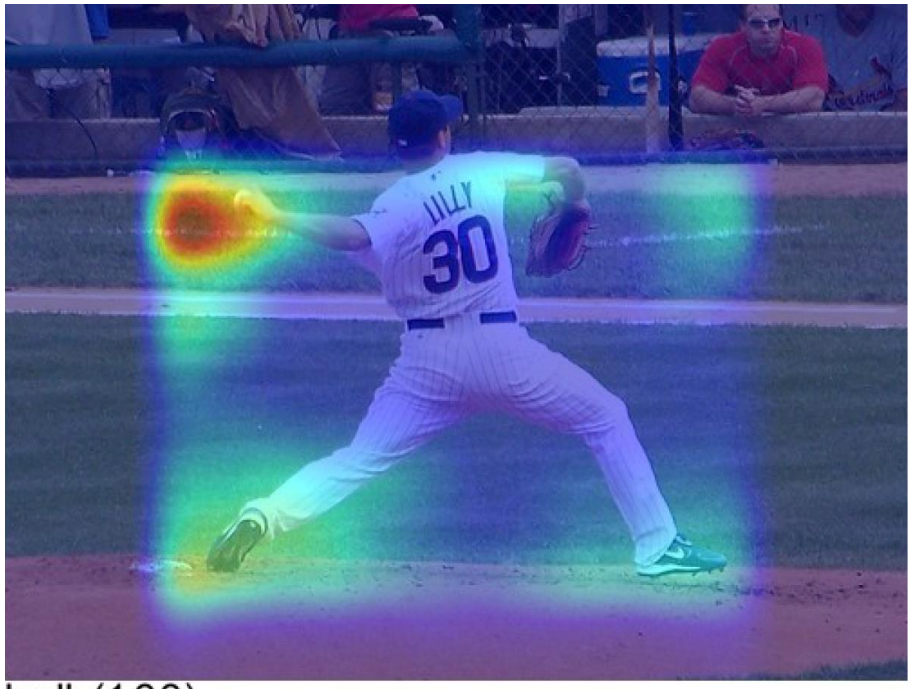
throwing (0.86)

a baseball player **throwing**

Provided grounded evidence for the generated caption.

Thanks to the MIL based image word detection

Interpretability



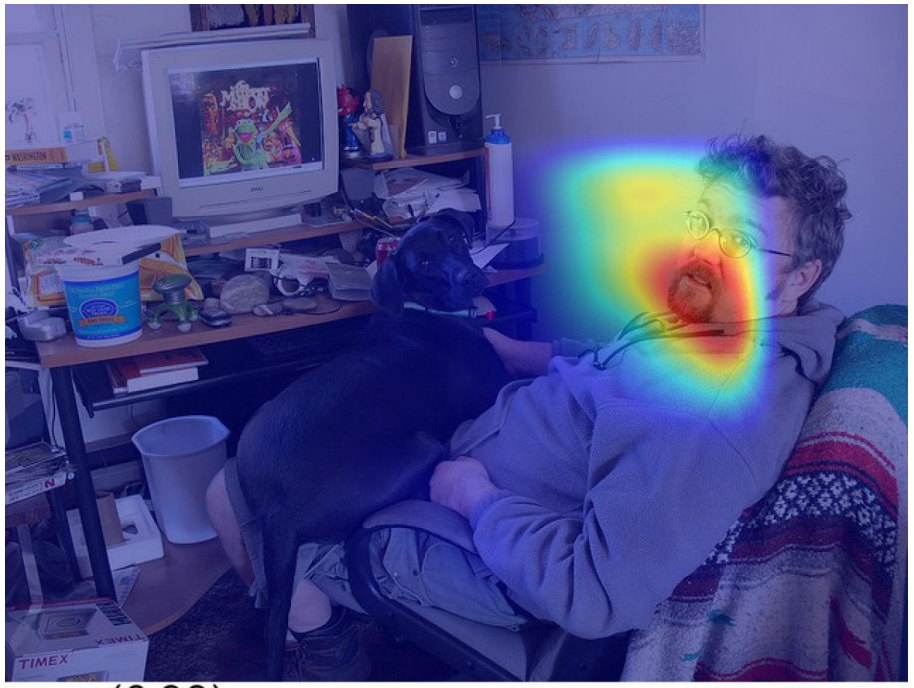
ball (1.00)

a baseball player throwing a **ball**

Provided grounded evidence for the generated caption.

Thanks to the MIL based image word detection

Interpretability



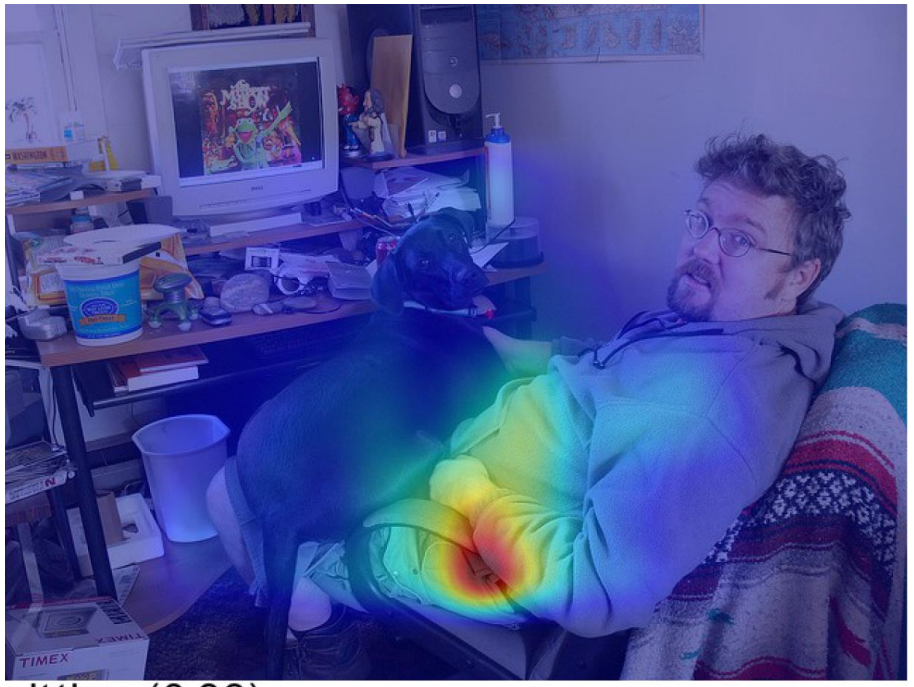
man (0.93)

a **man**

Provided grounded evidence for the generated caption.

Thanks to the MIL based image word detection

Interpretability



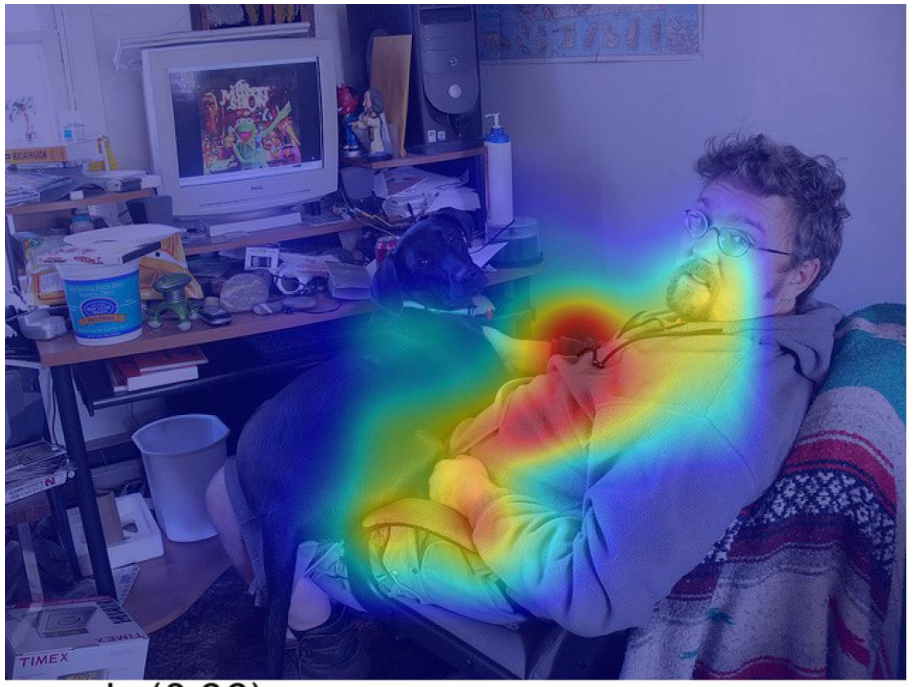
sitting (0.83)

a man **sitting**

Provided grounded evidence for the generated caption.

Thanks to the MIL based image word detection

Interpretability



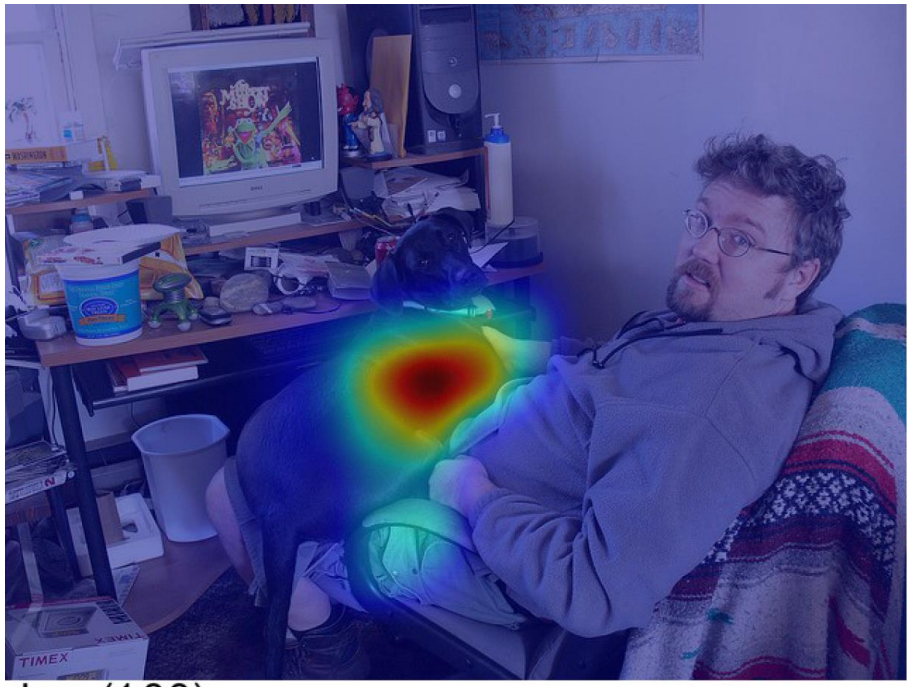
couch (0.66)

a man sitting in a **couch**

Provided grounded evidence for the generated caption.

Thanks to the MIL based image word detection

Interpretability



dog (1.00)

a man sitting in a couch with a **dog**

Thanks to the MIL based image word detection

Another popular paradigm: Multimodal Recurrent Neural Network (MRNN)

- Use fc7 as initial state in a recurrent neural network language model
- Words output in sequence

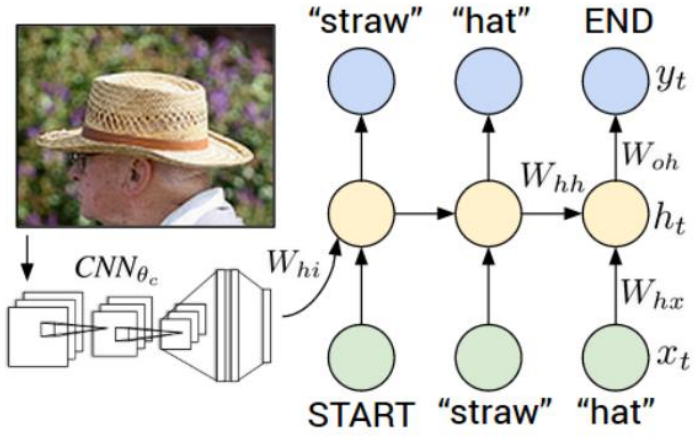
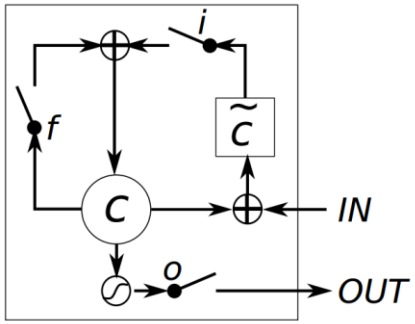


Image Credit: Karpathy and Fei-Fei 2015

LSTM (Hochreiter & Schmidhuber 97)



Gated Recurrent Neural Network (GRNN)

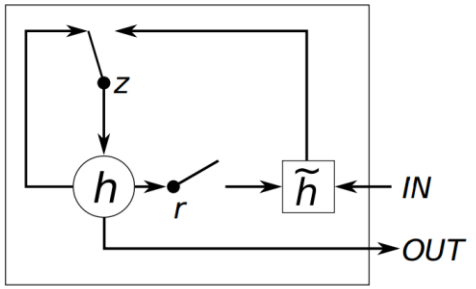


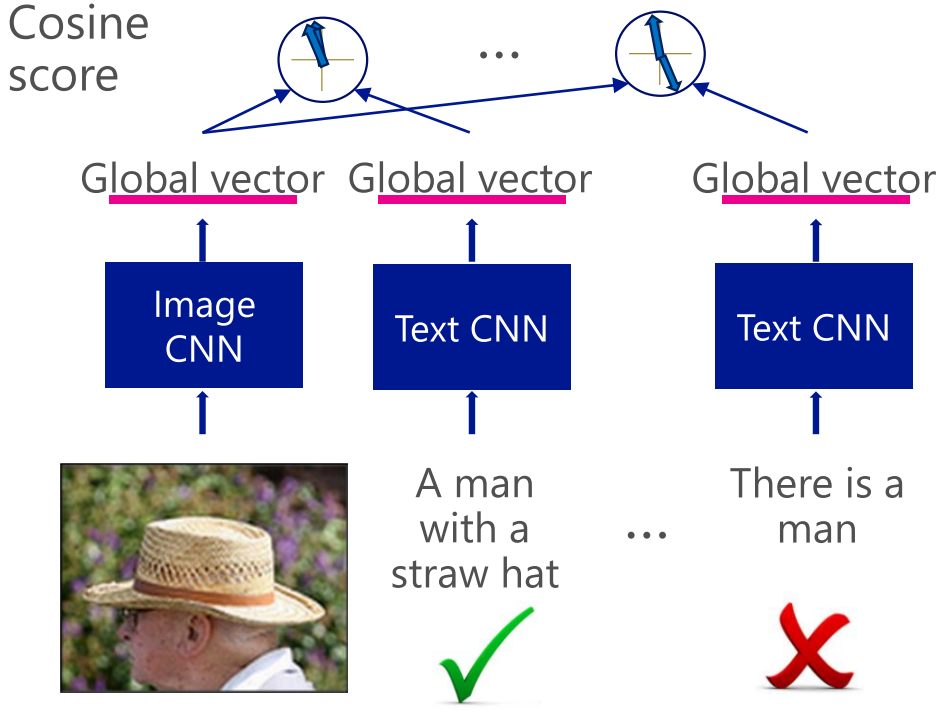
Image Credit: Cho et al. 2015

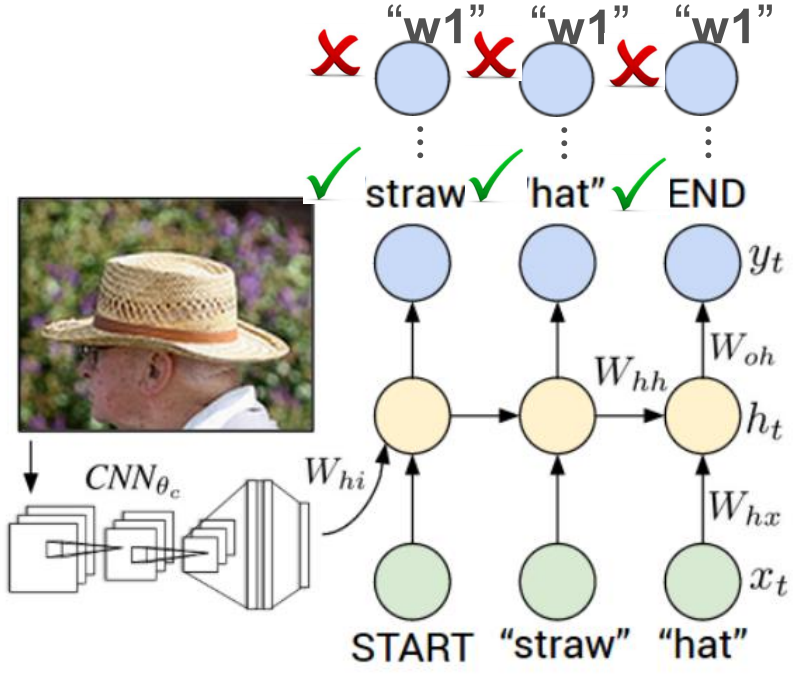
E.g., Hill+ 2014, Devlin+ 2015; Donahue+ 2015, Karpathy+ 2015; Kiros+ 2015; Mao+ 2015; Vinyals+ 2015; Xu+ 2015;

A comparison:

DMSM's objective:
the score of the reference to be higher than other generic captions.

MRNN's objective:
the score of the reference to be higher than arbitrary word sequences





DMSM focuses on semantics rather than syntax. E.g., ensures the reference (*semantically interesting*) scores higher than generic ones (grammatically correct but *semantically incorrect or boring*), while MRNN focus on syntax more.

Auto Evaluation & Human Evaluation

- MELM+DMSM and MRNN obtain same BLEU score
- But humans prefer MELM+DMSM's output more

System		BLEU %	Better or Equal to Human
Model 1:	**MELM + DMSM**	25.7	34.0%
Model 2:	**MRNN**	25.7	29.0%

Human judgers shown generated caption and human caption, choose which is "better", or equal.

Devlin, Cheng, Fang, Gupta, Deng, He, Zweig, and Mitchell **ACL 2015**, **Language Models for Image Captioning: The Quirks and What Works.**

Language Analysis

- MELM weakness: Long distance language modeling

MELM + DMSM	MRNN
a slice of pizza sitting on top of it	a bed with a red blanket on top of it
a black and white bird perched on top of it	a birthday cake with candles on top of it

- MRNN weakness: Repeated emissions

MELM + DMSM	MRNN
a large bed sitting in a bedroom	a bedroom with a bed and a bed
a man wearing a bow tie	a man wearing a tie and a tie

Language Analysis

- MRNN weakness: Repeated captions
  - Table 1: Systems produce same captions multiple times; MRNN does it the most
  - Table 2: MRNN repeat captions seen in training data verbatim more often

Table 1 **Number *Distinct* Captions in Testval (out of 20,244 instances)**

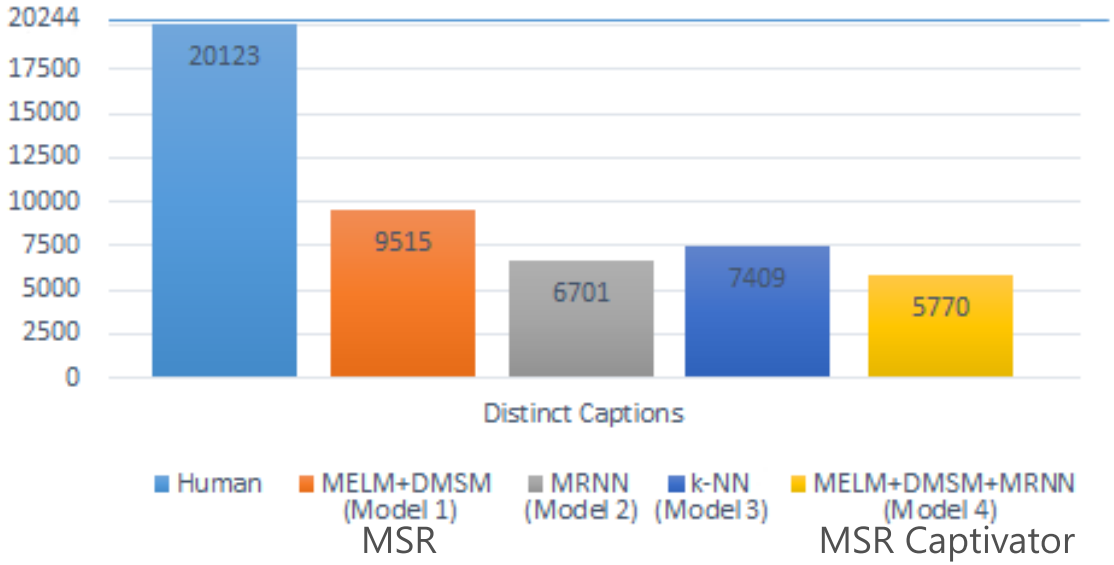
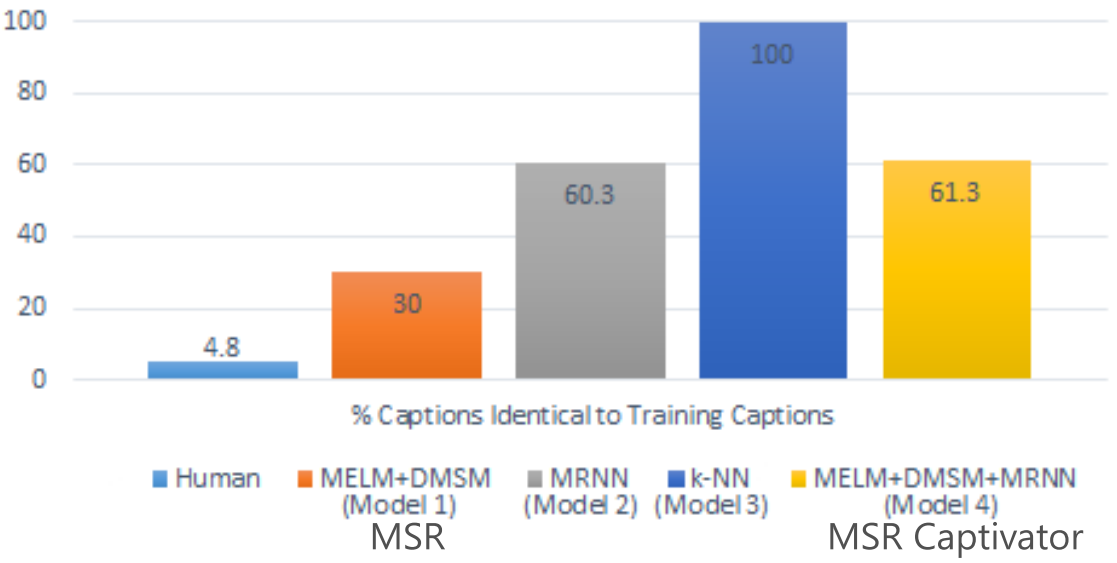


Table 2 **Percentage of Produced Testval Captions Found in Training Captions**



Example: MELM+DMSM: "A plate with a sandwich and a cup of coffee"

MRNN: "A close up of a plate of food" *(more generic)*

Image Diversity

- Bin test images based on visual overlap with training
  - MELM+DMSM does well on images with low overlap
  - MRNN does well on images with high overlap

Condition	Train/Test Visual Overlap BLEU		
	Whole Set	20% Least	20% Most
D-ME+DMSM	25.7	20.9	29.9
MRNN	25.7	18.8	32.0

BLEU scores based on visual overlap

Great progress, but what is missing?

e.g., could computer understand humors?

Image credit:
http://s122.photobucket.com/user/bmeuppls/media/stampede.jpg.html

nowadays:

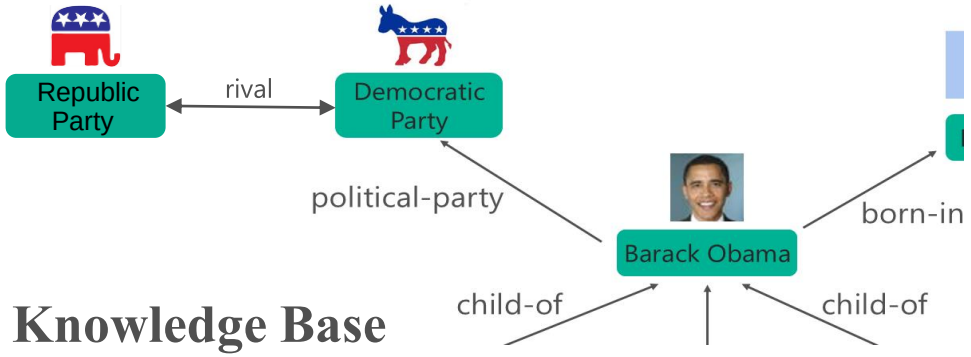
➡ a herd of elephants standing next to a man

Future: + face recognition etc.:

➡ a herd of elephants standing next to Obama

Future: + knowledge/common sense:

➡ Obama is chased by his republic rivals ☺



A knowledge base consists of digitalized common sense

Summary

- Language is (arguably) the best ***supervision*** for teaching machines to understand complex scenes *as humans do*.
- We generate ***interesting*** rather than generic captions
- We ***ground*** the generated caption on the salient content of the image
- In the future, need to incorporate knowledge base to give the machine the ***common sense*** (beyond an isolated image).

NOWADAYS:

→ a man standing on a boat

FUTURE:

(+ *face reco., Web, Knowledge Base, GPS, Map…*)

→ ~~a man~~ Xiaodong standing on a boat sailing on lake Washington ☺

Questions?

Learning semantic representations for *Text*, *Image*, and *Knowledge*:

DSSM/DMSM/Sent2Vec Tool kit available: http://aka.ms/sent2vec/

Relevant work

Donahue, Hendricks, Guadarrama, Rohrbach, Venugopalan, Saenko, and Darrell. **Long-term recurrent convolutional networks for visual recognition and description. CVPR, 2015.**

Devlin, Cheng, Fang, Gupta, Deng, He, Zweig, and Mitchell, **Language Models for Image Captioning: The Quirks and What Works. ACL 2015**

Fang, Gupta, Iandola, Srivastava, Deng, Dollar, Gao, He, Mitchell, Platt, Zitnick, Zweig, "**From Captions to Visual Concepts and Back,**" **CVPR, 2015**

Hill and Korhonen, **2014 Learning Abstract Concept Embeddings from Multi-Modal Data: Since You Probably Can't See What I Mean**

Karpathy and Fei-Fei, "**Deep Visual-Semantic Alignments for Generating Image Descriptions**". CVPR 2015

Kiros, Salakhutdinov, Zemel, "**Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models**". TACL 2015

Mao, Xu, Yang, Wang, Huang, Yuille. "**Deep Captioning with Multimodal Recurrent Neural Networks (m-RNN)**," ICLR 2015

Vinyals, Toshev, Bengio, and Erhan. **Show and tell: A neural image caption generator. CVPR, 2015.**

Xu, Ba, Kiros, Cho, Courville, Salakhutdinov, Zemel, Bengio, 2015. **Show, Attend and Tell: Neural Image Caption Generation with Visual Attention**.

Other relevant work on deep learning for KB, QA, SLU, Semantic representation

- W. Yih, M. Chang, X. He, and J. Gao, Semantic Parsing via Staged Query Graph Generation: Question Answering with Knowledge Base, ACL, July 2015
- X. Liu, J. Gao, X. He, L. Deng, Kevin Duh, and Y. Wang, Representation Learning Using Multi-Task Deep Neural Networks for Semantic Classification and Information Retrieval, NAACL, May 2015
- A. Elkahky, Y. Song, and X. He, A Multi-View Deep Learning Approach for Cross Domain User Modeling in Recommendation Systems, WWW, May 2015
- B. Yang, W. Yih, X. He, J. Gao, and L. Deng, Embedding Entities and Relations for Learning and Inference in Knowledge Bases, ICLR, May 2015
- G. Mesnil, Y. Dauphin, K. Yao, Y. Bengio, L. Deng, D. Hakkani-Tur, X. He, L. Heck, G. Tur, D. Yu, and G. Zweig, Using Recurrent Neural Networks for Slot Filling in Spoken Language Understanding, IEEE/ACM TASLP, March 2015
- H. Palangi, L. Deng, Y. Shen, J. Gao, X. He, J. Chen, X. Song, and R. Ward, Deep Sentence Embedding Using the Long Short Term Memory Network: Analysis and Application to Information Retrieval, *arXiv:1502.06922*, February 2015
- Y. Shen, X. He, J. Gao, L. Deng, and G. Mesnil, A Latent Semantic Model with Convolutional-Pooling Structure for Information Retrieval , CIKM, November 2014
- J. Gao, P. Pantel, M. Gamon, X. He, L. Deng, et al., Modeling Interestingness with Deep Neural Networks, EMNLP, October 2014
- W. Yih, X. He, and C. Meek, Semantic Parsing for Single-Relation Question Answering, ACL, June 2014
- J. Gao, X. He, W. Yih, and L. Deng, Learning Continuous Phrase Representations for Translation Modeling, ACL, June 2014
- P. Huang, X. He, J. Gao, L. Deng, A. Acero, and L. Heck, Learning Deep Structured Semantic Models for Web Search using Clickthrough Data, CIKM, October 2013
- G. Mesnil, X. He, L. Deng, and Y. Bengio, Investigation of Recurrent-Neural-Network Architectures and Learning Methods for Spoken Language Understanding, in *Interspeech*, August 2013
- G. Tur, L. Deng, D. Hakkani-Tur, and X. He, Towards Deeper Understanding Deep Convex Networks for Semantic Utterance Classification, ICASSP, March 2012

General References

- Auli, M., Galley, M., Quirk, C. and Zweig, G., 2013. Joint language and translation modeling with recurrent neural networks. In EMNLP.
- Auli, M., and Gao, J., 2014. Decoder integration and expected bleu training for recurrent neural network language models. In ACL.
- Bahdanau, D., Cho, K., and Bengio, Y. 2015. Neural machine translation by joingly learning to align and translate, in ICLR 2015.
- Bejar, I., Chaffin, R. and Embretson, S. 1991. Cognitive and psychometric analysis of analogical problem solving. Recent research in psychology.
- Bengio, Y., 2009. Learning deep architectures for AI. Foundumental Trends Machine Learning, vol. 2.
- Bengio, Y., Courville, A., and Vincent, P. 2013. Representation learning: A review and new perspectives. IEEE Trans. PAMI, vol. 38, pp. 1798-1828.
- Bengio, Y., Ducharme, R., and Vincent, P., 2000. A Neural Probabilistic Language Model, in NIPS.
- Berant, J., and Liang, P. 2014. Semantic parsing via paraphrasing. In ACL.
- Blei, D., Ng, A., and Jordan M. 2001. Latent dirichlet allocation. In NIPS.
- Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J. and Yakhnenko, O. 2013. Translating Embeddings for Modeling Multi-relational Data. In NIPS.
- Bordes, A., Chopra, S., and Weston, J. 2014. Question answering with subgraph embeddings. In EMNLP.
- Bordes, A., Glorot, X., Weston, J. and Bengio Y. 2012. Joint Learning of Words and Meaning Representations for Open-Text Semantic Parsing. In AISTATS.
- Brown, P., deSouza, P. Mercer, R., Della Pietra, V., and Lai, J. 1992. Class-based n-gram models of natural language. Computational Linguistics 18 (4).
- Chandar, A. P. S., Lauly, S., Larochelle, H., Khapra, M. M., Ravindran, B., Raykar, V., and Saha, A. (2014). An autoencoder approach to learning bilingual word representations. In NIPS.
- Chang, K., Yih, W., and Meek, C. 2013. Multi-Relational Latent Semantic Analysis. In EMNLP.
- Chang, K., Yih, W., Yang, B., and Meek, C. 2014. Typed Tensor Decomposition of Knowledge Bases for Relation Extraction. In EMNLP.
- Collobert, R., and Weston, J. 2008. A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning. In ICML.
- Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P., 2011. Natural language processing (almost) from scratch. in JMLR, vol. 12.
- Cui, L., Zhang, D., Liu, S., Chen, Q., Li, M., Zhou, M., and Yang, M. (2014). Learning topic representation for smt with neural networks. In ACL.
- Dahl, G., Yu, D., Deng, L., and Acero, 2012. A. Context-dependent, pre-trained deep neural networks for large vocabulary speech recognition, IEEE Trans. Audio, Speech, & Language Proc., Vol. 20 (1), pp. 30-42.
- Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T., and Harshman, R. 1990. Indexing by latent semantic analysis. J. American Society for Information Science, 41(6): 391-407
- Deng, L., Seltzer, M., Yu, D., Acero, A., Mohamed, A., and Hinton, G., 2010. Binary Coding of Speech Spectrograms Using a Deep Auto-encoder, in Interspeech.
- Deng, L., Tur, G, He, X, and Hakkani-Tur, D. 2012. Use of kernel deep convex networks and end-to-end learning for spoken language understanding, Proc. IEEE Workshop on Spoken Language Technologies.
- Deng, L., Yu, D. and Acero, A. 2006. Structured speech modeling, IEEE Trans. on Audio, Speech and Language Processing, vol. 14, no. 5, pp. 1492-1504.

General References

- Deng, L., Yu, D. and Acero, A. 2006. Structured speech modeling, IEEE Trans. on Audio, Speech and Language Processing, vol. 14, no. 5, pp. 1492-1504.
- Deng, L., Yu, D., and Platt, J. 2012. Scalable stacking and learning for building deep architectures, Proc. ICASSP.
- Deng, L. and Yu, D. 2014. Deeping learning methods and applications. Foundations and Trends in Signal Processing 7:3-4.
- Deoras, A., and Sarikaya, R., 2013. Deep belief network based semantic taggers for spoken language understanding, in INTERSPEECH.
- Devlin, J., Zbib, R., Huang, Z., Lamar, T., Schwartz, R., and Makhoul, J., 2014. Fast and Robust Neural Network Joint Models for Statistical Machine Translation, ACL.
- Duh, K. 2014. Deep learning for natural language processing and machine translation. Tutorial. 2014.
- Duh, K., Neubig, G., Sudoh, K., and Tsukada, H. (2013). Adaptation data selection using neural language models: Experiments in machine translation. In ACL.
- Fader, A., Zettlemoyer, L., and Etzioni, O. 2013. Paraphrase-driven learning for open question answering. In ACL.
- Fang, H., Gupta, S., Iandola, F., Srivastava, R., Deng, L., Dollár, P., Gao, J., He, X., Mitchell, M., Platt, J., Zitnick, L., Zweig, G., "From Captions to Visual Concepts and Back," arXiv:1411.4952
- Faruqui, M. and Dyer, C. (2014). Improving vector space word representations using multilingual correlation. In EACL.
- Firth, J. R. 1957. *Papers in Linguistics 1934–1951*, Oxford University Press, 1957
- Frome, A., Corrado, G., Shlens, J., Bengio, S., Dean, J., Ranzato, M., and Mikolov, T., 2013. DeViSE: A Deep Visual-Semantic Embedding Model, Proc. NIPS.
- Gao, J., He, X., Yih, W-t., and Deng, L. 2014a. Learning continuous phrase representations for translation modeling. In ACL.
- Gao, J., He, X., and Nie, J-Y. 2010. Clickthrough-based translation models for web search: from word models to phrase models. In CIKM.
- Gao, J., Pantel, P., Gamon, M., He, X., Deng, L., and Shen, Y. 2014b. Modeling interestingness with deep neural networks. In EMNLP
- Gao, J., Toutanova, K., Yih., W-T. 2011. Clickthrough-based latent semantic models for web search. In SIGIR.
- Gao, J., Yuan, W., Li, X., Deng, K., and Nie, J-Y. 2009. Smoothing clickthrough data for web search ranking. In SIGIR.
- Gao, J., and He, X. 2013. Training MRF-based translation models using gradient ascent. In NAACL-HLT.
- Getoor, L., and Taskar, B. editors. 2007. Introduction to Statistical Relational Learning. The MIT Press.
- Graves, A., Jaitly, N., and Mohamed, A., 2013a. Hybrid speech recognition with deep bidirectional LSTM, Proc. ASRU.
- Graves, A., Mohamed, A., and Hinton, G., 2013. Speech recognition with deep recurrent neural networks, Proc. ICASSP.
- He, X. and Deng, L., 2013. Speech-Centric Information Processing: An Optimization-Oriented Approach, in Proceedings of the IEEE.
- He, X. and Deng, L., 2012. Maximum Expected BLEU Training of Phrase and Lexicon Translation Models , ACL.
- He, X., Deng, L., and Chou, W., 2008. Discriminative learning in sequential pattern recognition, Sept. IEEE Sig. Proc. Mag.

General References

- Hermann, K. M. and Blunsom, P. (2014). Multilingual models for compositional distributed semantics. In ACL.
- Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T., and Kingsbury, B., 2012. Deep Neural Networks for Acoustic Modeling in Speech Recognition, IEEE Signal Processing Magazine, vol. 29, no. 6, pp. 82-97.
- Hinton, G., Osindero, S., and The, Y-W. 2006. A fast learning algorithm for deep belief nets. *Neural Computation*, 18: 1527-1554.
- Hinton, G., and Salakhutdinov, R., 2010. Discovering binary codes for documents by learning deep generative models. Topics in Cognitive Science.
- Hu, Y., Auli, M., Gao, Q., and Gao, J. 2014. Minimum translation modeling with recurrent neural networks. In EACL.
- Huang, E., Socher, R., Manning, C, and Ng, A. 2012. Improving word representations via global context and multiple word prototypes, Proc. ACL.
- Huang, P., He, X., Gao, J., Deng, L., Acero, A., and Heck, L. 2013. Learning deep structured semantic models for web search using clickthrough data. In CIKM.
- Hutchinson, B., Deng, L., and Yu, D., 2012. A deep architecture with bilinear modeling of hidden representations: Applications to phonetic recognition, Proc. ICASSP.
- Hutchinson, B., Deng, L., and Yu, D., 2013. Tensor deep stacking networks, IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 35, pp. 1944 - 1957.
- Jurgens, D., Mohammad, S., Turney, P. and Holyoak, K. 2012. SemEval-2012 Task 2: Measuring degrees of relational similarity. In SemEval.
- Kalchbrenner, N. and Blunsom, P. (2013). Recurrent continuous translation models., in EMNLLP
- Kiros, R., Zemel, R., and Salakhutdinov, R. 2013. Multimodal Neural Language Models, Proc. NIPS Deep Learning Workshop.
- Klementiev, A., Titov, I., and Bhattarai, B. (2012). Inducing crosslingual distributed representations of words. In COLING.
- Kocisky, T., Hermann, K. M., and Blunsom, P. (2014). Learning bilingual word representations by marginalizing alignments. In ACL.
- Koehn, P. 2009. Statistical Machine Translation. Cambridge University Press.
- Krizhevsky, A., Sutskever, I, and Hinton, G., 2012. ImageNet Classification with Deep Convolutional Neural Networks, NIPS.
- Landauer. T., 2002. On the computational basis of learning and cognition: Arguments from LSA. Psychology of Learning and Motivation, 41:43–84.
- Lao, N., Mitchell, T., and Cohen, W. 2011. Random walk inference and learning in a large scale knowledge base. In EMNLP.
- Lauly, S., Boulanger, A., and Larochelle, H. (2013). Learning multilingual word representations using a bag-of-words autoencoder. In NIPS.
- Le, H-S, Oparin, I., Allauzen, A., Gauvain, J-L., Yvon, F., 2013. Structured output layer neural network language models for speech recognition, IEEE Transactions on Audio, Speech and Language Processing.
- LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. 1998. Gradient-based learning applied to document recognition, Proceedings of the IEEE, Vol. 86, pp. 2278-2324.
- Levy, O., and Goldberg, Y. 2014. Linguistic Regularities in Sparse and Explicit Word Representations. In CoNLL.

General References

- Li, P., Hastie, T., and Church, K.. 2006. Very sparse random projections, in Proc. SIGKDD.
- Li, P., Liu, Y., and Sun, M. (2013). Recursive autoencoders for ITG-based translation. In EMNLP.
- Li, P., Liu, Y., Sun, M., Izuha, T., and Zhang, D. (2014b). A neural reordering model for phrase-based translation. In COLING.
- Liu, S., Yang, N., Li, M., and Zhou, M. (2014). A recursive recurrent neural network for statistical machine translation. In ACL.
- Liu, L., Watanabe, T., Sumita, E., and Zhao, T. (2013). Additive neural networks for statistical machine translation. In ACL.
- Lu, S., Chen, Z., and Xu, B. (2014). Learning new semi-supervised deep auto-encoder features for statistical machine translation. In ACL.
- Maskey, S., and Zhou, B. 2012. Unsupervised deep belief feature for speech translation, in ICASSP.
- Mesnil, G., He, X., Deng, L., and Bengio, Y., 2013. Investigation of Recurrent-Neural-Network Architectures and Learning Methods for Spoken Language Understanding, in Interspeech.
- Mikolov, T. 2012. Statistical Language Models based on Neural Networks, Ph.D. thesis, Brno University of Technology.
- Mikolov, T., Chen, K., Corrado, G., and Dean, J. 2013. Efficient estimation of word representations in vector space, Proc. ICLR.
- Mikolov, T., Kombrink,. S., Burget, L., Cernocky, J., Khudanpur, S., 2011. Extensions of Recurrent Neural Network LM. ICASSP.
- Mikolov, T., Yih, W., Zweig, G., 2013. Linguistic Regularities in Continuous Space Word Representations. In NAACL-HLT.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. 2013. Distributed Representations of Words and Phrases and their Compositionality. In NIPS.
- Mohamed, A., Yu, D., and Deng, L. 2010. Investigation of full-sequence training of deep belief networks for speech recognition, Proc. Interspeech.
- Mohammad, S., Dorr, Bonnie., and Hirst, G. 2008. Computing word pair antonymy. In EMNLP.
- Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., and Ng, A. 2011. Multimodal deep learning, Proc. ICML.
- Nickel, M., Tresp, V., and Kriegel, H. 2011. A three-way model for collective learning on multi-relational data. In ICML.
- Niehues, J. and Waibel, A. (2013). Continuous space language models using Restricted Boltzmann Machines, in IWLT.
- Reddy, S., Lapata, M., and Steedman, M. 2014. Large-scale semantic parsing without question-answer pairs. Transactions of the Association for Computational Linguistics (TACL).
- Sainath, T., Mohamed, A., Kingsbury, B., and Ramabhadran, B. 2013. Convolutional neural networks for LVCSR, Proc. ICASSP.
- Salakhutdinov R., and Hinton, G., 2007 Semantic hashing. in Proc. SIGIR Workshop Information Retrieval and Applications of Graphical Models
- Salton, G. and McGill, M. 1983. Introduction to Modern Information Retrieval. McGraw Hill.
- Sarikaya, R., Hinton, G., and Ramabhadran, B., 2011. Deep belief nets for natural language call-routing, in Proceedings of the ICASSP.
- Schwenk, H. 2012. Continuous space translation models for phrase-based statistical machine translation, in COLING.

General References

- Schwenk, H., Rousseau, A., and Attik, M., 2012. Large, pruned or continuous space language models on a gpu for statistical machine translation, in NAACL-HLT 2012 Workshop.
- Seide, F., Li, G., and Yu, D. 2011. Conversational speech transcription using context-dependent deep neural networks, Proc. Interspeech
- Shen, Y., He, X., Gao, J., Deng, L., and Mesnil, G. 2014. Learning Semantic Representations Using Convolutional Neural Networks for Web Search, in Proceedings of WWW.
- Shen, Y., He, X., Gao, J., Deng, L., and Mesnil, G. 2014. A convolutional latent semantic model for web search. CIKM
- Socher, R., Chen, D., Manning, C., and Ng, A. 2013. Reasoning With Neural Tensor Networks For Knowledge Base Completion. In NIPS.
- Socher, R., Huval, B., Manning, C., Ng, A., 2012. Semantic compositionality through recursive matrix-vector spaces. In EMNLP.
- Socher, R., Lin, C., Ng, A., and Manning, C. 2011. Learning continuous phrase representations and syntactic parsing with recursive neural networks, Proc. ICML.
- Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C., Ng A., and Potts. C. 2013. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank, Proc. EMNLP
- Son, L. H., Allauzen, A., and Yvon, F. (2012). Continuous space translation models with neural networks. In NAACL.
- Song, X. He, X., Gao. J., and Deng, L. 2014. Learning Word Embedding Using the DSSM. MSR Tech Report.
- Song, Y., Wang, H., and He, X., 2014. Adapting Deep RankNet for Personalized Search. Proc. WSDM.
- Songyot, T. and Chiang, D. (2014). Improving word alignment using word similarity. In EMNLP.
- Sundermeyer, M., Alkhouli, T., Wuebker, J., and Ney, H. (2014). Translation modeling with bidirectional recurrent neural networks, in EMNLP.
- Tamura, A., Watanabe, T., and Sumita, E. (2014). Recurrent neural networks for word alignment model. In ACL.
- Tran, K. M., Bisazza, A., and Monz, C. (2014). Word translation prediction for morphologically rich languages with bilingual neural networks. In EMNLP.
- Tur, G., Deng, L., Hakkani-Tur, D., and He, X., 2012. Towards Deeper Understanding Deep Convex Networks for Semantic Utterance Classification, in ICASSP.
- Turney P. 2008. A uniform approach to analogies, synonyms, antonyms, and associations. In COLING. Songyot, T. and Chiang, D. (2014). Improving word alignment using word similarity. In EMNLP.
- Vaswani, A., Zhao, Y., Fossum, V., and Chiang, D. 2013. Decoding with large-scale neural language models improves translation, in EMNLP.
- Wright, S., Kanevsky, D., Deng, L., He, X., Heigold, G., and Li, H., 2013. Optimization Algorithms and Applications for Speech and Language Processing, in IEEE Transactions on Audio, Speech, and Language Processing, vol. 21, no. 11.
- Wu, H., Dong, D., Hu, X., Yu, D., He, W., Wu, H., Wang, H., and Liu, T. (2014a). Improve statistical machine translation with context-sensitive bilingual semantic embedding model. In EMNLP.
- Wu, Y., Watanabe, T., and Hori, C. (2014b). Recurrent neural network-based tuple sequence model for machine translation. In COLING.

General References

- Yang, B., Yih, W., He, X., Gao, J., and Deng L. 2014. In NIPS-2014 Workshop Learning Semantics.
- Yang, N., Liu, S., Li, M., Zhou, M., and Yu, N. (2013). Word alignment modeling with context dependent deep neural network. In ACL.
- Yao, K., Zweig, G., Hwang, M-Y. , Shi, Y., Yu, D., 2013. Recurrent neural networks for language understanding, submitted to Interspeech.
- Yann, D., Tur, G., Hakkani-Tur, D., Heck, L., 2014. Zero-Shot Learning and Clustering for Semantic Utterance Classification Using Deep Learning, in ICLR.
- Yih, W., Toutanova, K., Platt, J., and Meek, C. 2011. Learning discriminative projections for text similarity measures. In CoNLL.
- Yih, W., Zweig, G., Platt, J. 2012. Polarity Inducing Latent Semantic Analysis. In EMNLP-CoNLL.
- Yih, W., He, X., Meek, C. 2014. Semantic Parsing for Single-Relation Question Answering, in ACL.
- Yih, W., Chang, M-W., He, X., Gao, J. 2015. Semantic parsing via staged query graph generation: question answering with knowledge base, In ACL.
- Zeiler, M. and Fergus, R. 2013. Visualizing and understanding convolutional networks, arXiv:1311.2901, pp. 1-11.
- Zhang, J., Liu, S., Li, M., Zhou, M., and Zong, C. (2014). Bilingually-constrained phrase embeddings for machine translation. In ACL.
- Zou, W. Y., Socher, R., Cer, D., and Manning, C. D. (2013). Bilingual word embeddings for phrase-based machine translation. In EMNLP.