Passive and Active Tweets

Ongoing work to improve quality of tweets (details in next three sections)

- Social media data streams, such as Twitter, are important sources of real-time and historical global information for science applications, e.g., augmenting validation programs of NASA science missions such as Global Precipitation Measurement (GPM).
- Determinant of output tweet quality from our tweet processing infrastructure is therefore
- **Determinant of output tweet quality from our tweet processing infrastructure is therefore**
- Deep learning models, particularly Convolutional Neural Networks (CNN), have been shown to be very effective for large-scale image recognition and classification.
- Because a large number of labeled images is required to develop CNN, doing so from scratch would be very costly in compute and time resources.
- Transfer Learning: takes advantage of pre-trained models as a starting point, thus mitigating the cost of model development for the current task. These reused models are, in effect, feature extractors, the outputs from which then become inputs for training smaller, more manageable classifiers.

- For the current task, we used VGG-16 as the feature extractor, by removing the final fully connected layers, and trained a linear support vector machine (SVM) to output the final classification (i.e., precipitation and “not-precipitation”).

Construct classifier to analyze images for precipitation-related information (e.g., is there rain in the image? is it a forecast map?)

- For the current task, we used VGG-16 as the feature extractor, by removing the final fully connected layers, and trained a linear support vector machine (SVM) to output the final classification (i.e., “precipitation” and “not-precipitation”).

Classifying Tweet-linked Documents

Use Hierarchical Attention Network (HAN)-based model to classify tweet documents, i.e., precipitation occurrence, type, and intensity, at given locations and times.

**Architecture: Hierarchical Attention Networks**

- **Word/Sentence Encoder** – Embed words to vectors through embedding matrix; apply bidirectional GRU to obtain representative, contextual hidden annotations of words/sentences.
- **Word/Sentence Attention** – Combine learned measure of importance with contextual word/sentence annotations; more relevant words have greater weighting, and vice versa.
- **Classification – Softmax**, categorical cross-entropy
- **Regression – Softmax**, mean squared error

**Features**

- Precipitation Occurrence
- Precipitation Type
- Precipitation Intensity

**Model Attention Weight Visualizations**

Model places greater importance on darker words during classification / regression.

**Results / Alternate Model Comparisons**

- Data Type / Algorithm
  - Precipitation Occurrence (Yes/No)
  - Precipitation Type (Rain/Snow/Crime/No)
  - Precipitation Intensity (mm/hr)
  - **83.9%**

**Confusion matrix of classification on a 90-image test set, split into two target categories, “precipitation” and “not-precipitation.”**

**Applications**

- Original Facebook posts scraped using Python program by minimai2
- More data in training set -> greater label accuracy
- 300+ training data -> about 80% correct classification

**Summary**

- Paradigm shift from a focus on recruiting citizen scientists to enriching the Twitter stream enables an unbounded approach to recruitment.
- Output tweet quality and quantity from our tweet processing infrastructure is increased by complementing “passive” tweets from the Twitter stream with tweets from “active” participants.
- Ongoing work to improve quality of tweets include (1) classifying documents and images that are endpoints of links in tweets, to extract additional information relevant to the tweets and (2) classifying Facebook weather pages and converting them to “active” tweets.

**References**

1. VGG-16 (Visual Geometry Group-16, also OxfordNet),
   https://github.com/kuunaka/keras-vgg16/
5. NASA/Goddard’s Earth Sciences Data and Information Services Center (GES DISC)
6. NASA/Goddard Space Flight Center; ADNET Systems, Inc.; Cornell University; University of Michigan Ann Arbor; CUNY Queens College
7. NASA Citizen Science for Earth Systems Program, for the project, “Mining Twitter data to augment NASA GPM validation.”

**Acknowledgments**

This work was supported, in part, by
- NASA Citizen Science for Earth Systems Program, for the project, “Mining Twitter data to augment NASA GPM validation.”