Experimenting with High Dimensional Vector Representations of Instagram Users

David Morales, Sumit Minocha
Department of Computer Science, Stanford University

Stanford | ENGINEERING
Computer Science

Motivation and Problem

Problem / Questions

- How can we best capture a users' 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

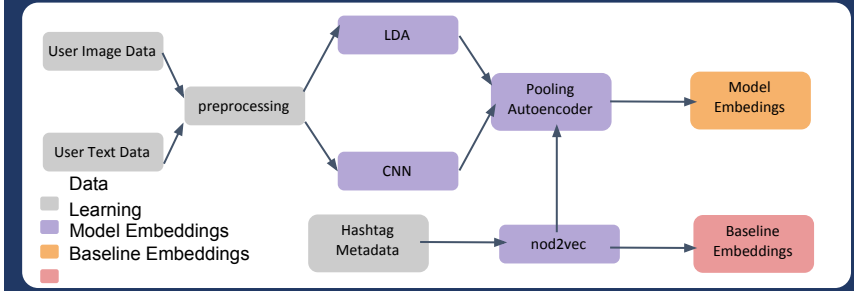
Approach

- **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:

```
Most similar user in node2vec embedding space has similarity 0.8319389820098877
Most similar user in lda embedding space has similarity 0.9286023974418864
Most similar user in cnn embedding space has similarity 0.3504593126577759
Most similar user in pooling embedding space has similarity 0.9019187922440836
```

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)

#whpflowerpower

link na bio. #introducaoalimentar #ia #gêmeos #Twins #mamaedegemeos #maedegemeos #maternidade #instababy #maedemenina

Fig 2: Closest user to 'sejkko' (left) was 'danrubin' (right) in CNN embedding space (from images)

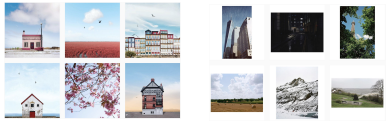


Fig 3: Closest user to 'sejkko' (left) was 'alexandreagarza' (right) in LDA space (text)

sejkko Thinking of beautiful Porto

alexandreagarza Already missing that Hawaii life, new vlog is live! Sharing all of our fave spots and things to do in Maui. Linking my swimsuit and shorts here http://liketk.it/2C8I and on the @liketoknow.it app #liketkit

Fig 4: Closest user to 'sejkko' was 'finn' (see example posts and captions on the left) in Autoencoder space (combined)

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

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

Sources

[1] Giovanni Alcantara. *Instagram Like Predictor. GitHub.* 2017.

[2] Grover et al. *node2vec: Scalable Feature Learning for Networks.* July 2016