Critique Report on the Paper: “Extracting Possessions from Social Media: Images Complement Language [2]”

Sumaiya Tabassum Nimi

Background Information and Purpose of the Research

Possession is the correspondence between two objects such that one object called the possessee is at the disposal of another object called the possessor. In this paper [2], possession relation was extracted from social media posts (in this case tweets) that contained both texts and images, making use of both these categories of data. In addition to detecting the possession relation, some associated information were also extracted, like the temporal characteristics of the possession and interest of the possessor in the possessee. It was also shown that both human and machine benefit from both texts and images while determining these aspects associated with possession.

Methods Proposed

A novel dataset consisting of 5000 tweets was constructed, that were labelled manually, using only texts as well as both texts and images, with the types, temporal characteristics and presence or absence of interest associated with the possession relations. Then a deep learning model was constructed for learning these associations. The model learned using the following three types of data that came from both texts and images.

- Text Component: LSTM [3], whose token is the concatenation of the following three-fold embedding of the words contained in the tweet.
  - Glove Word Embedding [4] pretrained with Common Crawl
  - Glove Word Embedding pretrained with Twitter
  - Fine-tuned Binary Embedding indicating whether the word indicates a potential possessee.
- Image Component-1: Weights of average pooling layer of pretrained InceptionNet [5].
- Image Component-2: Glove word embedding of top 5 tags of the image generated by Google Cloud Vision API [1], passed to LSTM. This proposed component was a novel contribution of the paper.

Strength of the Work

- Novel open-source dataset of 5000 tweets, manually annotated with information like types of possession relations, temporality of the found relations and interest of the possessor in the possessee was built for the work, that would facilitate research works conducted in the domain in future. Similar datasets were never annotated with the last two types of information.
- Statistically validated that both human and machine learning models can understand possession relations better when they have access to both texts and images, than when they are provided either.

- Novel idea of generating additional textual information by tagging the visual cues was proposed. Incorporating this information improved performance of the model employed for analyzing the possession relations.

Limitations of the Work

- Inferences involved getting images tagged by Cloud Vision API which could not be done offline, i.e. without internet connection. So the model would not work in standalone edge devices.
- Although not explicitly reported in the paper, the proposed model seemed memory intensive, as even a single inference would involve a large number of information gathered from both texts and images. Hence it would be difficult to load the model in memory constrained edge devices.
- Also the issue of inference time was not discussed in the paper. But it looked like inference would be very slow, as it would involve generating a large number of tokens and embedding from many different models, one of which involves a call to a remote API. Hence real-time inference would not be possible.
- Stop words were not pruned while generating tokes from the text, that led to unnecessary calculations done by the model, making the inference slower.

Suggested Future Studies

- Call to Vision API can be avoided by training a lightweight model to generate tags from images. That way it will be possible to infer offline also.
- Even not all, some stop words, especially the articles could be pruned, resulting in reduction of inference time without affecting the performance.
- InceptionNet can be replaced by a more lightweight counterpart, leading to reduced inference time and less load on memory.

References

[1] Cloud Vision API. https://cloud.google.com/vision/. Accessed: 09-21-2020.

[2] Dhivya Chinnappa, Srikala Murugan, and Eduardo Blanco. Extracting possessions from social media: Images complement language. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 663–672, 2019.

[3] Sepp Hochreiter and Jürgen Schmidhuber. Lstm can solve hard long time lag problems. In *Advances in neural information processing systems*, pages 473–479, 1997.

[4] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[5] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.