





Uncertainty Quantification of Multimodal Models

Ajibola Obayemi¹(✉)  and Khuong An Nguyen² 

¹ University of Brighton, Brighton, East Sussex BN2 4GJ, UK

e.a.obayemi@brighton.ac.uk

² Royal Holloway University of London, Egham, Surrey TW20 0EX, UK

khuong.nguyen@rhul.ac.uk

Abstract. Multimodal classification models, particularly those designed for fine-grained tasks, offer significant potential for various applications. However, their inability to effectively manage uncertainty often hinders their effectiveness. This limitation can lead to unreliable predictions and suboptimal decision-making in real-world scenarios. We propose integrating conformal prediction into multimodal classification models to address this challenge. Conformal prediction is a robust technique for quantifying uncertainty by generating sets of plausible classifications for unseen data. These sets are accompanied by guaranteed confidence levels, providing a transparent assessment of the model's prediction reliability. By integrating conformal prediction, our objective is to increase the reliability and trustworthiness of multimodal classification models, thereby enabling more informed decision-making in contexts where uncertainty is a significant factor.

Keywords: Uncertainty quantification · Multimodal models · Conformal Prediction · Fine-grained classification

1 Introduction

The key to adopting fine-grained classification models in real-world applications is their ability to handle uncertainty. These models need to be confident in their predictions and transparent about the uncertainty associated with each prediction. This is where uncertainty quantification becomes crucial. It allows the models to generate probabilistic outputs, expressing the most likely prediction and the associated confidence level. To this end, we explore conformal prediction, a technique for uncertainty quantification [1, 2]. Conformal prediction moves beyond point predictions. It generates sets of possible classifications for unseen data, guaranteeing the true class resides within the set with a predetermined confidence level. In low confidence scenarios, it might return an empty set, indicating the model's uncertainty [3, 4]. We evaluate our approach on the TAIMD-17k multimodal dataset [8]. The primary contributions of this work include the

integration of Conformal Prediction into multimodal models, an evaluation of the TAIMD-17k dataset with our method, and a proposed approach to enhance the reliability and trustworthiness of multimodal classification models.

2 Related Works

In their study, [5] developed a multimodal neural network (MMNN) comprising distinct feature extraction and classification phases. They employed this network to investigate a method for uncertainty estimation. To quantify model uncertainty, they incorporated dropout layers prior to each hidden layer within the feature extraction stages for every modality. As a result, when presented with an input (x) for model prediction, an ensemble of predictions is generated, and a probability distribution is constructed from the predicted probabilities. Ultimately, the standard deviation of this probability distribution is calculated to determine the model’s uncertainty. The dropout-based uncertainty estimation in the work described above uses the standard deviation of ensemble predictions from stochastic forward passes to gauge model uncertainty, offering a heuristic measure without formal guarantees. In contrast, the conformal prediction approach constructs prediction sets or intervals with a user-specified coverage probability by calibrating against a separate dataset, providing statistically valid uncertainty quantification that accounts for both aleatoric and epistemic uncertainty in a distribution-free manner. While dropout is integrated within the model architecture, conformal prediction is typically applied to the model’s output, offering stronger guarantees at the cost of requiring calibration data.

[6] and [7] emphasise the necessity of uncertainty quantification in multimodal models. Specifically, [6] proposed a healthcare decision system framework incorporating multimodal learning and uncertainty quantification. Similar to [5], their initial experiments employed Monte Carlo Dropout (MCD), dropout between network layers to produce a predictive distribution and calculate predictive entropy. However, this method suffers from the previously mentioned drawbacks associated with using dropout between layers, a limitation shared by both [6] and [7].

3 Methods

[9] offer a comprehensive analysis of uncertainty quantification techniques in their work and outline the various approaches used to measure uncertainty in machine learning models. Recognising the diverse landscape they present, we sought an uncertainty quantification method well-suited to multimodal classification. Inspired by prior research in this field, particularly the use of conformal prediction for quantifying uncertainty in unimodal models [10–12], we opted to explore conformal prediction as a way to investigate the predictive confidence and certainty of our text and image-based multimodal model.

3.1 Multimodal Model

In this section, we utilise the Multimodal Bi-Transformers (MMBT) architecture initially proposed by [13]. Multimodal Bi-Transformers (MMBT) is a vision-language model capable of learning a more encompassing relationship from a text and image multimodal dataset. MMBT combines pre-trained encoders for each modality (text and image) and then fine-tunes them jointly. The key idea is how it bridges the modalities: the image embedding is projected into the same space as the text tokens. To better understand this core component, Fig. 1, which visually dissects the MMBT architecture, offers valuable insights into how we adapted it for our specific application and its role in uncertainty quantification. Following the framework of [13], which implemented ResNet for visual feature extraction, we instead integrated Clip as our visual encoder, while preserving BERT for text encoding. The detailed multimodal configuration and hyper-parameters are provided in Table 1.

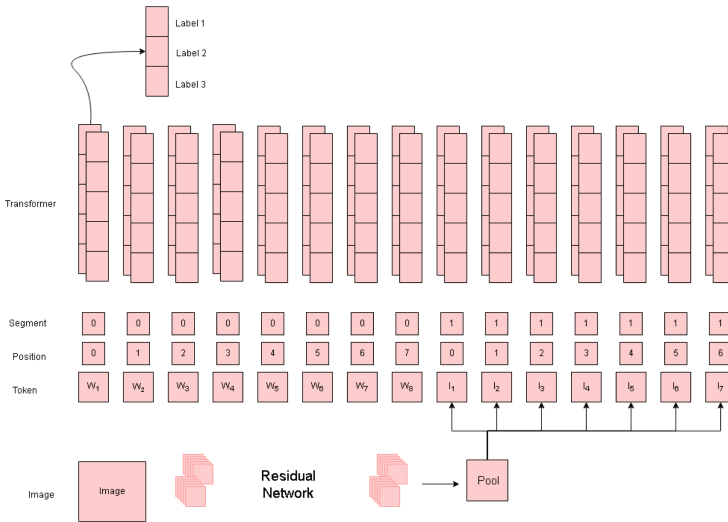


Fig. 1. Multimodal bi-transformer architecture depicting the image and text projections. Our implementation modified MMBT and replaced ResNet with CLIP as the image encoder.

3.2 Conformal Prediction

Conformal prediction, or conformal inference, offers a compelling approach to uncertainty quantification within machine learning. Unlike traditional point predictions, conformal prediction generates statistically valid prediction sets, called calibration sets or conformity regions. These sets possess a crucial property –

Table 1. Configuration and Hyperparameter for Multimodal model.

Feature	Value
Epoch	50
Optimiser	MADGRAD
Image encoder	CLIP
Image encoder size	288
Number of image embeddings	4
Text encoder	BERT
Token Sequence Length	120
Loss function	Cross Entropy
Learning rate	2e-8
Early stopping (patience)	5

they are calibration-invariant. This implies that conformal prediction guarantees, with a user-defined confidence level $(1 - \alpha)$, that the true label will reside within the predicted set, regardless of the underlying data distribution or model complexity. This feature is particularly advantageous for complex, non-parametric models like deep neural networks, where distributional assumptions may be challenging or misleading. Furthermore, conformal prediction avoids the need for specific loss functions typically employed in training, making it a flexible uncertainty quantification tool readily applicable to pre-trained models [3, 12, 14].

Mathematically, we can denote the target variable as Y and the prediction set as $I(x, \alpha)$. Where x represents the features of the test data point and α is the significance level ($\alpha \in [0, 1]$). The conformal predictor is considered valid when the condition expressed in 1 for all test data points (x) and significance levels (α) is true.

$$P(Y \in I(x, \alpha)) \geq 1 - \alpha \quad (1)$$

4 Experiments

To quantify uncertainty in our multimodal model Sect. 3.1, we followed and modified the implementation detailed in the work of [3]. Their work focuses only on unimodal architectures, but we adapted and modified this implementation to work with multimodal based architectures.

While [3] effectively demonstrated the implementation of conformal prediction (CP) for unimodal regression and classification tasks, their approach primarily focuses on handling unimodal datasets and pre-trained unimodal models. To address the inherent complexities of multimodal data, we present a modified conformal prediction implementation designed explicitly for multimodal-based architecture (detailed in Subsect. 3.1). This adaptation is necessary in our effort to explore how conformal prediction can be applied to quantify uncertainty in

multimodal models. We conducted our experiments using the TAIMD-17k multimodal dataset (publicly available at <https://github.com/multimodal-research/TAIMD-17k>). This dataset comprises 17,000 paired image and text samples of ink cartridge items, characterised by fine-grained classification features.

To initialise the conformal prediction framework for our experiments, we followed a multi-step process:

1. **Pre-trained Model Weight Loading:** We leveraged the pre-trained weights of our multimodal model. This step is crucial because conformal prediction relies on a pre-trained model to establish a baseline for non-conformity scores during calibration. The calibration process involves analysing how well the model's predictions align with the true labels within the calibration set. By leveraging a pre-trained model, we ensure the non-conformity scores effectively capture the model's behaviour and inherent uncertainties.
2. **Calibration Set Generation:** Recognising the crucial role of a well-calibrated dataset for valid conformal prediction guarantees, we meticulously constructed a calibration set. This set was derived from our existing evaluation split (`eval_split`). The evaluation split typically represents a portion of the training data reserved for assessing model performance during the training process.
3. **ConformalModel Instantiation:** We instantiated the *ConformalModel* class, which serves as the core component for the conformal prediction framework. The instantiation process involved specifying several key parameters:
 - (a) *nn_model*: This argument references the pre-trained multimodal model mentioned above. The *ConformalModel* class leverages this model to generate predictions for unseen data points during the conformal prediction process.
 - (b) *eval_dataloader*: This parameter represents a data loader object specifically designed for the evaluation split. Data loaders are crucial components in PyTorch (or similar deep learning frameworks) for efficiently managing and delivering batches of data during training and evaluation. By providing the *eval_dataloader*, we ensure the *ConformalModel* has access to the data points from the calibration set for the calibration stage.
 - (c) *alpha*: This parameter signifies the chosen significance level (α) for conformal prediction. A significance level of 0.05 translates to a 95% confidence level, meaning the constructed prediction sets aim to guarantee, with 95% confidence, that the true labels will reside within the sets for unseen data points.
 - (d) *lambda_criterion*: This parameter specifies the selection criterion used during the construction of prediction sets. In this case, we opted for the 'size' criterion, which prioritises the selection of classes with non-conformity scores less than or equal to the threshold, aiming for prediction sets with a minimum size while maintaining coverage guarantees.

5 Result

This section presents the experimental results of implementing conformal prediction on the multimodal classification model. Our evaluation focuses on key performance metrics: average coverage, expected calibration error, and conditional coverage. The average coverage measures the extent to which the prediction sets generated by conformal prediction contain the true class. A higher average coverage indicates that the model is generating more conservative predictions. Expected calibration error assesses the calibration of the prediction sets, evaluating how well the reported confidence levels align with the actual coverage. A lower expected calibration error signifies the accuracy of the model's confidence estimates. Conditional coverage is a more granular metric that examines coverage for different confidence levels, ensuring the model maintains consistent performance across varying degrees of uncertainty.

Figures 2 and 3 illustrate the error rates for two different confidence levels, 0.05 and 0.10, respectively. Additionally, Tables 2 and 3 provide a detailed breakdown of the performance of the model, including accuracy, uncertainty, emptiness, average prediction size, and overall error rate. These results offer insights into the effectiveness of conformal prediction in quantifying uncertainty and improving the reliability of the multimodal classification model.

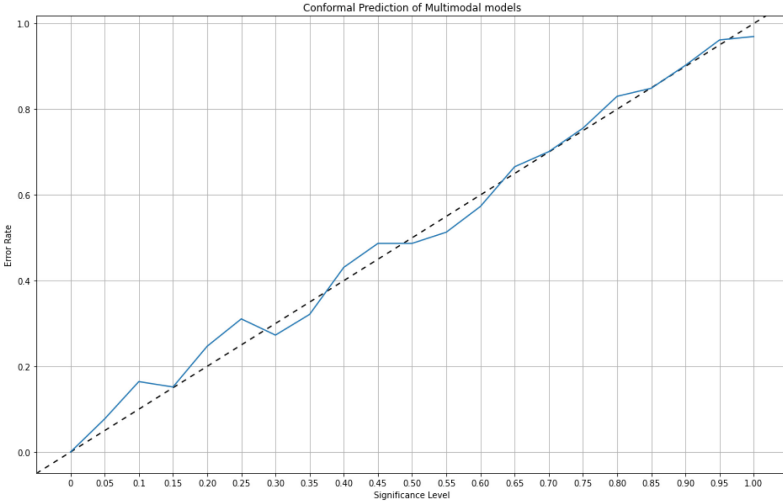


Fig. 2. The error rate for 95% confidence level, $\epsilon = 0.05$

6 Discussion

In this section, we offer a comprehensive evaluation of the results. By focusing on key metrics, we assess the performance of the approach and its ability to pro-

Table 2. Confidence level - 95%, $\epsilon = 0.05$

#	Accuracy	Uncertainty	Emptiness	Average Prediction Size	Error rate
1	1	100	0	30	0
2	0.9226	100	0	28	0.0774
3	0.8356	100	0	26	0.1644
4	0.8482	100	0	27	0.1518
5	0.7532	100	0	23	0.2468
6	0.6892	100	0	20	0.3108
7	0.7271	100	0	23	0.2729
8	0.6887	100	0	22	0.3213
9	0.5689	100	0	18	0.4311
10	0.5130	100	0	15	0.4870
11	0.5130	100	0	17	0.4870
12	0.4870	100	0	16	0.5130
13	0.4268	100	0	13	0.5732
14	0.3343	100	0	10	0.6657
15	0.2989	100	0	9	0.7011
16	0.2443	100	0	7	0.7557
17	0.1699	100	0	4	0.8301
18	0.1509	100	0	5	0.8491
19	0.0096	100	0	2	0.9024
20	0.0382	17.16	0	1	0.9617
21	0.0302	0	0	1	0.9697

Table 3. Confidence level - 90%, $\epsilon = 0.1$

#	Accuracy	Uncertainty	Emptiness	Average Prediction Size	Error rate
1	0.9990	100	0	30	0.0010
2	0.8370	100	0	27	0.1630
3	0.7426	100	0	22	0.2574
4	0.6358	100	0	19	0.3642
5	0.6262	100	0	21	0.3738
6	0.4865	100	0	15	0.5135
7	0.3616	100	0	9	0.6384
8	0.2960	100	0	9	0.7040
9	0.1955	100	0	6	0.805
10	0.8944	100	0	3	0.1056
11	0.0302	100	0	1	0.9698

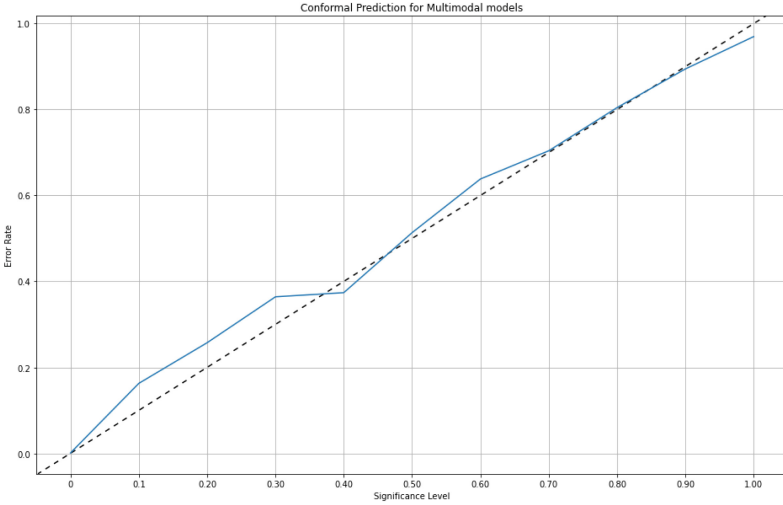


Fig. 3. The error rate for 90% confidence level, $\epsilon = 0.1$

vide reliable uncertainty quantification. Some of the key findings are highlighted below.

1. **Consistent Coverage and Calibration:** The conformal prediction method consistently achieves the desired coverage levels across a range of significance levels, indicating its strong calibration properties. This is evident in the error rate plots and the tabular results, where the empirical error rates closely align with the specified significance levels.
2. **Trade-off Between Accuracy and Uncertainty:** As expected, there is a trade-off between accuracy and uncertainty. The accuracy decreases as the significance level increases, but the uncertainty increases, reflecting a wider prediction interval. This behaviour is consistent with the fundamental principles of conformal prediction, which aims to provide guarantees on the coverage of the prediction intervals while maintaining a reasonable level of accuracy.
3. **Minimal Empty Intervals:** The results demonstrate that the prediction intervals are rarely empty, suggesting that the conformal prediction method can provide meaningful predictions even for challenging cases.
4. **Average Prediction Size:** The average prediction size indicates the informativeness of the predictions. While the exact interpretation of this metric may depend on the specific application, it can help understand the trade-off between accuracy and uncertainty.

7 Implications and Future Directions

This work highlights the effectiveness of conformal prediction in quantifying uncertainty in multimodal models, providing crucial prediction intervals for decision making. Future research should focus on improving the scalability and

efficiency for large datasets, customising the method for specific applications, comparing it with Bayesian approaches, and further exploring its theoretical properties. The findings establish a foundation for future research and a broader application of conformal prediction in various multimodal domains.

The results of this work carry substantial weight for the real-world use of conformal prediction with multimodal models. Delivering dependable prediction intervals with assured coverage is critical for applications demanding uncertainty quantification, like making choices when outcomes are uncertain.

References

1. Vovk, V., Gammerman, A., Shafer, G.: *Algorithmic Learning in a Random World*, vol. 29. Springer, New York (2005)
2. Gammerman, A., Vovk, V., Vapnik, V.: Learning by transduction. arXiv preprint [arXiv:1301.7375](https://arxiv.org/abs/1301.7375) (2013)
3. Angelopoulos, A.N., Bates, S.: A gentle introduction to conformal prediction and distribution-free uncertainty quantification. arXiv preprint [arXiv:2107.07511](https://arxiv.org/abs/2107.07511) (2021)
4. Zhang, D., Chatzimpampas, A., Kamali, N., Hullman, J.: Evaluating the utility of conformal prediction sets for AI-advised image labeling. In: *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1–19 (2024)
5. Brown, K.E., Bhuiyan, F.A., Talbert, D.A.: Uncertainty quantification in multimodal ensembles of deep learners, In: *The Thirty-Third International Flairs Conference* (2020)
6. Bezirganyan, G.: Data and decision fusion with uncertainty quantification for ML-based healthcare decision systems. In: *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pp. 5169–5172 (2023)
7. Bezirganyan, G., Sellami, S., Berti-Équille, L., Fournier, S.: LUMA: a benchmark dataset for learning from uncertain and multimodal data. arXiv preprint [arXiv:2406.09864](https://arxiv.org/abs/2406.09864) (2024)
8. Sheng, Q.Z., et al. (eds.) *Advanced Data Mining and Applications*, vol. 15392. Springer, Singapore (2025). <https://doi.org/10.1007/978-981-96-0850-8>
9. Abdar, M., et al.: A review of uncertainty quantification in deep learning: techniques, applications and challenges. *Inf. Fusion* **76**, 243–297 (2021)
10. Alvarsson, J., McShane, S.A., Norinder, U., Spjuth, O.: Predicting with confidence: using conformal prediction in drug discovery. *J. Pharm. Sci.* **110**(1), 42–49 (2021)
11. Vazquez, J., Facelli, J.C.: Conformal prediction in clinical medical sciences. *J. Healthc. Inf. Res.* **6**(3), 241–252 (2022)
12. Guha, E., Natarajan, S., Möllenhoff, T., Khan, M.E., Ndiaye, E.: Conformal prediction via regression-as-classification. arXiv preprint [arXiv:2404.08168](https://arxiv.org/abs/2404.08168) (2024)
13. Kiela, D., Bhooshan, S., Firooz, H., Perez, E., Testuggine, D.: Supervised multimodal bitransformers for classifying images and text. arXiv preprint [arXiv:1909.02950](https://arxiv.org/abs/1909.02950) (2019)
14. Nguyen, K.A., Luo, Z.: Reliable indoor location prediction using conformal prediction. *Ann. Math. Artif. Intell.* **74**, 133–153 (2015)