**Project Planning**

VizWiz Competition: <https://vizwiz.org/tasks-and-datasets/image-captioning/>

<https://eval.ai/web/challenges/challenge-page/739/overview>

Goal:

* Pick the best caption from a list of 5, given a picture and 5 captions.
* Pass pictures through CNN, and get latent representations.

Text Processing:

* Word2Vec, Embedding Layers, Tf-IDF

\*At the End\* Transformer

Next Steps:

* Understand Data
* CNN with images
* Research image captioning
  + We need to reference papers for our presentation
  + Literature Review
* Try to find past solutions

Meet Again: 11/3/2021 @ 8pm

**Resources:**

* Show and Tell: A Neural Image Caption Generator
  + <https://arxiv.org/pdf/1411.4555.pdf>
* <https://colab.research.google.com/drive/17rgnW8FemmDiCevDaiPTZ84RvXWFhjL5?usp=sharing>

------------------------------------------------------------------------------------------------------------------------

Questions:

* What is is\_rejected, is\_precanned? annotation.json
  + I know text\_detected means there is text in the photo.
* Anyone know how to batch data so we aren’t loading it into memory all at once?
  + Images we can do flow\_from\_directory, but what about text?
  + Don’t have to worry about text data

------------------------------------------------------------------------------------------------------------------------

November 5th Minutes:

Project planning:

* Preprocess json files into an easier format to work with.
  + E.g. convert to arrays or pandas df.
  + Process captions
    - Text processing (tf-idf, Word2Vec, etc.)
  + Image ID, Caption ID - Image ID
  + Dataframe
  + Image ID and List
  + Image ID and Caption ID
    - One of us creates this df, uploads it to google
* Pass training images through a model to retrieve feature vectors
  + Some of us can get feature vectors using different models
* Concatenate text data & images, feed to a single model
  + RNN

**Phase 1:** Before Nov 16 (11 days)

* Preprocessing Text Data (Vishal)
  + Dataframe with imaged IDs and captions - Done
* Process captions (Vishwa)
  + Vectorize/tokenize
  + N-grams, lemmatization, stemming
* Process Images (Multiple Models & Multiple People) (Leonid, Adam, Himayu)
  + VGG-16 (Adam)
  + Output should be defined size, e.g. (2048, )
  + Output layer is before the multi classification layer
    - Feature Vectors
* If done early, we can start researching how to combine models

**Phase 2:** Until Nov 30

* Combine text and image models into a single model.
  + Figure out how to turn that into a sentence.
* Presentation Slides
* Report (Might be able to submit after 30th)

I’ll send out a When2Meet so we can meet either 11/11 or 11/12

Group Project: Nov 30 - Presentations - Deadline for presentation slides & code