Stepwise AI Interpretive Approach for Mutimodal Data Fusion

Bowen Long, Enjie Liu, Renxi Qiu and Yanqing Duan

July 10, 2024

Stepwise AI Interpretive approach for Mutimodal Data Fusion

Bowen Long[1], Enjie Liu[1] SM IEEE, Renxi Qiu[1], Yanqing Duan[2]

[1]School of Computer Science & Technology, [2]Business School,
University of Bedfordshire, Luton, UK
bowen.long@study.beds.ac.uk, {Enjie.liu, Renxi.qiu, Yanqing.duan} @beds.ac.uk

Abstract* — In recent years, Artificial Intelligence technology has excelled in various tasks and is taking the world by storm. However, the various transformations in neural networks make it difficult to make sense of the reasons why decisions are made. For this reason, trustworthy AI techniques have started gaining popularity. AI interpretability serves as an anchor point in the field of data fusion for multimodal AI, providing in-depth insights. The paper proposed a Stepwise AI Interpretative (SAII) approach using different pairing methods of 'one-to-one' and 'many-to-many' in an attempt to illustrate/demonstrate the interpretability of the process of pairing images and text. A counterfactual instantiation method was used to compare the whole-local relationship between a set of images and their associated descriptive text. The approach was evaluated via 'task performance'.

Keywords—Trustworthy AI, Multimodal, Data fusion.

I. INTRODUCTION

Decisions made by sophisticated AI models usually entail a myriad of parameters and nonlinear transformations that are too complex for humans to understand and trust [1]. This trust is especially essential, where the decision makers need to understand the implications of a decision. Promoting the trust for AI-based solutions can realize the ultimate goal of responsible AI [2].

The field around the theme of Trustworthy AI (XAI) is growing like never before, and the number of studies in the field of XAI is growing by leaps and bounds. Since 2019, technology research around robustness, interpretability, and privacy protection has continued to grow. Moreover, as the industry began to implement AI, the exploration and practice of making AI trustworthy during its industrialization continued to mature. The number of related AI technologies increased year by year after integrating 'human-centric' computing elements [3].

As a result, the search for the meaning of the results produced is no longer satisfied by results based on the superior performance of AI systems, and it becomes more important to understand the reasoning process behind the model's decisions [4]. When important decisions are entrusted to an AI system that is not fully controllable, there are more serious risks due to the lack of responsibility allocation and misuse [5]. In this regard, the current industry on XAI focuses on transparent AI or explainable AI.

By gaining insights into these perspectives, it is easy to see that AI brings with it a controversy between the recognition of scientific and technological contributions and productivity gains on the one hand, and human fears of dependence, lag, and uncontrollable undesirable consequences on the other. These controversies are almost always closely related to the technological risks and ethical challenges involved in AI [6], such as AI bias, the impact of AI systems on ethics and human rights, cybersecurity, and the crisis of unemployment brought about by AI.

This paper is organized as: Section II presents a taxonomy and insight and the relevant research into the development methods for XAI. Section III present the proposed method of achieving explainability with in a multimodal system with data and image. Section IV provides a case study of the proposed method. Section V shows evaluation of the method for achieving XAI, followed by concluding remarks and future work in Section VI.

II. BACKGROUND AND LITERATURE REVIEW

A. Principles of Trustworthy AI

Based on due foresight and responsibility, EU AI High-level expert group (HLEG) may have anticipated that the governance of AI would be far from satisfied by merely sketching out a vague concept of Trustworthy AI in a multitude of terms. As a result, they have given more meanings to Trustworthy AI, summarised in the Fig. 1: (1) the core elements it should have; (2) the ethical principles it should protect; and (3) the specific requirements it should fulfil. These three scales have been expanded at a deeper level to find a way to prevent AI from bringing about a crisis of trust [7].

As listed below, seven specific needs to meet the ethical principles in place by reflecting on the stakeholders' positions.

- Human agency and supervision: includes fundamental rights, human agency and human supervision.
- Technical robustness and security: includes attack resistance and security, backup plans and general security, accuracy, reliability and repeatability.
- Privacy and data management, including respect for privacy, data quality and integrity, and access to data.
- Transparency Includes traceability, interpret-ability and communication.
- Diversity, non-discrimination and equity, includes avoidance of unfair bias, accessibility and universal design, and stakeholder engagement.
- Social and environmental well-being, includes sustainability and environmental friendliness, social impact, society and democracy.

- Accountability includes audit-ability, negative impact minimization and reporting, trade-offs and remediation.

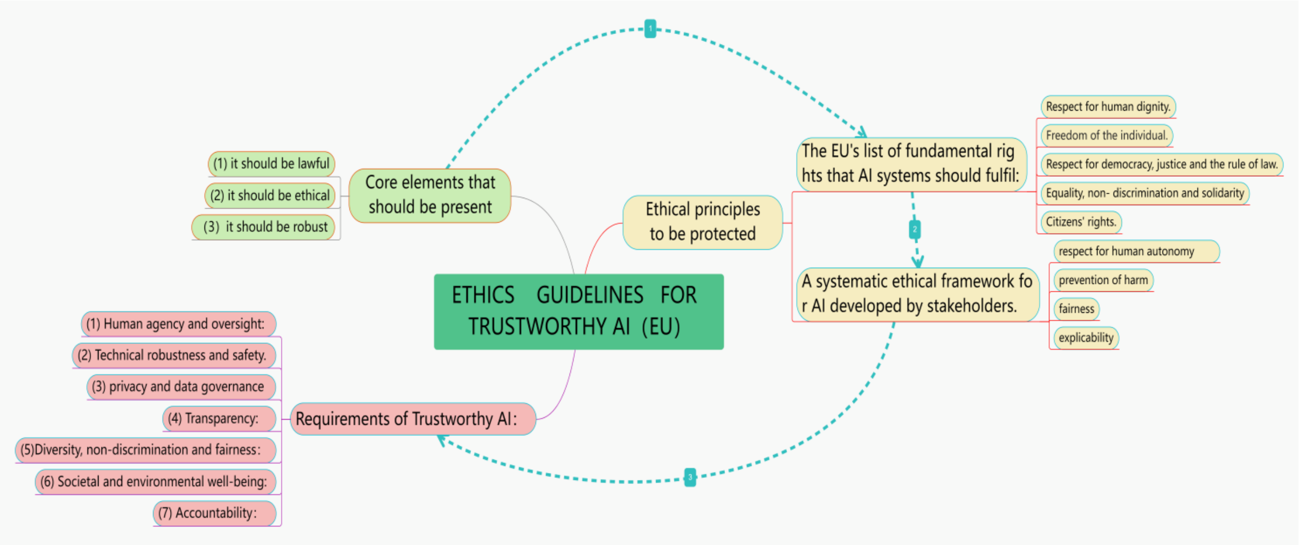

1
(1) it should be lawful
(2) it should be ethical
(3) it should be robust
Core elements that should be present
Ethical principles to be protected
The EU's list of fundamental rig hts that AI systems should fulfil:
Respect for human dignity.
Freedom of the individual.
Respect for democracy, justice and the rule of law.
Equality, non- discrimination and solidarity
Citizens' rights.
2
A systematic ethical framework fo r AI developed by stakeholders.
respect for human autonomy
prevention of harm
fairness
explicability
ETHICS GUIDELINES FOR TRUSTWORTHY AI (EU)
(1) Human agency and oversight:
(2) Technical robustness and safety.
(3) privacy and data governance
(4) Transparency:
(5)Diversity, non-discrimination and fairness:
(6) Societal and environmental well-being:
(7) Accountability:
Requirements of Trustworthy AI:
3

Figure 1. Summary of core elements in ethics guidance for Trustworthy AI

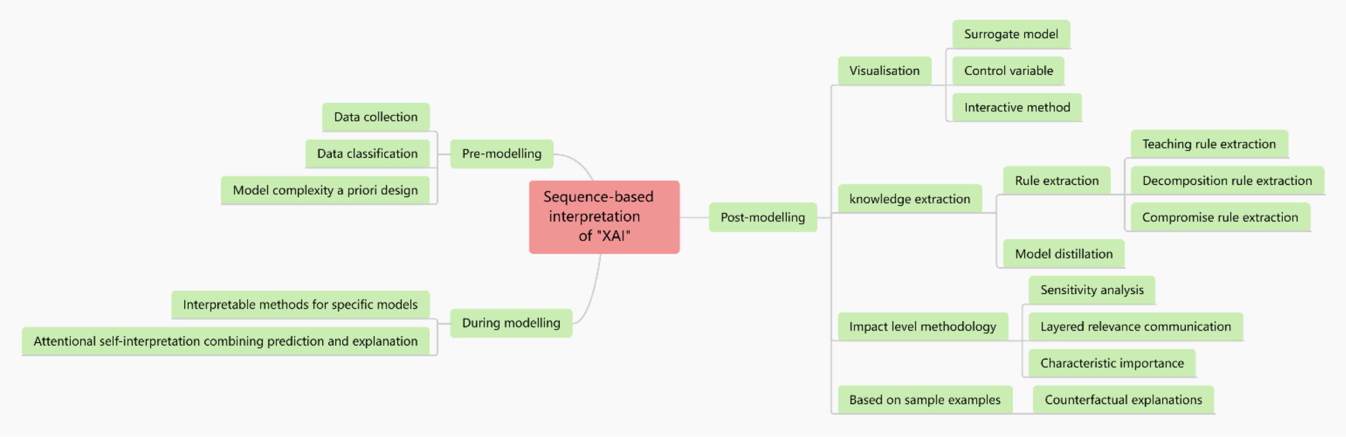

Data collection
Data classification
Model complexity a priori design
Pre-modelling
Sequence-based interpretation of "XAI"
Interpretable methods for specific models
Attentional self-interpretation combining prediction and explanation
During modelling
Post-modelling
Visualisation
Surrogate model
Control variable
Interactive method
knowledge extraction
Rule extraction
Teaching rule extraction
Decomposition rule extraction
Compromise rule extraction
Model distillation
Impact level methodology
Sensitivity analysis
Layered relevance communication
Characteristic importance
Based on sample examples
Counterfactual explanations

Figure 2. Interpretive methods in XAI

Explainability is a set of processes or methods that ensures the system is capable of allowing humans to comprehend its overall decision and reasoning. Through a review of the past literature, this study found that although many literatures have proposed a classification of interpretive methods, there is always a lack of uniformity in interpretive methods in XAI, this study adopts the following two scales:

- A sequence-based interpretive approach: pre-modelling, in modelling and post-modelling.

- Range-based interpretive approach: understanding the behaviour of the whole model - global interpretability; and understanding individual regional predictions - local interpretability.

B. *Sequence-based Explainable AI*

For the sequence-based interpretation approach, this study lists its specifics in Figure 2 and divides it into three phases: Pre-modelling, during modelling and Post-modelling.

C. *Past works in Sequence-based Explainable AI*

Pre-modelling: it occurs before the model development process, and this stage mainly involves data collection, data classification, and model a priori design of the model.

Many researchers use data augmentation for data complexity addition. In a series of work, such as in [9] and others, a new sample dataset was generated for training by performing a series of rotations, flipping, and adding noise in the image using self-supervised learning. Authors in [9] proposed the prototype network to measure data similarity. The approach has been widely adopted. Authors in [10] proposed a model called BRL based on Bayesian Rule Lists of Decision Trees, in the hope of gaining the trust of relevant practitioners through its simplicity and convincing ability.

During modelling: development of measures that inherently explain the model. This stage includes model-specific explanatory methods, prediction and explanation of attentional self - interpretation for joint modelling. Authors in [12] investigated gradients in convolution and inverse convolution to explore relationships in convolutional neural networks. Authors in [12] explored the VQA task of answering questions about images in free-form natural language, where the VQA model looks at the texts during this question and answer process.

Post-modelling: 'model-independent' or 'post-hoc interpretation' approaches, are currently a frequently cited classification approach in the field of XAI. Authors in [13] used a surrogate model approach to extract a decision tree representing the behaviour of the model. In [14], authors applied a virtual agent to improve level of trust. It also stated that multimodal explanations of the combination of vision and speech were more convincing. The authors in [15] used a quantitative method to rank feature importance when proposing a decision boundary-based model to explain data classification, calculating the increase in model prediction error to measure the importance of features.

D. Multimodal

With the advancement of data fusion technologies, data sources have become diverse. Examples include text, images, audio and video in the form of speech, sound and vision. These different modalities vary in size, representation, different predictive capabilities and contribution to the final task. Authors in [16] summarised previous multimodal tasks and concluded that the mixing of discourse and vision produced more competitive results comparing to the information produced by a single modal. Furthermore, the results of the study suggest that visual information plays a more important role in the results produced by the mixing of semantics and vision compared to textual information.

Since modalities are the form in which information is stored and represented, it is necessary to fuse the different modalities and then make predictions from the fused information. The most commonly used data fusion schemes are: early fusion, late fusion and intermediate (hybrid) fusion, which aim at fusing data, features, decisions and modalities in the intermediate hybrid layer.

In multimodal machine learning, data fusion is inevitably important for the generation of results, but it is also of enormous importance for a clear understanding of the link between inputs and results. Without understanding the connection between the two, there is no way to be confident that the results of a multimodal model are repeatable and trustworthy. For this reason, various stakeholders in the industry are actively working on the issue of trust in AI.

Pre- fusion

A multimodal data fusion method that integrates data prior to analysis. The fusion principle also relies on assumptions about the data preconditions: 1) the modalities are highly correlated and highly abstracted from each other; 2) each modality does not affect the original results even after processing. The data processing uses dimensionality reduction or cleaning to form feature vectors. With mutual fusion, the data are brought together through the activation mechanism of gate cells after using simple joins.

Post-fusion

Post-fusion uses a single modal source for fusion in the decision-making process. The approach is based on a fusion principle similar to human cognitive abilities, producing a common decision by integrating elements from different data sources and making judgments about their interactions. Data processing uses data dimensionality reduction or cleaning to form feature vectors. Data classification uses data lifting and repeated sampling based on the Bagging method. Data concatenation employs Bayes' rule, maximum fusion, and average fusion.

Hybrid fusion

The intermediate fusion is the fusion performed by the deep neural network module. It is also the most widely used method. During training, losses are passed back to the feature extraction network. After building the shared representation layer, fusion is performed using PCA and auto encoder. This fusion may be combined by slow or gradual fusion. It may lead to overfitting of the model as well as failure to learn the correlation of the respective modes.

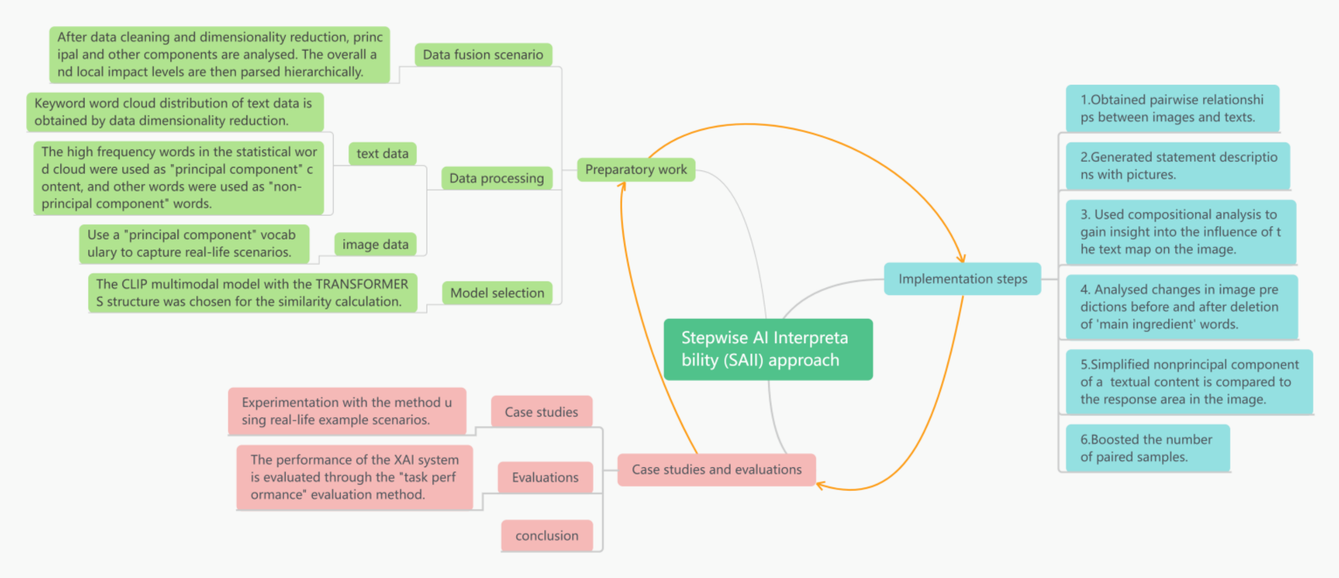

After data cleaning and dimensionality reduction, princ ipal and other components are analysed. The overall a nd local impact levels are then parsed hierarchically.
Data fusion scenario
Keyword word cloud distribution of text data is obtained by data dimensionality reduction.
The high frequency words in the statistical wor d cloud were used as "principal component" c ontent, and other words were used as "non-principal component" words.
text data
Data processing
Use a "principal component" vocab ulary to capture real-life scenarios.
image data
The CLIP multimodal model with the TRANSFORMER S structure was chosen for the similarity calculation.
Model selection
Preparatory work
Stepwise AI Interpreta bility (SAII) approach
Implementation steps
1.Obtained pairwise relationshi ps between images and texts.
2.Generated statement descriptio ns with pictures.
3. Used compositional analysis to gain insight into the influence of t he text map on the image.
4. Analysed changes in image pre dictions before and after deletion of 'main ingredient' words.
5.Simplified nonprincipal component of a textual content is compared to the response area in the image.
6.Boosted the number of paired samples.
Experimentation with the method u sing real-life example scenarios.
Case studies
The performance of the XAI system is evaluated through the "task perf ormance" evaluation method.
Evaluations
conclusion
Case studies and evaluations

Figure 3. The proposed Stepwise AI Interpretive approach

III. STEPWISE AI INTERPRETIVE APPROACH (SAII)

In this paper, we propose an interpretive approach to investigate the interpretability of the image-text pairing process using different pairing methods of 'one-to-one' and 'many-to-many'. A counterfactual instantiation method [17] was adopted to compare the global-local relationship between a set of images and their associated descriptive text. In the latter approach, a 'task-performance' method was used to assess the local effects [18]. Given that this interpretative method follows the progressive order of the user's insight events and scales from whole to local, the we refer to it as the Stepwise AI Interpretive Approach (SAII) and divide it into three parts, as shown in Figure 3:

- Preparatory work
- Implementation steps
- Case studies and evaluations

Preparatory work

Data fusion scenario: The significance of performing multimodal data fusion is to achieve better performance when performing in specific tasks. This study addresses the pre-fusion step for multimodal data. The impact of overall and local data on prediction is then analyzed hierarchically.

Data processing: The task is to obtain the keyword word cloud distribution of the text data through data dimensionality reduction and to use the high-frequency words in the word cloud as the 'principal component content' through statistical methods. Use the 'principal component' vocabulary to capture the real-life scene images.

Model selection: CLIP (Contrastive Language-Image Pre-Training) was developed by OpenAI to learn generic representations of multiple modalities [19]. Unlike traditional models, it is trained in a comparative learning manner on a large number of different datasets obtained from the Internet to understand images and associate images with relevant textual descriptions. Its visual coders typically use architectures such as Visual Transformers (ViTs) to process images, while text encoders are usually based on transformer architectures to process textual inputs. Cross-modal pairing is achieved by maximising the embedding similarity of the corresponding image-text pairs and minimising the cosine similarity of the non-matching pairs.

Implementation steps

1. Obtain pairwise relationships between images and texts: This study uses the term 'principal components' to investigate the relationship between each image itself and multiple texts. Meanwhile, during parallel processing, an image tensor is repeated several times to match the number of text inputs so as to keep the dimensionality of parallel processing consistent.

2. Generate statement descriptions and pair them with the pictures: The CLIP model takes input batches of pictures as well as text generated using a combination of 'principal component' words and 'non-principal component' words. The logarithms (i.e., scores) of the pictures and multiple texts are output before the activation function is applied. These logarithms are then converted to a probability density function using a Softmax function. The generated probability density function represents the model's understanding of the extent to which each piece of text is associated with a particular picture.

3. Use compositional analysis to gain insight into the influence of the text map on the image: After the model processing the image and the texts, a one-hot encoded tensor is created based on the maximum logit scores' indices. It can be used to determine the specific class or feature in the logits that should be focused on during the backward pass. In this step, for both the image and texts, it calculates gradients of the logits with respect to attention probabilities within the model using the one-hot vector. The gradients and attention probabilities are used to compute a Class Activation Map (CAM), which highlights the regions of the input (for both the image and the texts) that most relevant to the model's decision about the relationship mapping.

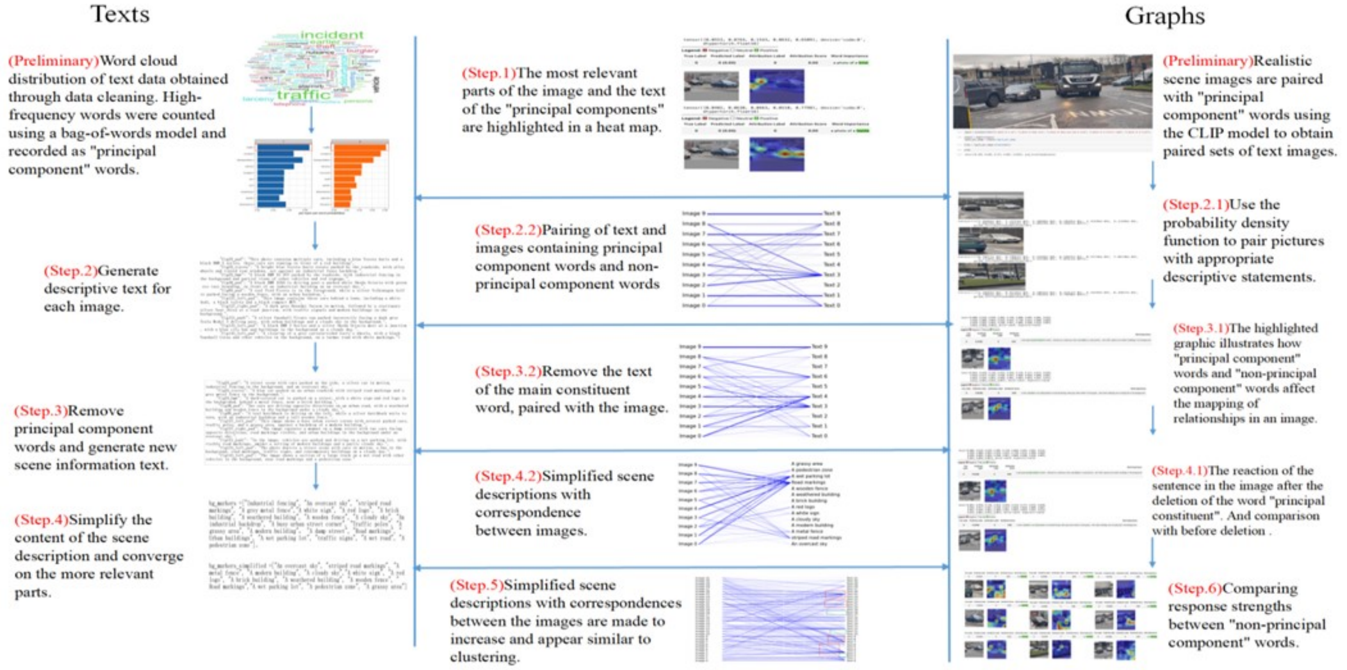

Texts
Graphs
(Preliminary)Word cloud distribution of text data obtained through data cleaning. High-frequency words were counted using a bag-of-words model and recorded as "principal component" words.
incident
traffic
(Step.2)Generate descriptive text for each image.
(Step.3)Remove principal component words and generate new scene information text.
(Step.4)Simplify the content of the scene description and converge on the more relevant parts.
(Step.1)The most relevant parts of the image and the text of the "principal components" are highlighted in a heat map.
(Step.2.2)Pairing of text and images containing principal component words and non-principal component words
(Step.3.2)Remove the text of the main constituent word, paired with the image.
(Step.4.2)Simplified scene descriptions with correspondence between images.
(Step.5)Simplified scene descriptions with correspondences between the images are made to increase and appear similar to clustering.
(Preliminary)Realistic scene images are paired with "principal component" words using the CLIP model to obtain paired sets of text images.
(Step.2.1)Use the probability density function to pair pictures with appropriate descriptive statements.
(Step.3.1)The highlighted graphic illustrates how "principal component" words and "non-principal component" words affect the mapping of relationships in an image.
(Step.4.1)The reaction of the sentence in the image after the deletion of the word "principal constituent". And comparison with before deletion .
(Step.6)Comparing response strengths between "non-principal component" words.

Figure 4. Processes of implementing the proposed approach in achieving XAI

4. Analyse changes in image predictions before and after deletion of 'principal component' words: In this step, it is assumed, by means of a counterfactual explanatory example, that the target can be identified equally well if the term 'non-principal component' is used. The 'principal component' words from the previous step are then removed and the responses to the ‘non-principal component’ words in the graphs are mapped using the same method as in the previous step.

5. Compare simplified ‘non-principal component’ textual content to the response area in the image: This step iteratively modifies the relevance map starting from the specified layer, which is usually the last layer. In our study, the specified layer was set to be the last layer. This is achieved by calculating the dot product of the current relevance map with the CAM, which effectively passes the relevance back to the layers of the network. This step is done using the respective attention probabilities of the image and text attention patches, respectively. Once this step is completed, the tensor representing the correlation of each region of the image or text is determined.

6. Boost the number of paired samples: The difference in match between the 'non-principal component' text and the image relevance is obtained by boosting the number of tasks, and the obtained tensor is used to represent the relevance of each region of the image to the text.

IV. CASE STUDY

In order to showcase the usability of the above methodology. The authors utilised a portion of the UK Police Department's open source police mobilisation records [20] as well as open source textual records of the types of police call out events in Montgomery County, USA [21]. This is then combined with real-time traffic control monitoring images from the UK Department of Transport [22] to construct a virtual traffic scenario. The approximate conceptualisation steps are shown in Figure 4. In a traffic system with many surveillance cameras on the road, there are many traffic photos and many descriptive texts for everyday events on the road, such as accidents or violations. Matching these photos to the descriptive texts helps to recognise the events occurring on the road.

The detailed implementation steps are as follows:

1. Use the CLIP model and the heat map visualisation tool, each image was paired with a phrase containing a 'principal component' word, highlighting areas of the image and each phrase to demonstrate the link between the image and the text of the phrase.

2. Generate an image description for each image using a combination of 'principal component' words and 'non-principal component' words. The term 'principal components' here refers to descriptions related to vehicle information, such as vehicle model and vehicle colour. The term 'non-principal component' refers to the scene context, such as background buildings or background sky and weather.

For each image and all descriptive texts, the CLIP model is used to generate probabilities between the image and each text. One text corresponds to the image itself and the other nine texts correspond to other traffic related images.

Utilise the CLIP model and visualisation, highlight the image regions and their descriptions generated in the previous step. The highlighted visualisation demonstrates how 'principal component' words and 'non-principal component' words affect the relational mapping of the CLIP model. If the highlighted portion of the textual description is primarily from the 'principal component' words, it indicates that the 'principal component' words are the most important to the CLIP model. Otherwise, it indicates that the term 'non-principal component' has an equally important impact.

3. Delete the 'principal component' words and use the 'non-principal component' words to generate a new image text description. The analysis was then repeated using the new text descriptions and the visualisations illustrated above. Comparing the two highlighted displays before and after deletion of the 'principal component' words to understand the relative importance between the 'principal component' words and the 'non-principal component' words.

4. Simplify the content of the 'non-principal component' from long sentences to phrases containing different contextual elements, such as municipal facility terms like 'road markings' and 'red logo'. The highlighted areas and phrases in the image are then visualised in conjunction with the CLIP model and heatmap.

5. Increase the number of image-text pairs containing 'non-principal component' to 32 pairs. Through the probabilities generated by the CLIP model, the relationships between them can be found.

V. EVALUATION

In order to evaluate and validate the validity and performance of the explanations made by the XAI system, different metrics have been implemented to achieve different explanatory goals. According to the evaluation method of 'task performance' proposed by in [18], evaluating the performance of the XAI system was carried out by adjusting and selecting elements was used for the evaluation.

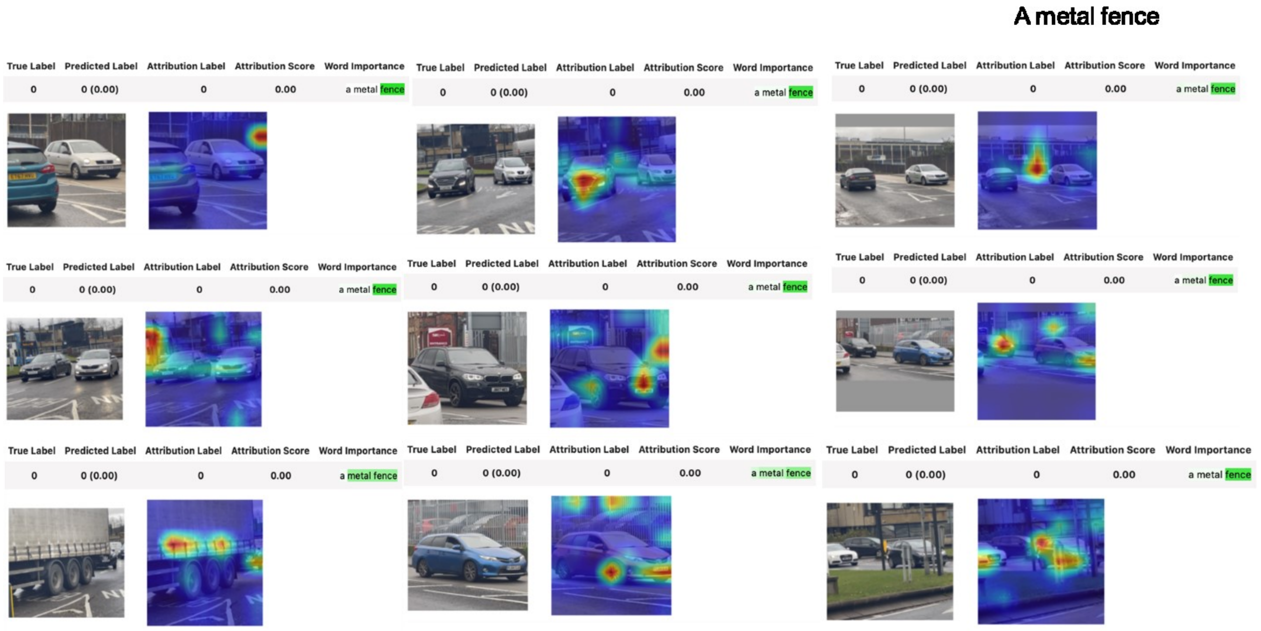
To facilitate the user's observation, we selected two phrases from Step 5 with high correlation: 'A metal fence', and low correlation: 'An overcast sky' for comparison.

As shown in Figure 5 and 6, the results clearly showed that, 'non-principal component' words that can produce strong correlations are recognised in the image with higher accuracy than 'non-principal component' words with weaker correlations.

VI. CONCLUSION AND FUTURE WORK

In this study, we designed and compared experiments using both realistic and counterfactual examples, benefiting from the CLIP model's ability to perform zero-shot learning. This approach can be rapidly deployed on small sample datasets, allowing for an interpretable study of model performance. A post-hoc interpretation method was chosen to explain the experimental results in the pre-fusion phase of the multimodal data in the AI model. Meanwhile, we evaluate the model at different scales in graph-text interaction tasks within vertical domains to gain insights into its performance.

The evaluation results show that through our interpretation approach, insights into the model's behavior can be obtained for a subset of the large dataset. In addition, some of the 'non-principal component' words affect image recognition in a manner similar to the 'principal component' words. There is also a significant difference in the effect of 'non-principal component' words on image recognition. This effect is not positively correlated with the number of features in the image but is perhaps linked to the neural units in the neural network.



True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
0	0 (0.00)	0	0.00	a metal fence

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
0	0 (0.00)	0	0.00	a metal fence

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
0	0 (0.00)	0	0.00	a metal fence

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
0	0 (0.00)	0	0.00	a metal fence

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
0	0 (0.00)	0	0.00	a metal fence

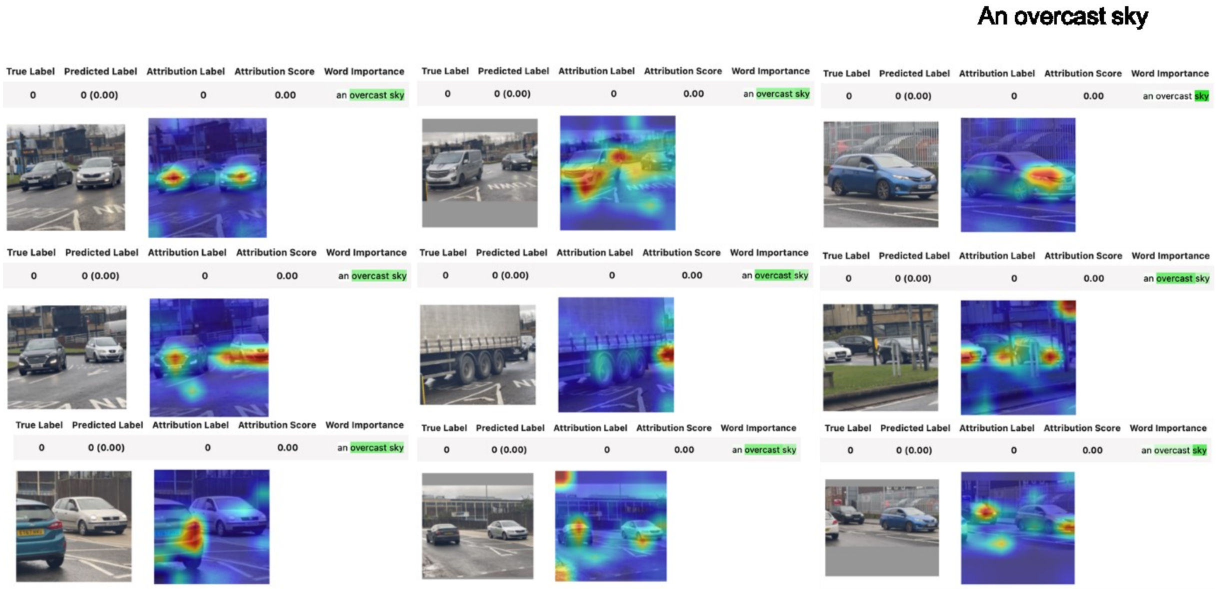
True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
0	0 (0.00)	0	0.00	a metal fence

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
0	0 (0.00)	0	0.00	a metal fence

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
0	0 (0.00)	0	0.00	a metal fence

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
0	0 (0.00)	0	0.00	a metal fence

Figure 5. Evaluation 1 – Pictures with metal fence



True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
0	0 (0.00)	0	0.00	an overcast sky

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
0	0 (0.00)	0	0.00	an overcast sky

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
0	0 (0.00)	0	0.00	an overcast sky

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
0	0 (0.00)	0	0.00	an overcast sky

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
0	0 (0.00)	0	0.00	an overcast sky

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
0	0 (0.00)	0	0.00	an overcast sky

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
0	0 (0.00)	0	0.00	an overcast sky

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
0	0 (0.00)	0	0.00	an overcast sky

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
0	0 (0.00)	0	0.00	an overcast sky

Figure 6. Evaluation 2 – Pictures with overcast sky

Only when a certain activation threshold is reached in the picture-text pairing can the two construct a link. We will continue our investigation in this direction.

In multimodal AI, data fusion aims to obtain better predictions when performing vertical tasks corresponding to the data. However, the dimensionality of the data in a given domain is complex, and the interaction between heterogeneous modal data is extensive. Therefore, The conclusions of this study need to be validated when dealing with large-scale data with higher dimensionality. In future research, we will focus on gaining insight into the local multimodal data embedding space to carry out interpretability studies, in an attempt to make the AI 'black box' more transparent.

ACKNOWLEDGEMENT

This project is funded by the Match Funding scheme from University of Bedfordshire.

REFERENCES

[1] Ali S., Abuhmed T., El-Sappagh S., Muhammad K., Alonso-Moral J.M., Confalonieri R., Guidotti R., Del Ser J., Díaz-Rodríguez N. and Herrera, F. (2023) 'Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence', Information

Fusion, 99, pp. 101805 Available at: 10.1016/j.inffus.2023.101805.

[2] Coeckelbergh M., ‘Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability’. Sci Eng Ethics 26, 2051–2068 (2020). https://doi.org/10.1007/s11948-019-00146-8

[3] Liu H., Wang Y, Fan W, Liu X, Li Y, Jain S, Liu Y, Jain A, and Tang J, ‘Trustworthy AI: A Computational Perspective’, ACM Trans. Intell. Syst. Technol. 14, 1, Article 4 (February 2023), 59 pages. https://doi.org/10.1145/3546872

[4] Kwon J.M., et al., ‘Artificial intelligence algorithm for predicting mortality of patients with acute heart failure’, PLoS One, vol. 14, no. 7, 2019.

[5] Bécue, A., Praça, I. & Gama, J., ‘Artificial intelligence, cyber-threats and Industry 4.0: challenges and opportunities’, Artif Intell Rev 54, 3849–3886 (2021). https://doi.org/10.1007/s10462-020-09942-2

[6] Du, S., & Xie, C., ‘Paradoxes of artificial intelligence in consumer markets: Ethical challenges and opportunities’, Journal of Business Research, 2021, 129, 961-974.

[7] A European approach to artificial intelligence (no date) Shaping Europe’s digital future. Available at: https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence (Accessed: 18 October 2023).

[8] Chen, Z., Fu, Y., Wang, Y.X., Ma, L., Liu, W. and Hebert, M., ‘Image deformation meta-networks for one-shot learning’, In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 8680-8689), 2019

[9] Snell, J., Swersky, K. and Zemel, R., ‘Prototypical networks for few-shot learning’, Advances in neural information processing systems, 30, 2017

[10] Letham B., Rudin C., McCormick T. H., and Madigan D., ‘Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model,’ Ann. Appl. Statist., vol. 9, no. 3, pp. 1350–1371, 2015.

[11] Simonyan K., Vedaldi A., and Zisserman A., ‘Deep inside convolutional networks: Visualising image classification models and saliency maps,’ 2013, arXiv: 1312.6034. [Online]. Available: http://arxiv.org/abs/1312.6034.

[12] Antol S., Agrawal A., Lu J., Mitchell M., Batra D., Zitnick C. L., and Parikh D., ‘VQA: Visual question answering,’ in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 2425–2433.

[13] Fong R. C. and Vedaldi A., ‘Interpretable explanations of black boxes by meaningful perturbation,’ in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 3429–3437.

[14] Weitz K., Schiller D., Schlagowski R., Huber T., and André E., ‘Let me explain!: Exploring the potential of virtual agents in explainable ai interaction design,’ J. Multimodal User Interfaces, 2020, pp. 1–12.

[15] Delaforge, A., Azé, J., Bringay, S., Mollevi, C., Sallaberry, A. and Servajean, M., 2022. Ebbe-text: Explaining neural networks by exploring text classification decision boundaries. IEEE Transactions on Visualization and Computer Graphics.

[16] Wu J., Mai S. and Hu H., ‘Interpretable Multimodal Capsule Fusion,’ in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 30, pp. 1815-1826, 2022, doi: 10.1109/TASLP.2022.3178236.

[17] Chou, Y., Moreira, C., Bruza, P., Ouyang, C. and Jorge, J. (2022) 'Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications', Information Fusion, 81, pp. 59-83 Available at: 10.1016/j.inffus.2021.11.003.

[18] Rawal A., McCoy J., Rawat D. B., Sadler B. M., and Amant R. S., ‘Recent Advances in Trustworthy Explainable Artificial Intelligence: Status, Challenges, and Perspectives,’ in IEEE Transactions on Artificial Intelligence, vol. 3, no. 6, pp. 852-866, Dec. 2022, doi: 10.1109/TAI.2021.3133846.

[19] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J. and Krueger, G., 2021, July. Learning transferable visual models from natural language supervision. In International conference on machine learning (pp. 8748-8763). PMLR.

[20] Welcome to data.police.uk (no date) Home. Available at: https://data.police.uk/ (Accessed: 11 May 2024).

[21] Montgomery County, M. Police Dispatched Incidents: Open Data Portal, Police Dispatched Incidents | Open Data Portal. Available at: https://data.montgomerycountymd.gov/Public-Safety/Police-Dispatched-Incidents/98cc-bc7d/about_data (Accessed: 11 May 2024).

[22] Live traffic cameras in London UK Live Traffic Cameras. Available at: https://uktraffic.live/england/london/ (Accessed: 11 May 2024).