Data challenge: link predictions in a graph of citations with abstracts and co-authors.

Hugo Simon,Enguerrand Chary, Samuel Gruffaz

Presentation

Tuesday, 15 Fabuary 2022

Source of informations:

- Citations between papers, $G = (V, E)$ (undirected graph).
- Abstract associated to each paper.
- coauthors relations.
- Surprisingly, we don't have time information ! It forces us to be agnostic of the underlying time process.

In real life, two types of reasons for citing a paper:

- We cite a paper because it treats the topic where we work. (information coming from the abstract)
- We cite a paper because it is the paper of a friend. (information coming from the authors network)

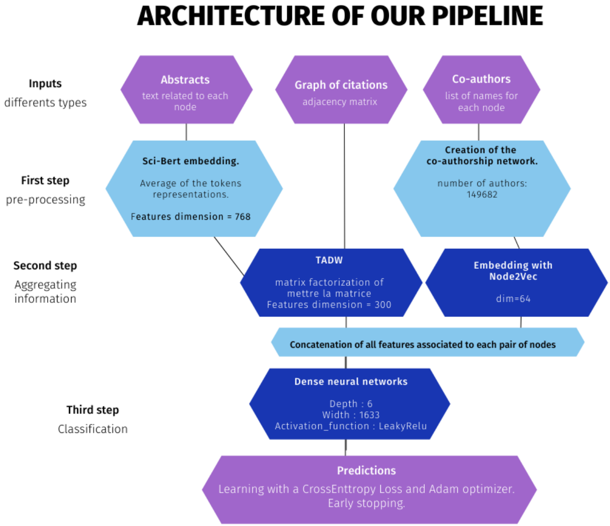

Figure: Summary scheme of our pipeline, mettre notre score final

Pre-processing:

- SciBERT is a BERT Architecture retrained on a corpus of scientific paper (with a big part of computer science paper). We load it 2 from hugging face. Each abstracts i was embbeded with a vector f_t of dimension 768.
- For the co-authorship graph, there is an edge between author i and j if they have written a paper together.

The principle of TADW is a matrix factorization:

$$\operatorname{argmin}_{W,H} ||M - W^T HT||_{frob}^2 + \lambda(||W||_{frob}^2 + ||H||_{frob}^2), \quad \lambda > 0$$

where:

- $M = \frac{1}{2}(\bar{A} + \bar{A}^2)$, $\bar{A}$ the normalized adjacency matrix of $G = (V, E)$.
- $T \in \mathbb{R}^{n_v \times 768}$, the abstracts features.
- $W \in \mathbb{R}^{n_v \times k}$, the structure features, k an hyper-parameter.
- $H \in \mathbb{R}^{k \times d}$ a transformation to create an interaction between W and T.

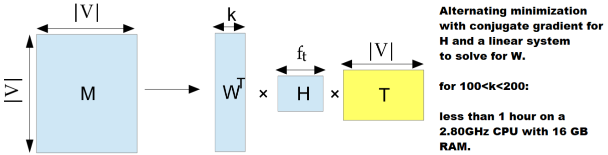

Figure: Matrix factorization.

We used Node2VEc to have a numerical representation of authors proximity.

- 2nd Order random Walk
- applies a bias factor alpha to reweigh the edge weights depending on the previous state.
- alpha is a function that takes as inputs the current node and the potential next node.
- gives a good and homogeneous exploration of information

Putting together

Finally, we put all together.

- Article author features are made by concatenating average author embeddings and the embeddings of the 2 closest authors for each pair of articles. The idea is that the likelihood of citation between 2 articles depends on the proximity of its closest authors.

- Article final features are made by concatenating article TADW features and article author features.

Training set (x_i, y_i) for a classification task:

- All existing edges are added with label $y = 1$.
- We sample the same number of non-existing edges with label $y = 0$.
- We concatenate the article features associated to the two nodes of an edge to get the final features (x_i). We double the size of our dataset by permuting nodes features in the concatenation to enforce the invariance due to undirected edges.

The classifier:

- MLP classifier with 5 hidden dense layers, LeakyReLU activation function, and 0.3 dropout rate.

Results

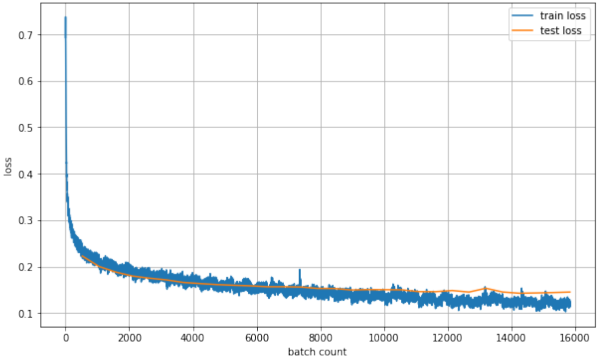

Figure: Loss plot of our final training

We reach a validation loss of 0.1434 but only score 0.1613 submitting prediction to the challenge, which means our dataset is not exactly representative of the scoring set.

Interpretation of some phenomenons:

- When we add the features related to co-authorship, we decrease the loss by 0.04. There is a real source of information in the relations information.
- Little improvement between $k = 150$ and $k = 300$, we can reduce the dimension of text features !

If we want to do better:

- Adding features in the co-authorship graph with topic modeling and process it with TADW or something else.
- Work differently with pluridisciplinary papers and cross-domain collaboration [1].
- Using a directed graph to compute fisher information metrics related to citations.

Jie Tang, Sen Wu, Jimeng Sun, and Hang Su.
Cross-domain collaboration recommendation.
In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1285–1293, 2012.

David Liben-Nowell and Jon Kleinberg.
The link-prediction problem for social networks.
Journal of the American society for information science and technology, 58(7):1019–1031, 2007.

Robin Brochier, Adrien Guille, and Julien Velcin.
Link prediction with mutual attention for text-attributed networks.
In *Companion Proceedings of The 2019 World Wide Web Conference*, pages 283–284, 2019.

Cheng Yang, Zhiyuan Liu, Deli Zhao, Maosong Sun, and Edward Chang.
Network representation learning with rich text information.

In *Twenty-fourth international joint conference on artificial intelligence*, 2015.

Peng Wang, BaoWen Xu, YuRong Wu, and XiaoYu Zhou.
Link prediction in social networks: the state-of-the-art.
Science China Information Sciences, 58(1):1–38, 2015.

Iz Beltagy, Kyle Lo, and Arman Cohan.
Scibert: A pretrained language model for scientific text.
In *EMNLP*. Association for Computational Linguistics, 2019.

Hsiang-Fu Yu, Prateek Jain, Purushottam Kar, and Inderjit Dhillon.
Large-scale multi-label learning with missing labels.
In *International conference on machine learning*, pages 593–601. PMLR, 2014.