The authors describe an image caption generating system that combines recent ideas and successes from machine translation and image classification. Their system is designed and trained explicitly to go from input image to an output description, whereas most previous models have joined together models trained at the sub-tasks of object recognition, language generation, and more. Explicitly, the model is a single neural network that can be trained end-to-end. They draw an analogy to machine translation, and in fact use the very same model, replacing the source sentence encoder RNN with a CNN applied to the input image. Despite the model being trained end-to-end for captioning, the CNN is first separately pre-trained on ImageNet for image classification.

The paper's strength lies mostly in their strong results and evaluation methods. The model itself is relatively familiar once one is familiar with CNNs and LSTMs. The novelty is in treating the task as a form of machine translation. This distinguishes them from a prior work (Mao et. al) which uses neural networks, but not LSTMs, and does not feed CNN output directly to the language RNN. Their results are state of the art by a wide margin compared to previous results. Their decision to use only a generative model and not focus on ranking makes sense and is well defended. They dedicate much of the paper to discussing the various datasets they trained and tested on, and the evaluations performed, to ensure a rigorously tested baseline for future work. Their motivation of accessibility is also sound.

There are some weaknesses in their methods and evaluation, though. As for their methods, assigning only two turkers per image might result in high variance. With three or more, it would be easier to have a "majority vote". However, it's a net positive that they do use human evaluation to corroborate their BLEU scores at all. Also, though they mention briefly the chosen CNN model's performance with respect to transfer learning, there is no comparison to other possible architectures or motivation for why they were not chosen. Finally, the vocabulary size is not shared, although this could influence their qualitative word similarity results later.

The evaluation is strong, but there are further ways to explore. Though they mention the data is smaller than that of, say, ImageNet, they don't provide training times or hardware capabilities necessary to replicate their results. It's good that they use METEOR and CIDER, but they don't devote any of the writing to analyzing why their system which beat humans on the MSCOCO development set by the BLEU metric failed to do so by these metrics. Finally, some quantitative testing on the word embedding similarities, or a quantitative approach to comparing them to embeddings trained on other corpora, would likely be illuminating. I suspect there would be some interesting differences owing to the visual nature of the encoded data.

Overall, this is an important paper is it is one of the first to show how a single NN architecture can combine vision and language for compelling results. One relation and possible future research direction is in training the discriminative ranking task, and transferring those gains to the generative task, as done for visual dialog. One could also tune the network to do well on objects/words not seen in training images or captions using some nearest neighbors method with transfer learning from ImageNet and some language corpora.
Most Interesting Thought

Where else can we apply this encoder-decoder model like this? It seems very general, and one could imagine generating sentence descriptions of anything given enough training data! It could be tried on video, as the paper mentions early works focused on, audio recordings of scenes, or any sort of structured business data. One could imagine this method eclipsing current slot-filling methods for article summarizing sports games from statistics, and being more "natural" in the process.