Anchor Attention for Hybrid Crowd Forecasts Aggregation

Extended Abstract

Yuzhong Huang
USC Information Sciences Institute
Marina Del Rey, California
yuzhongh@isi.edu

Andrés Abeliuk
USC Information Sciences Institute
Marina Del Rey, California
aabeliuk@isi.edu

Fred Morstatter
USC Information Sciences Institute
Marina Del Rey, California
fredmors@isi.edu

Pavel Atanasov
Pytho, LLC.
Brooklyn, New York
pavel@pytho.io

Aram Galstyan
USC Information Sciences Institute
Marina Del Rey, California
galstyan@isi.edu

ABSTRACT

Forecasting the future is a notoriously difficult task. One way to address this challenge is to "hybridize" the forecasting process, combining forecasts from a crowd of humans, as well as one or more machine models. However, an open challenge remains in how to optimally aggregate inputs from these pools into a single forecast. We proposed anchor attention for this type of sequence summary problem. Each forecast is represented by a trainable embedding vector. An anchor attention score is used to determine input weights. We evaluate our approach using data from a real-world forecasting tournament, and show that our method outperforms the current state-of-the-art aggregation approaches.

CCS CONCEPTS

• **Human-centered computing → Computer supported cooperative work**;

KEYWORDS

Aggregation; Crowd Sourcing; Embedding; Attention Model

ACM Reference Format:
Yuzhong Huang, Andrés Abeliuk, Fred Morstatter, Pavel Atanasov, and Aram Galstyan. 2020. Anchor Attention for Hybrid Crowd Forecasts Aggregation. In *Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), Auckland, New Zealand, May 9–13, 2020,* IFAAMAS, 3 pages.

1 INTRODUCTION

The "wisdom of crowds" effect has been demonstrated as a successful approach in diverse domains [7], extending to complex problem-solving tasks such as reconstructing gene regulatory networks [4] or geopolitical forecasting [8]. Previous work [1, 6, 9] has identified linear combinations among forecaster estimates. However, such an approach is not optimal, since it assigns a single weight to each forecaster towards a variety of forecasting problems.

With our proposed anchor attention method, the combined weight is conditioned on a question, forecaster, and time, and is more flexible than a single weight per forecaster at any point in time. The key insight to this method is the anchor attention models learn a representation that could recall the best human and machine forecasters for a given question.

Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), B. An, N. Yorke-Smith, A. El Fallah Seghrouchni, G. Sukthankar (eds.), May 9–13, 2020, Auckland, New Zealand.

Table 1: Summary statistics of the questions and forecasts in our dataset. STD is the sample standard deviation.

Statistic	Min	Median	Mean (STD)	Max
Forecasts per Question	9	192	265.2 (204.6)	1029
Forecasts per User	1	32	44.4 (63.2)	1339
Users per Question	8	108	173.6 (148.9)	669
Days Question is Open	1	42	53.3 (42.2)	184

2 CROWD FORECASTING PLATFORM

In this work, we study forecasts about geopolitical events. These forecasts are created on a forecasting platform we developed, Synergistic Anticipation of Geopolitical Events (SAGE)[5][1]. Our hybridized forecasting platform contains both human forecasts and forecasts generated by machine models. Human forecasts come from recruited participants from Amazon Mechanical Turk. Machine models include: 1. AutoRegressive Integrated Moving Average (ARIMA) [2]; 2. M4-Meta [3]; and 3. Arithmetic Random Walk (RW). There are 375 questions in this dataset and 2240 human participants. We will release an anonymized version of this dataset[2].

3 METHODS

3.1 Baseline Methods

We compare our approach to [1], who proposed an aggregation method using temporal decay, differential weighting based on past performance, and extremization. The approach outlined by [1] provides a simple yet strong performance baseline.

We consider three variants based on [1]:

- **M0:** Unweighted average with temporal decay.
- **M1:** Weighted average with temporal decay and extremization.
- **M2:** Weighted average with temporal decay, differential weighting based on past performance and extremization.

[1] https://sage-platform.isi.edu/
[2] https://github.com/YuzhongHuangCS/AnchorAttention

3.2 Anchor Attention

3.2.1 Issue with Self Attention. Following the notation in [10], we denote the input vector as X, and three trainable weight matrices as W^Q, W^K, W^V. The common practice in self-attention models is to use the last hidden state H_i as a representation of information in X until time step i.

$$Q = X \cdot W^Q, K = X \cdot W^K, V = X \cdot W^V. \tag{1}$$

$$\alpha_{ij} = align(Q_i, K_j) = \frac{exp(Q_i \cdot K_j^T / \sqrt{d_k})}{\sum_{j'} exp(Q_i \cdot K_{j'}^T / \sqrt{d_k})}. \tag{2}$$

$$H_i = (V_1, V_2, \cdots, V_i) \cdot (\alpha_{i1}, \alpha_{i2}, \cdots, \alpha_{ii})^T, \tag{3}$$

We can see H_i is most influenced by X_i (the last input), more specifically its projection Q_i, because Q_i will be used as the query vector to compute the alignment score with input from previous time steps. For language tasks, this is desired, as the last token usually contains important information (e.g., "?" implies a sentence is a question). In the task of forecast aggregation, this is *not* desired. Any forecast can be the last forecast at the moment of making an aggregated forecast. The output of the aggregation system should not be oversensitive to the last forecast.

3.2.2 Anchor Attention. Anchor attention use an anchor vector A, which is independent of input sequence X, to replace the query vector Q. Here A is the sentence embedding of question text that captures the semantics of question. Now the alignment score is calculated as:

$$\alpha_{ij} = align(A, K_j) = \frac{exp(A \cdot K_j^T / \sqrt{d_k})}{\sum_{j'} exp(A \cdot K_{j'}^T / \sqrt{d_k})}. \tag{4}$$

With this modification, we didn't give forecasts at a particular position any advantages.

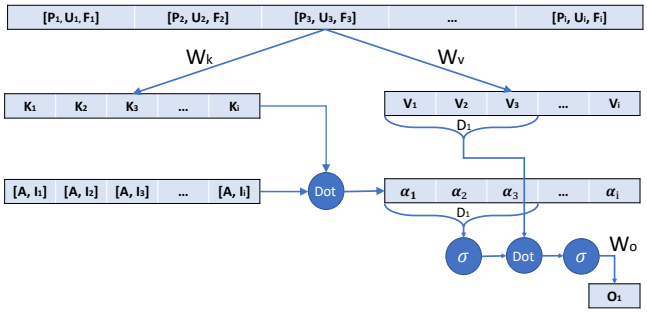

Figure 1: Overview of proposed model.

4 EXPERIMENTS

4.1 Average Brier score comparison

We perform 5-fold cross-validation. Cross-validation is done at the question level, meaning that all of the forecasts for a particular question are constrained within their respective fold. We record the mean and variance of the Brier score across folds, and the quantile range of all Brier scores in Table 2. Note that the range of Brier score is $[0, 2]$. Our proposed model has the lowest average Brier score and quantile scores, along with the lowest variance, suggesting our model is very stable.

Table 2: Results of 5-fold cross-validation across different methods. We report the mean Brier score as well as the quantile range (e.g., for our proposed Attention-based model, the forecast at the 25th percentile has a Brier score of 0.037).

Method	Mean (STD)	25%	50%	75%
M0	0.321 (0.274)	0.126	0.249	0.432
M1	0.319 (0.316)	0.088	0.212	0.422
M2	0.304 (0.321)	0.073	0.192	0.419
Attention	**0.251** (0.265)	0.037	0.173	0.382

4.2 Analysis of Attention Scores

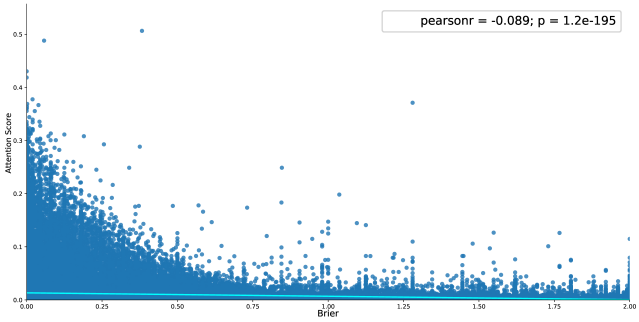

Figure 2: Relationship between attention score and Brier score. The trend line shows a downward trend, indicating that higher attention scores are given to forecasts with lower Brier scores.

Figure 2 shows forecasts' attention scores against their Brier scores. We observe that attention score is negatively correlated (coef=-0.089, $p < 10^{-10}$) to Brier score, which means good forecasts (low Brier) have higher weights in aggregation. Also, we observe there are no points in the upper right corner, which means forecasts with a high Brier score do not receive a high attention score. This suggests that the model is learning representations that are useful to distinguish good forecasts from bad ones.

Conclusion: Anchor Attention has shown to be able to identify high quality forecasts. A more detailed version is available on arXiv.

ACKNOWLEDGMENTS

This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via 2017-17071900005. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

REFERENCES

[1] Pavel Atanasov, Phillip Rescober, Eric Stone, Samuel A Swift, Emile Servan-Schreiber, Philip Tetlock, Lyle Ungar, and Barbara Mellers. 2017. Distilling the wisdom of crowds: Prediction markets vs. prediction polls. *Management science* 63, 3 (2017), 691–706.

[2] Rob Hyndman and Yeasmin Khandakar. 2008. Automatic Time Series Forecasting: The forecast Package for R. *Journal of Statistical Software, Articles* 27, 3 (2008), 1–22. https://doi.org/10.18637/jss.v027.i03

[3] Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. 2019. The M4 Competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting* (2019). https://doi.org/10.1016/j.ijforecast.2019.04.014

[4] Daniel Marbach, James C Costello, Robert Küffner, Nicole M Vega, Robert J Prill, Diogo M Camacho, Kyle R Allison, Andrej Aderhold, Richard Bonneau, Yukun Chen, et al. 2012. Wisdom of crowds for robust gene network inference. *Nature methods* 9, 8 (2012), 796.

[5] Fred Morstatter, Aram Galstyan, Gleb Satyukov, Daniel Benjamin, Andres Abeliuk, Mehrnoosh Mirtaheri, KSM Tozammel Hossain, Pedro Szekely, Emilio Ferrara, Akira Matsui, Mark Steyvers, Stephen Bennet, David Budescu, Mark Himmelstein, Michael Ward, Andreas Beger, Michele Catasta, Rok Sosic, Jure Leskovec, Pavel Atanasov, Regina Joseph, Rajiv Sethi, and Ali Abbas. 2019. SAGE: A Hybrid Geopolitical Event Forecasting System. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*. International Joint Conferences on Artificial Intelligence Organization, 6557–6559. https://doi.org/10.24963/ijcai.2019/955

[6] Ville A Satopää, Jonathan Baron, Dean P Foster, Barbara A Mellers, Philip E Tetlock, and Lyle H Ungar. 2014. Combining multiple probability predictions using a simple logit model. *International Journal of Forecasting* 30, 2 (2014), 344–356.

[7] James Surowiecki. 2004. The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business. *Economies, Societies and Nations* 296 (2004).

[8] Philip E Tetlock. 2017. *Expert political judgment: How good is it? How can we know?* Princeton University Press.

[9] Lyle Ungar, Barbara Mellers, Ville Satopää, Philip Tetlock, and Jon Baron. 2012. The good judgment project: A large scale test of different methods of combining expert predictions. In *2012 AAAI Fall Symposium Series*.

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Lawrence Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *NIPS*.