Reproducibility Report on *Concept-Attention Whitening for Interpretable Skin Lesion Diagnosis*

Junlin Hou[1], Jilan Xu[2], and Hao Chen[1,3]
[1] The Hong Kong University of Science and Technology, Hong Kong, China
[2] Fudan University, Shanghai, China
[3] HKUST Shenzhen-Hong Kong Collaborative Innovation Research Institute, Shenzhen, China
csejlhou@ust.hk[1], jhc@cse.ust.hk[3]

February 20, 2025

Abstract

This report presents an attempt to reproduce the results of the paper *Concept-Attention Whitening for Interpretable Skin Lesion Diagnosis* [3]. The study implements the Concept-Attention Whitening (CAW) framework for interpretability in deep learning-based skin lesion classification. Our reproduction includes dataset preprocessing, model implementation, training, and evaluation on the Derm7pt and SkinCon datasets. While we successfully implemented CAW, we observed minor deviations in performance metrics compared to the original paper. This report outlines our methodology, challenges encountered, and suggestions for future improvements. Code available at GitHub Repository.

1 Introduction

Automated skin lesion diagnosis is a crucial application of deep learning in medical imaging. The paper introduces CAW, a whitening transformation technique [2] that replaces Batch Normalization (BN) in deep neural networks to enhance feature decorrelation and interpretability. The study evaluates CAW on dermoscopic datasets (Derm7pt, SkinCon) [1, 6] and claims improvements in classification accuracy and robustness.

This reproducibility study aims to:

- Preprocess the Derm7pt and SkinCon datasets [5].
- Implement CAW using ResNet-based architectures.
- Train and evaluate models on both datasets.
- Compare reproduced results with those from the original paper.

2 Methodology

The CAW framework consists of:

- **Disease Diagnosis Branch:** Uses CAW layers in CNNs for classification.
- **Concept Alignment Branch:** Ensures features align with predefined clinical concepts.

2.1 Mathematical Formulation

The Concept-Attention Whitening (CAW) framework consists of two key transformations applied to feature representations:

- **Whitening Transformation** – Removes correlations between feature dimensions.
- **Orthogonal Transformation** – Aligns the transformed features with predefined medical concepts.

2.1.1 Whitening Transformation

Given an input feature map $Z \in \mathbb{R}^{b \times d \times h \times w}$, where:

b is the batch size, d is the feature dimension, h, w are spatial dimensions.

We reshape Z into $d \times n$, where $n = b \times h \times w$. The whitening transformation is then applied as:

$$\psi(Z) = W(Z - \mu 1_{1 \times n}) \tag{1}$$

where:

- μ is the mean feature value over n samples,
- $W \in \mathbb{R}^{d \times d}$ is the whitening matrix.

The whitening matrix W is computed using the ZCA [4] algorithm:

$$W = U\Lambda^{-\frac{1}{2}}U^T \tag{2}$$

where:

- U is the eigenvector matrix of the feature covariance,
- Λ contains the corresponding eigenvalues.

This transformation removes feature correlations and standardizes feature distributions.

2.1.2 Orthogonal Transformation

After whitening, an orthogonal transformation is applied to align features with predefined clinical concepts:

$$Z' = Q^T\psi(Z) \tag{3}$$

where:

- $Q \in \mathbb{R}^{d \times d}$ is an **orthogonal matrix** containing concept-aligned feature vectors.
- Each column q_k of Q represents a concept c_k.

2.1.3 Optimization of Q via Concept Alignment

To estimate Q, the model leverages weakly-supervised learning. Given a concept dataset $X_c = \{X_{c_k}\}_{k=1}^{K}$, where each X_{c_k} consists of images containing concept c_k, we optimize:

$$\max_{q_1,q_2,\ldots,q_K} \sum_{k=1}^{K} \frac{1}{|X_{c_k}|} \sum_{x_{c_k} \in X_{c_k}} q_k^T \text{AvgPool}(\tilde{M}_{c_k} \odot \psi(f(x_{c_k}))) \tag{4}$$

subject to:

$$Q^T Q = I_d \tag{5}$$

where:

- $\tilde{M}_{c_k}$ is the binary concept mask obtained from the concept activation map M_{c_k}.
- $\odot$ denotes element-wise multiplication.
- **AvgPool** performs global spatial pooling over the feature map.

2.1.4 Optimization of Q using Cayley Transform

Since direct optimization of Q is computationally intractable, we update it iteratively using the Cayley transform:

$$Q^{(t+1)} = (I + \frac{\eta}{2}A)^{-1}(I - \frac{\eta}{2}A)Q^{(t)} \tag{6}$$

where:

- η is the learning rate,
- $A = G(Q^T) - QG^T$ is a skew-symmetric matrix,
- G is the **gradient of the loss function**.

This ensures that Q remains an orthogonal matrix during training.

2.1.5 Final Disease Prediction Loss

The final disease classification loss is computed as:

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} L_{\text{ce}}(g(Q^T \psi(f(x_i))), y_i) \tag{7}$$

where:

- $f(x)$ represents extracted features,
- $g(\cdot)$ is the classifier,
- L_{ce} is the cross-entropy loss for skin disease classification.

Pseudocode for CAW

Require: Disease dataset $D = \{(x_i, y_i)\}_{i=1}^{N}$, Concept dataset $X_c = \{X_{c_k}\}_{k=1}^{K}$
Require: Pretrained ResNet model, learning rate η, whitening matrix W, orthogonal matrix Q
Initialize: Load pretrained ResNet, replace Batch Normalization (BN) with CAW layer
Train Disease Diagnosis Branch:
for each epoch **do**
  for each mini-batch $(x, y) \in D$ **do**
    Extract feature map $Z = f(x)$
    Apply Whitening Transformation:
      Compute mean μ and subtract from Z
      Compute whitening matrix W and apply $Z' = W(Z - \mu 1_{1\times n})$
    Apply Orthogonal Transformation:
      Compute transformed feature $Z'' = Q^T Z'$
    Compute classification loss $L_{ce} = \text{CrossEntropy}(g(Z''), y)$
    Update model parameters using gradient descent
  end for
end for
Train Concept Alignment Branch:
for each epoch **do**
  for each concept batch X_{c_k} **do**
    Extract feature map $Z_{con} = f(X_{c_k})$
    Generate Weakly-Supervised Concept Mask:
      Compute concept activation map $M_k = W_k^T Z_{con}$
      Normalize and threshold to get binary mask $\tilde{M}_k(i, j)$
    Optimize Orthogonal Matrix Q:
      Compute gradient G using concept alignment loss
      Update Q using Cayley transform:
      $Q^{(t+1)} = (I + \frac{\eta}{2}A)^{-1}(I - \frac{\eta}{2}A)Q^{(t)}$
  end for
end for
Final Model Evaluation:
Evaluate model performance on test dataset
Compute interpretability metrics using concept alignment
Return: Trained CAW model with optimized W and Q

3 Experimental Setup

3.1 Hardware and Software

- **Hardware:** NVIDIA GPU (T4), Intel Xeon CPU, 64GB RAM
- **Software:** Python 3.11, PyTorch 2.5.1, CUDA 11.8, NumPy, pandas, scikit-learn

3.2 Dataset Preparation

- **Derm7pt:** 827 images with 12 clinical concepts. [5]
- **SkinCon:** 3,221 images with 22 clinical concepts. [6]

Preprocessing steps:

- Image resizing to 224×224

- Data augmentation (random flip, cropping, rotation)
- Dataset split: 70% training, 15% validation, 15% test for SkinCongroh2021evaluating and original split for dataset Derm7pt as in metafiles [**?**]itemize

3.3 Hyperparameters

- **Backbone:** ResNet-18 for Derm7pt, ResNet-50 for SkinCon
- **Learning Rate:** 2×10^{-3}
- **Batch Size:** 64
- **Epochs:** 100

4 Results

The table below compares our reproduced results with those of the original paper. We were able to run it 1 time due to limitation for GPU on Google Colab.

Table 1: Comparison of disease diagnosis results (mean_{std} over three runs).

Method	**Derm7pt**			**SkinCon**		
	AUC	ACC	F1	AUC	ACC	F1
CAW (Original)	$\mathbf{88.60}_{0.10}$	$\mathbf{84.79}_{0.79}$	$\mathbf{81.34}_{0.85}$	$\mathbf{80.47}_{0.60}$	$\mathbf{79.00}_{0.19}$	$\mathbf{77.76}_{0.57}$
CAW (Reproduced)	82.30	81.56	80.99	72.55	69.03	71.89

5 Challenges and Discussion

During the reproduction of the Concept-Attention Whitening (CAW) framework, we encountered several challenges that affected the implementation and performance outcomes. These difficulties are categorized as follows:

5.1 Implementation Details and Ambiguities

One of the primary challenges was the lack of detailed implementation instructions in the original paper. Key hyperparameters and certain training configurations were not explicitly stated, requiring us to make assumptions and conduct multiple experiments to fine-tune the model.

5.2 Dataset Issues

- **Dataset Availability**: Some images from the SkinCon dataset were missing, and data preprocessing steps were not fully explained in the original work.
- **Label Variations**: There were minor discrepancies in concept annotations between our dataset and those reported in the paper. Metafile for SkinCon does not have 22 concepts information, getting right metafile from author took a little bit time.

5.3 Concept Alignment Sensitivity

The concept alignment branch, which optimizes the orthogonal matrix Q, was highly sensitive to initialization and learning rates. Achieving stable convergence required careful tuning, and small changes in initialization sometimes led to significant variations in the results.

5.4 Computational Resources

The CAW model, particularly the whitening transformation and concept alignment, increased computational demands compared to standard batch normalization. Training required higher memory usage and longer convergence times, especially for large-scale datasets such as SkinCon (33 h on cpu).

5.5 Result Discrepancies

As you can see in results table above our reproduced results did not match the original findings. These variations may be due to hyperparameters and loss functions.

6 Conclusion

Although our reproduced results did not match the exact performance metrics reported in the original paper, we were able to successfully reproduce the CAW framework and validate its key contributions. The methodology provided a clear and detailed description of the algorithm, ensuring reproducibility. Our implementation followed the original pipeline, including dataset preprocessing, model training, and evaluation. Further refinements in parameter tuning and dataset preprocessing may help bridge the gap between the original and reproduced results.

References

[1] Yizhe Bie, Li Luo, and Hao Chen. Mica: Towards explainable skin lesion diagnosis via multi-level image-concept alignment. *arXiv preprint arXiv:2401.08527*, 2024.

[2] Zhi Chen, Yilun Bei, and Cynthia Rudin. Concept whitening for interpretable image recognition. *Nature Machine Intelligence*, 2(12):772–782, 2020.

[3] Junlin Hou, Jilan Xu, and Hao Chen. Concept-attention whitening for interpretable skin lesion diagnosis. *arXiv preprint*, arXiv:2404.05997, 2024.

[4] Lang Huang, Yitong Zhou, Feng Zhu, Li Liu, and Ling Shao. Iterative normalization: Beyond standardization towards efficient whitening. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4874–4883, 2019.

[5] Jeremy Kawahara, Shayne Daneshvar, Giuseppe Argenziano, and Ghassan Hamarneh. Seven-point checklist and skin lesion classification using multitask multimodal neural nets. *IEEE Journal of Biomedical and Health Informatics*, 23(2):538–546, Mar 2019.

[6] C. Patrício, J.C. Neves, and L.F. Teixeira. Coherent concept-based explanations in medical image and its application to skin lesion diagnosis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3798–3807, 2023.