Ikhlas Attarwala
04-30-19 / Week #5
Applications of NNLMs / Paper 2

DenseCap: Fully Convolutional Localization Networks for Dense Captioning
by Johnson, Justin et al.

There are 4 common subtasks to image understanding: **classification** (single label to whole image), **captioning** (description of whole image), **detection** (find objects/regions in whole image and assign labels to each), and **dense captioning** (similar to detection but descriptions in place of labels). DenseCap is a **dense captioning** task that uses computer vision (CV) to localize & describe notable regions in images, into natural language. DenseCap has 4 subcomponents:

1. a VGG-16 CNN
2. a localization layer that extracts features around anchor points, generates attributes of boxes, and then extracts the features of these boxes
3. a recognition network that takes region, converts them into 4096-length vectors, and feed it into an RNN
4. an RNN language model that takes the vectors, applies a fully connected layer and ReLU to it, feeds it into an LSTM that generates a word, and repeats until the LSTM hits an ending token.

I found this paper interesting, and the captions their model generates per region is pretty great. It also appears the captions include lots of color-related vocabulary, and best of all (in my opinion), the regions and captions that are generated per image enable someone to search for specific objects within the images using text queries. Their model also tends to outperform EdgeBoxes on the Visual Genome dataset and find lots of noteworthy regions. One “flaw” that’s not really a flaw that I would point out is that the regions that are produced don’t seem to mimic the average regions humans tend to pay attention to. For example, in an image of an elephant playing soccer, there are dozens of regions DenseCape creates, but some regions that a human might pay attention to such as the tusks, are not boxed, but other regions such as a portion of the **ground** below the soccer ball are, which we probably wouldn’t pay much attention to.