Transformers for Electricity Price Forecasting

Oscar Llorente, Jose Portela,

***Abstract*—The abstract goes here.**

***Index Terms*—IEEE, IEEEtran, journal, LATEX, paper, template.**

I. Introduction

II. Attention in Deep Learning

This section will show the evolution of Attention mechanisms throughout time, starting in the NLP field where they were created.

There are many problems in Artificial Intelligence, as classification based on features, that do not need any temporal notion. In the field of Deep Learning the same happens. An example in this area would be Image Classification or Image Segmentation. However, there are other areas, as NLP, where this type of knowledge is needed. In language the order of words matters. For this reason from the beginning the field has used a different type of algorithms than Computer Vision. The first deep learning approach was Recurrent Neural Networks [1]. This type of Neural Network ha a structure that allows it to remember information of past events. Theoretically, this type of Neural Networks is able to learn very long term dependencies, something fundamental for NLP. However, in practice this is not the case, as explored in [2]. For solving these, another variant from RNNs was created: Long Short Term Memory Networks [3].

One of the main problems approached by the NLP community is translation. This is a problem of sequence to sequence type, where the output is not only a label, as in Image or Text Classification, but a multiple output, in this case a complete sentence. Another example in NLP domain would be Question Answering. For this type of problems in NLP is used a structure of encoder-decoder. To improve the encoder-decoder architecture explained before, another type of layer was presented, an Attention layer [4]. This layer is, in fact, the basis of the Transformer model. The Attention layer allows the decoder to focus its attention in a specific word or group of words from the input of the encoder (the original sentence in the case of translation). This helped improved the State of the Art of many NLP problems. However, in 2017, Google developed a model based only on the Attention layer, without any Recurrent layer, the Transformer.

One of the advantages of the Transformer architecture is that it enables the generation of much bigger Deep Learning models. An example of this would be the BERT model [5], used for many purposes by retraining it with new data. There is, in fact, an entire python library dedicated to this purpose which is used widely called transformers. Then, in recent years, extraordinarily large models that have accomplished really difficult tasks, like generating very realistic stories (GPT3) [6] or even writing programming code (Codex) [5], were created. Beside the NLP domain, where the effectiveness of the Transformer is undeniable, recently the Transformer architecture has achieved great success in other areas, as Autonomous Vehicles [7] or Image Classification [8]. This is, as mentioned before, one of the main reasons that motivates this article because it proofs that this type of model can be used in other areas successfully. As an example of how other data can be treated as words to use a Transformer one can look into the structure of Vision Transformer (ViT), the Transformer applied to Image Classification.

O.Llorente is with Ericsson, e-mail: oscar.llorente.gonzalez@ericsson.com.
J. Doe and Jane. Doe are with Anonymous University.
Manuscript received April 19, 2005; revised August 26, 2015.

III. Proposed Transformer model for Electricity Price

In this section the model proposed will be explained. Here the structure proposed is based on a Transformer Encoder only, as in the case of BERT [9] or ViT [8].

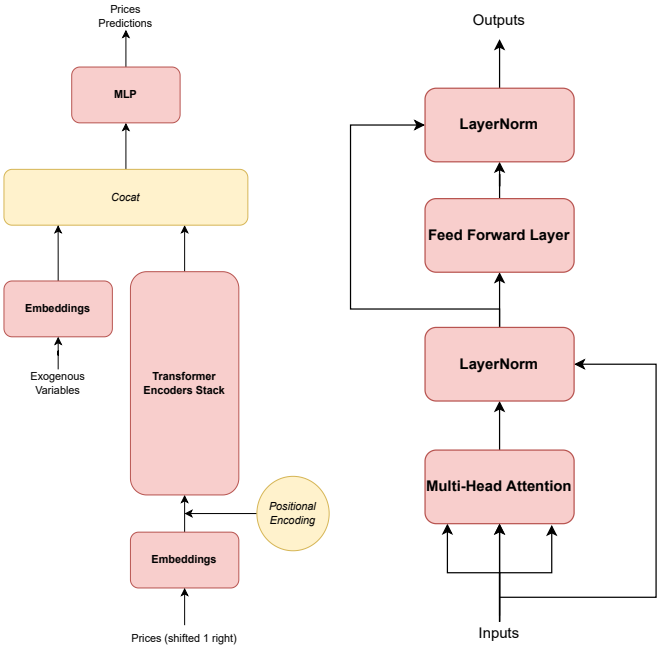
First, it is important to take into account that the model has two different sources of data, the prices and the exogenous variables. Therefore, some modifications to the encoder architecture have been made to include that. The architecture can be observed in Figure 1a.

The model has four main structures, two types of Embeddings, a stack of Transformer Encoders and a final Multi-Layer Perceptron for making the final predictions. Due to the importance of the Transformer Encoder, its structure can also be observed in Figure 1b.

In this case, since the objective will be to predict the 24-hour price for the next day, each day with its 24 hours will be treated as a single unit. Therefore, following an analogy with NLP, the original domain for this type of model, each day can be considered as a word.

Next, it is also important to clarify the Embeddings layers. Even they are not really Embeddings as in NLP, this name has been adopted following the example from language models. These layers are in fact a Linear layer and a ReLU non-linearity to transform the dimension of the inputs. Then, this layers will transform each word in a higher order vector. It is also important to take into account that there are two Embeddings layers, one for prices and another for exogenous variables.

At this point is important to clarify that not all the elements will be passed on to the Transformer layers stack. Only the prices shifted one place to the right (since the objective is to predict the next 24h only looking at the past) will be introduced in the Transformer Encoders layers. Then, the output of the stack will be concatenated with the embeddings of the exogenous variables (without any shift) and introduced in a multilayer perceptron to make the final prediction.



(a) Model Architecture (b) Encoder Architecture

Fig. 1: Model and Transformer Encoder Architectures

IV. Case Study

A. Set Up

V. Conclusion

The conclusion goes here.

Appendix A
Proof of the First Zonklar Equation

Appendix one text goes here.

Appendix B

Appendix two text goes here.

Acknowledgment

The authors would like to thank...

References

[1] D. E. Rumelhart, J. L. McClelland, Learning Internal Representations by Error Propagation, in: Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations, MIT Press, 1987, pp. 318–362.

[2] Y. Bengio, P. Simard, P. Frasconi, Learning long-term dependencies with gradient descent is difficult, IEEE Transactions on Neural Networks 5 (2) (1994) 157–166. doi:10.1109/72.279181.

[3] S. Hochreiter, J. Schmidhuber, Long Short-Term Memory, Neural Computation 9 (8) (1997) 1735–1780. doi:10.1162/neco.1997.9.8.1735.

[4] D. Bahdanau, K. Cho, Y. Bengio, Neural Machine Translation by Jointly Learning to Align and Translate (May 2016). arXiv:1409.0473, doi:10.48550/arXiv.1409.0473.

[5] M. Chen, J. Tworek, H. Jun, Q. Yuan, H. P. d. O. Pinto, J. Kaplan, H. Edwards, Y. Burda, N. Joseph, G. Brockman, A. Ray, R. Puri, G. Krueger, M. Petrov, H. Khlaaf, G. Sastry, P. Mishkin, B. Chan, S. Gray, N. Ryder, M. Pavlov, A. Power, L. Kaiser, M. Bavarian, C. Winter, P. Tillet, F. P. Such, D. Cummings, M. Plappert, F. Chantzis, E. Barnes, A. Herbert-Voss, W. H. Guss, A. Nichol, A. Paino, N. Tezak, J. Tang, I. Babuschkin, S. Balaji, S. Jain, W. Saunders, C. Hesse, A. N. Carr, J. Leike, J. Achiam, V. Misra, E. Morikawa, A. Radford, M. Knight, M. Brundage, M. Murati, K. Mayer, P. Welinder, B. McGrew, D. Amodei, S. McCandlish, I. Sutskever, W. Zaremba, Evaluating Large Language Models Trained on Code (Jul. 2021). arXiv:2107.03374, doi:10.48550/arXiv.2107.03374.

[6] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language Models are Few-Shot Learners (Jul. 2020). arXiv:2005.14165, doi:10.48550/arXiv.2005.14165.

[7] Tesla, Tesla AI Day (Aug. 2021).

[8] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, in: International Conference on Learning Representations, 2022.

[9] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (May 2019). arXiv:1810.04805, doi:10.48550/arXiv.1810.04805.