EXTRACTION OF CAUSAL RELATIONSHIP USING THE BASE LINE LIGHTWEIGHTING MODEL AND CROSS ATTENTION

CHAEBYEOL LEE[1], SEUNGWOO WOO[2], JIHOON SEO[3]

[1]Kangnam University, ICT Engineering, Yongin-si, Republic of Korea

[2]Kangnam University, Artificial Intelligence Convergence Engineering, Yongin-si, Republic of Korea

[3]Kangnam University, Artificial Intelligence Convergence Engineering, Yongin-si, Republic of Korea

E-mail: [1]202104310@kangnam.ac.kr, [2]woow0708@kangnam.ac.kr, [3]jihoon@kangnam.ac.kr

ABSTRACT

Although research on causal inference has been actively conducted in recent years in the field of natural language processing, research on this has been insufficient in the field of image processing. Therefore, in this paper, we propose a new methodology to solve the problem of visual causal inference based on input images by utilizing the Vision Transformer (ViT) structure. After lightening the existing baseline model provided with the causal relationship-based inference dataset, the causal relationship between images is extracted using cross-attention. This method reduces the complexity of the model, improves the efficiency, and can effectively understand the complex relationships embedded in visual data.

Keywords: *Vision Transformer, Visual Causal Inference, Data Analysis*

1. INTRODUCTION

The Causal Inference is the process of understanding and inferring the impact of a specific event on another event. The Causal Inference in the field of Natural Language Processing (NLP) aims to predict future scenarios that may arise due to observed behavior. For example, if an input image represents a fire situation, then from the observed behavior "a fire starts" it can predict the future state "its surroundings become unsafe" or the future behavior "firefighters arrive". It is the core of the Causal Inference. It occurs, for example, in a variety of natural language generation tasks, such as news, story, and conversation generation [1].

However, the issue of the Causal Inference is a challenging task that can be solved by language models by understanding the concepts of context and causal relationships. In recent research, the language models with various structures, such as the BERT (Bidirectional Encoder Representations from Transformers) Series Models and EGCER (Effect Generation based on Causal Event Reasoning), are being applied to solve this issue [2].

The Causal Inference between images is the process of inferring causal relationships based on similarities and logical relationships between input images. For research into Strong AI Technology, the need for logical thinking models that go beyond simple classification and generation has increased.

By looking at an image, humans can immediately recognize objects and their properties and use their knowledge to answer questions and make inferences. However, if we close our eyes and imagine that we can build a scene only through touch, we can realize that reasoning without vision is not easy. It is because inference is entirely possible in humans, but it is independent of visual perception [3].

The inference based on the Causal Relationships is to train visual inference skills by analyzing relationships between input image data on a category basis. This visual reasoning ability requires the Model to be able to effectively extract global context and detailed features. Therefore, the overall performance of the Model is highly dependent on the context and feature extraction ability.

Therefore, this paper seeks to solve the issue of the Visual Causality Inference based on input images by utilizing the ViT (Vision Transformer) structure. Since ViT makes it easy to understand global relationships through the Self-Attention, by leveraging the strengths of the Model, it will solve the issue of the Visual Inference by setting event-based causal relationships as a knowledge background without any additional information. In other words, the purpose of this paper is to extract causal relationships using the Cross-Attention after lightweighting the Base Line Model provided with the causal relationship-based inference dataset.

2. RELATED WORK

2.1 CNN & ViT(Vision Transformer)

Sections CNN is mainly used in the field of computer vision and is a neural network structure for detecting local patterns and features in images. ViT is the model that applies the Transformer Structure, which has been successful in the field of natural language processing, to the field of computer vision. Previously, CNN models were used, but recently, global image information can be considered and processed by using the Transformer Structure.

CNN and ViT have the same goal of creating the Feature Map that well expresses image features, but there is a big difference in their basic structure and operating principles. Figure 1 compares the embedding methods of CNN and ViT models. CNN passes the image through the Kernel to learn by selecting local features and partial patches of the image and extracts features of the entire image, while ViT partitions the image into small patches and learns the correlation between each patch. By using the Self-Attention to extract the entire features of the image by considering the influence of all image patches on each other, all image patches can participate in learning and provide a high-level image representation [4].

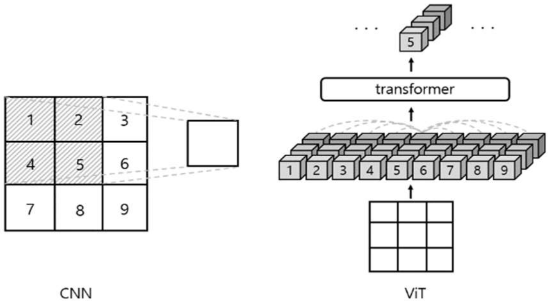

Figure 1: Embedding of CNN and ViT Model

Vanilla ViT (Vanila Vision Transformer) refers to the basic form of ViT used without any special modifications or additions. It can be seen as the cornerstone of the Transformer Structure utilized in the image processing field. It partitions the image into patches and adds the location information of the image patches through the Positional Encoding. The embedded patches pass through the encoder layer and are normalized. It ultimately inputs the CLS Token into the MLP Head and outputs what class it has [5].

2.2 Resolving Causal Inference issues in NLP

2.2.1 Causal-BERT

Recent research has focused on improving the NLP systems by introducing pre-trained models such as the BERT. It is a model that trains a general-purpose language understanding model using Wikipedia. It learns to mask certain words in the entire sentence and analyze the entire sentence in both directions to predict it. It allows to create models for a variety of tasks, such as question answering and verbal reasoning, by tuning just one additional output layer.

Despite the great success of pre-trained language models in the NLP systems, models trained with unsupervised learning methods have difficulty capturing causal relationships and do not show satisfactory results in the issue of the NLCI(Natural Language Causal Inference). Causal-BERT, proposed to solve this issue, is an effective model for various text-level causal inference. It solves the Causal Inference Issue by transfer learning BERT series models [6].

2.2.2 EGCER

EGCER is based on the Causal Event Inference to generate a result sentence for a given input sentence. It uses causal events to connect causal relationships between cause and effect sentences, allowing for graph-based understanding. It constructs the event-causality network in which semantically similar events are grouped together with sufficiently extracted causal relationship pairs so that cause-effect pairs can be eventized in each sentence pair. Research has shown that this performs better than the Causal-BERT, which is based on word-level causal inference [7].

2.2.3 CausalNet

CausalNet is a graph-based representation that has emerged to solve the problem that existing methods lack coverage for causal knowledge. After automatically collecting a network of causal terms extracted from a large web corpus, we present a new indicator that appropriately models the causal strength between terms based on this network, and provide a data-driven approach to aggregate common-sense causal reasoning between short texts (words, sentences). That is, through CausalNet, the degree to which the causal strength representing the causal relationship between the two words can be calculated, and through this, the causal relationship between the texts can be evaluated. There is a research result that if the data is sufficient, the causal resonance performance can be further improved [8].

2.3 Visual question Answering

There is the Neuro-Symbolic Model, the visual question-and-answer (VQA) model that derives answers through image recognition and question understanding. It is the Neural Symbolic VQA approach that separates inference from visual recognition and language understanding. It extracts the properties of objects from images and performs inference by splitting the question through preset conditions in the query [9]. However, the visual causality-based models were not found in existing studies. Therefore, in this paper, it proposes a robust visual causal inference model by applying existing models, and compare the inference ability of the proposed model through performance evaluation.

3. PROPOSED METHOD

3.1 Dataset

The causal relationship-based inference dataset consists of 240,000 image data and 60,000 labeled data collected and de-identified from raw data through Getty Images and crowdsourcing. The crowd workers classify 10 detailed categories that fit the image: Processing, Performance, Growth, Consumption, Pollution, Operation, Cutting, Arrangement, Extraction and Damage. Each category represents a method of presenting images that can be inferred through the cause image. It calculates the correct answer rate by showing one cause image and guessing the result image, and based on this, the difficulty level for each image is selected.

In the dataset, it creates several question folders for each category, and within the folders there are cause and answer images. As shown in Figure 2, the issue of the Causal Relationship Inference can be solved by selecting the result image that establishes a context with the cause image.

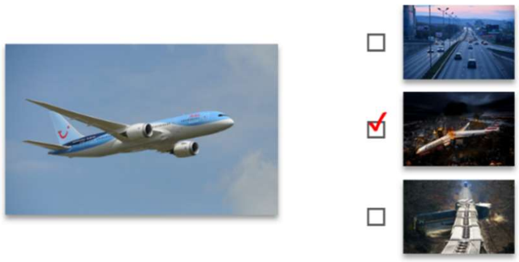

Figure 2: Inference Question Format based on Causal Relationship

3.2 Dataset

AI Hub provides several model examples with datasets. We chose to build on a ViT-based model that outperformed the CNN-based models despite its large size. The baseline model follows the structure of the vanilla ViT model with a few minor differences. First, the embedding size is very small, a quarter of the size of the vanilla ViT model, and it assigns CLS tokens to the embedded images and performs self-attention. The embedded CLS tokens, along with a new CLS token for outcome prediction, are then input to another transformer model for computation. Finally, the MLP head estimates the outcome based on the CLS token for prediction [10].

3.3 Approach

In this paper, we propose a method to reduce the model's complexity by using cross attentions to extract causal inference based on the existing baseline. As previously mentioned, the baseline has many parameters and FLOPs(FLoating Point Operations Per Second) to capture detailed features between multiple images and identify causal relationships. This approach can achieve high accuracy. However, images provide intuitive representations of clear situations, making the 'causal inference' task, which is the purpose of the model, simpler than the task in NLP. Therefore, a lighter model is appropriate.

3.3.1 Model Logic

Its goal is to infer causality between input images, identify causal images belonging to each category, and select one of the candidate answer images with the causal inference.

The proposed models are based on the Vision Transformer, and their common structure is shown in Figure 3. The input of the proposed model is several images consisting of the cause and candidate answer images. Because images are input to the Model as a batch, it has limitations with the commonly used image processing model structure alone. Therefore, based on the Base Line Model provided with the dataset, it is to propose a new model structure consisting of the Feature Extractor and Causal Inference Extractor.

The Feature Extractor identifies input images before inferring causal relationships between images. By accurately identifying and analyzing the features of an image, the Model can capture important information within the image and understand patterns or relationships based on this. It identifies the relationship between images through features extracted from the image causing the question and the candidate answer images.

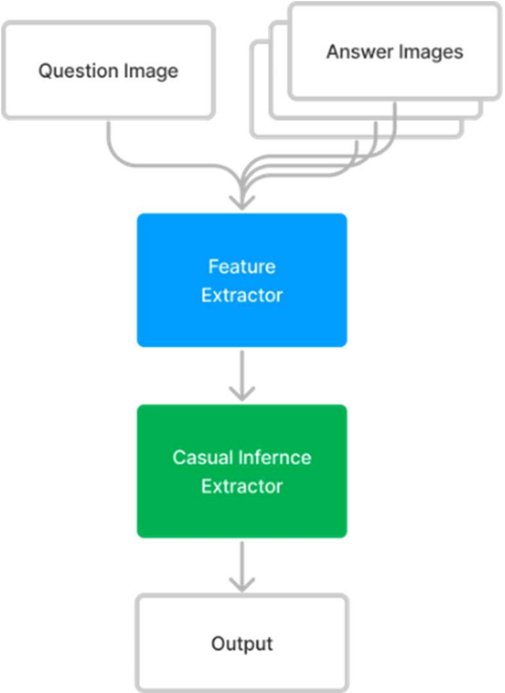

Figure 3: Common Structure of the Proposed Model

The Causal Inference Extractor shall determine the relationship between the causal and the candidate answer images based on the features extracted from the image. The Model analyzes similarities, differences, and interactions between images to identify the answer image that best matches the causal image. For example, there is a way to evaluate how elements such as the presence and location of objects in the image are connected to the context of the causal image. It allows the Model to generate answers to more complex visual questions, providing the understanding needed to identify subtle causal relationship between images. It identifies the causal relationship of all input images and then outputs the ID of the answer image with the largest causal relationship.

It develops the Model that is fairly lightweight and converges quickly by minimizing the accuracy drop in the existing Base Line Code. The loss value is measured through the Cross Entropy, and the F1 Score is used as the performance indicator. It also reduces unnecessary calculations in the Causal Extractor of the base line and at the same time focuses more on intensively identify causal relationships between images. To achieve this, it implements a module that simplifies the existing Multi-head Self-Attention and another module that utilizes cross-attention in the causal extraction process.

3.3.2 Model Formula; Feature Extractor

(a) Use of ViT

This Model does not aim to perform precise tasks such as 'Detection' and 'Segmentation' that utilize the detailed features of the image as much as possible. It also requires less detail than traditional image processing models such as 'Pattern', 'Outline', or 'Texture', and must learn flexibly from all inputs, it must operate with minimal Inductive Bias. Since, it is judged that it would be difficult to understand the context of the image because the CNN-based model has a relatively large Inductive Bias.

The Feature Extractor follows the ViT structure to extract features of the image. Similar to natural language processing models, the ability to focus on the overall context and key parts is important for this model task. In other words, the goal is to use the feature of the image to find other images with the highest correlation. It requires a certain level of feature extraction ability, but it is not a task that necessarily requires deep structural feature extraction ability. The advantage of the Transformer Structure, which originated in the field of natural language processing, is that it effectively extracts context information from sentences by focusing on key words. The ViT structure can understand the global context by dividing the image into patches and discovering the global relationship between each patch. It also allows focusing on important areas based on the relationships between patches. Focusing on local information can interfere with the causal inference because the amount of information about the image itself becomes excessive rather than the interrelationship between images. To solve this issue, the ViT Structure with low Inductive Bias is used.

(b) Lightweight Method

Unlike traditional baselines, the proposed model embeds via 2D convolution, which reduces the input image size and scales the patch size while reducing the embedding size. Although the size of the Base Line patch is very small, it allows the ViT model to focus more on detailed features, improving the model's feature extraction ability. However, as the number of Num Patches increases due to the small patch size, FLOPs. To solve the issue of FLOPs becoming too large and inference time becoming very long, the number of patches is reduced by increasing the patch size.

Because many detailed features may be lost in this process, the patch embedding method is changed to an embedding method through the Convolution. It is intended to minimize the loss of local feature information caused by increasing the patch size by including the local feature extraction ability of the Convolutional Layer in the embedding. It is expected to further improve the ability to extract features suitable for the Causal Inference by extracting and embedding local features with the embedding through the Convolution, and then utilizing the global context extraction ability of ViT.

The embedding size is also set larger than the model proposed by the Base Line. As the embedding increases, the ability to extract features from the image improves, but at the same time, the number of parameters also increases linearly. By increasing the patch size, the amount of information contained in one patch increases, so if it is selected to extract more detailed features, the embedding size must be maintained or increased. However, the embedding size is reduced because it is judged to be more appropriate to understand the overall context rather than the detailed features of the image when performing the task.

Due to the CLS token added to each image, representative information and global feature of the entire input image can be obtained. This information is very important for understanding the feature of the image, and the amount of computation can be greatly reduced by using only CLS token in subsequent calculations. Additionally, it is easy to identify causal relationships based on the global context of the image.

Additionally, it increases the size of the Base Line patch from 4 to 16, reducing the number of patches the model learned from 1,024 to 64. It requires a smaller patch because the size of the input image is 128x128, but the patch size of 4x4 is judged to be too small. Small patches focus more on local features of the image and can capture fine features such as texture or edges of objects that may be missed in large patches. On the other hand, larger patches focus on global features compared to small patches, allowing the overall context to be understood based on global features and relationships between images, so it is judged that there would be no problem using a large patch size.

Finally, it reduces the embedding dimensionality. The embedding and hidden dimension in the Base Line are defined as 256 and 512, respectively. As the number of patches decreases significantly, the embedding size must increase. Therefore, it decides that it may create a relatively smaller model by setting the embedding dimension to 128.

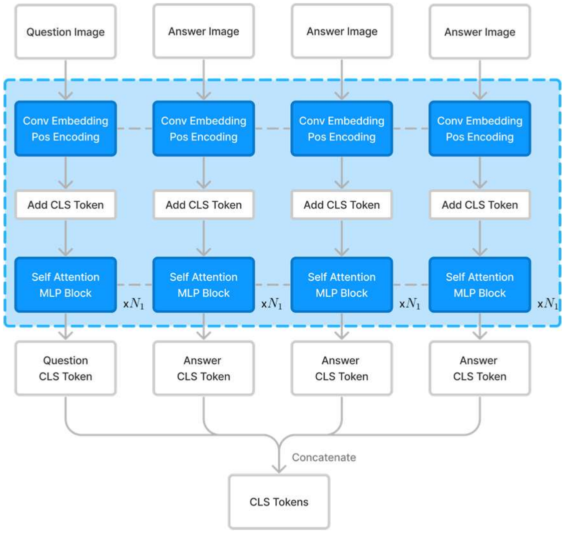

Figure 4: Feature Extractor

The images, namely $\mathrm{I_Q, I_{A1}, I_{A2}, I_{A3}}$ undergo feature extraction using the ViT block. This block utilizes convolutional embedding and self-attention mechanisms to extract features from the images. It is noteworthy that the trainable weights are shared for all images, enabling the model to learn similar features from different images and enhance its generalization ability.

After processing each image, the model generates feature maps for $\mathrm{F_Q, F_{A1}, F_{A2}, F_{A3}}$ as follows:

$$F_Q = FeatureExtractor(I_Q) \quad (1)$$

$$F_{A1} = FeatureExtractor(I_{A1}) \quad (2)$$

$$F_{A2} = FeatureExtractor(I_{A2}) \quad (3)$$

$$F_{A3} = FeatureExtractor(I_{A3}) \quad (4)$$

The feature map is used to extract CLS tokens that summarize the overall information of the images.

$$CLS_Q = ExtractCLS(F_Q) \quad (5)$$

$$CLS_{A1} = ExtractCLS(F_{A1}) \quad (6)$$

$$CLS_{A2} = ExtractCLS(F_{A2}) \quad (7)$$

$$CLS_{A3} = ExtractCLS(F_A 3) \quad (8)$$

These tokens are then concatenated and sent to the casual inference extractor.

$$CLS\ Tokens = Concat(CLS_Q, CLS_{An}) \quad (9)$$

3.3.3 Model Formula; Casual Inference Extractor

The most important thing in extracting the Causal Relationship is to identify the causal and logical relationships between causal images that are input to the Model. It goes beyond simply assessing the similarity of two data points and aims to determine how one event affects the other. Traditional similarity measurement methods such as the Cosine Similarity do not sufficiently capture the inherent complex causal relationships, so a module that can extract the Causal Relationship is needed. In this module, finding the relationship between the cause and the candidate answer images should be given priority. There is also a need to understand what relationships exist across inputs. By combining CLS tokens, it designs the Causal Inference Extractor inspired by the Self-Attention and the Cross-Attention Method used in the Transformer's Decoder Structure.

(a) Base Line Model using the Self-Attention

Figure 4 shows the structure of the Causal Inference Extractor using the Self-Attention. Because computational complexity may increase significantly if all inputs are used, the Model is constructed using only CLS tokens.

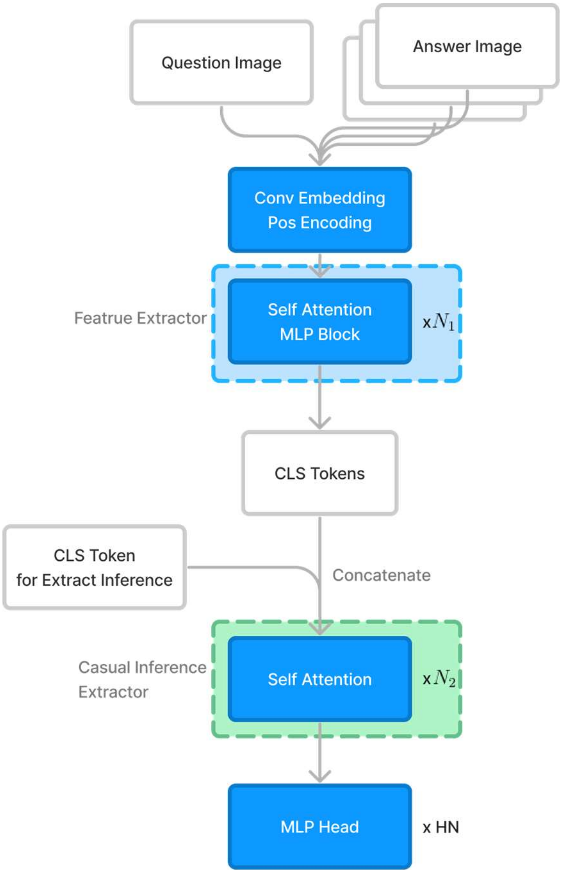

Figure 5: Model using the Self-Attention

It combines the CLS tokens of all input images into one, then adds a new CLS token and performs the Self-Attention. It allows the Q&A relationship and A&A relationship to be identified through the Attention between all CLS tokens. By combining CLS tokens and performing the Self-Attention, the relationship between the cause and the candidate answer images can be identified. After the self-attention operation is completed, only the newly added CLS token is extracted and passes through the FFN. At this time, only the core features for causal extraction can be extracted while reducing parameters and FLOPs. Since the relationship between the cause and the candidate answer images and the relationship between the candidate answer images can be known, meaningful information can be delivered to 'MLP_Head' before final output.

(b) Proposed Model using the Cross-Attention

The Self-Attention, the key element of the Transformer Model, calculates the values for 'Query', 'Key', 'Value' by multiplying different weight matrices to learn the relationship between input vectors.

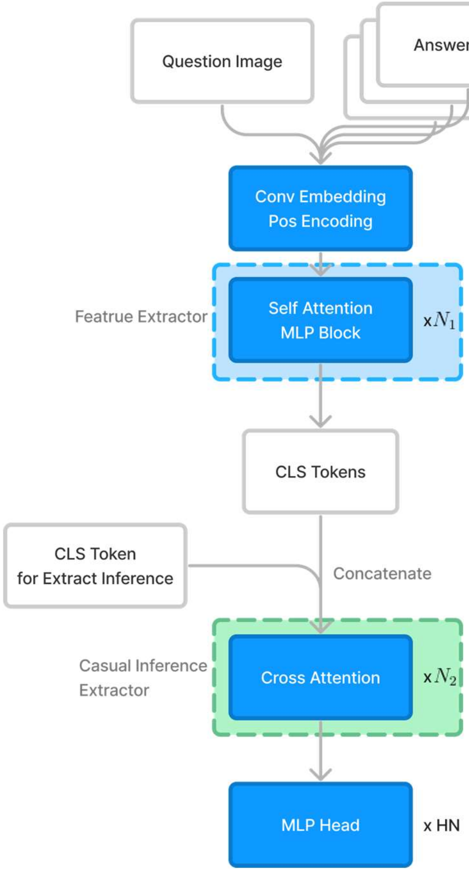

Figure 6: Model using the Cross-Attention

The Cross-Attention calculates the output value using 'Query' generated by the Model undergoing fine tuning and the values for 'Key' and 'Value' generated by the model trained in advance.

ʻQuery' is adjusted to be compatible with the values for 'Key' and 'Value' of the model learned in advance during training. It can be interpreted as referring to the knowledge of the model learned in advance.

$$SA(Q,K,V) = Softmax(\frac{QK^T}{\sqrt{d_K}})V \qquad (10)$$

$$CA(Q_f,K_{f_0},V_{f_0}) = Softmax(\frac{Q_f K_{f0}^T}{\sqrt{d_K}})V_{f_0} \qquad (11)$$

The Self-Attention is calculated from input vectors where $Q, K, and\ V$ are all the same. The Cross-Attention is calculated from K_{f_0}, V_{f_0} from the same input, but Q_f is calculated from different inputs [11].

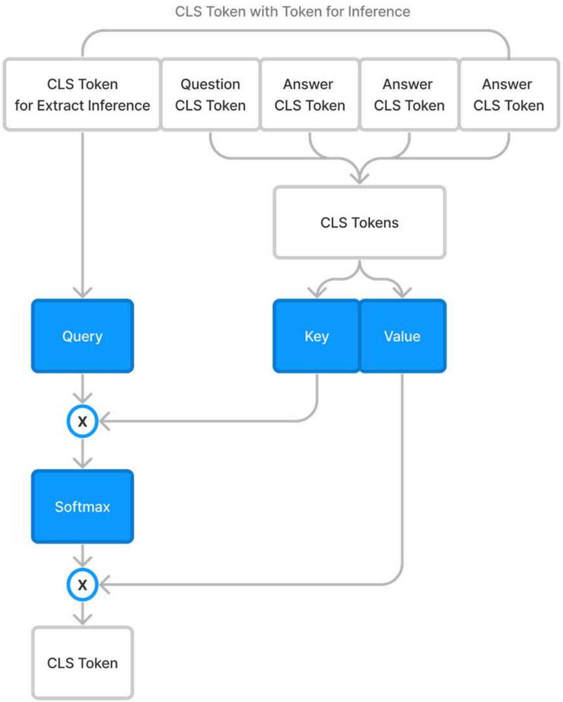

Figure 7: Calculation Method of the Cross-Attention

Therefore, it does not use attention to the data itself like the Self-Attention, but performs attention that focuses on the relationships between different data. In the proposed model, as can be seen in Figure 5, the Causal Inference Extractor is designed using the Cross-Attention. Figure 6 structurally shows the calculation method of the Cross-Attention. Use the new CLS token as a query, and use the Q&A CLS token as 'Key' and 'Value'. The 'Query' learns the relationship between the question and each answer image from 'Key' and 'Value', and learns to pay attention to answer images with a high causal relationship. It is to enable the CLS token to understand the relationship between the Q&A more deeply than the Self-Attention, and to extract answers that reveal the clearest causal relationship.

The approach of directly applying the Cross-Attention between existing tokens, rather than adding new CLS tokens in the Cross-Attention Mechanism, can reduce the number of parameters in the Model. However, while performing the Cross-Attention between existing tokens, the unique characteristic information of the tokens may be distorted, reducing the overall performance of the Model.

Although there may be a slight increase in parameters by adding new CLS tokens, the information of each token can be effectively integrated without distorting the characteristics of existing Q&A CLS tokens.

3.4 Differences from Prior Research

As confirmed in related studies, many studies have already been conducted on the problem of causal reasoning in the field of NLP. However, it is known that such research is insufficient in the field of image processing. In this paper, we tried to learn visual reasoning skills using input images. In addition, instead of using Self-Attention, a key element of the transformer model, we introduced Cross-Attention to improve causal reasoning. This method enables causal reasoning more efficiently by focusing on the relationships between different data. Finally, by applying the lightweight technique to the existing baseline model, the complexity of the model was reduced and the efficiency was improved. This was able to reduce the computational cost while maintaining the performance of the model.

4. RESULTS AND DISCUSSION

As can be seen from the BERT and Causal-BERT, the large-scale natural language processing models mentioned earlier, in order to infer subtle causal relationships between sentences or words, models with many parameters inevitably perform well. Accordingly, it is assumed that image processing, which requires more expression and calculation than text, would require large-scale models and parameters. However, images contain much more information than text, and experiments have proven that even a small number of parameters are sufficient to understand the context of the image and perform causal-based inference. The performance was measured using the F1 score as the evaluation metric using the validation data provided by the aforementioned dataset. The following parameters, FLOPs, and inference time were measured on an M3 PRO processor with 11 CPU cores and 14 GPU cores.

Table 1: Experimental Result of Feature Extractor – Cross Attention (CLS token added).

	FE – MHA (Base Line)	**FE – MHA (Light)**	**FE – CA (Ours)**
Parameters	5.349M	1.712M	1.713M
FLOPs	160,593,927,204	360,214,052	441,550,228
FLOPs	10^{11}	10^{8}	10^{8}
Inference Time	3.844sec	0.007sec	0.013sec
F1 Score	0.92	0.90	0.89
Embedding Size	256	128	128
Input Size	(224,224)	(128,128)	(128,128)

Table 1 shows the results of changing the model from the baseline, the lightweight model, and the cross-attention model with the addition of CLS tokens for inference. The lightweight model with the baseline is the Light model, and the causal inference module with the new CLS token to perform cross-attention and pass through FFN is the CCA module. The baseline has the highest number of parameters with 5.349M. The lightweight model has 1.712M, which is 3.637M fewer parameters than the baseline. The FLOPs are significantly reduced. While the baseline has about 10^{11} FLOPs, our models have about 10^{8} FLOPs. During the light weighting process, we made a number of changes that resulted in a significant reduction in the number of parameters. First, we reduced the size of the input image, which is very much related to the purpose of our model as mentioned earlier. By reducing the image size, the embedding size, and the patch size, we were able to reduce the inference time from 3.844 seconds to 0.007 seconds because the total number of patches was reduced. With this level of light weighting, there is a tradeoff to be made, as a reduction in accuracy is inevitable. Our model scored 0.02 points less despite the significant reduction in parameters, so our light weighting approach is successful. We also changed the causal inference part of the baseline to cross-attention. This is the module labeled CCA in the table above. After replacing it with our new module, we checked the results and found that the parameters were slightly higher than the lightweight model due to the addition of CLS tokens for inference. However, this allowed the model to converge faster, resulting in a 0.01 increase in accuracy over the lightweight model.

Table 2: Experimental Result of Feature Extractor – Cross Attention (CLS token not added).

	FE – MHA (Base Line)	**FE – MHA (Light)**	**FE – CCA (Ours)**
Parameters	5.349M	1.712M	1.810M
FLOPs	160,593,927,204	360,214,052	359,613,224
FLOPs	10^{11}	10^{8}	10^{8}
Inference Time	3.844sec	0.007sec	0.007sec
F1 Score	0.92	0.90	0.91
Embedding Size	256	128	128
Input Size	(224,224)	(128,128)	(128,128)

Table 2 shows the same causal inference module (CA) using cross-attention in the causal inference part, but without adding CLS tokens for inference. This shows that adding CLS tokens for inference to our cross-attention model was effective, as it reduced the parameters, but increased the FLOPs and inference time, and decreased the accuracy. There is another point to note from these results: the CA model used CLS tokens from the question image as queries. Since there are multiple layers of cross-attention instead of a single one, the integrity of the aforementioned CLS token was compromised as the model learned, which is why the score dropped.

Table 3: Experimental Result of Feature Extractor – Cosine Similarity.

	FE – MHA (Light)	FE - COS
Parameters	1.712M	1.447M
FLOPs	360,214,052	358,756,976
FLOPs	10^8	10^8
Inference Time	0.007sec	0.012sec
F1 Score	0.90	0.81
Input Size	(128,128)	(128,128)

Table 3 shows a causal inference module (COS) using Cosine Similarity. To ensure the identification of causal relationships, we equalized the feature extraction process and predicted the final output using cosine similarity, which measures the similarity between two pieces of data. The resulting F1 score was 0.81, indicating that the model was not extracting enough causal relationships between images. The score is high because the model's feature extraction measures the cosine similarity of the CLS tokens representing the question-and-answer candidates' features. This indicates that the feature extraction accurately extracts the necessary features for causal inference from the input question and answer candidate images. However, additional processing is required to predict causal relationships.

5. CONCLUSION

In this paper, it proposes the new methodology to lightweight the Base Line Model and extract causal relationships using the Cross-Attention to effectively solve the visual causal inference issue based on input images. In the feature extraction part, we reduced the input image size and increased the patch size to improve the baseline and reduce the computation. We also lightened the model by reducing the embedding size. We used the ViT structure to capture features related to the global context, and applied a light weighting method using convolutional embedding to ensure the weak feature extraction capability. In the causal relationship extraction part, we used a cross-attention mechanism with additional CLS tokens to preserve the extracted features of the cause and answer candidate images and accurately identify the relationship between the question and answer candidates.

In future follow-up studies, it will build a more sophisticated and comprehensive visual causal inference model by focusing on estimating hidden causal relationships between images based on the accelerated causal inference speed. In addition, it plans to expand the research to explore the practical applicability of the model by building a causal inference model under various input environments and confirm its performance and effectiveness in real environments.

REFERENCES:

[1] Gunhee Cho, et al. "Analyzing Natural Language Processing Model BART Suitable for Solving Causal Inference Problems" kiise, pp. 359-361, 2022.

[2] Zhongyang Li, et al. "CausalBERT: Injecting Causal Knowledge into Pre-trained Models with Minimal Supervision," arXiv preprint arXiv:2107.09852, 2021.

[3] Kexin Yi, et al. "Neural-Symbolic VQA: Disentangling Reasoning from Vision and Language Understanding" arXiv:1810.02338, 2018.

[4] Jin-gi Lee "Hello, Transformer" Acon Publishing, 2022.

[5] Alexey Dosovitskiy, et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale" arXiv:2010.11929, 2010.

[6] Zhongyang Li, et al. "CausalBERT: Injecting Causal Knowledge Into Pre-trained Models with Minimal Supervision" arXiv:2107.09852, 2021.

[7] Feiteng Mu, et al. "Effect Generation Based on Causal Reasoning" Findings of the Association for Computational Linguistics: EMNLP 2021, 2021.

[8] Zhiyi Luo, et al. "Commonsense Causal Reasoning between Short Texts" Proceedings, Fifteenth International Conference on Principles of Knowledge Representation and Reasoning, 2016.

[9] Hyungzun Mun and Sung-Bae Cho, "Causal Counterfactual Reasoning for Efficient Neuro-Symbolic Visual Question Answering" kiise, pp. 1241–1243, 2022.

[10] https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&dataSetSn=71286

[11] Seungwon Seo, et al. "A Novel Transfer Learning Method Using Cross-Attention" Academic Presentation Papers of the Korean Society for Information Science, 2023.