Paper Citation Classification

Chung Bui, Sophia Liu, Camilo Saavedra
{csb41, sophliu, csaavedr}@stanford.edu

Motivation

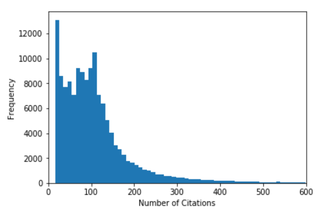
- Choosing a research paper topic can be a difficult task
- We would like to predict the number of citations a paper will receive so researchers can more easily find relevant fields and topics
- Various models were used to predict the binned citation number given inputs from the paper, such as the title
- Model performance was evaluated with average F1 Score

Dataset

We originally attempted to gather a dataset using Crossref, but found it difficult to collect enough papers. Instead, an existing dataset was used from a study on the advantages of short paper titles [1].

In total, there are 140,000 records of general academic papers from 2007 to 2013. Each paper includes:

- Year of publication
- Paper's title
- Journal's title
- Length of paper title
- Number of citations



Features

- Input features include title, year, title length, and journal
- Word Embeddings: Transform title to input feature
  - Tried several pretrained models
  - Froze the weights to account for limited dataset

Results

Naive Bayes

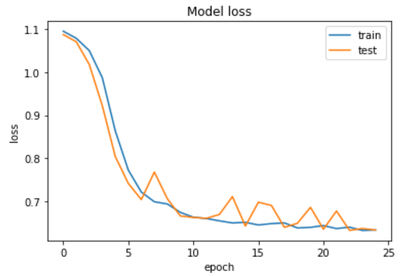
Citation Bins	<60	60-120	>120
<60	0.78	0.27	0.00
60-120	0.14	0.48	0.28
>120	0.04	0.34	0.67

Random Forest

Citation Bins	<60	60-120	>120
<60	0.82	0.11	0.02
60-120	0.11	0.61	0.38
>120	0.03	0.38	0.55

Neural Network

Citation Bins	<60	60-120	>120
<60	0.83	0.13	0.00
60-120	0.12	0.59	0.38
>120	0.03	0.24	0.68



Model	Train F1 Score	Test F1 Score	Num. Train Samples	Num. Test Samples
Naive Bayes	0.4796	0.4783	118915	20986
Random Forest	0.9999	0.6336	118915	20986
Neural Network	0.7027	0.7026	118915	20986

Future Work

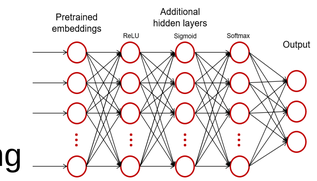
- Collect more data in relevant fields
- Train encoder to perform encoding from large datasets of academic titles
- Include abstract ideas and author in input data
- Explore more advanced machine learning techniques to generate relevant paper features, such as titles

Models

- Multinomial Naïve Bayes Classifier
  - Baseline Model

$$\mathcal{L}(\phi_y, \phi_{k|y=0}, \phi_{k|y=1}) = \prod_{i=1}^{n} p(x^{(i)}, y^{(i)})$$
$$= \prod_{i=1}^{n} \left(\prod_{j=1}^{d_i} p(x_j^{(i)} | y; \phi_{k|y=0}, \phi_{k|y=1}) \right) p(y^{(i)}; \phi_y)$$

- Random Forest
  - High Variance Model
  - Fitted 100 decision trees onto dataset sub-samples
- Neural Network
  - Best Results
  - Pretrained embedding layer followed by 3 fully connected layers



Discussion

- Many model options were considered
- Best performance was achieved with pretrained embeddings and neural networks with an average F1 score of 0.7
- The performance is invariant to the pretrained embedding, need to add more features to improve

References

Ibanez, A., et al. “Predicting Citation Count of Bioinformatics Papers within Four Years of Publication.” *Bioinformatics*, vol. 25, no. 24, 2009, pp. 3303–3309., doi:10.1093/bioinformatics/btp585.

Letchford, Adrian, et al. “The Advantage of Short Paper Titles.” Royal Society Open Science, vol. 2, no. 8, 2015, p. 150266., doi:10.1098/rsos.150266

Yan, Rui, et al. "Citation count prediction: learning to estimate future citations for literature."Proceedings of the 20th ACM international conference on Information and knowledgemanagement. ACM, 2011.