Dean&Francis
ISSN 2959-6157

Counterfactual Causal Attention Learning: Enhancing Fine-Grained Visual Recognition via Indirect Effect Optimization

Hangyu Peng

Undergraduate Student (Year 4)
Dalian Neusoft University of Information
Dalian, China
penghangyu80@gmail.com

Abstract:
Fine-grained visual recognition (FGVR) aims to distinguish subtle differences among visually similar categories. However, conventional attention mechanisms lack quantitative approaches to evaluate the quality of the learned attention during training, which limits their effectiveness. To address this limitation, we propose a novel Counterfactual Causal Attention Learning (CCAL) framework for fine-grained image classification and person re-identification. In our approach, the attention map is modeled as a confounding variable within a causal graph, and counterfactual interventions are employed to assess its impact on model predictions. By optimizing the indirect effect (IE), CCAL enhances the reliability of attention and improves overall recognition performance. Extensive experiments on multiple FGVR benchmarks demonstrate consistent improvements, including a 1.3% Top-1 accuracy gain on the CUB-200-2011 dataset.

Keywords:—fine-grained visual recognition, attention mechanism, counterfactual attention learning, causal inference, indirect Effect

I. Introdution

This work builds upon the experimental methodology proposed in [1]. Attention mechanisms play a vital role in human visual perception by enabling focus on relevant regions within complex scenes, which enhances recognition efficiency. This principle has been widely applied in computer vision, particularly for fine-grained visual recognition, where capturing subtle inter-class differences is crucial. Attention mechanisms help mitigate challenges caused by complex backgrounds, occlusions, and pose variations [17,18], and have become fundamental components of many state-of-the-art recognition models [2,3,4].

However, most current approaches rely heavily on weak supervision, focusing only on the final prediction without explicitly considering the causal relationship between attention and prediction. For example, in datasets containing birds or airplanes, attention models sometimes mistakenly focus on irrel-

evant features like the sky or background foliage, which can mislead the model's understanding of truly discriminative features [5,6,7]. Additionally, models that attend to only part of an object's attributes may suffer from limited generalization. As shown in [1], traditional attention learning methods are suboptimal and often fail to produce sufficiently discriminative attention maps, even for well-trained models, leading to occasional misclassifications [11]. This suggests that relying solely on final loss signals under weak supervision is insufficient to ensure meaningful and robust attention [12].

To address these issues, we propose a novel training framework based on counterfactual interventions within a causal inference setting. By treating the attention map as a confounder, we perform counterfactual analyses to quantify the effect of attention on the model's predictions. Through optimizing the indirect effect, our method encourages the model to focus on genuinely discriminative regions, reducing reliance on spurious cues. We implement various counterfactual intervention strategies and demonstrate consistent improvements on several fine-grained recognition benchmarks [1].

In summary, our proposed Counterfactual Causal Attention Learning (CCAL) approach significantly enhances classification accuracy in fine-grained visual tasks. By integrating multi-level counterfactual information, our model more effectively captures key visual features. On the CUB200 bird classification dataset, CCAL achieves 89.6% accuracy, surpassing the state-of-the-art API-NET [19] by 0.5%, and improving over our baseline by 1.3%, thereby validating the effectiveness of the counterfactual learning paradigm.

II. Relate work

A. Attention Mechanisms

In recent years, attention mechanisms have emerged as a crucial technique for enhancing representation in fine-grained visual recognition tasks, leading to significant performance improvements. Unlike traditional global feature learning methods, attention modules can adaptively identify highly discriminative local regions, effectively addressing the challenges posed by subtle inter-class variations. Sermanet et al. [8] pioneered the integration of visual attention into fine-grained image classification by introducing a recursive framework that guides models to focus on semantically salient regions, laying the foundation for spatially selective modelling. Building on this, Liu et al. employed reinforcement learning to dynamically enhance the relevance and responsiveness of attention extraction. Subsequently, approaches such as MA-CNN [9], MAMC [10], and WS-DAN [13] explored bottom-up attention designs, combining local cues with global context to automatically discover discriminative regions and achieve multi-scale feature collaboration, achieving state-of-the-art results on several benchmark datasets. Moreover, attention-based models have been widely adopted in cross-view recognition tasks—including person re-identification, vehicle retrieval, and video understanding—where they effectively mitigate misalignment and background noise, thereby improving the semantic coherence and discriminative power of visual features.

Despite their empirical successes, conventional attention mechanisms remain fundamentally data-driven and are susceptible to exploiting spurious correlations between input and attention representations, which can lead to incorrect attributions and diminished generalisation capabilities.

B. Causal Inference

Recently, causal inference has been proposed as a principled framework to tackle these challenges by distinguishing structural dependencies from non-causal statistical associations [14]. By constructing Structural Causal Models (SCM) and leveraging key tools such as interventions (do-operations) and counterfactual reasoning, causal analysis has demonstrated significant value across various visual tasks, including image classification, visual question answering, domain generalisation, and multimodal reasoning [15,16].

Notably, the Causal Attention Learning (CAL) framework [1] represents a major advance by modelling attention maps as latent confounders in a causal graph and quantifying their effect on predictions through structural interventions, thereby fostering more robust and interpretable attention learning. However, CAL is limited to observational-level causal analysis and lacks the capability to explicitly model or intervene upon the indirect pathways through which inputs influence outputs via the attention mechanism.

To overcome this limitation, we propose the Counterfactual Causal Attention Learning (CCAL) framework, which incorporates counterfactual interventions along the attention pathway. By substituting the learned attention with counterfactual samples and estimating the resultant changes in predictions, CCAL explicitly quantifies the indirect effect (IE) of inputs mediated by attention and minimizes this effect during training. This enables suppression of spurious pathways while amplifying genuine causal contributions. Compared to existing regularisation-based attention strategies such as entropy constraints, dropout, or normalisation, CCAL offers a theoretically principled

ISSN 2959-6157

and causally complete modelling paradigm. Experimental results on the CUB-200-2011 dataset demonstrate that our method not only improves classification accuracy but also produces attention distributions with greater semantic coherence and discriminative focus [1].

III. Approach

Our model takes as input a color image X with dimensions $H \times W \times C$, where H represents the image height, W the image width, and C the number of channels. We employ an RGB colour model with three channels. The model's output is the predicted image category Y, where $Y ?\mathrm{O}$, and O denotes the set of all fine-grained categories in the dataset. Formally, our model can be represented as $y = f(X)$, where $f(\cdot)$ denotes the proposed function.

A. Feature Extraction

This method extracts relevant features from the image using a convolutional neural network (CNN). Specifically, we design a three-layer CNN for feature extraction, which can be expressed as follows:

$$A_1 = CNN(X) \tag{1}$$

Here, A_1 denotes the feature map obtained through the convolutional neural network, where $A_1 \in \mathbb{R}^{H \times W \times C}$, with HHH representing the height of the image, W the width, and C the number of channels in the feature map.

Simultaneously, to obtain the corresponding features of the image, the feature map undergoes a pooling operation, which can be expressed by the following formula:

$$h_i = \phi(X * A_i) = \frac{1}{HW} \sum_{h=1}^{H} \sum_{w=1}^{W} X^{h_i w} A_i^{h_i w}, \tag{2}$$

i=$\{1,2,3,,,, C\}$ represents different channels, A_i denotes the feature map of the *i*-th channel, and *hi* represents the features of the input image *X* in the *i*-th channel. Ultimately, the extracted feature $\mathcal{F}_x$ for the input image *X* can be represented as:

$$\mathcal{F}_x = Norm(\| h_1, h_2, \ldots, h_c), \tag{3}$$

$\|$ denotes the concatenation in the feature dimension, and Norm represents the normalization of the features.

B. Counterfactual Intervention

First, we construct the prediction process as a causal graph $G = (V, E)$, where V represents the nodes and E represents the edges of the causal graph. In this model, there are three nodes: the input image X, the extracted feature map A, and the final output Y. The prediction process of the model can be represented as shown in the figure below. The direction of the arrows indicates the order of inference in the model, and the red crosses represent where we intervene to break the path from X to A, simultaneously constructing a new feature map $\bar{A}$ through counterfactual intervention.

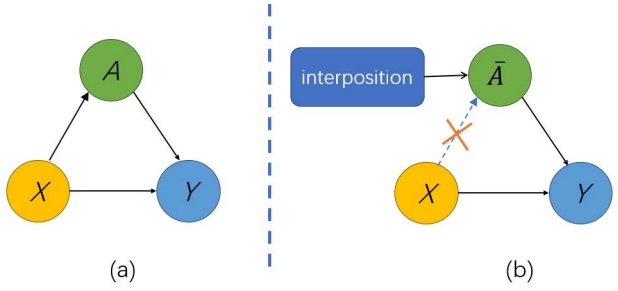

Figure 1: Causal learning schematic. (a) represents the conventional model structure. (b) represents the model structure with counterfactual intervention applied

As shown in Figure 1(a), for the prediction result Y, we consider both the input X and the extracted feature map A as factors influencing the outcome, where the feature map A is also affected by the input X. To explore the causal relationship between the input X and output Y, we treat A as a confounding variable.

The outcome without intervention is denoted as $Y(A = A, X = X)$, while the feature map after counterfactual intervention is denoted as $do(A = A)$. The counterfactual prediction result is expressed as $Y(do(A = \bar{A}), X = X)$. This formulation enables us to investigate the causal effect between X and Y.

Inspired by causal learning, we uncover the causal relationship by minimizing the indirect effect (IE) between X and Y, which can be formulated as:

$$Y_{IE} = \mathbb{E}_{\bar{A}\gamma}\left[Y(A = A, X = X) - \left(do\left(A = \bar{A} \right), X = X \right) \right] \tag{4}$$

where $\mathbb{E}_{\bar{A}\gamma}$ denotes the expectation over the entire training set. In practice, this expectation is minimized by optimizing the model's loss across the full dataset.

C. Loss Function

Based on the above analysis, the loss function of our model consists of two parts. The first part is the classification loss after counterfactual intervention, denoted as $\mathcal{L}_{CE}$ and the second part is the loss from other subtasks, denoted as $\mathcal{L}_{other}$. The overall loss function of the model can be expressed as:

$$\mathcal{L}_{Total} = \alpha \mathcal{L}_{CE} + \beta \mathcal{L}_{other}, \quad (5)$$

where α and β are hyperparameters used to balance these two loss components.

IV. Experiments

A. Implementation details

We adopt fine-grained image recognition as the classification task and conduct experiments on the CUB200-2011 dataset. In equation (2), we set $H = 448$, $W = 448$, and $C = 16$. Since there are no other subtasks, the hyperparameter β is set to 0. For the classification task, we use the cross-entropy loss function to optimize the model. Experiments are performed on an NVIDIA® GeForce RTX™ 4090 GPU. The learning rate is set to 0.001, with 160 training epochs, and the hidden dimension is set to 65,536.

B. Result

We evaluated the effectiveness of the proposed Counterfactual Causal Attention Learning (CCAL) method on the fine-grained bird visual recognition task. Using traditional spatial attention as the baseline, we compared our method against this baseline as well as the CAL approach. The experimental setup, implementation details, and results for this task are described as follows.

table 1: Comparison of top-1 classification accuracy (%) with state-of-the-art fine-grained image classification methods on the CUB200-2011 dataset

Method	CUB
RA-CNN[20]	86.1
MAMC[9]	86.7
DCL[21]	87.6
API-NET[19]	90.1
Baseline	89.3
Baseline+CAL	90.6
Baseline+CCAL	89.6

table 2: Quantitative analysis of attention. We compare the classification accuracy (%) of our method with three other attention regularization strategies, including attention dropout, entropy regularization, and attention normalization, and evaluate the quality of the learned attention maps via mIoU (%) using the ground-truth bounding boxes on the CUB dataset.

Method	CUB	mIoU
Baseline	89.3	54.2
Baseline+CAL	90.6	67.4
Baseline+CCAL	89.6	68.1

C. Analysis

We analyzed the impact and sensitivity of the main parameters mentioned above and obtained the following results. Parameter analysis experiments were conducted on the fine-grained visual recognition task using the CUB-200-2011 dataset. We compared the performance of the CAL method with several state-of-the-art fine-grained image classification methods. The experimental results show that CAL achieved a Top-1 classification accuracy of 90.6% on the CUB dataset, outperforming existing methods such as RA-CNN (86.1%), MAMC (86.7%), DCL (87.6%), and API-NET (90.1%). The baseline model achieved an accuracy of 89.3%, which was further improved to 90.6% by incorporating CAL, indicating that CAL significantly enhances the model's capability in fine-grained image classification tasks. In comparison, although the CCAL method also improved performance, its accuracy reached only 89.6%, slightly lower than CAL, suggesting that CAL is more effective in boosting classification accuracy.

D. Visualizatio

ISSN 2959-6157

In this section, we conduct an in-depth analysis of our trained CCAL model's decision-making process on the CUB200-2011 dataset for the "Yellow-headed Blackbird" class using model attention heatmaps, such as Grad-CAM. We specifically focus on whether the model effectively captures the head features of the Yellow-headed Blackbird.

Figure 2 The heatmap results shown above indicate that the model has effectively captured the head features of the Yellow-headed Blackbird

E. Conclusion

This study aims to address the limitations of conventional attention mechanisms in fine-grained visual recognition (FGVR), specifically the insufficient evaluation of attention quality and weak supervisory signals. Inspired by previous work, we propose and implement a counterfactual causal attention learning method designed to enhance the model's focus on discriminative regions while mitigating reliance on spurious features.

By modeling the attention map (A) as a confounding variable in the causal path from input to prediction (X → Y) and employing counterfactual interventions, our framework effectively evaluates attention quality and generates robust supervisory signals through optimization of the indirect effect (IE). We explore various intervention strategies and attention methods to improve this approach.

Experimental results on multiple FGVR tasks, including fine-grained bird classification (CUB-200-2011), demonstrate the effectiveness of our method. Compared to traditional baselines, our method significantly improves recognition accuracy. Although its classification accuracy is slightly lower than some state-of-the-art methods (e.g., CAL), our Counterfactual Causal Attention Learning (CCAL) achieves superior mIoU performance (68.1% vs. 67.4%), indicating its effectiveness in generating more semantically coherent and discriminative attention distributions critical for model interpretability.

Essentially, this work successfully integrates causal inference into fine-grained visual attention learning, providing a robust solution to enhance the interpretability and reliability of model focus regions. Future work may extend this approach to other challenging visual tasks.

Acknowledgment

I sincerely thank my parents for supporting me with the necessary resources during this research. I also appreciate Dalian Neusoft University of Information for providing the inspiration that motivated this project.

References

[1] Y. Rao, G. Chen, J. Lu, and J. Zhou, "Counterfactual attention learning for fine-grained visual categorization and re-identification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Virtual, 2021.

[2] M. H. Guo, T. X. Xu, J. J. Liu, Z. N. Liu, P. T. Jiang, T. M. Mu, S. M. Martin, M. X. Ma, H. C. Zhang, Y. Q. Cui, Y. R. Xu, Z. P. Zhou, S. S. Zhou, R. B. Liang, B. F. Ding, J. H. Li, and M. M. Cheng, "Attention mechanisms in computer vision: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 3, pp. 1489-1513, 2022

[3] M. H. Guo, C. Z. Lu, Z. N. Liu, M. M. Cheng, and S. M. Martin, "Visual attention network," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, pp. 7832-7839, 2022.K. Elissa, "Title of paper if known," unpublished.

[4] M. H. Guo *et al.*, "Attention mechanisms in computer vision: A survey," *J. Adv. Res.*, vol. 38, pp. 215-249, 2022, doi: 10.1016/j.jare.2021.11.006Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[5] D. Zhang, H. Zhang, J. Tang et al., "Causal intervention for weakly-supervised semantic segmentation," in Adv. Neural Inf. Process. Syst., vol. 33, 2020, pp. 2225-2236.

[6] H. Dong, F. Han, L. Si, W. Qiang, and L. Zhang, "Background Debiased SAR Target Recognition via Causal Interventional Regularizer," arXiv preprint arXiv:2308.06606, 2023.

[7] P. T. Jiang, L. H. Han, Q. Hou, M. M. Cheng et al., "Online attention accumulation for weakly supervised semantic segmentation," IEEE Trans. Image Process., vol. 32, pp. 586-599, 2022

[8] B. Zhang, J. Xiao, Y. Wei, M. Sun, K. Huang, "Reliability does matter: An end-to-end weakly supervised semantic segmentation approach," in Proc. AAAI Conf. Artif. Intell., 2020, vol. 34, no. 7, pp. 12711–12718.

[9] Y. Zhang, Z. Zhou, Y. Cao, G. Li, S. Wu, and X. Zhang, "MAMC-Optimal on Accuracy and Efficiency for Automatic Modulation Classification with Extended Signal Length," 2024 International Conference on Machine Learning and Cybernetics (ICMLC), vol. 1, pp. 272–276, 2024, doi: 10.1109/ICMLC61633.2024.10705364.

[10] Z. Yang, Z. Wang, L. Luo, H. Gan, and T. Zhang, "SWS-DAN: Subtler WS-DAN for fine-grained image classification," Journal of Visual Communication and Image Representation, vol. 79, p. 103233, Aug. 2021.

[11] J. Ruan, G. Liang, J. Zhao, H. Zhao, J. Qiu, et al., "Deep learning for cybersecurity in smart grids: Review and perspectives," IET Smart Grid, vol. 5, no. 1, pp. 2–16, 2022.

[12] B. Zhang, J. Xiao, Y. Wei, M. Sun, K. Huang, "Reliability does matter: An end-to-end weakly supervised semantic segmentation approach," in Proc. AAAI Conf. Artif. Intell., 2020, vol. 34, no. 7, pp. 12711–12718.

[13] P. Sermanet, A. Frome, E. Real, "Attention for fine-grained categorization," arXiv preprint arXiv:1412.7054, 2014.1

[14] A. Komanduri, X. Wu, Y. Wu, and F. Chen, "From identifiable causal representations to controllable counterfactual generation: A survey on causal generative modeling," *arXiv preprint arXiv:2310.11011*, 2023.

[15] L. De Lara, A. González-Sanz, N. Asher, and L. Risser, "Transport-based counterfactual models," *J. Mach. Learn. Res.*, vol. 25, pp. 1-62, 2024.

[16] H. Li, X. Wang, Z. Zhang, and W. Zhu, "Out-of-distribution generalization on graphs: A survey," *arXiv preprint arXiv:2202.07987*, 2022.

[17] D. Wu, M. Ye, G. Lin, and X. Gao, "Person re-identification by context-aware part attention and multi-head collaborative learning," *IEEE Trans. Image Process.*, vol. 30, pp. 4843–4856, 2021.

[18] M. Liu, J. Zhao, Y. Zhou, H. Zhu, and R. Yao, "Survey for person re-identification based on coarse-to-fine feature learning," *Multimed. Tools Appl.*, vol. 81, no. 12, pp. 17099–17124, 2022.

[19] Peiqin Zhuang, Yali Wang, and Yu Qiao. Learning attentive pairwise interaction for fine-grained classification. arXiv preprint arXiv:2002.10191, 2020

[20] Jianlong Fu, Heliang Zheng, and Tao Mei. Look closer to see better: Recurrent attention convolutional neural network for f ine-grained image recognition. In CVPR, 2017.

[21] Yue Chen, Yalong Bai, Wei Zhang, and Tao Mei. Destruction and construction learning for fine-grained image recognition. In CVPR, pages 5157–5166, 2019.