Predicting Fatalities with Pre-trained Temporal Transformers: A Time Series Regression Approach*

Luca Macis[1], Marco Tagliapietra[1], Elena Siletti[1], and Paola Pisano[1]

[1]Department of Economics and Statistics ”Cognetti de Martiis”, University of Turin

June 28, 2024

*This paper documents a contribution to the VIEWS Prediction Challenge 2023/2024. Financial support for the Prediction Challenge was provided by the German Ministry for Foreign Affairs. For more information on the Prediction Challenge please see Hegre et al. (Forthcoming) and https://viewsforecasting.org/research/prediction-challenge-2023

Introduction

The ability to predict with precision events, such as the number of battle-related fatalities, is not only of academic interest, but it holds significant implications for policy-making and conflict prevention too. This broad interest has enlightened our research, which, specifically, delves into the use of temporal transformers as a new approach to predict the number of battle-related deaths, at the country-level and over a forecast temporal horizon spanning from 3 to 14 months.

Our Artificial Intelligence - Early Warning System (AI-EWS), proposed for the 2023/24 VIEWS prediction competition [Hegre et al., Forthcoming], leverages a multi-headed attention mechanism as outlined by Vaswani et al. [2017]. We chose the temporal transformers due to their proven efficacy in time series representation learning, as demonstrated in Zerveas et al. [2021]. The model incorporates residual connections from input to output, preserving linear activation, a method supported by empirical evidence for its effectiveness in time-series forecasting [Zeng et al., 2023]. The following section details the methodology used to harness this model for time series regression.

0.1 The Temporal Transformers Model

The proposed AI-EWS employs a transformer model to predict fatalities across all countries over twelve weeks. Inspired primarily by the Time-series Dense Encoder (TiDE) model, known for its proficiency in long-term forecasting [Das et al., 2023], our approach replaces the traditional dense encoder with an attention-based encoder, improving our results. Consistent with the original TiDE implementation, our model includes residual connections from input to output, preserving linear activation, a method validated for its efficiency in time-series forecasting [Zeng et al., 2023].

Past data processing involves segmenting into covariates (the independent features) and target values (the dependent feature, i.e. the number of fatalities), and dimensionality reduction of the covariates using a residual block, with linear activation, to keep foundational linearity [Das et al., 2023]. Merging this processed data with positional en-

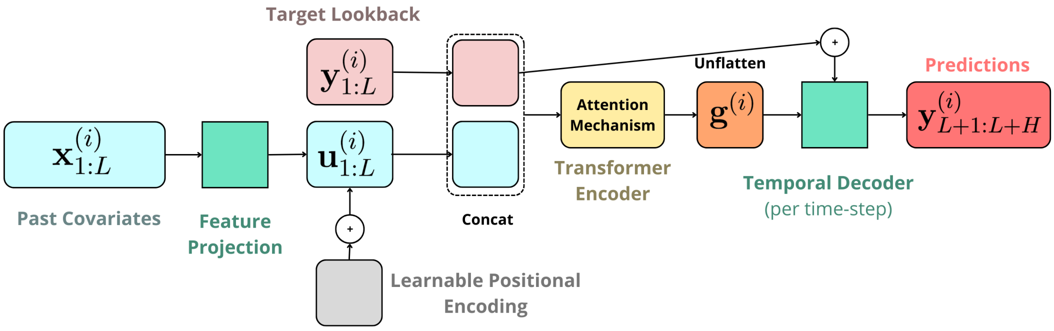

Figure 2: Model Architecture

coding, it's then channeled through the attention mechanism crucial in the transformer framework. The outputs are structured to effectively interface with the temporal decoder, which processes the data incrementally per forecast horizon, using direct residual connections, which integrate information from past values.

Overall, the model's design is geared towards robust long-term prediction, while maintaining a prime simplicity in its architecture and balancing advanced modeling techniques with practical forecasting reliability. The model is trained using the Negative Log Likelihood (NLL) loss function to optimize its probabilistic forecasts, assuming a negative binomial distribution of the outcome and fitting the model effectively to the multivariate nature of the problem.

$$\text{NLL Loss} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{H}\log P(y_{ij}|y_{i-L:i}, x_{i-L:i})), \tag{1}$$

where N is the number of training samples, H is the number of forecasted horizons, y_{ij} is the actual value for the j-th horizon of the i-th sample, $P(y_{ij}|x_i)$ is the predicted probability distribution of the j-th horizon given the input sample $(y_{i-L:i}, x_{i-L:i})$, and L is the lookback length of the sample.

For the purpose of code reproducibility, the parameters used are shown in the table below:

feat_proj	**transf_enc_dim**	**head_size**	**num_heads**	**num_transf_enc**	**mlp_dec_dim**	**dropout**
32	32	32	3	1	1024	0.1

Data Implementation

To address the challenge of forecasting battle-related deaths over a time horizon of 3 to 14 months, we incorporated lag features with an initial gap of 2 months:

target	**lag_3**	**lag_4**	**lag_n**	**lag_13**	**lag_14**
1	4.0	5.0	...	14.0	15.0
2	5.0	6.0	...	15.0	*15.0*
3	6.0	7.0	...	*15.0*	*15.0*
4	7.0	8.0	...	*15.0*	*15.0*
5	8.0	9.0	...	*15.0*	*15.0*
6	9.0	10.0	...	*15.0*	*15.0*
7	10.0	11.0	...	*15.0*	*15.0*
8	11.0	12.0	...	*15.0*	*15.0*
9	12.0	13.0	...	*15.0*	*15.0*
10	13.0	14.0	...	*15.0*	*15.0*
11	14.0	15.0	...	*15.0*	*15.0*
12	15.0	*15.0*	...	*15.0*	*15.0*
13	*15.0*	*15.0*	...	*15.0*	*15.0*
14	*15.0*	*15.0*	...	*15.0*	*15.0*
15	*15.0*	*15.0*	...	*15.0*	*15.0*

In the table, italicized values represent unknown future values not available in the dataset. For these, we used the last known historical value.

To ensure robust model training and evaluation, we partitioned the data to exclude any entries with unknown future values. Our data partitioning strategy for each country was as follows:

1. **Dropped Data (last 14 months)**: All entries with unknown future values were excluded during training and validation.
2. **Validation (6-10 months before testing period)**: The 6-10 months immediately preceding the testing period were used for validation.
3. **Training (remaining data)**: All remaining data before the validation period were allocated to the training set.

Results

The tables below present a comparison between the results of our "Transformer" model and the benchmark models from the competition.

CRPS

Model	*Y2018*	*Y2019*	*Y2020*	*Y2021*	*Y2022*	*Y2023*	mean
Transformer	**12,16763**	10,4524	23.76307	76,88208	**119,0455**	51,11477	**48.90424**
benchmark_conflictology_12m	14.48288	**9.146306**	**21.33933**	**76.84948**	123.9952	**50.35671**	49.02849
benchmark_last_with_poisson	20.17346	9.480041	23.69811	85.60546	131.0171	678.9598	158.8226
benchmark_boot_240	23.57732	22.45758	31.41744	86.62632	120.2492	52.72215	56.50867
benchmark_exactly_zero	24.13045	23.01876	32.04058	87.33901	120.9682	53.54319	56.50636

IGN

Model	*Y2018*	*Y2019*	*Y2020*	*Y2021*	*Y2022*	*Y2023*	mean
Transformer	0,841352	0,877735	0,849775	0,802891	1,344082	1,067288	0,963854
benchmark_conflictology_12m	**0.640281**	**0.610132**	**0.566535**	**0.685623**	**0.694711**	**0.682261**	**0.646257**
benchmark_last_with_poisson	1.198439	1.045585	1.110316	1.227781	1.124429	1.124699	1.138875
benchmark_boot_240	1.123216	1.111029	1.115448	1.152036	1.154555	1.154135	1.135403
benchmark_exactly_zero	1.55813	1.55813	1.549433	1.614664	1.632058	1.614664	1.587513

MIS

Model	*Y2018*	*Y2019*	*Y2020*	*Y2021*	*Y2022*	*Y2023*	mean
Transformer	**121,3819**	102,1785	381,7671	1436,088	2284,74	**841,821**	**861,3294**
benchmark_conflictology_12m	186.554	**89.0579**	**344.964**	**1435.55**	**2142.13**	1042.92	873.529
benchmark_last_with_poisson	380.623	172.686	455.806	1690.71	2599.28	13523.5	3120.434
benchmark_boot_240	454.09	426.006	606.003	1708.3	2380.74	1030.99	1101.355
benchmark_exactly_zero	482.609	460.375	640.812	1746.78	2419.36	1070.86	1363.466

Conclusions

In this study, we presented our predictive model for forecasting battle-related fatalities over a forecast horizon ranging from 3 to 14 months.

The results demonstrate that the Transformer model outperforms on average the benchmark models in CRPS and MIS, but not in IGN, where the Conflictology benchmark model surpasses our Transformer model. Overall, the Transformer model shows superior probabilistic forecasting and prediction accuracy compared to the benchmarks. Importantly, it should be noted that no optimization of the Transformer model was performed in this evaluation. Future optimization efforts are planned to further enhance the model's performance, potentially leading to even better forecasting results and more accurate predictions.

References

A. Das, W. Kong, A. Leach, S. Mathur, R. Sen, and R. Yu. Long-term forecasting with tide: Time-series dense encoder, 2023.

H. Hegre et al. The 2023/24 views prediction competition. *Journal of Peace Research*, Forthcoming. XXX.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

A. Zeng, M. Chen, L. Zhang, and Q. Xu. Are transformers effective for time series forecasting? *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(9): 11121–11128, Jun. 2023. doi: 10.1609/aaai.v37i9.26317. URL `https://ojs.aaai.org/index.php/AAAI/article/view/26317`.

G. Zerveas, S. Jayaraman, D. Patel, A. Bhamidipaty, and C. Eickhoff. A transformer-based framework for multivariate time series representation learning. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD '21, page 2114–2124, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383325. doi: 10.1145/3447548.3467401.