Exploring Multi-Modal Transformers for Monetary Policy Calls

Maximilian Böhm

September 26, 2023

Abstract

The rapid advancement of transformer architectures has reshaped the landscape of artificial intelligence and machine learning, with applications spanning various domains. This paper delves into the exploration of a multimodal transformer model designed to process information from three distinct modalities: text, audio, and video. The primary objective is to understand how the integration of these modalities influences the model's predictive capabilities. The real-world application of this model is in the financial sector, where it leverages a dataset consisting of Monetary Policy Calls (MPCs) from six major economies. These calls play a crucial role in influencing market volatility and the trajectory of financial instruments. The model's predictive capability is tested against forecasting the volatility and price movements of six financial instruments post these policy announcements. This study builds upon previous work, aiming not only to replicate results but also to experiment with different embeddings and assess the influence of each modality. The findings suggest that the fusion of modalities and the use of finetuned embeddings can enhance the model's performance, with potential implications for both technical and financial sectors.

1 Introduction

The emergence of transformer architectures [1] has significantly revolutionized the realm of artificial intelligence and machine learning, with applications spanning from natural language processing to computer vision. Their adaptability and capability to process sequential data have made them particularly alluring for various multimodal tasks. As data generation sees an ever-increasing diversity across modalities, including text, audio and video, the importance of constructing models that can merge and process this information increases [2].

This project endeavors to explore these transformer architectures by replicating and experimenting with a multimodal transformer model, an architecture that synthesizes information from three distinct modalities: text, audio and video. The central focus lies in exploring how different modalities, coupled with varying architectural choices behave and influence the model's predictive capabilities.

Through a pipeline of cross-modal-attention, self-attention mechanisms and fusion layers, the model seeks to harness the power of each modality, ultimately aming for heightened accuracy and performance.

While the primary drive is technical exploration, the real-world application of this innovation is rooted in a vital sector: financial markets. Specifically, this project leverages the Monopoly-dataset[3], a comprehensive compilation of Monetary Policy Calls (MPCs) conducted by the central banks of six major economies over more than a decade. These calls, pivotal in the realm of finance, influence market volatility and the trajectroy of numerous financial instruments. Therefore, the model's predictive capability extends to forecasting the volatility and price movements of six financial instruments post these monetary policy announcements.

Building upon foundations established by Mathur et al.[3], this project not only seeks to reproduce their results but also delves into experimentation, fine-tuning embeddings, and assessing modality-specific influences.

This project finds that while fusing multiple data modalities can enhance transformer model performance in finance, the choice and quality of modalities are pivotal. Despite their potential, transformers remain a "black box," with challenges in interpretability.

The following chapters will unravel the methodology, experiments and consequent findings with a concluding segment that encapsulates the broader implications and potential future trajectories of multimodal transformers in both technical and financial domains.

2 Literature

In the realm of deep learning, especially the domain of natural language processing (NLP), the Transformer architecture has emerged as a groundbreaking model, influencing a multitude of research areas and applications.

Historically, sequence modeling in NLP relied heavily on recurrent architectures, such as Recurrent Neural Networks (RNNs)[4] and their variants, like Long Short-Term Memory (LSTM)[5] networks and Gated Recurrent Units (GRUs)[6]. These models process sequences element by element, making them inherently sequential and thereby limiting their ability to be parallelized during training.

The Transformer model, introduced by Vaswani et al. in the seminal paper "Attention is All You Need"[1], deviated from this sequential approach. Instead of relying on recurrence, the Transformer model utilizes a mechanism called "attention" to draw global dependencies between input and output. This allows the model to consider other words in the input sentence, irrespective of their distance from the current word, facilitating a deeper understanding of context.

The core innovation behind the Transformer is the "self-attention" mechanism, which weighs input tokens differently, allowing the model to focus on (or "attend to") specific tokens when producing an output. Mathematically, this attention mechanism computes a weighted sum of input tokens based on their relevance to the current token. The model uses queries, keys, and values to determine these relevancies and weights in the attention computation[1].

Following the introduction of the Transformer architecture, numerous variants and adaptations have been proposed. Models such as BERT[7], Wav2vec2[8] or BEiT[9], which are used in this project to generate embeddings, are built upon the Transformer backbone. These models, while differing in training methodology and application, underscore the versatility and capability of the Transformer architecture.

Transitioning from the foundational understanding of transformers in NLP, it is evident that their application is not confined to mere language processing tasks but extends to various interdisciplinary areas, including the financial sector.

In the vast landscape of financial markets, risk prediction and price movement classification remain pivotal. One important source of information aiding these predictions are the insights gathered from Monetary Policy Calls (MPC). These periodic video conferences are forums where central banks' governors, such as the Federal Reserve Bank in the United States, discuss their monetary policies, actions taken for the financial welfare of the country, and risk assessments related to to the country's economic growth. The content of these calls, encompassing both a prepared press speech and a subsequent question-answer session with journalists, offers valuable announcements regarding policy decisions and illuminates the anticipated trajectory of the economy[3].

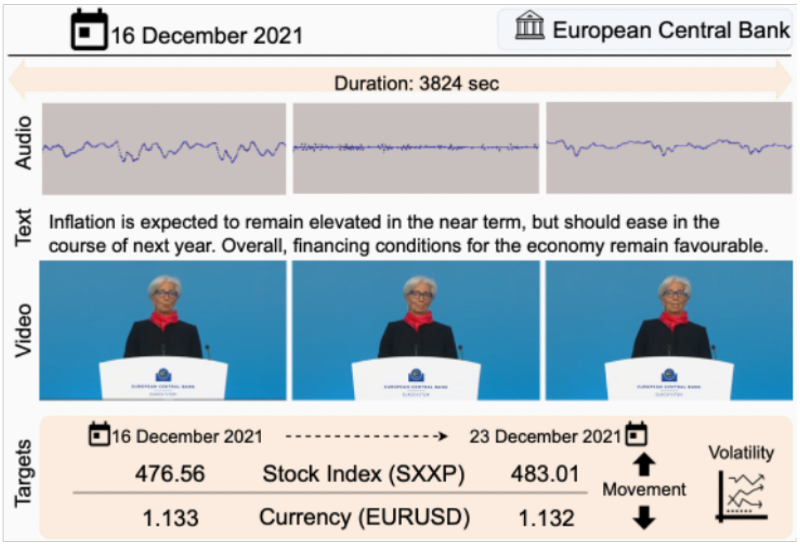
Historically, MPCs have been observed to significantly influence the financial stock markets. An illustrative example is the observed increased volatility of the S&P 500 index on the days the Federal Reserve Bank holds its MPCs, wherein the volatility is roughly three times greater than on non-MPC-days[3].

To combine the information that lies in these MPCs with the predictive power of transformer architectures Mathur et al. provide a dataset and a multimodal transformer model to analyze the MPCs and make financial predictions.

While existing research has leveraged text and audio modalities from these calls for predictions [10, 11], there remains a largely untapped potential in the realm of visual cues. Herein lies the novelty and richness of data: not just focusing on what is being said, but how it is conveyed. Non-verbal cues, such as the complexity of language used, the tone of voice, facial expressions, and other behavioral indicators, can significantly influence trading activities in the financial markets[3]. Such visual aspects provide insights into the human behaviours and emotions displayed, which might sometimes be in contrast with the verbal content presented. For instance, while the textual content of an MPC might show optimism, the speakers' tone, facial expressions, and hesitance might suggest otherwise, leading to market reactions like declining currency value and increasing stock price volatility[3]. This project largely relies on the work of Mathur et al.. They provide not only the dataset but also a model called MPCNet, a model that synergizes the power of cross-modal transformer blocks and modality-specific attention fusion. By harnessing the concurrent information from visual, vocal and verbal modalities, this architecture has shown promising results in forecasting financial risks and price movements associated with MPCs, outperforming other competitive deep learning methodologies[3].

Figure 1 illustrates the general idea behind utilizing the Monopoly-dataset in conjunction with the MPCNet model to predict financial metrics. Complementing this, figure 2 delves deeper into the architecture of the MPCNet. In their experiments Mathur et al. use a baseline architecture consisting of nine baseline models (e.g. a historic Price model, a basic MLP, LSTM, ...). MPCNets architecture will now be explained in more detail. Given the three modalities: video, audio, and text, the Monopoly-dataset encapsulates each modality in sequences that are synchronized and correspond with one another[3]. The dataset is segmented on an utterance level. Feature representations for each utterance are computed using BERT for texts, wav2vec2 for audios, and BEiT for videos. These representations are then fed into a locally aware position encoding layer, which amalgamates a 1D temporal convolutional layer, designed to capture the local sequence structure and sinusiodal positional encoding.

To discern correlations and inter-dependencies between modalities, the model employs cross-modal transformers. For two distinct modalities α and β where $\alpha, \beta \in$ {Video, Audio, Text} and $\alpha \neq \beta$, the cross-modal attention layer integrates cross-modal in-



Inflation is expected to remain elevated in the near term, but should ease in the course of next year. Overall, financing conditions for the economy remain favourable.

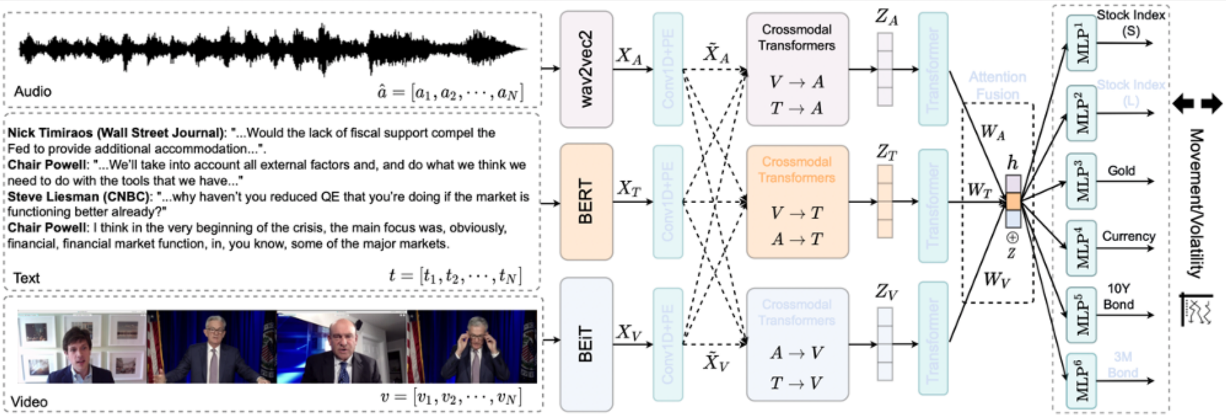
Figure 1: This figure presents the concept of the MPCNet model and the Monopoly dataset. It emphasizes the integration of three modalities (video, audio, and text), to forecast financial metrics, specifically Price Movement and Price Volatility. Displayed as a representative example is a segment from an ECB MPC[3].

formation through latent adaption. Subsequently, the latent adaption representation $Z_{\alpha\rightarrow\beta}$, sharing the same target modality β, is concatenated and channeled through self-attention transformers to aggregate temporal information. These temporal encodings are then concatenated and processed through a feed-forward layer, resulting in the ensemble temporal representation Z. To emphasize the significance of a specific target modality representation Z_α in relation to its sibling representation Z_β, where $\alpha \neq \beta$, the model employs a specific attention fusion mechanism[3]. The ensemble temporal representation Z is then amalgamated with the fused temporal representation Z_{fused} using a feed-forward layer equipped with a residual block, culminating in the final representation. This ultimate hidden representation undergoes processing through six Multi Layer Perceptrons (MLPs) to produce the final prediction. A linear activation function is employed for volatility prediction, while a sigmoid activation is utilized for price movement prediction. Additionally, Mean Squared Error (MSE) and Binary Cross-Entropy (BCE) serve as the loss functions for these tasks, respectively[3].

The labels are calculated from the price data in the following manner. For a given target variable $u \in$ {Stock Index (Small), Stock Index (Large), Gold Price, Currency Exchange Rate, Long-term bond yield (10-years), Short-term bond yield (3-months)} with price p_i on day i, the volatility is the natural log of the standard deviation of return prices r in a window of r days, given as,

$$\upsilon^u_{d,d+\tau} = \ln\left(\frac{\sqrt{\sum_{i=d}^{d+\tau}(r_i-\bar{r})^2}}{\tau}\right), \upsilon \in \mathbb{R} \tag{1}$$

where $r_i = \frac{p_i - p_{i-1}}{p_{i-1}}$ is the return price on day i of the target m, and $\bar{r}$ is the average of these returns over a period of τ days [3].
The price movement labels are calculated also over a period of τ days. For a given target, whose price p can either rise or fall on a day $d + \tau$ compared to a previous



Nick Timiraos (Wall Street Journal): "...Would the lack of fiscal support compel the Fed to provide additional accommodation...".
Chair Powell: "...We'll take into account all external factors and, and do what we think we need to do with the tools that we have..."
Steve Liesman (CNBC): "...why haven't you reduced QE that you're doing if the market is functioning better already?"
Chair Powell: I think in the very beginning of the crisis, the main focus was, obviously, financial, financial market function, in, you know, some of the major markets.

Figure 2: This illustration depicts the architecture of the MPCNet, designed to accept three modalities (video, audio, and text) as input sources[3].

day d, the classification task is the following,

$$y^u_{[d,d+\tau]} = \begin{cases} 1 & p_{d+\tau} \geq p_d \\ 0 & p_d \geq p_{d+\tau} \end{cases} \tag{2}$$

Another relevant transformer model that focuses on the financial field is the FinBERT[12] model. Financial sentiment analysis is pivotal for understanding market dynamics. Addressing the unique nuances of financial language, FinBERT was developed as an adaption of the BERT model, specifically fine-tunded for financial texts. Unlike BERT, which was trained on general English, FinBERT captures the specialized meanings of terms in the financial context, such as "bullish" or "bearish". By leveraging BERT's pre-trained representations and refining them on financial datasets, FinBERT offers enhanced sentiment prediciton accuracy for financial texts[12].

3 Data

Mathur et al.[3] unveil the Monopoly-dataset, which aggregates videos, audios, and transcripts from monetary policy meetings of six major economies: United States, United Kingdom, European Union, Canada, New Zealand, and South Africa. The selection was based on English-language content titled "Monetary Policy".
Data was sourced by Mathur et al. from official central bank sites using BeautifulSoup Python package. Videos and transcripts were extracted using Urllib5 and PDFPlumber6, respectively. The dataset also comes with a synchronisation file that can be used to match the utterances of the transcripts with the subclips of the video files. Daily price time series from January 2000 to March 2022 were compiled via the Bloomberg Terminal, with a primary focus on MPCs between January 2009 and March 2022, since conference calls started being reliably released post 2009. Of the 464 acquired MPCs, 340 were retained, by Mathur et al. after addressing alignment issues. These span 15.729 minutes, with an average call lasting 53 minutes. The mean number of audio utterances across the calls is 587.54 ± 38.32, with a maximum of 2462 utterances. After checking the completness of the data again for the scope of this project (e.g. text, audio and video data available for every MPC), 319 valid MPCs remained from the 340. The dataset is split into training, validation,

and test sections at a 70:10:20 ratio to prevent future data from influencing past predictions.
This study additionally uses the MELD[13] (Multimodal EmotionLines Dataset) to finetune embeddings. The MELD dataset is designed for emotion recognition in conversations, derived from dialogues of the TV show ”Friends”. Enhanced with audio and visual modalities, MELD provides a rich source for multimodal emotion analysis. It uniquely offers conversational context, capturing a range of emotions in a real-world setting, making it usefull for multimodal machine learning tasks.

4 Methodology

The primary objective of this project lies in replicating a multimodal transformer model and experimenting with it. This chapter explains the conducted experiments and the modifications to the original model.
The project's inception was marked by the replication of the MPCNet model from the study of Mathur et al.[3]. The overarching ambition was to emulate the results presented in the paper, harnessing the Monopoly dataset for this purpose.
To replicate the MPCNet model, this study follows the methodology of the paper from Mathur et al. as previously portrayed in the Literature chapter.

As the paper of Mathur et al. lacks implementation details for parts of the model, the following implementation choices are made for this study: The final MLPs have each two linear layers where the first linear layer transformes the input embedding size to a hidden dimension and the second linear layer reduces the hidden dimension to a single output value. The paper does not contain details about the padding and flattening of the final fused representation before it is passed to the final MLPs, thus in this study a zero-padded tensor is used and the final representation is masked to selectively keep or zero out elements. The masked data is then flattened, by reshaping the tensor form 2D into 1D, and placed in the zero-padded tensor. Also, after problems with the multihead-attention part of the presented implementation of the cross-transformer layer, the torch implementation of the multihead-attention is used.
Upon the successful replication of the model, the focus shifts to a series of experiments aimed at exploring the behaviour and adaptability of the transformer architecture. Therefore the model's architecture is modified in several ways to facilitate experiments with diverse combinations of modalities. These encompass pairings of audio and video, audio and text, and video and text, as well as standalone modalities of text, video and audio.
The Monopoly-dataset comes equipped with a trio of embeddings: BERT (text), Wav2vec2 (audio), BEiT (video), which have a size of 768 for each utterance. While these embeddings are useful for a broad spectrum of use cases, embeddings tailored for financial contexts might improve the model's performance. To address this, FinBert is used to generate the text embeddings. For the audio and video modalities no existing finetuned models were used. Instead the base models (e.g. Wav2vec2 and BEiT) were finetuned on a emotion classification task using the MELD dataset. Here the Wav2vec2 and BEiT models were primed to predict the emotion labels of MELD subclips. After the training phase of the emotion classification task, the models states were archived. Subsequently, embeddings for the Monopoly-dataset were generated by channeling the subclips for each utterance through the archived models hidden states. The extraction of these subclips was conducted based on their

temporal markers in the video file, employing torch.audio for audio segments and open-cv for video frames.

After generating the finetuned embeddings, the single modality versions of the replicated MPCNet model were trained with their respective finetuned embedded modality. In the final lag of experiments, modalities that demonstrated an increased performance were used in the tri-modality replicated MPCNet model.

In undertaking these experiments, this study endeavors to provide insight into the capabilities of a multimodal transformer model within the realm of financial prediction. It seeks to unravel the intrinsic value and influence of each individual modality, while simultaneous exploring the potential that emerges when these modalities are merged within a singular model. Furthermore, the experiments are designed to shed light on the pivotal role of embeddings, highlighting not only their impact on the model but also the effects of fine-tuning these embedding generation models. Collectively, this research underscores the implications of architectural design decisions in shaping the efficacy and adaptability of multimodal transformers.

5 Results

In this chapter, the empirical findings derived from the experiments explained in the methodology section are presented. While this study closely follows the original experiments of Mathur et al., certain deviations in hyperparameters were necessary due to hardware differences and were the result of hyperparameter tuning. Specifically, this study employs a hidden dimension $H = 256$, dropout $\delta = 0.1$, number of attention heads $n_h = 2$, and a learning rate of $1e^{-5}$. The replicated model, developed using PyTorch, is optimized with the AdamW algorithm over 25 epochs, with a patience setting of 10, on a Tesla A100 GPU. For evaluating the predicted volatility, the mean squared error (MSE) metric is utilized.

The experiments revealed that the replicated model could not yield meaningful results for the price movement prediction. Consequently, this chapter will focus solely on the results pertaining to volatility prediction, where the model aims to forecast the volatility two days post the MPC. As dipicted in Table 1, a comparison suggests that the replicated model's performance in predicting volatility aligns closely with that of the original MPCNet model.

While the experiments aimed to provide a comprehensive analysis across all modalities, it is important to note that the generation of video embeddings proved to be particularly time-intensive. Due to these computational constraints, the results for this specific experiment could not be completed within the designated timeframe and are, therefore, not included in the presented findings.

The results presented in Table 1 detail the experimental outcomes for predicitons made two days post-call. Given this context, these findings are compared with the results from Mathur et al., which consider predictions one and three days after the MPC. The results show that the replicated model yields outcomes close to the range of the ones from the original. The other experiments were designed to assess the performance of the transformer architecture under various modifications, such as the utilization of single or dual modalities and the incorporation of fine-tuned embeddings. For these specific experiments, the focus is on comparing the results of the replicated model versions amongst themselves, rather than against the original MPCNet.

The conducted experiments reveal a consistent trend: the fusion of modalities tends

MSE for every Model Version - Asset Class Combination						
Model	Stock Index (Small)	Stock Index (Large)	Currency Exchange Rate	Gold Price	10-Year Bond Yield	3-Month Bond Yield
Video	2.881	2.856	3.047	3.290	2.952	2.749
Audio	2.779	2.814	2.742	2.913	2.850	2.769
Text	2.806	2.938	2.950	3.129	2.929	2.592
Text_Audio	2.275	2.301	2.640	2.487	2.434	2.331
Audio_Video	2.379	2.809	2.463	2.849	2.592	2.660
Video_Text	1.902	3.037	2.150	2.944	2.961	1.898
Video_Audio_Text	2.074	2.651	2.290	2.255	2.138	1.812
Video-finetuned	-	-	-	-	-	-
Audio-finetuned	2.450	2.794	2.628	2.891	2.635	2.518
Text-finetuned	2.267	2.238	2.263	2.227	2.301	1.861
V._A._T.-finetuned	2.306	2.149	1.857	1.935	1.978	1.791

Table 1: Performance comparison of the replicated model for volatility prediction in terms of MSE two days after the call. The Results are averaged over 5 independent runs. V._A._T. is short for Video_Audio_Text.

to enhance model performance across nearly all Model-Asset-Class combinations. Integrating all three modalities results in performance improvements in half of the tested scenarios. In instances where it doesn't achieve the top performance, the tri-modal fusion still holds its ground, often ranking second in terms of efficacy. A notable observation is the significant performance boost when utilizing FinBERT for text embeddings. This contrasts with the results from the finetuned wav2vec2 embeddings, which appear to diminish the model's effectiveness, when it is used as the only modality. But when the finetuned wav2vec2 embeddings are used with the finetuned FinBert embeddings in a tri-modal model, the performance is imporved. Several factors could account for this discrepancy. The architecture employed for finetuning might either lack the capacity to extract more valuable insights or might be overfitting to its training data. Another possibility is that the data used for finetuning diverges considerably from the characteristics of the Monopoly-dataset. When using the finetuned embeddings in the tri-modal model, the additional information provided by all the embeddings seems to improve the models performance and the problems with the finetuned wac2vec2 embeddings seem to be balenced out by the other modalities.

To account for a qualitative analysis of the replicated model used for this study Figure 3, Figure 4 and Figure 6 can be observed. These show excerpts of the attention weights during the model training. Figure 3 shows the patterns of the attention weights in the self-attention mechanism of the model for a batch with four observations. Patterns are noticeable but hard to understand. Figure 4 takes a closer look into this mechanism for one observation and only the first 60 utterances. When we look at the attention weights for the video modality and examine the original video at utterance 18 and 19 the governors body language is hardly changing, but in this special case the video contains a woman translating the governors speech into sign language and is showing more emotions that relate to the content of the speech. Figure 5 shows this part of the video. When we look at the audio attention

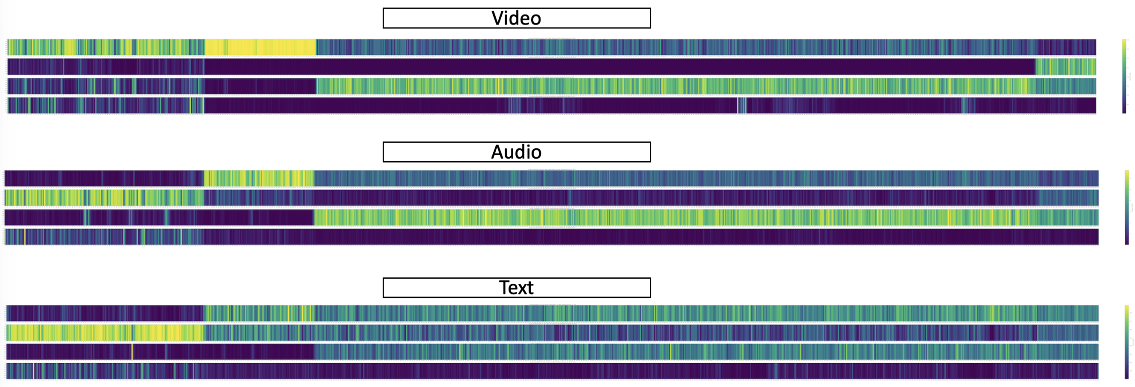


Figure 3: This figure shows heatmaps illustrating the attention weights of the base model across three modalities: video, audio, and text. Each heatmap represents attention patterns for an entire video, audio or text file respectively, segmented into its subclips, with each subclip corresponding to an individual utterance. The displayed data encompasses one batch with a size of four. A more pronounced yellow signifies higher attention weights.

weights and compare these with the audio at utterance 17, we observe a long pause of the governor after talking about the risk of continuing decreasing currency exchange rates, which leads to higher inflation. For the text modality the attention weights for utterance 23 are very high. In this utterance the governor also talks about inflation. It is not directly observable why the attention on all modalities attends to different utterances, which on the one hand can show the importance of fusing modalities to get a fused representation that includes all the valuable information from the fused modalities together and on the other hand indicates that the attention process is complex. Figure 6 supports this, because there the attention weights of the cross attention are shown and almost no pattern is observable. The only observable thing is, that the attention seems to understand where the actual MPC ends and that the rest is only padded (in the Figure after utterance 291 and 266).

6 Conclusion

In the rapidly evolving landscape of machine learning and artificial intelligence, the fusion of modalities has emerged as a promising approach to enhance the performance of models. This study underscores the potential of combining different data types or modalities to achieve improved results. Notably, transformer models, renowned for their versatility, have demonstrated their abilities across many fields, including finance.

However, there are challenges. An important decision lies in the selection of modalities to fuse. On the one hand adding more modalities might result in more valuable information by combining the information of multiple modalities that complement each other. But adding modalities might not always improve, but even decrease the performance of a multimodal transformer model. This underscores the importance of choosing the right modalities. Thus the data quality of a modality is important. As we have seen with the price movement prediction, multi modal transformer models can be a complex task and be hard to replicate.

We also see that embeddings which are not inherently part of the model, can have a significant influence on models performance. To generate embeddings, either exist-

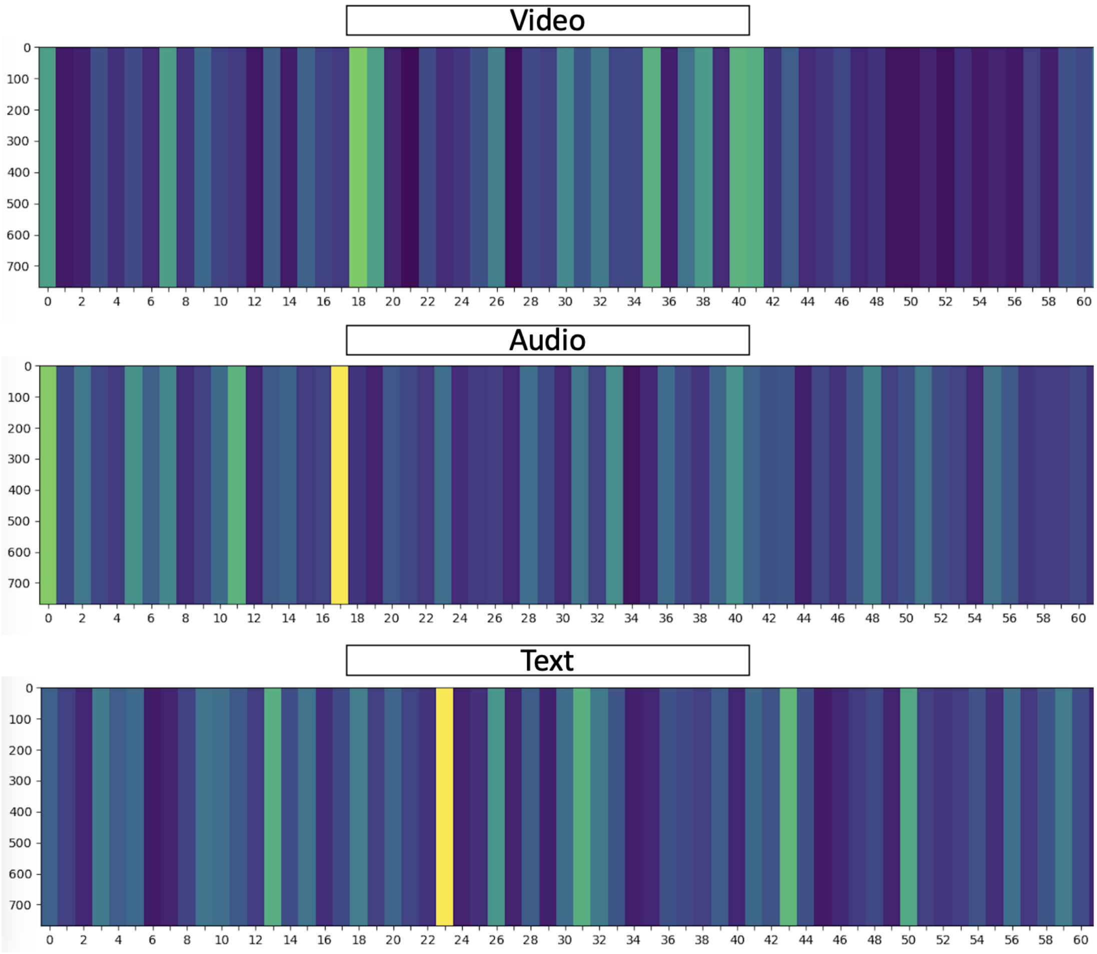


Figure 4: This figure presents a section of Figure 3, focusing on the self-attention weights of the first 60 utterances across all three modalities for a single MPC call from the Bank of South Africa. The x-axis enumerates the utterance numbers, while the y-axis represents the embedding index. A more pronounced yellow signifies higher attention weights.

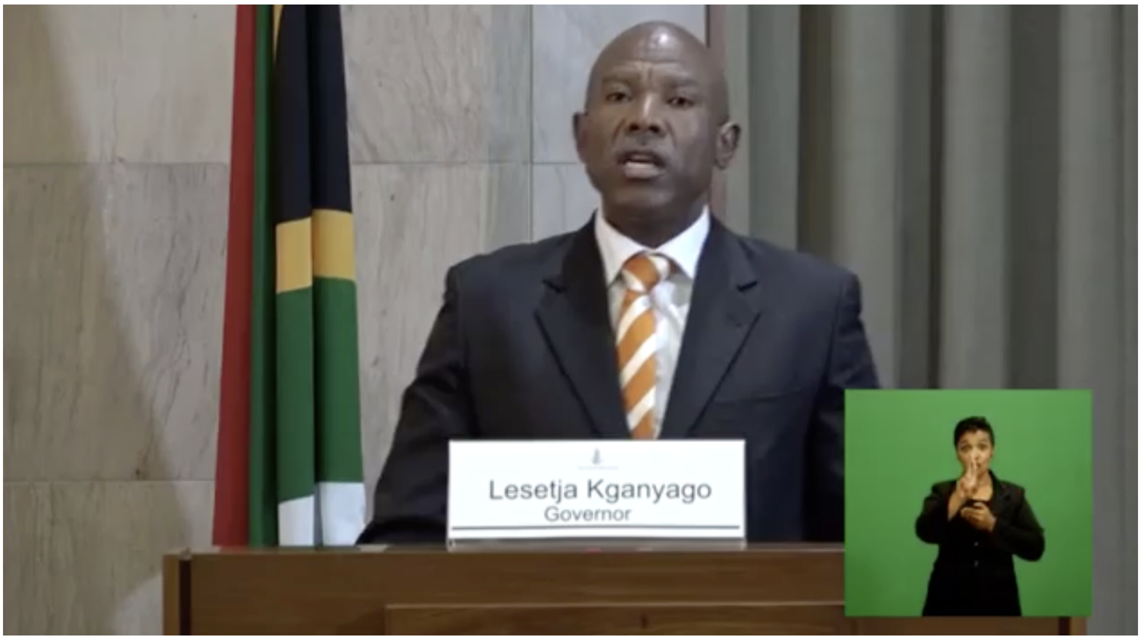


Figure 5: This figure displays a frame corresponding to the 18th utterance of the MPC call, as represented in the heatmaps of Figure 4. In the bottom right corner, a woman ca be observed translating the spoken content into sign language. The video is part of the Monopoly-dataset and has, for memory reasons, a reduced resolution.

ing finetuned models can be used or existing models can be finetuned on a different task to improve the informational value of the embeddings.
A notable feature of the transformer architecture is the attention mechanism. However, this mechanism, while mathematically elegant, remains largely hidden to human interpretation. This opacity extends to the broader transformer model, being a ”black box” in terms of human understanding. The reasons behind specific predictions can remain elusive. In the study of Mathur et al. all the baseline models are ”black box”-models. While there exist models from economics- and marketmicrostructure-theory that offer more transparency and are rooted in economic principles, they often come with high costs for the needed market data. But MPC videos and transcripts are freely available from central banks. Thus data costs can be a potential driver to the adoption of transformer models in finance.
Future research might try to incorporate more modalities. It would be interesting to see if there is a saturation point beyond which the addition of modalities ceases to benefit, or even decreases model performance. However, practical constraints, such as hardware limitations, might cap the number of modalities that can be integrated. In the context of this study, the inclusion of historical price data could offer valuable insights, given the profound influence of past prices on future volatility.
Lastly, it is imperative to acknowledge the interconnectedness of the global financial ecosystem. This study, along with that of Mathur et al., has not delved deep into the cross-relationships between asset prices and central banks across nations. Monetary policies in one country can have effects across borders, influencing asset prices of other countries.

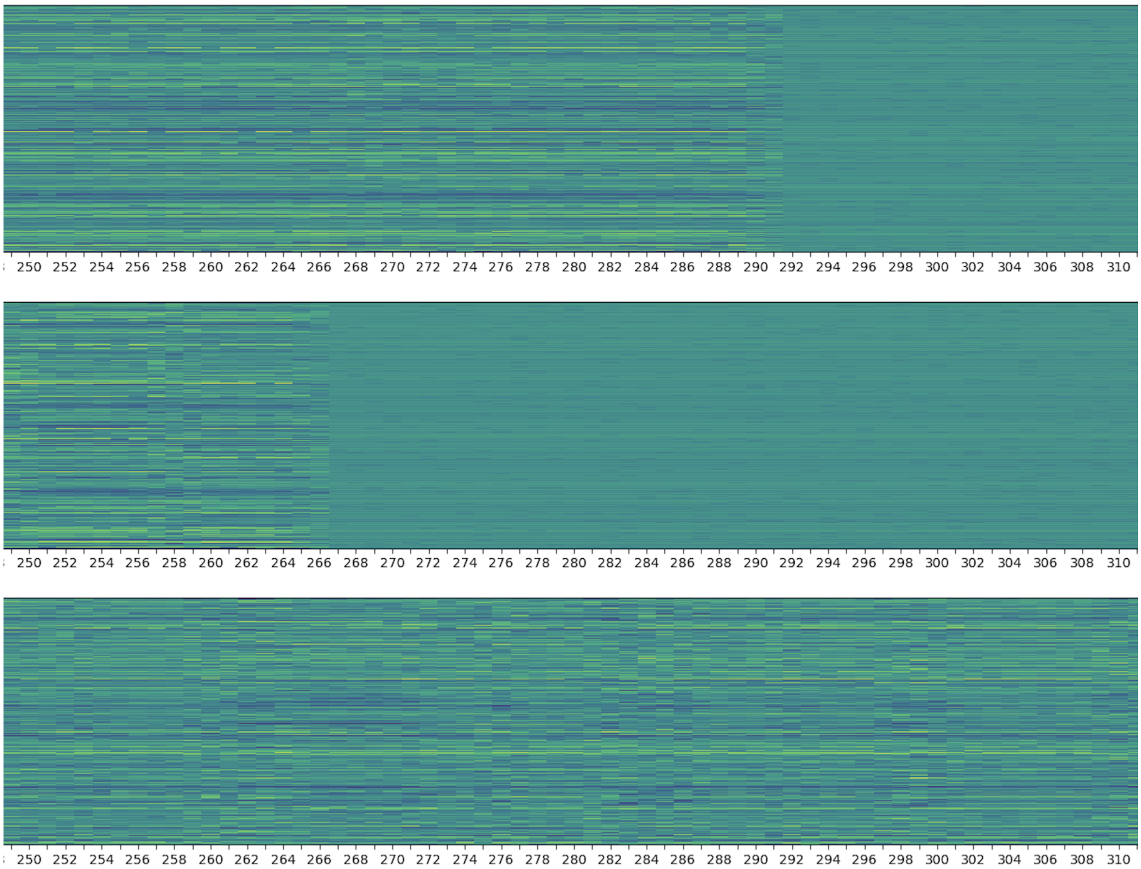


Figure 6: This figure presents a segment of the heatmaps representing the cross-attention weights of the base model across three modalities: video, audio, and text. The displayed data encompasses three observations from a batch consisting of four observations. A more pronounced yellow signifies higher attention weights.

References

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[2] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.

[3] Puneet Mathur, Atula Neerkaje, Malika Chhibber, Ramit Sawhney, Fuming Guo, Franck Dernoncourt, Sanghamitra Dutta, and Dinesh Manocha. Monopoly: Financial prediction from monetary policy conference videos using multimodal cues. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 2276–2285, 2022.

[4] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

[5] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[6] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[8] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

[9] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

[10] Ramit Sawhney, Mihir Goyal, Prakhar Goel, Puneet Mathur, and Rajiv Shah. Multimodal multi-speaker merger & acquisition financial modeling: A new task, dataset, and neural baselines. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6751–6762, 2021.

[11] Yu Qin and Yi Yang. What you say and how you say it matters: Predicting stock volatility using verbal and vocal cues. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 390–401, 2019.

[12] Dogu Araci. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*, 2019.

[13] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*, 2018.