Reading your paper:  
second pass

Submitted by Ekin Akyürek

# OVERVIEW & PURPOSE

Reading a technical paper is unlike reading fiction: prepare to spend a lot more time, do multiple passes, take notes, survey literature, figure out the math, understand the algorithms. You are just starting machine learning, so many concepts will be unfamiliar. However you will have a whole semester with your paper and we will follow a process in which your understanding will improve gradually.

# HOW TO READ

To help you with this task please first read a couple of guides on how to read:

* [How to read a paper](http://www.sigcomm.org/sites/default/files/ccr/papers/2007/July/1273445-1273458.pdf) by S. Keshav.
* [How to read a technical paper](https://www.cs.jhu.edu/~jason/advice/how-to-read-a-paper.html) by J. Eisner.

# SECOND PASS

Go over your paper more carefully using Keshav’s second pass. This should take you about an hour. Please give the reference to your paper and answer the questions below. The paper reference should have the format “Author, Year, Title, Howpublished” (you can optionally include a link).

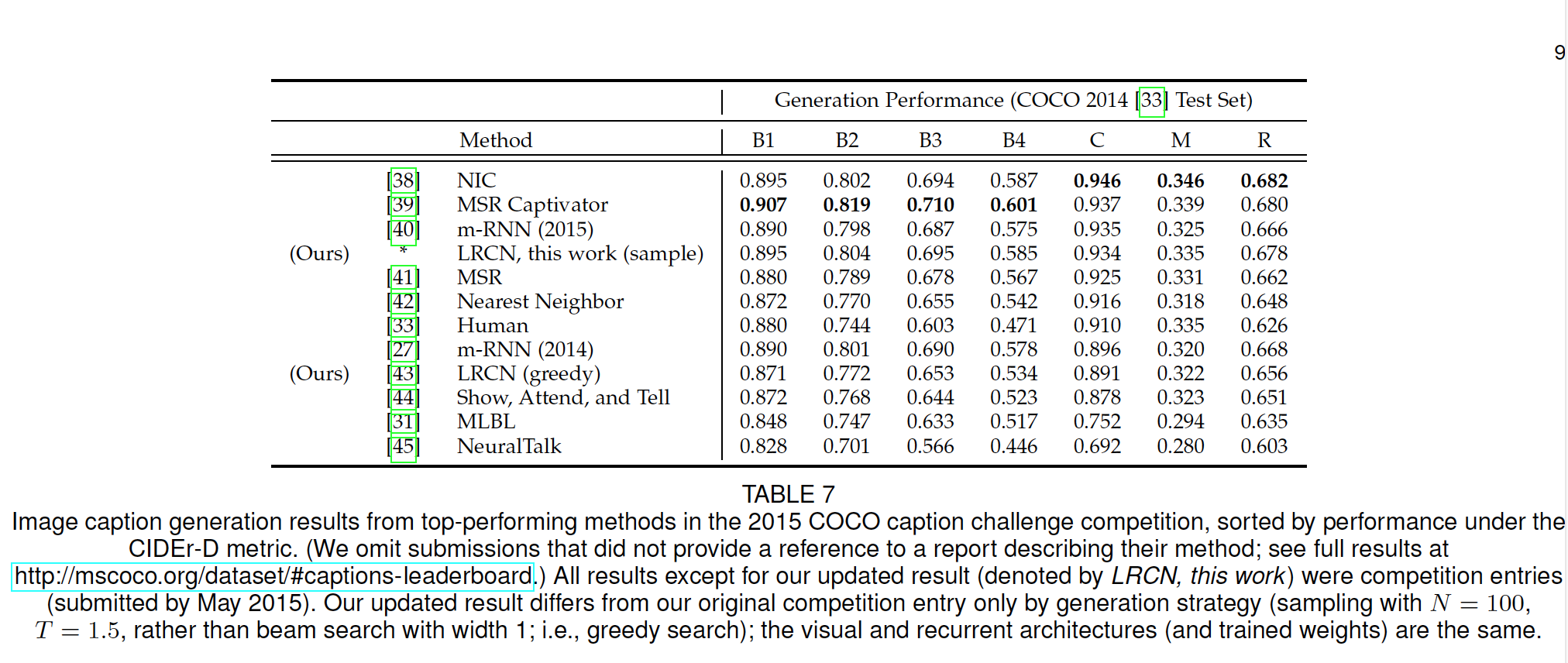
**Paper reference: Donahue, Jeffrey, et al. "Long-term recurrent convolutional networks for visual recognition and description." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.**

1. Data:Flickr30k 30k images with different sizes. We should resize images before usage.. The input to the convolutional network is the center cropped 224x224 image. It contains 5 caption for an image. Ms CoCo 300k images with different sizes and 5 caption for an image.

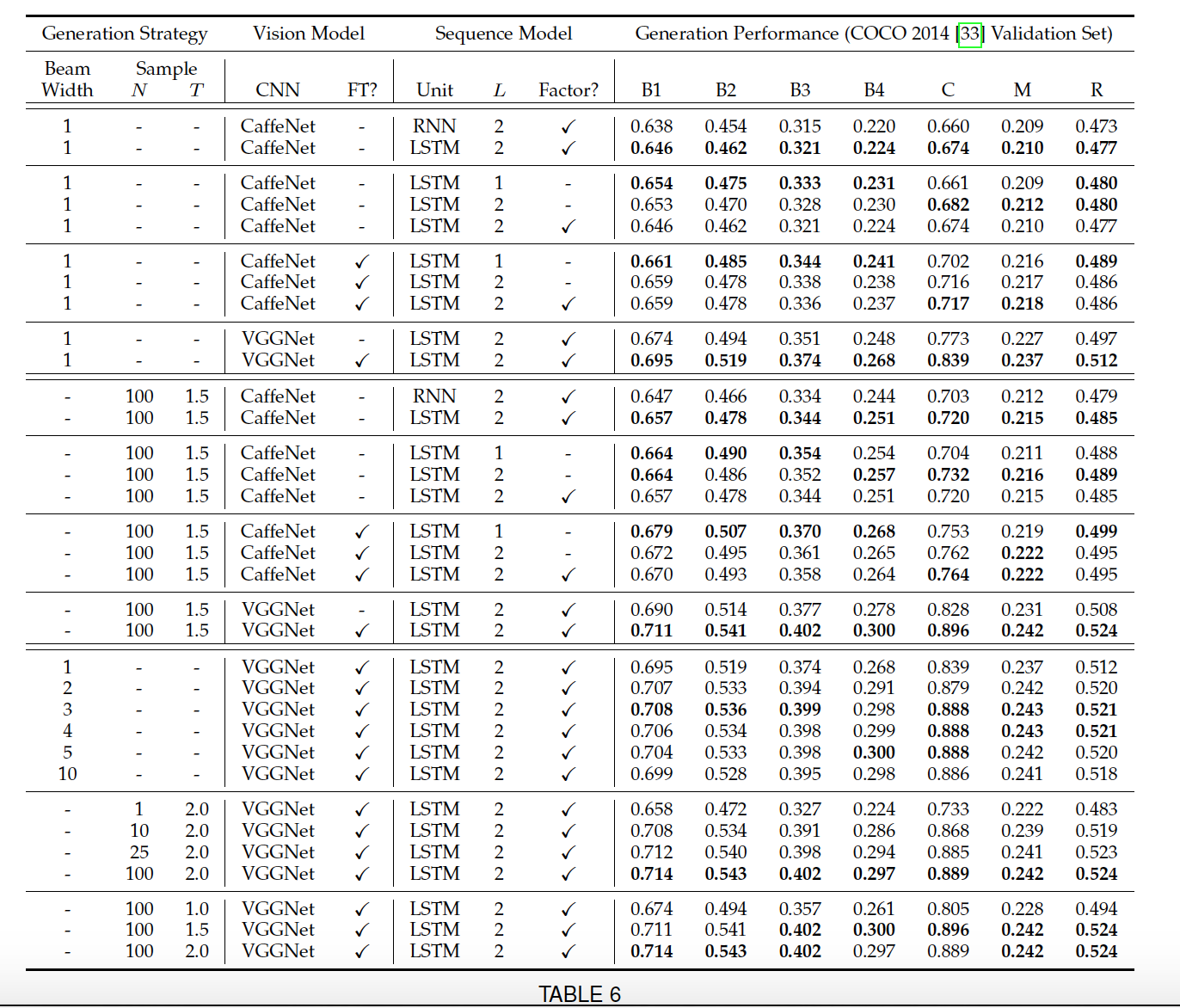
2. Baselines: Choosing most probable word in each time step is the baseline for caption generation. In the art of state of their model, they use beam search technique when generating words. In terms of model, they use simpler CNN models such AlexNet and CaffeNet to test their model instead of VggNet. They also try RNN’s instead of LSTM’s to compare results. Before VggNet+LSTM a CaffeNet+RNN model can be tested as baseline.

3. Upper bounds: Article don’t talk about upper bounds, however since it is image captioning and there is 5 alternative caption for image, there is no one correct answer and class. The ranking is a bit subjective. Thus, I cannot estimate upper bounds for which my model’s captions are in the top 5 captions in the datasets

4. State-of-the-art: In caption generation task BLUE-1 score is 0.714, BLUE-2 score is 0.543, BLUE-3 0.402, BLUE-4 0.297. CIDEr-D is 0.889, METEOR is 0.242, ROGUE-L is 0.524. In image generation task, they didn’t improve state of the art but they are very close the state of the art. The article “Show and Tell: A Neural Image Caption Generator” O. Vinyals et al. was the best results in METEOR(0.346), ROGUE-L(0.682), CIDEr-D(0.946) metrics. They use more complex CNN network (Inception model) in their article. The results are shown in the below figure.



5. Ablation study: Article tries RNN instead of LSTMs and simpler CNN models instead of VggNet as baseline. They also generate captions by looking most probable word or beam search strategy. As we can see from the below table, switching from CaffeNet to VggNet and RNN to LSTM improved a lot.



# DATA

Find and download all necessary data used in your paper. This is the most important part of this assignment, without data you can’t replicate the work. Please consult the class TA if you think you cannot obtain the data.

I downloaded all images and captions.

# UNKNOWN TERMS

Take the list of unknown terms you have prepared in the last assignment and look for definitions on the web. You can use the [deep learning glossary](http://www.wildml.com/deep-learning-glossary/) ([here are some more](https://www.google.com.tr/search?q=deep+learning+glossary)), [deeplearning.net](http://deeplearning.net/), a [textbook](http://www.deeplearningbook.org/), [wikipedia](https://en.wikipedia.org/wiki/Main_Page), [scholarpedia](http://www.scholarpedia.org/article/Main_Page), etc. (Please let me know if you find any other resources so I can add them to the list).

● **term1:** **LSTM**s are used for sequence learning tasks. We unroll same LSTM units and cascade them to create sequence without knowing their length. They have hidden unit, memory cell, forget gate, input gate, output gate. Sigmoid function used all of the gates. It determines which portion of data should be omitted. We train them by backpropagation by determining a time depth(T). Memory cell and hidden cells are the inputs of next LSTM unit.

● **term2: Semantic video role tuple predictors:** I couldn’t find exact definition but as I understand it’s a computer vision method not a deep learning concept.

● **term3**: **Conditional random fields** are a class of [statistical modelling method](https://en.wikipedia.org/wiki/Statistical_model) often applied in [pattern recognition](https://en.wikipedia.org/wiki/Pattern_recognition) and [machine learning](https://en.wikipedia.org/wiki/Machine_learning), where they are used for [structured prediction](https://en.wikipedia.org/wiki/Structured_prediction). Whereas an ordinary [classifier](https://en.wikipedia.org/wiki/Statistical_classification) predicts a label for a single sample without regard to "neighboring" samples, a CRF can take context into account; e.g., the linear chain CRF (which is popular in [natural language processing](https://en.wikipedia.org/wiki/Natural_language_processing)) predicts sequences of labels for sequences of input samples.

# RELATED WORK

Go over the paper’s related work, both past and future, you found in the last assignment. Apply [Keshav’s first pass](http://www.sigcomm.org/sites/default/files/ccr/papers/2007/July/1273445-1273458.pdf) to each, this should take 5-10 minutes per paper. Please give the citations and your understanding of these papers (and whether/why they matter to your paper) below:

* Paper1: Xu, Kelvin, et al. "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention." *ICML*. Vol. 14. 2015.

It uses similar model with my article. CNN decoder+LSTM encoder. However, in encoder part it uses attention mechanism. They define soft and hard attention mechanism in the article. Hard attention changes the stochastic gradient descent algorithm a bit, however, soft attention is similar to optimizating of the stochastic gradient descent with a lamda factor..

* Paper2: Karpathy, Andrej, and Li Fei-Fei. "Deep visual-semantic alignments for generating image descriptions." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015.

Their model segmentate images and describe each segment independently. They use BRNN(Bidirectional RNN) with Regional Convolutional Neural Networks AS baseline. Then, they put it to mrnn form.

* Paper3: Mao, Junhua, et al. "Deep captioning with multimodal recurrent neural networks (m-rnn)." *arXiv preprint arXiv:1412.6632* (2014).

MRnn models contains two word embedding layers, the recurrent layer, the multimodal layer, and the softmax layer. Word embedding layers convert one-hot word vector to more dense vectors. Rnn layers has 256 hidden unit and uses ReLU activation function. Then multimodal layers has three inputs: the word-embedding layer II, the recurrent layer and the image representation. For the image representation they use CNN network.

Then m(t) = g2(Vw · w(t) + Vr · r(t) + VI · I);