Experimenting with High Dimensional Vector Representations of Instagram Users
David Morales, Sumit Minocha
Department of Computer Science, Stanford University

Motivation and Problem

Problem / Questions
- How can we best capture a user's interests via their Instagram likes, posts, and captions?
- How can we quantify and interpret a person from their social media presence?
- Can we improve the representative power of a user embedding (on top of a naively constructed linkage graph) by additionally incorporating image and text data?

Dataset
- Instagram dataset of 17 posts each for 972 influencers (Iconosquare Index) [1]
  - Includes handle, caption, tags, mentions metadata.
  - Scraped post urls to get image data as well.
- Preprocessing involved group by key operation (by username), translation and character filtering of captions, and image scaling.

Approach

Approach Components:
- Learning Components: Node2Vec, LDA, CNN used to transform tag similarities, text and image data into embeddings for each user.
- Autoencoder: Embedding outputs from previous steps concatenated and pooled by autoencoder to yield final, combined vectors for each user.

Experimental Setup
- Baseline: Construct linkage graphs where links exist between users sharing tags. Running node2vec on this undirected graph gives embeddings based on high-level user similarity.
- Evaluation: If we were to build a graph from resulting embeddings, would it be meaningful?
  - Quant: Measurements of projected network structure (communities, clustering coeff, connected components)
  - Qual: Closest neighbor to a user in each embedding space

Overall Architecture

More Results

Quantitative Summary
We randomly sample a user and find its nearest neighbors (based on cosine similarity) in each of the constructed embedding spaces:

From the table of similarities above, we can see that the space between embeddings in each space are comparable, but a lot closer in the visual, CNN embeddings.

Attributes of Similar Users
- Fig 1: Closest user to ‘sejkko’ (left) was "faby_mamaedegemeos" (right) in LDA space (text)
- Fig 2: Closest user to ‘sejkko’ (left) was ‘danrubin’ (right) in CNN embedding space (from images)
- Fig 3: Closest user to ‘sejkko’ (left) was ‘alexandreagarza’ (right) in LDA space (text)

Future Work
- K-means: perform on each embedding space (use #communities from results tab to make informed choice of k), then can further analyze users in each cluster or even try to interpret / compare the ‘average user’ or centroid of 2 neighboring clusters

Analysis

Qualitative
- Users baseline considered similar were marginally so, only some select examples illustrated embeddings’ competence.
- Visual embeddings seem to overvalue the sky, maybe due to our pre-trained CNN we used?
- LDA output recognizes similar travel theme
- Pooling AE extracts similar very similar users

Future Work
- K-means: perform on each embedding space
  - Use #communities from results tab to make informed choice of k, then can further analyze users in each cluster or even try to interpret / compare the ‘average user’ or centroid of 2 neighboring clusters

Sources

Conclusion

Embedding Space Capacity
- Each embedding space clearly captures different aspects of similarity between users (as expected).
- Our final autoencoder result actually was quite promising and outputted similarly successful results as the one pictured for each of the top 5 users considered most similar to ‘sejkko’
- Industry applications span product recs, content filtering, or targeted advertising