TableCNN: Deep Learning Framework for Learning Tabular Data *

Pranav Sankhe, Elham Khabiri, Bhavna Agrawal, and Yingjie Li

[1] University at Buffalo, Buffalo, NY
[2] IBM Research, Yorktown Heights, USA
pranavgi@buffalo.edu
{khabiri,bhavna,yingjie}@us.ibm.com

Abstract. Databases and tabular data are among the most common and rapidly growing resources. But many of these are poorly annotated (lack sufficient metadata), and are filled with domain specific jargon and alpha-numeric codes. Because of the domain specific jargon, no pretrained language model could be applied readily to encode the cell content. We propose a deep learning based framework, TableCNN, that encodes the semantics of the surrounding cells to predict the meaning of the columns. We propose application of Byte Pair Encoding (BPE)[5] to create tokens for each cell and treat each cell as a phrase of existing tokens. Once tokenized, we process it with a CNN network to develop a classifier.

1 Introduction

Tables are rich in data and can provide vital information about the object due to the virtue of its structure. Extracting useful insights from tabular data may require domain expertise, especially if the information is comprised of domain specific jargon's or alpha-numeric codes. Our method provides a supervised learning solution to classify an unknown column in such tables into predefined column classes which can also come from a knowledge graph. Existing methods[2, 3, 1] cannot accommodate such data.

2 Methodology and Results

Cell entries are tokenized using Byte-Pair Encoding with a stopping condition defined by the token frequency threshold. Cell embedding is generated using Word2Vec[4]; each row across the tokenized table is treated as a sentence for Word2Vec model learning. We extract micro tables from the table, with a target column and surrounding columns having set number of rows; which are model parameters. Micro tables are then processed through TableCNN to classify which

* Supported by IBM Research

class it belongs to. Class is defined as the header of a column in the table used for training.

Network: TableCNN Fig 1 is an abstract view of the TableCNN network architecture. We extract column and row features in separate networks and combine them to regress final output. Row and column features are computed by a convolution operation over the first row and the target column of micro table respectively. Outputs from row and column features are concatenated and fed to a fully connected layer with a SoftMax layer to make final prediction.

For our experiment, we use a manufacturing database table that contains 112 columns and 115 thousand rows. Fig 2 shows that we obtain correct column predictions, except for column 49. Upon further inspection, we found that most of the cells in column 49 were empty, and thus the loss of classification accuracy. This shows that network is able to learn features from the surrounding columns along with the target column entries.

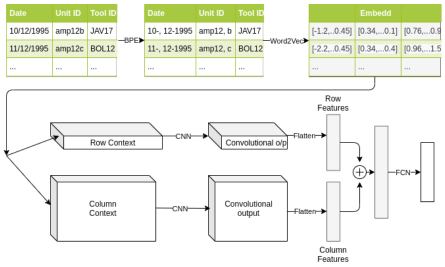

Fig. 1: TableCNN Architecture.

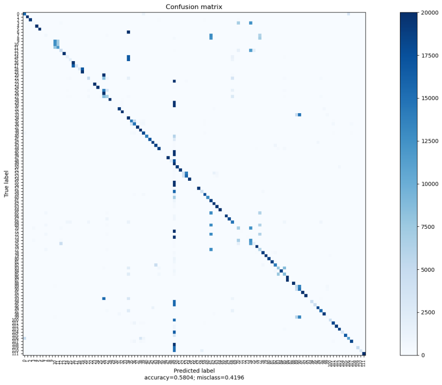

Fig. 2: Confusion Matrix.

3 Conclusions

In this poster, we present a supervised learning framework that can classify columns of a table with arbitrary alpha-numeric data. The arbitrary alphanumeric nature of data prevents us from using pre-trained language models.

References

1. Chen, J., Jiménez-Ruiz, E., Horrocks, I., Sutton, C.A.: Learning semantic annotations for tabular data. CoRR **abs/1906.00781** (2019)
2. Efthymiou, V., Hassanzadeh, O., Rodriguez-Muro, M., Christophides, V.: Matching web tables with knowledge base entities: From entity lookups to entity embeddings. In: International Semantic Web Conference (2017)
3. Limaye, G., Sarawagi, S., Chakrabarti, S.: Annotating and searching web tables using entities, types and relationships. PVLDB **3**, 1338–1347 (09 2010)
4. Mikolov, T., Chen, K., Corrado, G.S., Dean, J.: Efficient estimation of word representations in vector space. CoRR **abs/1301.3781** (2013)
5. Sennrich, R., Haddow, B., Birch, A.: Neural machine translation of rare words with subword units. ArXiv **abs/1508.07909** (2016)