Project – Fine Tuning CLIP

William Eriksson Linköping University
Karl Schelin Linköping University
Oscar Ljungdahl Linköping University
Hugo Dahlquist Linköping University

Arian Bethoui Linköping University

Abstract

The project aimed to improve the OpenCLIP vision-language model by OpenAI for remote sensing imagery using the RSICD dataset. Different parameters were investigated, for example, batch size, temperature and number of epochs. Two different image encoders were also compared (ViT-B/14 vs. ViT-B/32), ultimately chosing ViT-B/14.

The solution uses both supervised contrastive learning and semi-supervised pre-training with consistency loss on the image and text encoders using augmentations. To enable robust zero-shot classification, multiple prompt templates were deployed for each class. When training the model on the supervised dataset the parameters were locked for our image encoder and only trained the text-encoder. This achieved an F1-score of 0.80 and accuracy of 0.83.

1 Introduction

Large-scale vision-language models like CLIP (Contrastive Language–Image Pretraining) Radford et al. [2021] have shown strong performance on many image understanding tasks by aligning image and text embeddings in a joint space. OpenCLIP [Ilharco et al., 2021], an open-source reimplementation of CLIP, makes this framework more accessible for fine-tuning and domain adaptation.

In this project, the focus was on adapting OpenCLIP to remote sensing imagery using the RSICD dataset, which consists of aerial images and associated captions. The primary goal is to fine-tune a CLIP model to improve performance on tasks such as zero-shot classification and image-text retrieval.

Several aspects of CLIP training are known to influence performance. These include the batch size, temperature parameter in the contrastive loss, number of training epochs, and choice of model architecture [Chen et al., 2020, He et al., 2020].For instance, larger batch sizes provide more negative pairs for contrastive learning but may reduce performance in some domains [Radford et al., 2021]. The temperature parameter, often treated as a trainable scalar, controls the sharpness of the similarity distribution [Chen et al., 2020]. Choosing the right model size also involves a trade-off between performance and computational cost [Yao et al., 2021].

In this work, how these parameters affect CLIP fine-tuning on remote sensing data were systematically explored. The aim was to answer the question: *How can OpenCLIP be best fine-tuned for domain-specific classification to maximize accuracy and F1-score*

Preprint. Under review.

2 Problem Formulation

Given an image-text pair (x_i, y_i), where $x_i \in \mathcal{X}$ is an image and $y_i \in \mathcal{Y}$ is a caption, the model learns two embedding functions:

- Image encoder: $f_\theta : \mathcal{X} \to \mathbb{R}^d$ (ViT-B/32 or ViT-B/14, $d = 512$)
- Text encoder: $g_\phi : \mathcal{Y} \to \mathbb{R}^d$ (Transformer)

Cosine similarity between two vectors $\mathbf{u}$ and $\mathbf{v}$ is defined as:

$$\text{sim}(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u}^\top \mathbf{v}}{||\mathbf{u}||_2 ||\mathbf{v}||_2}$$

Grade 3: Supervised Contrastive Training
In the basic Grade 3 version, the model is trained on N labeled image-caption pairs from Split B using a symmetric contrastive loss:

$$\mathcal{L}_{\text{CLIP}} = \frac{1}{2}\big(\mathcal{L}_{\text{image-to-text}} + \mathcal{L}_{\text{text-to-image}}\big)$$

$$\mathcal{L}_{\text{image-to-text}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(\text{sim}(f_\theta(x_i), g_\phi(y_i))/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(f_\theta(x_i), g_\phi(y_j))/\tau)}$$

$$\mathcal{L}_{\text{text-to-image}} = -\frac{1}{N}\sum_{j=1}^{N} \log \frac{\exp(\text{sim}(g_\phi(y_j), f_\theta(x_j))/\tau)}{\sum_{i=1}^{N} \exp(\text{sim}(g_\phi(y_j), f_\theta(x_i))/\tau)}$$

where $\tau > 0$ is a learnable temperature parameter.

Grade 5: Semi-supervised Training and Encoder Freezing
In the advanced Grade 5 version, we extend the supervised CLIP loss with an additional consistency regularization term using M unlabeled images from Split A:

$$\mathcal{L}_{\text{consist}} = \frac{1}{M}\sum_{u=1}^{M} \|f_\theta(\xi(x_u)) - f_\theta(\xi'(x_u))\|_2^2$$

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CLIP}} + \lambda \cdot \mathcal{L}_{\text{consist}}, \quad \lambda = 0.5$$

Here, ξ and ξ' are two random augmentations of the same image (e.g., crop, color jitter).

After training, once the image encoder f_θ produces robust representations, it is frozen. The second phase consists of fine-tuning only the text encoder g_ϕ (and optionally the temperature τ and a projection head) using a cosine similarity loss:

$$\mathcal{L}_{\text{align}} = 1 - \frac{1}{N}\sum_{i=1}^{N} \text{sim}(g_\phi(y_i), f_\theta(x_i))$$

This forces the caption embeddings to align tightly with the fixed image feature space.

Zero-shot Inference
For zero-shot classification with K classes, we define M prompt templates $\mathcal{T} = \{T_1, \ldots, T_M\}$. For each class k, we construct M textual prompts:

$$t_k^{(m)} = T_m(\text{class } k), \quad m = 1, \ldots, M$$

The final class embedding is the average of its template encodings:

$$\bar{g}_\phi(t_k) = \frac{1}{M}\sum_{m=1}^{M} g_\phi\left(t_k^{(m)}\right)$$

Given an image x, we compute cosine similarity scores between $f_\theta(x)$ and all $\bar{g}_\phi(t_k)$ and predict based on the highest score.

The prediction probability for class k given an image $\mathbf{x}$ is:

$$p(y = k|\mathbf{x}) = \frac{\exp(\text{sim}(f_\theta(\mathbf{x}), \bar{g}_\phi(t_k))/\tau)}{\sum_{j=1}^{K} \exp(\text{sim}(f_\theta(\mathbf{x}), \bar{g}_\phi(t_j))/\tau)}$$

3 Method

In this section, the practical steps and techniques used to address the image-text matching and zero-shot classification tasks were described. These descriptions build directly on the formal objectives defined in Section 2.

3.1 Zero-Shot Classification with Prompt Template Aggregation

To enable robust zero-shot classification, multiple prompt templates were deployed for each class, as described in Section 2. For each class, M class-specific prompts were generate and encoded using the CLIP text encoder. The resulting embeddings are averaged to obtain a single class representation. At inference time, the cosine similarity between the L2-normalized image embedding and each class's aggregated text embedding was computed. Afterwards the softmax function (see Section 2) was applied to obtain class probabilities. The class with the highest probability is selected as the prediction for each test image.

3.2 Supervised Training

The CLIP model was fine-tuned on the labeled subset (Split B) of the RSICD dataset, which was split into 80% training and 20% validation using the symmetric contrastive loss $\mathcal{L}_{\text{CLIP}}$ defined in Section 2. Early stopping was implemented with 5 epochs of patience if validation error was not improved. Both image and text embeddings are L2-normalized prior to similarity computation. The model is trained to maximize the similarity between matching image-text pairs and minimize it for non-matching pairs. We use the Adam optimizer with a learning rate of 1×10^{-5}.

3.3 Unsupervised training

A training of a an unsupervised model was conducted, where only the contrastive loss between augmentations was considered. This was evaluated using a KNN classifier on the training data. An accuracy of 95% was reached, however, the text-encoder could not match these results while intergrating a semi-supervised

3.4 Semi-Supervised Pre-training with Consistency Loss

For the advanced (Grade 5) version, unlabeled images from Split A were incorporated using the consistency loss $\mathcal{L}_{\text{consist}}$ (see Section 2). For each unlabeled image, two random augmentations (e.g., random cropping, color jitter) were generated. The aim was to minimize the L2 distance between their embeddings, encouraging the model to produce stable representations under augmentation. The total loss is a weighted sum of the supervised contrastive loss and the consistency loss, with the unsupervised weight $\lambda = 1$.

3.5 Align text encoder

After the semi-supervised training was finished the image encoder was frozen as it already had discovered robust image representations. This according to KNN-classifier results and the careful consideration of the projected image-encodings. Thereafter, the model was trained with the align loss function to align the text encodings of all the captions to the image representations.

3.6 Implementation Details

Experiments are conducted using the ViT-B/32 or ViT-B/14 image encoder and the Transformer-based text encoder from the CLIP model. ViT-B/32 splits images into larger 32x32 pixel patches, yielding lower-resolution embeddings that are faster to compute but sacrifice fine-grained details. ViT-B/14 uses smaller 14x14 pixel patches, capturing higher-resolution spatial features at the cost of increased compute and memory usage. Data augmentations for consistency training include random resized cropping and color distortion.

3.7 Evaluation of model

The evaluation of the result was conducted by using the metrics accuracy and F1-score on dataset C. For comparison a defined benchmark model is the semi supervised model with a batch size = 32, ViT-B/14, Learning rate = 1×10^{-5} and 5 epochs of training.

Abolation studies was conducted by alternating the used augmentations, the use of multiple templates for zero shot classification and the removal of text-to-image or image-to-text retrival from the loss function. Parameter analysis was used by alternating the hyperparameters batch size, temperature and number of epochs.

4 Result

4.1 Supervised

Model	F1-Score	Accuracy
ViT-B/14	0.75	0.77
ViT-B/32	0.66	0.64

Table 1: Benchmark supervised model using ViT-B/14 in comparison with ViT-B/32. Temperature = 0.07, Batch size = 32, Epochs = 5.

Model	F1-Score	Accuracy
Removal of Image-to-text	0.68	0.72
Removal of text-to-image	0.71	0.75
Removal of multiple templates	0.65	0.69

Table 2: Ablation study based on F1-Score and Accuracy

Model	F1-Score	Accuracy
Temperature = 0.12	0.73	0.76
Batch size = 64	0.77	0.80
Epochs = 10	0.80	0.83

Table 3: Parameter analysis based on F1-Score and Accuracy

The supervised model using ViT-B/14 achieved an F1-score of 0.75 and an accuracy of 0.77 using the benchmark settings described in the method section (see Section 3). These results serve as a baseline for comparison with the ablation and parameter studies that follow.

As shown in Table 2, removing any single component led to a drop in both F1-score and accuracy compared to the benchmark. This shows that each part of the model setup contributes to its performance.

The largest decrease in performance occurred when multiple templates were removed. It indicates that using a variety of textual prompts helps the model better understand and match different class descriptions, which improves generalization.

Removing the image-to-text and text-to-image components also led to performance drops. This confirms that both directions of the contrastive loss are important for building strong connections between image and text features.

A parameter analysis was also performed to explore how changes in training settings affect the model's performance. Table 3 shows the impact of changing temperature, batch size, and number of epochs.

Increasing the number of epochs from 5 to 10 led to the best improvement, with an F1-score of 0.80 and an accuracy of 0.83. This shows that the model benefits from longer training. A larger batch size also helped slightly, while a higher temperature value reduced performance, suggesting that the original setting (0.07) may be closer to optimal for this task.

4.2 Semi-supervised

The Semi-supervised model resulted in an F1 score of 0.76 and an accuracy of 0.80 using the benchmark settings stated in the method under the section "Evaluation of model". These results are used to set a baseline that is further used in comparission with the abolation study below.

To understand the individual contributions of different components to the overall model performance, an ablation study was conducted. This systematic approach involves selectively removing or altering specific elements of our proposed methodology and observing the resulting impact on the F1-score and accuracy. By isolating these components, we can ascertain their importance and gain an understanding of how each contributes to the model's effectiveness in zero-shot classification and image-text retrieval tasks on the RSICD dataset.

Model	**F1-Score**	**Accuracy**
ViT-B/14	0.76	0.80
ViT-B/32	0.72	0.70

Table 4: Benchmark semi-supervised model

Model	**F1-Score**	**Accuracy**
Removal of Image-to-text	0.62	0.65
Removal of text-to-image	0.65	0.68
Removal of multiple templates	0.58	0.61
Removal of augmentation (rotation)	0.60	0.63
Removal of augmentation (color jitter)	0.61	0.64

Table 5: Ablation study based on F1-Score and Accuracy

As depicted in Table 5, the systematic removal of individual components generally resulted in a discernible decrease in both F1-score and accuracy compared to the established benchmark. This empirically underscores the synergistic contribution of each element to the model's overall performance.

The most substantial decline in performance was observed with the removal of multiple templates for zero-shot classification. This outcome suggests that the aggregation of diverse textual prompt formulations is critical for constructing robust class representations, thereby enhancing the model's ability to generalize across various descriptive contexts within remote sensing imagery.

Similarly, the individual removal of the image-to-text and text-to-image loss components consistently led to a notable reduction in performance. This highlights the indispensable role of maintaining a strong, reciprocal alignment between visual and textual embeddings. Subtle variations in the performance impact between these two components may indicate a nuanced asymmetry in how the model prioritizes or leverages either image or text information for cross-modal alignment within this particular dataset.

The exclusion of data augmentations, such as rotation and color jitter, also negatively affected the scores. This indicates that data augmentation plays an important role in the model's generalization capabilities and improving its robustness to common variations in input data. These findings suggest that while the core contrastive learning objectives are foundational, thoughtfully applied augmentations are instrumental for achieving effective domain adaptation, particularly in the inherently diverse and complex landscape of remote sensing scenarios.

5 Conclusion

This study has provided valuable insights into fine-tuning OpenCLIP for remote sensing imagery classification. Our experiments with the RSICD dataset revealed several key findings about optimizing vision-language models for domain-specific applications.

The higher-resolution ViT-B/14 image encoder consistently outperformed ViT-B/32 across both supervised and semi-supervised approaches, confirming the importance of spatial resolution in capturing fine-grained features of remote sensing imagery. This aligns with findings that more detailed visual representations improve cross-modal alignment.

Our ablation studies demonstrated that each component of the model architecture contributes meaningfully to overall performance. Multiple text templates proved particularly crucial for zero-shot classification, supporting the observation that prompt diversity enhances model robustness. The bidirectional nature of the contrastive loss (image-to-text and text-to-image) was also validated as essential for effective cross-modal alignment.

Training parameters significantly influenced model performance. Increasing training epochs from 5 to 10 yielded the most substantial improvement (F1-score of 0.80, accuracy of 0.83), suggesting that longer training enables better adaptation to domain-specific features. Larger batch sizes also improved performance, aligning with findings on the importance of negative sample diversity in contrastive learning.

For future work, we recommend exploring:

- Alternative augmentation strategies specifically designed for aerial imagery
- Integration of self-supervised pretraining methods similar to momentum contrast approaches
- Investigation of fine-grained alignment techniques for improved performance
- Adaptation of temperature scheduling during training to optimize similarity distribution

6 Contribution Statement

All members participated in regular group meetings, jointly discussed project design and results, and contributed to the writing and revision of the final report. Below we tried to capture what everyone was responsible for but most of the work was done together and we all helped with everything. Most of the coding was done in sessions together on multiple computers, so we could ask each other for help.

- **William** teamed up with Hugo to build both our fully-supervised and semi-supervised training pipelines, merging labeled and unlabeled data. They experimented with data augmentations to see how performance changed. William then collaborated with Arian to interpret those results and wrote the problem formulation section.
- **Hugo** worked with William to make sure our labeled/unlabeled splits were set up correctly and to build the data-loading pipeline. He ran hyperparameter tests over learning rates and batch sizes, tested different augmentations, and drafted the methodology section, updating it as we discovered new tweaks.
- **Arian** developed our zero-shot classification system by designing text prompts and combining their embeddings. He and Karl ran evaluations across model versions and generate results. Arian wrote the introduction and results sections.
- **Karl** implemented the consistency loss for semi-supervised learning and partnered with Hugo on image augmentations for unlabeled data. He compared supervised vs. semi-supervised performance. Together with Arian, he explored variations in the consistency loss, and Karl wrote both the results and ethics sections.
- **Oscar** set up the initial contrastive learning pipeline—including the projection head and nearest-neighbor evaluation—with William and Hugo. He helped refine our semi-supervised approach with Karl, edited the report, and wrote the conclusion section.

7 Ethical considerations

There are several situations where the use of the project could be considered harmful or unethical. Some of the situations that fit these descriptions are:

- **Discrimination and spreading of misinformation:** The capabilities of CLIP can be exploited for malicious purposes, including the generation or classification of hate speech and misinformation. In contexts like surveillance or military applications, biases can lead to discriminatory profiling or targeting errors, with severe consequences for individuals and communities
- **Black-box nature:** The intricate and often inscrutable nature of CLIP's classification process presents a significant hurdle for user comprehension. This difficulty in understanding how the model arrives at its outputs makes the identification and subsequent reduction of inherent biases considerably more challenging. Consequently, in critical domains such as military contexts, where an erroneous classification could lead to severe and irreversible repercussions, this lack of transparency in the model's decision-making introduces a crucial ethical concern regarding accountability, reliability, and the potential for unintended harmful consequences arising from these poorly understood processes. .

References

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*, 2020. URL `https://arxiv.org/abs/2002.05709`.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. URL `https://arxiv.org/abs/1911.05722`.

Gabriel Ilharco, Mitchell Wortsman, et al. Openclip: An open-source reimplementation of clip, 2021. URL `https://github.com/mlfoundations/open_clip`.

Alec Radford, Jong Wook Kim, Chris Hallacy, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021. URL `https://arxiv.org/abs/2103.00020`.

Zhen Yao et al. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*, 2021.