2024 5th International Conference on Information Science, Parallel and Distributed Systems (ISPDS) | 979-8-3503-5271-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/ISPDS62779.2024.10667569

Deception Detection from Linguistic and Physiological Data Streams Using Bimodal Convolutional Neural Networks

Panfeng Li*
Department of Electrical and Computer Engineering
University of Michigan
Ann Arbor, USA
* Corresponding author: pfli@umich.edu

Mohamed Abouelenien
Department of Computer Science
University of Michigan
Ann Arbor, USA
zmohamedra@umich.edu

Rada Mihalcea
Department of Computer Science
University of Michigan
Ann Arbor, USA
mihalcea@umich.edu

Zhicheng Ding
Fu Foundation School of Engineering and Applied Science
Columbia University
New York, USA
zhicheng.ding@columbia.edu

Qikai Yang
Department of Computer Science
University of Illinois Urbana-Champaign
Urbana, USA
qikaiy2@illinois.edu

Yiming Zhou
Department of Engineer Sciences
Saarland University of Applied Science
Saarland, Germany
yiming.zhou@htwsaar.de

***Abstract*—Deception detection is gaining increasing interest due to ethical and security concerns. This paper explores the application of convolutional neural networks for the purpose of multimodal deception detection. We use a dataset built by interviewing 104 subjects about two topics, with one truthful and one falsified response from each subject about each topic. In particular, we make three main contributions. First, we extract linguistic and physiological features from this data to train and construct the neural network models. Second, we propose a fused convolutional neural network model using both modalities in order to achieve an improved overall performance. Third, we compare our new approach with earlier methods designed for multimodal deception detection. We find that our system outperforms regular classification methods; our results indicate the feasibility of using neural networks for deception detection even in the presence of limited amounts of data.**

Keywords—Neural Networks; Multimodal Deception Detection

I. INTRODUCTION

Deception detection has been a topic of interest across many research fields – ranging from psychology [1] to computer science [2]. With an ever-growing accessibility to multimodal media, for instance social media like YouTube and Snapchat, the detection of deceit based on multimodal data becomes increasingly necessary.

While deception detection is widely used in police interrogation, law enforcement, and employee security screening, the methods used often have a large time-requirement and rely highly upon physiological sensors and human experts, leading to bias and poor accuracy [4]. There have been efforts to eliminate the need of human experts and introduce automated approaches. Machine learning methods have been used for the purpose of deception detection in the past, and efforts have been made to leverage multiple modalities to make predictions on the truthfulness of unseen data [2].

These previous studies relied either on a single modality or on integrated multiple modalities in order to detect deceit using regular classification methods. The usage of a single modality might not provide enough information in order to detect deceit. On the other hand, the usage of multiple modalities means more information, and accordingly provides improved performance in many cases, reaching approximately 60-70% accuracy [3, 5].

This implies that there is still room for improvement and provides the opportunity to take advantage of the availability of multiple modalities to apply advanced learning techniques. Recent studies have shown that convolutional neural networks (CNNs) [12-25] can improve the state-of-the-art performance on various tasks, including image analysis [6, 7], toxicity detection [8], data mining [9], which most recently inspired researchers' interests in utilizing deep learning into the deception detection problem. For instance [10], implemented a fake review detection model using CNNs. However, a single modality was used to construct the network. An additional concern with the usage of multimodal data is the difficulty of collecting such data compared to a single modality. This fact causes the size of multimodal datasets to be limited, which may negatively affect the performance of deep learning methods, which do traditionally use very large datasets for training.

This paper addresses the problem of deception detection using multimodal neural networks. The paper makes three important contributions. First, we use neural networks to learn

from two separate modalities, namely the linguistic and physiological modalities. Second, we construct a fused neural network that learns from both modalities, which to our knowledge has not been attempted before. Third, we compare our approach with earlier approaches that used regular machine learning techniques. Furthermore, we address the issues that arise using a CNN with a small training dataset by using a simple approach to solve the overfitting and large variance problems, namely using majority voting. We additionally devise a new procedure to deal with small datasets, including choosing an appropriate number of parameters as well as fixing the previous trained network weights to form a modality-wise training process.

II. Dataset

Our dataset includes two scenarios, namely "Abortion" and "Best Friend". The subjects were asked to sit comfortably on a chair in a lab and were connected to four physiological sensors including blood volume pulse, skin conductance, skin temperature, and abdominal respiration sensors. The participants were informed of the topic matter before each individual recording. In the two scenarios, subjects were allowed to speak freely first truthfully and then deceptively.

Subjects. The multimodal dataset includes recordings collected from 104 students, including 53 females and 51 males. All subjects expressed themselves in English, had several ethnic backgrounds, and had an age range between approximately 20 and 35 years.

Abortion. In this scenario participants were asked to provide first a truthful and then a deceptive opinion about their feelings regarding abortion and whether they think it is right or wrong. The experimental session consisted of two independent recordings for each case.

Best Friend. In this scenario subjects were instructed to provide an honest description of their best friend, followed by a deceptive description about a person they cannot stand. In the deceptive response, they had to describe an individual they cannot stand as if he or she was their best friend. Hence, in both cases, the person was described positively.

III. Bimodal CNNs

Our data is from two sources, namely the transcripts of the participants' responses, and the physiological data collected during the recordings. Accordingly, we utilize a linguistic CNN (LingCNN), a physiological CNN (PhysCNN), and a BiModal CNN network. The latter one fuses the previous two networks. In addition, a word2vec model devised by [11] is used to transfer the transcripts to vectors as the input to our LingCNN.

PhysCNN. We construct a 1-Dimensional (1-D) CNN for the physiological modality. The inputs of the neural net consist of preprocessed physiological data with dimension of 32, the outputs are the classification results of the input samples.

Firstly, the input data goes through the convolutional layer. We set three different filter sizes as 3, 4, 5, which are the same with the ones in [9]. ReLU (Rectified Linear Unit) activation and max pooling are applied after convolution. All the pooled features are saved, concatenated, and flattened at the end. We pass the flattened output through an added fully-connected layer, with a maximized activation, which provides our final prediction. Cross-entropy is used for training.

LingCNN. We construct a convolutional model for our linguistic module, which is simplified from [9]'s TextCNN. In contrast to the PhysCNN model, this is a two-dimensional model. Similar to the cited paper, we chose filter sizes to be 3×3, 4×4, 5×5.

BiModal CNN. The BiModal CNN represents a modality-wise fashion by first training the PhysCNN and LingCNN models. The relationship among them is shown in Figure 1.

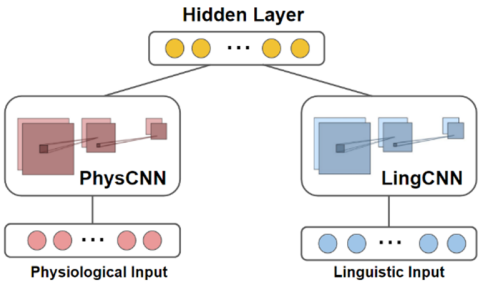

Fig. 1: BiModal CNN

IV. Experimental Setup

A. Data Preprocessing

The physiological measurements are extracted at a rate of 2,048 samples per second using the Biograph Infinity Physiology suite. These features contain raw physiological measurements of the heart rate, skin conductance, respiration rate, and skin temperature using four different sensors. Additionally, we compute their statistical descriptors including maximum and minimum values, means, power means, standard deviations, and mean amplitudes (epochs). The final physiological measurements set include a total of 59 physiological features that contain 40 features extracted from the raw measurement of the heart rate sensor, five skin conductance features, five skin temperature features, and seven respiration rate features. Furthermore, two measurements are extracted from the heart rate and the respiration rate sensors combined, namely, the mean and heart rate max-min difference, which represents a measure of breath to heart rate variability.

We then simply average the values of the physiological data over the whole time period. The dimensions of the feature vectors are reduced from 59 to 32 following the application of Principal Component Analysis (PCA). PCA was used in order to reduce the features dimensions as well as the number of required weights in the network. Furthermore, our preliminary results indicated better performance following dimensionality reduction.

Sentences in the transcripts were converted into word vectors in order to process the linguistic modality. To learn the representations of words, namely "word embeddings", we use the word2vec model devised by [11], where the training dataset is from Matt Mahoney. We set the embedding size, namely the length of word vectors as 32, similar to that of the physiological modality, and only keep the top 500 words with highest

frequency in the text documents. Finally, we obtain a 500 × 32 word embedding matrix and a word dictionary, where each word corresponds to a unique value.

For each text transcript, we delete all non-verbal and non-numerical items and save the results as a transcript string, which is then transferred to a transcript vector through the dictionary described above. To unify the length of all the vectors for batch implementation in training and testing, we firstly identify the transcript vector(s), which have the maximum length M, and accordingly pad the remaining vectors with zeros. If a word does not exist in the dictionary, we replace it with a special notation as "UNK", which also corresponds to the value zero. Furthermore, each value in the transcript vectors is transferred to a word vector through a lookup operation on the previous embedding matrix. Hence, the transcripts are represented as arrays with dimension M × 32.

B. Training and Testing Procedures

We randomly shuffle and split our dataset for training and testing with a ratio of 9 : 1 and save the shuffled and split index. By using the same index, we are able to match the features from the two modalities, when integrated together.

For the linguistic and physiological modalities, the final predictions are obtained after applying a maximization function on the output scores of the network. The integrated network takes the output scores from linguistic and physiological modalities as input, and concatenates them as a single feature vector. The details of training and testing for the overall framework one-time are as follows:

- Train linguistic and physiological modalities once using all the training data.
- Fix the weights for the linguistic and physiological modalities and input the training and testing data to obtain the corresponding linguistic and physiological features.
- Use the above training features as inputs for training the overall framework, and record the test results on testing features.

Specifically, we apply the majority voting method to determine the final predictions in order to address the overfitting and variance problem of the network. We record all the prediction results among a certain number of running times and decide the label for each test sample using the mode value. We also perform a stability analysis in Section 5, which shows the majority voting method is effective and stable.

V. Experimental Results and Discussion

Our entire dataset consists of 416 samples including the "Abortion" and "Best Friend" topics. We evaluate the performance of the features extracted from each of the two topics as well as both topics combined using the overall accuracy and class recall. Furthermore, we compare the performance of our proposed networks to that of learning using regular classifiers such as Decision Tree, Support Vector Machine (SVM), and Logistic Regression.

A. Individual and integrated modalities

Figure 2 shows the deception and truthfulness recall in addition to the overall accuracy using different modalities for the "Abortion" topic. The figure indicates that overall the combination of linguistic and physiological modalities improves the performance as compared to the physiological modality. Specifically, while the physiological modality achieves the highest accuracy for the deception class, it attains the lowest truthful class accuracy, and is not performing as well as the linguistic modality considering the overall accuracy. The linguistic features exhibits close performance to the integrated modality. We may state that, for the "Abortion" topic, the combination of physiological modality with the linguistic one does not benefit our model.

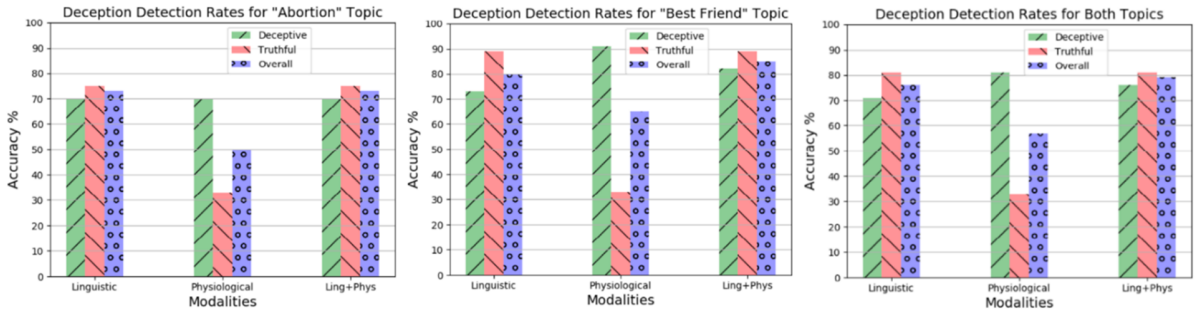

Fig. 2 Deception recall, truthfulness recall, and overall accuracy percentages for individual and integrated modalities using features extracted from the "Abortion", "Best Friend" and both topics

The performance of the features extracted from the "Best Friend" topic is significantly better than the first topic using different modalities as can be seen in Figure 2. The overall accuracy using the linguistic modality reaches nearly 80% as compared to the approximately 70% achieved in the "Abortion" topic. Using the physiological modality, we compare 65% achieved for the "Best Friend" topic with 50% for "Abortion" topic on overall accuracy. The overall accuracy using both modalities indicates noticeable improvement compared to using individual modalities.

Combining the two topics provides lower performance across all three modes of evaluation for all modalities. This may be rationalized by considering the fact that our model performed relatively poorly on the "Abortion" topic. As a result, the overall performance is slightly worse than that of "Best Friend"

topic but better than “Abortion” topic. This can be seen in Figure 2.

In all three cases, we see that the detection rate of deceptive responses is better than that of the truthful one for the physiological modality. The reason behind this difference may be because the deceptive scenarios triggered more emotional arousal for the subjects, resulting in physiological patterns that were beneficial in training the networks. On the other hand, since the linguistic modality extracts semantic relations present in the same topic, the comparable performance for deceptive and truthful responses might be reasonable, as we train and test on data from the same topic.

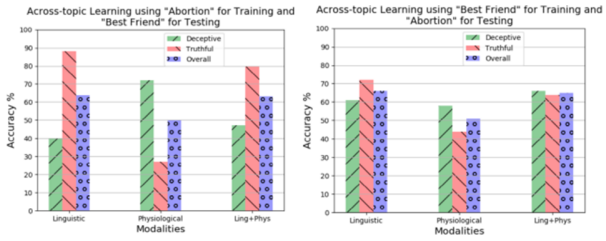

Fig. 3: Deception, truthfulness, and overall accuracy percentages for individual / integrated modalities using across-topic learning

B. *Cross-Topic Learning*

We analyze how well our model works on cross-topic deception detection. We train the model using the data from the “Abortion” topic and test on data from “Best Friend” topic. The results are presented in Figure 3. The linguistic modality outperforms the physiological and the combined modalities on detecting truthful responses, but performs the worst of the three on deceptive responses. The overall accuracy of the integrated modality is similar with the one of linguistic and they both exceed 60%.

This performance is flipped for the physiological modality, where we see the best performance is achieved using deceptive responses. Once again, this is likely because the physiological markers for deceptive responses are more indicative than those of the truthful responses.

In Figure 3, we can notice that while the trends are the same for linguistic and physiological modalities, the gaps between the recall figures for deceptive and truthful responses are significantly lower than the previous one. The results in this case indicate more stability regarding the truthful and deceptive classes performance compared to their performance in Figure 2. This can be explained by having more domain-specific words in the “Abortion” topic, which affects the learning process.

We may further compare these results with the ones discussed in subsection 5.1. For the linguistic modality, the overall accuracy is lower for cross-topic learning, which indicates that the linguistic features are topic-dependent.

We also observe that the physiological modality, regardless of the topic used to train and testing consistently provides skewed results. Furthermore, training on “Best Friend” topic and testing on “Abortion” topic decreases the overall performance as compared to training and testing on the same “Best Friend” topic, but shows a very slight improvement as compared to training and testing on the same “Abortion” topic.

The combination of the two modules also does not perform as well regarding the overall accuracy for cross-topic learning as compared to the results in subsection 5.1.

C. *Stability Analysis*

Here we analyze the stability of our modalities. Since we determine our final predictions using majority voting among results from different running times, it is important to find the relationship between the accuracy and the number of running times. We tested the overall accuracy, deceptive recall, and truthful recall on individual topics and both topics combined.

The results are shown in Figure 4.

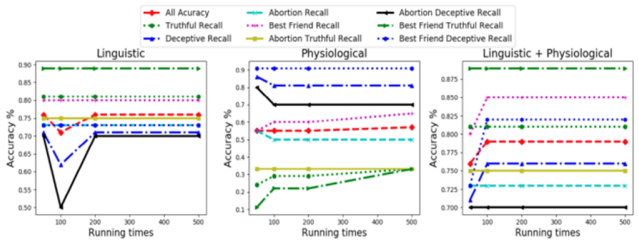

Fig. 4: Accuracy results among different running times

From Figure 4, we can notice that the deceptive recall of the “Abortion” topic using the linguistic modality firstly decreases at 100 running times. The deceptive recall and overall accuracy also decrease accordingly. However, they quickly return to the normal level at 200 running times and stay consistent till 500 running times.

For the physiological modality, the truthful recall on the “Best Friend” topic and both topics combined is increasing when the running time goes from 50 to 500, while the deceptive recall on “Best Friend” and both topics is slightly decreasing. The “All Accuracy” of the physiological modality stays consistent from 50 to 200 running times and increases slightly at 500 running times.

For the integrated modalities, due to the increase of best friend deceptive and truthful recalls in the beginning, the recalls of the “Best Friend” topic and both topics increase. After 100 running times, all the accuracy figures remain unchanged, which indicates that the integrated modality is stable over running times despite the observed changes with the linguistic and physiological modalities. In conclusion, our models are stable after running times of 200.

D. *Compared with the Regular Models*

We used the best multimodal systems for deception detection reported in a previous work [5] and compare their performance with ours. In those models, psycholinguistic lexicons and unigrams were used for linguistic features, while the paper used the same types of physiological features we utilized. In the end, the linguistic and physiological features were concatenated, and decision tree classifiers were used to give the final results. Here we also use SVM and logistic regression for classification. We compare results on the two

topics combined – "Abortion" and "Best Friend", and use both linguistic and physiological data, as shown in Figure 5.

In our experiments, decision trees also used majority voting after a running them of 200. SVM and logistic regression did not need majority voting as their results remain stable over

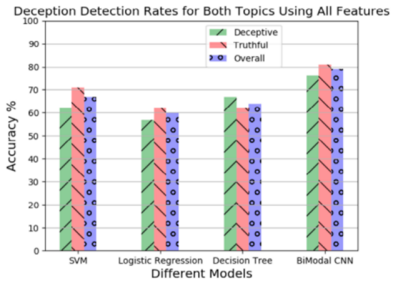

Fig. 5: Comparison among regular models and our BiModal CNN

different running times. We note that, for all the different detection rates (deceptive, truthful and overall), our model performs better.

VI. Conclusion and Future Work

From the experimental results, we observed that the linguistic modality worked significantly better than the physiological modality. One of the reasons is that the linguistic modality used all the information in the transcripts, while the physiological modality simply averaged the data over the whole time period, which could result in loss of some physiological patterns in the learning process.

It can also be noticed that in the majority of the cases, the bimodal network achieved better performance that the unimodal ones. This indicates that the proposed fused neural network can integrate and learn discriminative features from multimodal data, which results in improved and more reliable performance.

For training and testing on the same topic, we note that by combining both modalities, the overall accuracy is higher than that obtained using the individual modalities. The same trend is observed for cross-topic learning, as well. We can therefore conclude that bimodal fusion has an overall advantageous effect over using individual modalities. This may be explained by considering that the fused network was provided by richer information using the two modalities.

Our experiments also indicated that cross-topic learning leads to a decrease in the performance for our model especially for the linguistic modality, which indicates that the performance is topic-dependent. For future work, we will consider performing a time-series analysis to potentially discover time-dependent relationships among the data. For the BiModal CNN, we will also extract different sizes of hidden layers from the LingCNN and PhysCNN, and then concatenate them to form new feature vectors.

References

[1] B. M. DePaulo, J. J. Lindsay, B. E. Malone, "Cues to deception," Psychological Bulletin, vol. 129, no. 1, pp. 74–118, 2003.

[2] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock, "Finding deceptive opinion spam by any stretch of the imagination," in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1, 2011, pp. 309–319.

[3] M. Abouelenien, V. Pérez-Rosas, R. Mihalcea, and M. Burzo, "Deception detection using a multimodal approach," in Proceedings of the 16th International Conference on Multimodal Interaction, 2014, pp. 58–65.

[4] C. F. Bond and B. M. DePaulo, "Accuracy of deception judgments," Personality and Social Psychology Review, vol. 10, pp. 214–234, 2006.

[5] M. Abouelenien, V. Pérez-Rosas, R. Mihalcea, and M. Burzo, "Detecting deceptive behavior via integration of discriminative features from multiple modalities," IEEE Transactions on Information Forensics and Security, vol. 12, pp. 1042–1055, 2017.

[6] H. Liu, Y. Shen, W. Zhou, Y. Zou, C. Zhou, and S. He, "Adaptive speed planning for unmanned vehicle based on deep reinforcement learning," arXiv preprint arXiv:2404.17379, 2024.

[7] Y. Shen, H. Liu, X. Liu, W. Zhou, C. Zhou, and Y. Chen, "Localization through particle filter powered neural network estimated monocular camera poses," arXiv preprint arXiv:2404.17685, 2024.

[8] W. Dai, J. Tao, X. Yan, Z. Feng, and J. Chen, "Addressing unintended bias in toxicity detection: An lstm and attention-based approach," in ICAICA, 2023, pp. 375–379.

[9] Y. Li, X. Yan, M. Xiao, W. Wang, and F. Zhang, "Investigation of creating accessibility linked data based on publicly available accessibility datasets," in Proceedings of the 2023 13th International Conference on Communication and Network Security, 2024, p. 77–81.

[10] C. Sun, Q. Du, and G. Tian, "Exploiting product related review features for fake review detection," Mathematical Problems in Engineering, 2016.

[11] T. Mikolov, K. Chen, and J. Dean, "Efficient estimation of word representations in vector space," arXiv preprint arXiv:1301.3781, 2013.

[12] M. Xiao, Y. Li, X. Yan, M. Gao, and W. Wang, "Convolutional neural network classification of cancer cytopathology images: taking breast cancer as an example," arXiv preprint arXiv:2404.08279, 2024.

[13] X. Yan, W. Wang, M. Xiao, Y. Li, and M. Gao, "Survival prediction across diverse cancer types using neural networks," arXiv preprint arXiv:2404.08713, 2024.

[14] Z. Zeng, D. Wang, F. Yang, H. Park, Y. Wu, S. Soatto, B.-W. Hong, D. Lao, and A. Wong, "Wordepth: Variational language prior for monocular depth estimation," arXiv preprint arXiv:2404.03635, 2024.

[15] R. Zhang, Z. Zeng, Z. Guo, X. Gao, K. Fu, and J. Shi, "Dspoint: Dual-scale point cloud recognition with high-frequency fusion," arXiv preprint arXiv:2111.10332, 2021.

[16] R. Zhang, Z. Zeng, Z. Guo, and Y. Li, "Can language understand depth?" in Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 6868–6874.

[17] L. Ma and M. Li, "A quantitative way to utilise the social network in social status: A study using cbdb song dynasty data," in 10th International Conference of Digital Archives and Digital Humanities, 2019.

[18] Y. Tian, H. Zhang, Y. Jiang, P. Li, and Y. Li, "A fusion feature for enhancing the performance of classification in working memory load with single-trial detection," IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 27, no. 10, pp. 1985–1993, 2019.

[19] M. Liu, H. Zhang, J. Song, and M. Lu, "Using generative model for intelligent design of dielectric resonator antennas," Microwave and Optical Technology Letters, vol. 66, no. 1, p. e34013, 2024.

[20] Y. Ge, Z. Xu, Y. Xiao, G. Xin, Y. Pang, and L. Itti, "Encouraging disentangled and convex representation with controllable interpolation regularization," in WACV, 2023, pp. 4761–4769.

[21] J. Yao, T. Wu, and X. Zhang, "Improving depth gradient continuity in transformers: A comparative study on monocular depth estimation with cnn," arXiv preprint arXiv:2308.08333, 2023.

[22] C. Zhou, Y. Zhao, J. Cao, Y. Shen, X. Cui, and C. Cheng, "Optimizing Search Advertising Strategies: Integrating Reinforcement Learning with Generalized Second-Price Auctions for Enhanced Ad Ranking and Bidding," arXiv preprint arXiv:2405.13381, 2024.

[23] J. Yuan, L. Wu, Y. Gong, Z. Yu, Z. Liu, and S. He, "Research on intelligent aided diagnosis system of medical image based on computer deep learning," arXiv preprint arXiv:2404.18419, 2024.

[24] Y. Gong, H. Zhang, R. Xu, Z. Yu, and J. Zhang, "Innovative deep learning methods for precancerous lesion detection," International Journal of Innovative Research in Computer Science & Technology, vol. 12, no. 2, pp. 81–86, 2024.

[25] N. Quach, Q. Wang, Z. Gao, Q. Sun, B. Guan, and L. Floyd, "Reinforcement learning approach for integrating compressed contexts into knowledge graphs," arXiv preprint arXiv:2404.12587, 2024.