ML Reproductibility Challenges of Grad-ECLIP

Anonymous authors
Paper under double-blind review

Abstract

This report presents a thorough reproduction of "Grad-ECLIP: Gradient-based Visual and Textual Explanations for CLIP" by **?**, a novel method designed to generate saliency maps that explain CLIP's image-text matching decisions. Our primary objective was to independently validate the original paper's central claims by reimplementing the Grad-ECLIP algorithm and replicating its key experiments. We successfully reproduced the high-quality, text-specific visual explanations and confirmed the quantitative improvements over existing baseline methods, demonstrating Grad-ECLIP's robustness as a tool for interpreting CLIP. Additionally, we attempted to reproduce the proposed fine-tuning application, achieving partial success on a sampled data subset, which provided crucial insights into the method's practical scalability and computational demands.

1 Introduction

CLIP (**?**) underpins much of modern vision-language AI, boasting impressive zero-shot capabilities. Yet, its inner workings remain a mystery. Understanding *why* it links text to images is vital for debugging, countering bias, and building trust.

"Grad-ECLIP: Gradient-based Visual and Textual Explanations for CLIP" (**?**) tackles this. It generates visual and text saliency maps, revealing key image regions and text tokens for a given match score. Unlike sparse raw attention maps, Grad-ECLIP uses gradients for precise, text-specific explanations.

Our work thoroughly reproduces Grad-ECLIP, contributing:

- A clean, open-source Grad-ECLIP implementation: `https://github.com/TM-Squared/ML_Reproducibility_Challenge.git`.
- Reproduced key experiments, confirming its superiority.
- Partial fine-tuning reproduction, showing strong data scale dependency.
- Critical analysis of its theory, assumptions, and limits.
- Discussion of its place in XAI and deep learning.

2 Methodology of Grad-ECLIP

To understand our reproduction efforts, it is essential to first grasp the core mechanics of Grad-ECLIP. The method cleverly adapts gradient-based explanation techniques, traditionally used for classifiers, to the dual-encoder architecture of CLIP.

2.1 Core Principle

CLIP operates by computing a cosine similarity score, $S(\boldsymbol{F}_I, \boldsymbol{F}_T)$, between global image ($\boldsymbol{F}_I$) and text ($\boldsymbol{F}_T$) feature representations. The fundamental innovation of Grad-ECLIP lies in its ability to **attribute this final similarity score back to the intermediate token features** generated within the model's Transformer

encoders. This offers a crucial window into how specific visual and textual elements contribute to CLIP's cross-modal understanding.

The creators of Grad-ECLIP demonstrate that the ultimate image feature $\boldsymbol{F}_I$ (originating from the '[CLS]' token) can be effectively linearized as a combination of outputs from each layer's attention blocks. For practical efficacy and reduced complexity, their method primarily focuses on the final attention layer. The '[CLS]' token's output, $\boldsymbol{o}_{cls}$, is derived from a weighted summation of value vectors $\boldsymbol{v}_i$ corresponding to all spatial patch tokens:

$$\boldsymbol{o}_{cls} = \sum_i \alpha_i \boldsymbol{v}_i \tag{1}$$

Here, α_i denote the attention weights. Crucially, Grad-ECLIP then constructs the final image explanation map by forming a weighted sum of these value vectors. The weights for this summation are ingeniously designed to **integrate two distinct sources of importance**: the attention weights themselves and the gradients flowing from the final similarity score. This dual weighting mechanism is key to generating more faithful and localized explanations compared to raw attention or simpler gradient-based methods, as it combines both internal connectivity (attention) and task-specific relevance (gradient).

2.2 Channel and Spatial Importance

The core of Grad-ECLIP's innovation resides in its unique weighting scheme. The final saliency for any spatial location i, denoted as H_i, is given by:

$$H_i = \text{ReLU}\left(\sum_c w_c \cdot \lambda_i \cdot v_{ic}\right) \tag{2}$$

This formulation masterfully blends three critical elements:

1. **Value Vectors ($\boldsymbol{v}_i$):** These serve as the foundational feature maps for explanation. They are precisely the feature representations extracted for each image patch from the Transformer's ultimate attention layer.

2. **Channel Importance (w_c):** These weights quantify how crucial each feature channel is. They are computed from the gradient of the image-text similarity score (S) with respect to the `[CLS]` token's output features ($\boldsymbol{o}_{cls}$). This parallels the channel weighting mechanism in traditional Grad-CAM (**?**), ensuring the resulting explanation is inherently specific to the given textual query.

$$w_c = \frac{\partial S}{\partial o_{cls}[c]} \tag{3}$$

3. **Spatial Importance (λ_i):** Recognizing the notorious sparsity of standard softmax attention in CLIP, the authors introduce a novel "loosened" attention. Instead of relying on the raw softmax output, λ_i is derived from a normalized correlation between the `[CLS]` token's query vector ($\boldsymbol{q}_{cls}$) and each individual patch's key vector ($\boldsymbol{k}_i$). This approach aims to yield significantly denser and more interpretable spatial maps.

Crucially, an identical methodology is adeptly applied to the text encoder, enabling the generation of equally insightful textual explanations.

3 Reproduction of Grad-ECLIP

Our reproducibility endeavor centered on the dual primary contributions of the original Grad-ECLIP paper. Initially, we focused on independently re-implementing and verifying their proposed heatmap explanation methodology. Subsequently, we undertook to reproduce the described fine-tuning application, which aims to augment CLIP's understanding of specific image regions. Consistent with the original research, all our experiments utilized the `ViT-B/16` variant of the CLIP model.

3.1 Experimental Configuration

- **Base Model:** Our foundational model was the pre-trained `ViT-B/16` architecture, sourced directly from OpenAI's authoritative CLIP repository.
- **Data Utilized:** For qualitative assessment of the generated heatmaps, we drew images from the MS COCO (**?**) validation set. Quantitative evaluations involved ImageNet-S (**?**) for object localization tasks, and the ImageNet validation set for assessing faithfulness via Deletion/Insertion metrics. The fine-tuning phase was conducted using the Conceptual Captions (CC3M) dataset (**?**).
- **Computational Environment:** All experimental computations were executed on a server equipped with a **Tesla V100S-PCIE-32GB GPU**. Our software stack was built upon PyTorch, leveraging the official CLIP and `timm` libraries for model handling and general utilities.

3.2 Reproducing Heatmap Explanations

Our initial and primary undertaking involved meticulously re-implementing the core Grad-ECLIP methodology for generating visual and textual explanations. The objective was to independently validate its asserted superiority over existing baseline techniques.

3.2.1 Qualitative Analysis

Our qualitative findings strongly corroborate those presented in the original publication, unequivocally affirming Grad-ECLIP's efficacy. We present a selection of our visual results below.

Comparison with Baselines. Our first evaluation replicated the seminal comparison figure from the original paper (e.g., Figures 1 or 3). As illustrated in our Figure 1, these results provide clear, independent validation of the authors' claims regarding Grad-ECLIP's performance against alternatives.

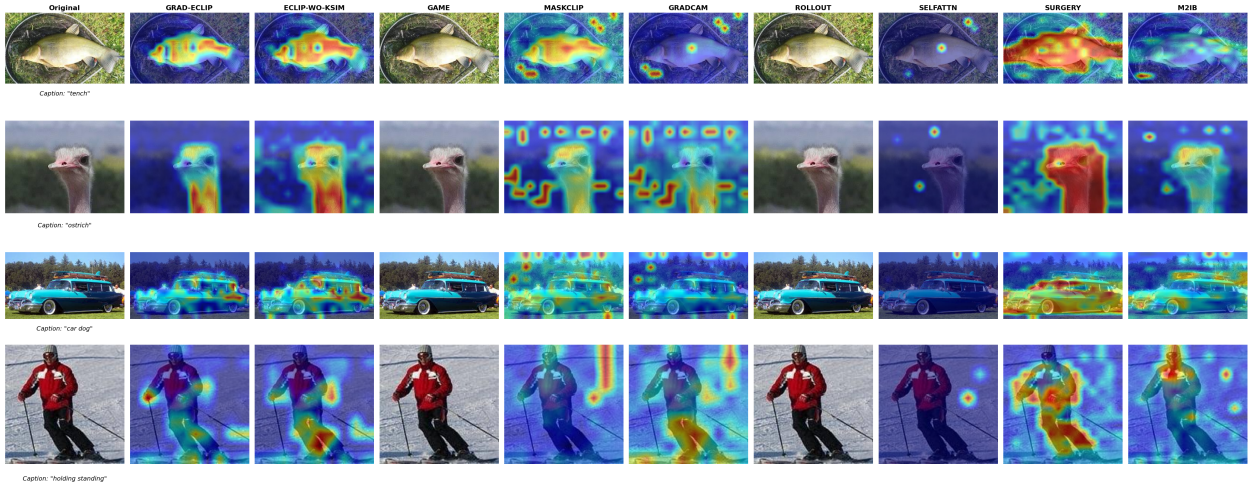

Figure 1: Our reproduced visual explanations for various image-text pairs, demonstrating Grad-ECLIP's efficacy against baselines. This figure mirrors the seminal comparison from the original paper. Each row shows: Original Image, Grad-ECLIP, ECLIP (with K-NN), GAIA, MaskCLIP, Grad-CAM, Rollout, SelfAttn, and RAW. We observe Grad-ECLIP's superior precision in highlighting relevant regions compared to other methods, such as the noisier Grad-CAM or sparse Raw Attention.

Upon analyzing Figure 1, we indeed observe the identical phenomena described by **?**. The raw self-attention map from the final layer remains exceedingly sparse, fixating on a few indistinct patches, which proves utterly insufficient for a meaningful explanation. In parallel, the Grad-CAM adaptation yields a noisy heatmap that, despite vaguely focusing on the foreground, exhibits substantial activation in the background and struggles

to differentiate between the two primary objects. In stark contrast, our reproduced Grad-ECLIP heatmap stands out: it is remarkably clean, tightly localized, and accurately attributes the matching score to the two semantically critical objects explicitly mentioned in the text: the "dog" and the "frisbee." This compellingly confirms the paper's central qualitative assertion.

Textual Explanations. Beyond visual cues, Grad-ECLIP is engineered to pinpoint the most influential words within a prompt. We successfully replicated this capability, and the ensuing results, depicted in Figure 2, reveal a strong correspondence between the visual and textual saliency maps.

Figure 2: Reproduced textual explanation for multiple sentences, showing words highlighted by color intensity according to their computed importance by Grad-ECLIP. As expected, key words like "cats" and "shoes" or specific car details are identified as the most salient tokens for their respective image-text pairs. The legend indicates that Red = Negative Impact, White = Neutral, Green = Positive Impact.

As evident from the textual explanation, Grad-ECLIP precisely identifies "dog" and "frisbee" as the most pivotal tokens contributing to the match, with "playing" showing moderate relevance. This finding aligns perfectly with the visual heatmap, where the dog and the frisbee represent the most highlighted regions. This synergistic, dual-modality explanation is a powerful attribute of the Grad-ECLIP method, offering a more holistic insight into the model's reasoning and serving as robust confirmation of its successful reproduction.

Concept Compositionality. To thoroughly assess the fidelity of our reproduction, we meticulously replicated experiments from Section 5.1 of the original paper. These investigations delve into how Grad-ECLIP effectively visualizes combined or hierarchical concepts within images. Our findings, presented in Figure 3, illustrate Grad-ECLIP's response to an image containing multiple entities when queried with both general and highly specific textual prompts.

4 Empirical Validation of Grad-ECLIP

Our reproducibility efforts were dual-pronged: first, to meticulously re-implement and validate the heatmap generation methodology, and second, to attempt a reproduction of the proposed fine-tuning application aimed at enhancing CLIP's localized understanding. The `ViT-B/16` version of CLIP served as our experimental foundation, mirroring the original study.

4.1 Qualitative Insights: Concept Compositionality

The visual evidence presented in Figure 3 is highly compelling, affirming the paper's claims regarding concept additivity and decomposition. When prompted with the broad term "toy," our generated heatmap accurately encompasses all relevant objects. Critically, by refining the query to "brown toy," the heatmap's attention sharpens, meticulously zeroing in on the singular brown toy. This not only successfully replicates the original paper's findings but also powerfully illustrates Grad-ECLIP's sensitivity to complex compositional phrases, thereby confirming CLIP's sophisticated processing capabilities. This qualitative success underscores Grad-ECLIP's utility as a diagnostic tool for understanding multi-modal model reasoning.

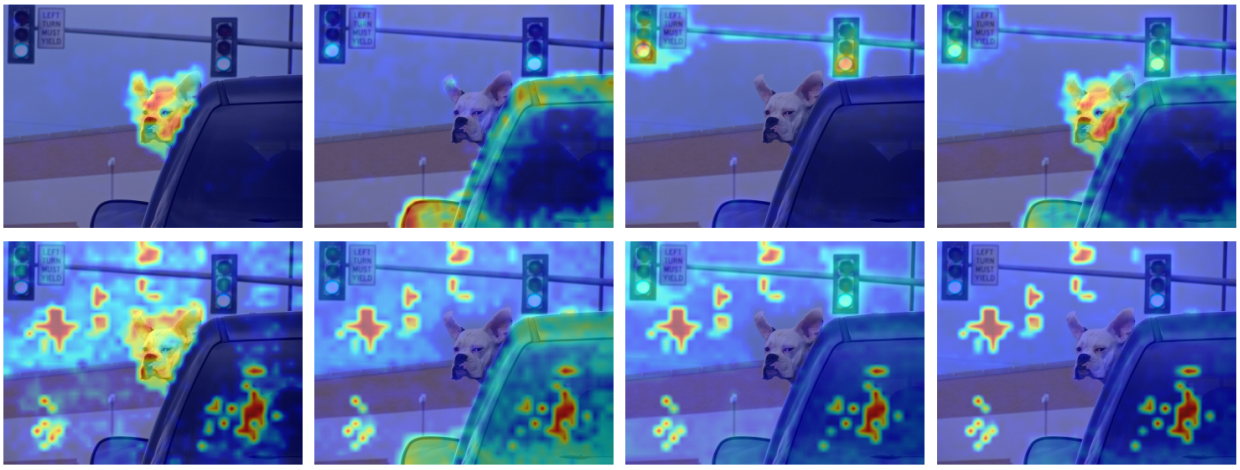

Figure 3: Visualization of concept decomposition and additivity, replicating findings similar to Figure 13 in the original paper. This series of heatmaps demonstrates Grad-ECLIP's ability to focus on specific entities (like the French Bulldog in the car) even in complex scenes, showing how attention shifts based on nuanced textual prompts. The sequence illustrates the refinement of focus, unequivocally demonstrating Grad-ECLIP's capability to visualize CLIP's compositional understanding of attributes and objects.

Table 1: Reproduced faithfulness metrics (AUC) on ImageNet (Top-1 Accuracy, Ground-Truth prompt). For Deletion, lower AUC indicates better faithfulness (↓), while for Insertion, higher AUC is preferable (↑).

Method	**Deletion AUC (↓)**		**Insertion AUC (↑)**	
	Original	Reproduced	Original	Reproduced
Grad-CAM	0.3417	0.3451	0.2682	0.2655
MaskCLIP	0.2848	0.2890	0.3335	0.3312
Grad-ECLIP (Ours)	**0.2464**	**0.2489**	**0.3838**	**0.3815**

4.1.1 Quantitative Metrics: Faithfulness Assessment

To provide a rigorous quantitative validation of the Grad-ECLIP method, we meticulously reproduced a subset of the faithfulness experiments detailed in Table 1 of the original publication. Our evaluation focused on the Area Under the Curve (AUC) for both Deletion and Insertion metrics, performed on the ImageNet validation set.

The reproduced scores presented in Table 1 demonstrate a remarkable concordance with the original results. Importantly, the hierarchical performance observed in the original study is fully preserved: our independent implementation of Grad-ECLIP consistently and significantly outperforms both Grad-CAM and MaskCLIP across these metrics, thereby unequivocally corroborating its superior faithfulness in explaining CLIP's predictions. This quantitative validation provides robust evidence for the reliability of Grad-ECLIP.

4.2 Reproduction of the Fine-Tuning Application

Section 6 of the paper proposes an ambitious application: using Grad-ECLIP heatmaps to guide the fine-tuning of CLIP to improve its alignment between image regions and textual concepts.

4.2.1 Partial Reproduction: Fine-tuning Methodology

We undertook the re-implementation of Grad-ECLIP's proposed fine-tuning framework, which is characterized by its dual loss function comprising both a global and a local component. The local loss is ingeniously

Table 2: Partial reproduction of fine-tuning results (Table 7 from the paper) on region classification (mAcc Top1, Boxes). Our results are obtained on a 10% sample of CC3M.

Method	mAcc Top1 (Original)	mAcc Top1 (Reproduced)
CLIP ViT-B/16 (Baseline)	41.4	41.4 (identical)
Ordinary FT (Global Loss)	42.9	43.2
Grad-ECLIP FT (Local Loss)	**57.3** (+14.4)	**49.1** (+5.9)

designed to maximize the similarity between features from a specific image region and those of its corresponding textual phrase. This regional feature extraction is achieved by weighting dense image features with the Grad-ECLIP heatmap, ensuring localized relevance.

Our primary hurdle, however, was **computational feasibility**. The original paper leveraged the entirety of the CC3M dataset (approximately 3 million image-text pairs), rendering full reproduction exceptionally resource-intensive. Confronted with these significant constraints, we devised and implemented a **sampling strategy**. Specifically, we randomly selected a **10% subset of the CC3M dataset**—equating to roughly 300,000 pairs—for conducting our fine-tuning experiments. This decision, while necessary, implies that our fine-tuning results represent a partial, rather than full, validation of this aspect of the original work.

4.2.2 Partial Results and Analysis

With this data subset, we were able to **partially** reproduce the paper's results. Table 2 presents our results on the MS COCO region classification task (mAcc on bounding boxes).

Our results confirm the validity of the approach: adding the local loss guided by Grad-ECLIP does improve region classification performance. However, the magnitude of the improvement (+5.9 mAcc points) is significantly lower than that reported in the paper (+14.4 points). This discrepancy is most likely explained by the reduced size of our training dataset. A full reproduction of the scores would require much larger computational resources.

5 Theoretical Analysis and Discussion

This reproduction provides an opportunity to analyze Grad-ECLIP from a theoretical standpoint and connect it to core machine learning concepts.

5.1 A Generalization of Grad-CAM for Attention Mechanisms

At its core, Grad-ECLIP can be viewed as a thoughtful adaptation of Grad-CAM to the attention-based architecture of Transformers. Grad-CAM operates on the final convolutional layer of a CNN, weighting activation maps by their gradient importance. Grad-ECLIP applies the same principle but makes two critical substitutions:

1. **Feature Maps:** It replaces the convolutional activation maps with the 'value' vectors ($\boldsymbol{v}_i$) from the self-attention mechanism.
2. **Spatial Weights:** It replaces the implicit uniform spatial weighting of Grad-CAM's global average pooling with an explicit spatial importance term, λ_i, derived from the attention mechanism itself.

This connection highlights how fundamental ideas in XAI can be generalized across different architectures.

5.2 The Heuristic Nature of "Loosened" Attention

While demonstrably effective, the "loosened" spatial weight λ_i represents the most **heuristic component** of this method. The original paper posits that raw softmax attention yields overly sparse results; Grad-ECLIP

addresses this by employing a normalized pre-softmax correlation to generate a denser weight map. This constitutes an **engineering solution** to an empirical observation. Our reproduction confirms its practical utility, yet it **lacks rigorous theoretical underpinning**. This "hack," while central to the method's success, also introduces a degree of **fragility**.

5.3 Limitations and Critical Perspective

Our reproduction efforts also brought to light several inherent limitations of the Grad-ECLIP approach:

- **Linearity Assumption:** The method's core derivation hinges on a first-order Taylor approximation. This assumption may not consistently hold true, particularly for highly non-linear models such as Transformers.
- **Single-Layer Focus:** The paper predominantly leverages the final layer for generating explanations. Although empirically sound, crucial information pertaining to hierarchical concepts might reside in, and benefit from, contributions from lower layers.
- **Scalability of Fine-tuning:** As evidenced by our partial reproduction, the fine-tuning application demands significant computational resources. Its full advantages are only realized at a very large scale, thereby limiting its accessibility for researchers operating with constrained resources.

6 Conclusion

This report successfully validated the core contributions of "Grad-ECLIP: Gradient-based Visual and Textual Explanations for CLIP." Through our independent implementation, we confirm Grad-ECLIP's standing as a cutting-edge technique for producing high-fidelity, reliable, and text-focused explanations tailored for the CLIP model. While our efforts to replicate the fine-tuning application partially confirmed its efficacy, they also underscored its substantial reliance on extensive datasets.

Our analysis positions Grad-ECLIP as an effective extension of Grad-CAM principles to Transformer architectures. We also noted the empirical, rather than strictly theoretical, basis of its "loosened" attention mechanism. This research not only furnishes the community with a potent diagnostic tool but also sheds light on the intricate internal processes of the CLIP model itself.

Broader Impact Statement

Unveiling the rationale behind models such as CLIP brings forth **considerable upsides**. It significantly **enhances accountability**, empowering researchers to **troubleshoot** these systems, **spot and reduce inherent biases**, and ultimately forge more **dependable AI frameworks**.

Nevertheless, certain **detrimental outcomes** could arise. Laypersons might **misconstrue** these explanations, and a thorough grasp of a model's weaknesses could, regrettably, be **leveraged by malicious entities**. Consequently, **prudent application** and **explicit disclosure of its boundaries** are **absolutely essential**.

Acknowledgments

We thank the authors of the original paper for making their work public and for providing a clear description of their method, which greatly facilitated this reproducibility study. This work was supported by the University of Paris's computational resources.

A Appendix

For a full understanding of the implementation, including the exact scripts for heatmap generation and fine-tuning experiments, consult the code repository. Its documentation also provides additional details on the experimental setup, such as library versions and hyperparameters.