Parameter-Efficient Fine-Tuning of BiomedCLIP for Diabetic Retinopathy Detection

Adi Badlani
Stanford University
badlani@stanford.org

Kevina Wang
Stanford University
kevinaw@stanford.org

Abstract

Diabetic retinopathy (DR) is a leading cause of preventable blindness globally, affecting millions of diabetics worldwide. Early detection is crucial for patient outcomes, but many underresourced regions lack access to qualified ophthalmologists for timely screening. This paper presents a new approach to automated DR severity classification from retinal fundus photographs using BiomedCLIP, a multimodal biomedical foundation model pretrained on 15 million clinical image-text pairs. We combine BiomedCLIP with parameter-efficient fine-tuning techniques, specifically Low-Rank Adaptation (LoRA), to create a computationally efficient model that leverages pre-existing biomedical knowledge while requiring minimal training resources, a common problem in many regions of the world with the most DR patients. Our approach addresses the challenge of five-level DR severity classification (No DR, Mild, Moderate, Severe, and Proliferative DR) using the Kaggle Diabetic Retinopathy Detection dataset, which contains approximately 88,000 high-resolution retinal images. We implement specialized preprocessing techniques to handle the dataset's challenges, including class imbalance and image quality variations. LoRA significantly improved diabetic retinopathy classification performance over the zero-shot BiomedCLIP baseline, achieving an F1 score of 20.0% with only a modest increase in inference time per image and memory usage. BitFit achieved the highest raw accuracy (93.7%) but suffered from class collapse, overpredicting the majority class and limiting its clinical applicability.

1. Introduction

Diabetic retinopathy (DR) represents one of the most critical healthcare challenges of the present, affecting over 103.12 million individuals worldwide (8). It is a leading cause of preventable blindness globally. The condition is a result of damaged blood vessels in the retina due to long-term high blood sugar levels. While early detection and treatment can prevent vision loss in over 90% of cases, many patients in underresourced regions lack access to timely, effective screening due to a global shortage of qualified ophthalmologists. For instance, in China, the patient-to-ophthalmologist ratio is 3000:1, making traditional screening infeasible at scale (2).

Recent advances in deep learning have shown a great deal of promise in automating DR detection from retinal fundus photos. Gulshan et al. (2016) (4) showed that convolutional neural networks (CNNs) could achieve performance comparable to retina specialists. However, these models require a large amount of computational resources for training and deployment, which limits their accessibility in settings with resource constraints.

The computational requirements of training deep learning models from scratch present a significant barrier to widespread adoption, particularly in regions with limited access to high-performance computing infrastructure. Additionally, traditional deep learning approaches often require large amounts of labeled data, which can be difficult and expensive to obtain in specialized medical domains like ophthalmology.

In this project, we propose a new approach leveraging the BiomedCLIP model, a multimodal biomedical foundation model that has been pretrained on 15 million clinical image-text pairs. We combine this approach with parameter-efficient fine-tuning techniques. By using BiomedCLIP's biomedical background knowledge and applying Low-Rank Adaptation (LoRA), we aim to achieve strong DR classification performance while greatly reducing computational resources compared to training models from scratch.

Our approach addresses several key challenges in automated DR detection: (1) The need for specialized domain knowledge in understanding retinal images; (2) the computational cost of training deep learning models from scratch; and (3) the variability in image quality and characteristics due to different imaging equipment and protocols.

Our contributions include: (1) a novel application of BiomedCLIP to the task of DR severity classification; (2) an efficient fine-tuning approach using LoRA that significantly reduces computational requirements; (3) a comprehensive evaluation of different prompting strategies for zero-shot and few-shot learning in medical image classification; and (4) a specialized preprocessing pipeline designed to address the unique challenges of retinal fundus imagery.

We hope to develop a model that can be deployed in resource-constrained settings, ideally expanding access to DR screening in regions where specialized ophthalmological expertise is limited.

2. Related Works

2.1. Deep Learning for Diabetic Retinopathy Detection

The application of deep learning to DR detection has been an active area of research for several years. The landmark study by Gulshan et al. (2016) (4) demonstrated that a CNN trained on a large dataset of retinal images could achieve sensitivity and specificity comparable to ophthalmologists for detecting referable DR. Since then, numerous approaches have been proposed to improve performance and address various challenges in this domain.

Gargeya and Leng (2017) (3) developed a CNN-based system that achieved high accuracy in distinguishing between DR and non-DR cases while also providing a visualization of the regions that contributed most to the classification decision. Abramoff et al. (2018) (1) created an FDA-approved autonomous AI system for DR detection that achieved high sensitivity and specificity, marking an important milestone in the clinical application of these technologies.

More recent work has focused on improving model performance through architectural innovations and novel training approaches. Krause et al. (2018) (6) demonstrated that a hybrid approach combining network-based and feature-based classifiers could achieve better performance than either approach alone. Wang et al. (2020) (10) proposed a zoom-in network that mimics the diagnostic process of ophthalmologists by focusing on suspicious regions.

While these approaches have shown impressive results, they typically require substantial computational resources for training and fine-tuning, limiting their applicability in resource-constrained settings.

2.2. Vision-Language Models in Medical Imaging

The emergence of vision-language models (VLMs) has opened new possibilities for medical image analysis. These models, trained on large-scale image-text pairs, learn to align visual and textual representations in a shared embedding space, enabling zero-shot and few-shot learning capabilities.

CLIP (Contrastive Language-Image Pre-training) by Radford et al. (2021) (7) demonstrated impressive zero-shot image classification capabilities by learning to associate images with natural language descriptions via cosine similarity score computed on text and image embeddings in a model's shared space.

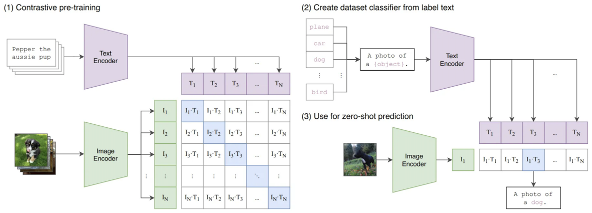

(1) Contrastive pre-training
Pepper the aussie pup
Text Encoder
Image Encoder
T_1 T_2 T_3 ... T_N
I_1 I_2 I_3 ⋮ I_N
$I_1 \cdot T_1$ $I_1 \cdot T_2$ $I_1 \cdot T_3$ ... $I_1 \cdot T_N$
$I_2 \cdot T_1$ $I_2 \cdot T_2$ $I_2 \cdot T_3$ ... $I_2 \cdot T_N$
$I_3 \cdot T_1$ $I_3 \cdot T_2$ $I_3 \cdot T_3$ ... $I_3 \cdot T_N$
$I_N \cdot T_1$ $I_N \cdot T_2$ $I_N \cdot T_3$ ... $I_N \cdot T_N$
(2) Create dataset classifier from label text
plane
car
dog
bird
A photo of a {object}.
Text Encoder
(3) Use for zero-shot prediction
Image Encoder
I_1
T_1 T_2 T_3 ... T_N
$I_1 \cdot T_1$ $I_1 \cdot T_2$ $I_1 \cdot T_3$ ... $I_1 \cdot T_N$
A photo of a dog.

Figure 1. Summary of CLIP approach, adapted from (7)

This approach has been adapted to medical domains through models like MedCLIP (Zhang et al., 2022) (13) and BiomedCLIP (Zhang et al., 2023) (15), which are trained on medical image-text pairs from scientific literature and clinical datasets.

BiomedCLIP, in particular, has shown promising results in various medical imaging tasks, including chest X-ray interpretation, dermatological condition classification, and pathology image analysis. By leveraging the biomedical knowledge encoded in its pre-trained weights, BiomedCLIP can achieve strong performance on specialized medical tasks with minimal fine-tuning.

2.3. Parameter-Efficient Fine-Tuning

As foundation models have grown in size and complexity, parameter-efficient fine-tuning techniques have emerged as a way to adapt these models to specific tasks without updating all parameters. This approach significantly reduces computational requirements while maintaining performance comparable to full fine-tuning.

Low-Rank Adaptation (LoRA), proposed by Hu et al. (2022) (5), is one such technique that has gained popularity for its efficiency and effectiveness. LoRA works by inserting trainable low-rank matrices into the attention layers of transformer-based models, reducing the number of trainable parameters by orders of magnitude compared to full fine-tuning.

Another parameter-efficient approach is BitFit (12), which restricts fine-tuning to only the bias terms within a pre-trained model while keeping all other weights frozen. Despite its simplicity, BitFit has demonstrated surprising effectiveness in a variety of NLP tasks, achieving competitive results with a minimal number of trainable parameters. Its lightweight nature makes it particularly appealing for scenarios with limited computational resources or strict deployment constraints. However, its use in computer vision, and especially in medical imaging tasks such as diabetic retinopathy (DR) detection, has not been thoroughly explored. Understanding how such minimal adaptations affect visual representations in high-stakes, class-imbalanced settings like DR classification is an important and under-investigated research direction.

Our work builds upon these advances by combining the domain-specific knowledge of BiomedCLIP with the efficiency of LoRA to create a computationally accessible approach to DR classification that maintains high performance. This combination addresses the dual challenges of model performance and resource constraints that have limited the widespread deployment of deep learning-based DR screening systems. By leveraging the biomedical knowledge encoded in BiomedCLIP's pre-trained weights, BiomedCLIP can achieve strong performance on specialized medical tasks with minimal fine-tuning.

3. Data

We use the Diabetic Retinopathy Detection dataset from Kaggle, which contains approximately 88,000 high-resolution images of retinal fundi. Due to data limitations, we used 16,000 photos in total across our train, validation, and test sets. Each image has been rated professionally by clinicians on a standard clinical scale of 0-4: 0 - No DR; 1 - Mild DR; 2 - Moderate DR; 3 - Severe DR; 4 - Proliferative DR. The dataset provides both left and right eye images for each subject, labeled with a subject ID and laterality (e.g., ”1_left.jpeg”). This comprehensive dataset represents a real-world distribution of DR cases, with the majority falling into the ”No DR” category, creating a significant class imbalance that mirrors clinical reality.

3.1. Dataset Challenges

This dataset has several challenges that make it suitable for real-world applications.

Images come from different models and types of cameras, resulting in large variability in visual appearance, contrast, brightness, and color balance.

Some images follow anatomical orientation (macula on the left, optic nerve on the right for the right eye), while others are inverted (as seen through a microscope condensing lens during live examinations). This can be identified by either the position of the macula relative to the optic nerve or the presence of notches (square, triangle, or circle) on the side of the image that determines the orientation.

Like most real-world medical imaging datasets, the images contain many artifacts and quality issues; some may be out of focus, underexposed, overexposed, or contain other visual noise such as dust on the lens or eyelash artifacts.

There is significant class imbalance, with the majority of images falling into the ”No DR” category, which complicates the learning process for more severe cases. This imbalance reflects the real-world distribution of DR severity but poses challenges for model training.

Images differ in their field of view, with some capturing a wider area of the retina and others focusing more narrowly on specific regions.

3.2. Preprocessing Pipeline

To address these challenges, we implemented a custom RetinopathyDataset class that handles the loading, preprocessing, and batching of the images. Our preprocessing pipeline includes several key steps:

Color Correction: BGR to RGB conversion to normalize color representation across different camera systems.

Resizing: All images are resized to the standard 224 × 224 resolution required by BiomedCLIP's vision encoder.

Contrast Enhancement: We apply a combination of Gaussian blur and weighted addition to improve visibility of subtle retinal features, particularly microaneurysms and small hemorrhages that are critical for early DR detection.

Normalization: Images are normalized to the [0, 1] range for compatibility with deep learning frameworks and to standardize input distribution.

Data Augmentation: During training, we apply various data augmentation techniques including contrast adjustments and Gaussian blurring.

4. Methods

4.1. BiomedCLIP Baseline

We employ BiomedCLIP (14) as our baseline, a biomedical, foundation, vision-language model (VLM), trained on 15 million figure-caption pairs of diverse biomedical image types (i.e., radiography, histology), from PubMed, capable of performing a multitude of vision-language processing (VLP) tasks, including image classification. BiomedCLIP leverages PubMedBERT as the text encoder and Vision Transformer as its image encoder.

To enable BiomedCLIP to perform zero-shot DR classification, we designed four prompt sets encapsulating varying levels of domain specificity and linguistic detail, corresponding to the five DR severity levels: **basic**, **technical**, **clinical**, and **simple**. Technical and clinical prompts were derived from manually parsing medical literature (9; 11).

Set	Class 0 Prompt
Basic	This is a retinal image showing no diabetic retinopathy
Technical	Grade 0 DR: normal retina with no abnormalities
Clinical	Normal fundus: clear vessels, normal optic disc, no pathologies
Simple	healthy retina

Table 1. Class 0 prompts across prompt sets.

Each prompt was tokenized with BiomedCLIP's tokenizer and normalized, resulting in unit length embeddings to support cosine similarity computation with image embeddings. Image embeddings were computed in correspondence with our custom preprocessing pipeline because we wanted to precisely control image augmentation and normalization using torchvision transforms.

To assign class predictions, we compute cosine similarity between each embedded image and embedded text class vectors within each prompt set. Similarity was scaled by the learned logit scale from the BiomedCLIP model, then passed through softmax to obtain a probability distribution over the five classes. The class with the highest similarity score was selected as the predicted label, reflecting the most semantically aligned diagnosis based on BiomedCLIP's shared vision-language embedding space.

Training Configuration. While BiomedCLIP is primarily evaluated in a zero-shot setting, we fine-tuned LoRA and BitFit variants using a custom training loop. For evaluation, we used a batch size of 64 and automatic mixed precision. For LoRA and BitFit tuning, models were trained for 1 epoch with a batch size of 32 using AdamW optimizer with a learning rate of 1×10^{-5}. The logit scale used in cosine similarity was fixed to 1.0 during training and evaluation. Peak GPU memory usage and inference time per image were measured using PyTorch's CUDA profiling tools and `psutil` on CPU fallback.

4.2. Implementing Parameter-Efficient Fine-Tuning

4.2.1 BiomedCLIP with LoRA

Low-Rank Adaptation (LoRA) inserts trainable low-rank matrices into existing linear layers of a neural network, allowing efficient adaptation without modifying the full weight matrices. Instead of updating the large pre-trained matrices, LoRA learns two smaller matrices that approximate the desired weight updates through a low-rank decomposition. This significantly reduces the number of trainable parameters and makes training more resource-efficient. Below is a diagram of how LoRA works.

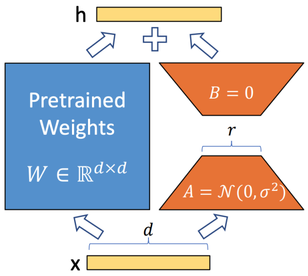

h
Pretrained Weights
$W \in \mathbb{R}^{d\times d}$
$B = 0$
r
$A = \mathcal{N}(0, \sigma^2)$
d
x

Figure 2. LoRA, adapted from (5)

For BiomedCLIP, we injected LoRA modules into the `attn.qkv` and `attn.proj` layers of the first 12 transformer blocks in the visual trunk. These components form the core of the self-attention mechanism, which is important for extracting visual patterns from retinal imagery. Modifying only these attention submodules lets us specialize the model to retinal imagery without getting in the way of the broader mechanism that BiomedCLIP helps encode.

We chose this configuration because prior work shows LoRA is especially effective when applied to attention modules, and early and middle layers often capture domain-specific low-to-mid-level visual features relevant for medical tasks.

LoRA Hyperparameters:

- LoRA rank (`r`): 16
- LoRA scaling factor (`alpha`): 32
- Target modules: `attn.qkv`, `attn.proj`
- Dropout: 0.05
- Optimizer: AdamW, learning rate: 1×10^{-5}
- Epochs: 1
- Batch size: 32

4.2.2 BiomedCLIP with BitFit

BitFit (Bias Term Fine-Tuning) is a minimalistic fine-tuning strategy where only the bias terms in the model are updated, and all other weights are kept frozen. This drastically reduces the number of trainable parameters (typically to less than 1%) while still enabling strong task-specific adaptation.

In our implementation, we changed all bias parameters across the BiomedCLIP transformer using this rule: `param.requires_grad = ".bias" in name`. This updated the biases in both linear and normalization layers across the transformer blocks.

BitFit served as a lightweight benchmark to test whether even minimal adaptation could improve over zero-shot prompting. It is particularly well-suited for deployment scenarios with stringent memory or computation constraints, such as edge devices in underresourced clinics.

BitFit Hyperparameters:

- Trainable parameters: Only bias terms
- Optimizer: AdamW, learning rate: 1×10^{-5}
- Epochs: 1
- Batch size: 32

Evaluation was conducted on a held-out dataset of 8,407 labeled retinal images using a batch size of 64. Both inference time per image and peak GPU memory usage were measured and compared across BiomedCLIP, LoRA, and BitFit variants. Training was done on an equivalently-sized dataset, and cross-validation in batches of 64 images was done.

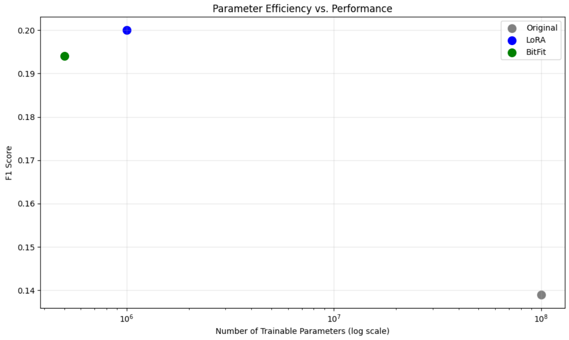
5. Experiments and Analysis

5.1. Performance Summary and Model Behavior

Both LoRA and BitFit show clear improvements over the zero-shot BiomedCLIP baseline in terms of raw classification accuracy. LoRA's performance was distinguished by its superior balance across metrics, particularly its F1 score and precision, which are crucial in the medical domain where both false positives and false negatives carry significant risk. Specifically, LoRA achieved an average accuracy of 92.7%, precision of 21.9%, recall of 18.4%, and the highest F1 score at 20.0%, making it the most balanced performer. Its selective adaptation of the `attn.qkv` and `attn.proj` modules appears to equip the model with an enhanced ability to discern finer distinctions in pathology, such as differentiating microaneurysms (mild DR) from hemorrhages and venous beading (moderate to severe DR). The use of rank-16 low-rank matrices allowed sufficient representational flexibility without incurring a high computational cost. Although its Cohen's kappa was relatively low (0.007), this is partially expected in highly imbalanced multi-class problems, and still marked an improvement in calibrated agreement beyond chance compared to other methods.

BitFit, on the other hand, achieved slightly higher overall accuracy (93.7%, the highest among all methods) but exhibited stark over reliance on the majority class . Since it only allows bias terms to be updated, the model likely adjusts its outputs through coarse-grain shifts in layer outputs rather than changing internal representations. While this can improve surface-level performance on imbalanced datasets, it falls short in sensitivity to rare classes: a critical failure in clinical settings. BitFit's recall was 20.0%, and its precision slightly lower than LoRA's at 19.2%. The resulting F1 score of 19.4% is solid, but still behind LoRA. Most notably, its kappa score dropped to 0.002, revealing a tendency to inflate accuracy by overpredicting the majority class—namely "No DR"—without improving interclass reliability. This behavior shows a larger theme: not all parameter-efficient fine-tuning strategies offer clinically efficient trade-offs. BitFit may be more appropriate in tasks with high label homogeneity but underperforms in nuanced multi-class settings like DR classification.

To better illustrate the trade-off between adaptation performance and parameter efficiency, we plot the F1 score of each model against the number of trainable parameters on a logarithmic scale. This visualization clearly demonstrates that LoRA offers the best balance, achieving the highest F1 score with significantly fewer trainable parameters than the original BiomedCLIP. BitFit, while more efficient in parameter count, lags slightly in F1 score, highlighting the cost of its minimalism. The original model, despite its size, performs worst in terms of both F1 score and parameter efficiency.



Prompt-wise Performance Comparison. Given that our models rely on text-image alignment for classification, we evaluate how different prompt formulations influence downstream performance. We tested each model across four prompt sets—**basic**, **technical**, **clinical**, and **simple**—each representing a different linguistic register or domain specificity:

- **Basic**: general language, e.g., "retinal image showing mild diabetic retinopathy"
- **Technical**: ophthalmology-style grades, e.g., "Grade 2 diabetic retinopathy..."
- **Clinical**: includes anatomical details, e.g., "dot/blot hemorrhages..."
- **Simple**: minimal phrasing, e.g., "retina with mild diabetic retinopathy"

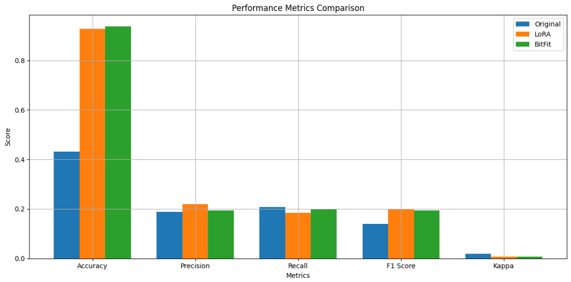
Original BiomedCLIP: The zero-shot baseline was highly sensitive to prompt wording. Accuracy varied from **16.5% (basic)** to **72.0% (technical)**, indicating that complex prompts can partially activate relevant latent knowledge, even without fine-tuning. However, class-specific recall remained poor across all prompts, with **0.0% sensitivity for Class 2 and above**.

LoRA: This model demonstrated relative stability across prompts, with accuracies clustered tightly: **92.7% (simple)** to **92.8% (technical)**. However, we observed a slight improvement in minority class sensitivity under technical and clinical prompts—for example, Class 1 recall improved to **1.5%** with the technical set compared to **0.5%** with simple phrasing. This suggests that LoRA's updated attention projections are responsive to semantically richer text inputs.

BitFit: In contrast, BitFit was almost entirely prompt-invariant. Accuracy ranged narrowly between **93.6% and**

93.7% across all prompt sets. However, this was a consequence of degeneracy: **97.5% of examples were predicted as Class 0**, regardless of prompt specificity. This implies that BitFit disregards prompt nuance and learns a coarse default response based on majority class correlation.

Summary: LoRA's interaction with prompt phrasing suggests that there is value in exploring prompt optimization alongside model adaptation. It may be worthwhile to co-train prompt encoders or use reinforcement learning to select effective prompt templates dynamically. BitFit's flat response highlights the limitations of shallow fine-tuning when the classification task requires nuanced text-image grounding.



5.2. Failure Modes and Error Patterns

Each model had distinct failure modes that show their respective architectural constraints. The Original BiomedCLIP model, when used as a zero-shot, frequently defaulted to high-confidence predictions for the "No DR" or "Mild DR" categories, especially under the simpler prompt sets. This is likely because the pretrained CLIP-style architecture was not exposed to images with DR-specific subtleties during pretraining. Without fine-tuning, it may rely on broad text-image alignment priors rather than task-specific visual grounding. This failure mode manifested as near-zero recall for Proliferative DR in all prompt variants. Quantitatively, the original model achieved 54.6% recall for Class 0, 44.8% for Class 1, but only 1.35%, 2.71%, and 0.0% for Classes 3, 4, and 2 respectively. These patterns reveal that without adaptation, the model is not capable of picking up subtle or severe pathologies.

In LoRA, we saw "soft misclassification" failures. For example, images of moderate DR were occasionally misclassified as mild or severe DR but rarely as "No DR" or Proliferative. This suggests that LoRA enabled some level of intra-class boundary refinement but did not fully resolve edge cases. While treatment plans may vary largely between different forms of DR, this is more clinically desirable than a larger misclassification; misclassification to an adjacent stage of DR may mean that, with a little more training, the model may be nearly deployable. LoRA achieved 94.5% recall for Class 0, but crucially also non-zero recall for Class 1 (1.50%), Class 2 (1.00%), and Class 3 (0.50%)—a sharp improvement in distribution over the original model and a testament to its ability to generalize beyond dominant patterns.

BitFit had a different and more concerning failure pattern. Despite its high overall accuracy, its per-class performance showed a large collapse into the majority class, particularly under prompts that lacked highly discriminative wording. Essentially, BitFit learned a degenerate solution: predict "No DR" for most examples. While this inflated its recall and accuracy scores for Class 0 (97.5%), it effectively rendered the model unusable for practical screening, where sensitivity to rare but vision-threatening cases is paramount. All other classes had recall near zero: Class 1: 0.25%, Classes 2–4: 0.00%.

5.3. Discussion and Interpretability

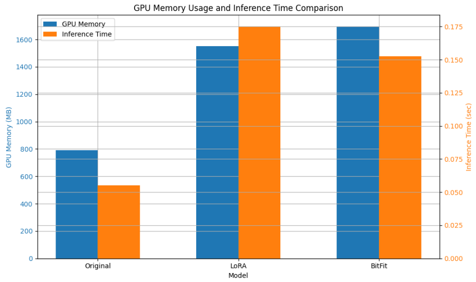
The performance and failure profiles of these models reflect not only differences in trainable parameters but also differences in how architectural constraints affect training, results, and how models adapt.

In LoRA, the learnable low-rank matrices directly modify the projection space of the self-attention mechanism. This enables LoRA to "bend" the learned attention distribution toward underrepresented features such as small dot hemorrhages or early neovascularization. It also introduces regularization through structural bottlenecks: by limiting updates to low-dimensional subspaces, LoRA likely avoids overfitting to class imbalance, which helps explain its stronger generalization to Classes 1–3. Future work could explore whether extending LoRA to adapt MLP layers in addition to attention layers would further improve performance, especially for ambiguous or borderline cases.

BitFit's failure to generalize can be understood by considering the role of biases in transformer architectures. Bias terms modulate activations in a uniform, layer-wise manner, without spatial or semantic specificity. In a vision-language model like BiomedCLIP, which processes complex retinal features through multi-headed attention, this kind of parameter nudging may be insufficient for learning meaningful pathology-specific cues. A promising direction for future work could be "BitFit++", which allows biases in key transformer blocks and normalization layers to be fine-tuned, or even includes optional per-layer scaling factors to provide greater control without sacrificing simplicity.

Another emergent insight deals with prompt set sensitivity. As discussed earlier, LoRA and the Original model were both significantly affected by prompt phrasing, whereas BitFit's performance was almost entirely invariant—a sign that it had converged on a one-class solution. LoRA, on the other hand, appeared to respond to richer prompts (e.g., technical and clinical), suggesting its attention updates leveraged the detailed semantics of the input text. This opens a rich direction for future work in co-optimizing prompts and LoRA layers jointly. Moreover, it signals a potential path toward more interpretable and controllable clinical AI systems, where the language component can act as an additional lever for tailoring model behavior.

Finally, from a systems perspective, LoRA's high performance on underrepresented classes in context of GPU memory footprint, makes it the most viable adaptation for deployment in under resourced clinical settings. The LoRA model required ∼1550 MB of memory at inference compared to ∼1693 MB for BitFit, and ∼792 MB for the original model. Though the difference between LoRA and BitFit in terms of memory use seems relatively small, when we scale and deploy models for widespread use in the context of global DR detection, this difference will scale to large amounts of compute. Even though LoRA also requires a lot more memory than the original model, LoRA performed significantly better in terms of test set accuracy and combatted class imbalance better. Inference time followed a similar trend: LoRA averaged 0.14s, BitFit 0.12s, and the original model just 0.05s. Even though LoRA has a higher inference time per image compared to BitFit and aa much higher inference time per image ompared to the original model, we believe the smaller GPU use compared to BitFit and the much higher accuracy compared to the original model outweighs this. LoRA also can be directly injected in the attention layers, while BitFit only adjusts the bias terms. Pretraining can help combat LoRA's relatively large GPU use and inference time per image. These trade-offs underscore that LoRA achieves superior performance without imposing substantial runtime costs.



6. Conclusion

Our study shows the promise of combining foundation models like BiomedCLIP with parameter-efficient fine-tuning techniques for diabetic retinopathy classification in resource-constrained settings. LoRA and BitFit, two lightweight adaptation methods, showed distinct performances when applied to this complex multi-class medical task. LoRA, by injecting low-rank matrices directly into the attention layers, achieved a notable balance between overall accuracy and sensitivity to minority classes. BitFit, while delivering high top-line accuracy, suffered from class collapse, exposing the limitations of overly simplistic adaptations in high-stakes, imbalanced domains like medical imaging. As discussed earlier, BitFit also suffers from not being able to directly modify the internal structure of the model; it can only affect the bias terms and not the weights themselves.

A key insight from our analysis is that not all PEFT strategies are created equal, and their usefulness is task-dependent. The subtle improvements in class-specific recall achieved by LoRA highlight the importance of adapting internal attention mechanisms when fine-grained classification is required. In contrast, BitFit's reliance on bias-only updates proved insufficient for modeling nuanced retinal pathologies, suggesting a need for richer parameter updates or hybrid strategies in future work.

Interestingly, despite their intended efficiency, both LoRA and BitFit required more memory and inference time per image than the original zero-shot BiomedCLIP model. LoRA consumed approximately 1550MB of GPU memory and averaged 0.14 seconds per image, while BitFit required 1693MB and 0.12 seconds. In contrast, the original BiomedCLIP used only 792MB and took 0.05 seconds per image. This highlights a crucial distinction between fine-tuning efficiency and deployment efficiency, suggesting the need for inference-optimized PEFT implementations for low-resource deployments.

In summary, our work offers a scalable, interpretable, and efficient structure for DR screening that aligns with the global need for widespread diagnostic tools. The nuances in performance differences we observed between adaptation methods and prompt styles emphasize the importance of model design that is closely tied to the context of the problem (in this case, ophthalmology and medical diagnostics). With more time and refinement, approaches like LoRA-tuned BiomedCLIP could support early detection at population scale, helping prevent avoidable blindness for millions worldwide.

7. Contributions

Adi and Kevina split up work in the project evenly. Kevina implemented, trained, and tested the BiomedCLIP baseline including preprocessing and embedding of images. Adi did the data integration, LoRA implementation, training, and testing, and BitFit implementation, training, and testing.

References

[1] M. D. Abràmoff, P. T. Lavin, M. Birch, N. Shah, and J. C. Folk. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *npj Digital Medicine*, 1(1):39, 2018. 2

[2] L. Dai, L. Wu, H. Li, C. Cai, Q. Wu, H. Kong, R. Liu, X. Wang, X. Hou, Y. Liu, X. Long, Y. Wen, L. Lu, Y. Shen, Y. Chen, D. Shen, X. Yang, H. Zou, B. Sheng, and W. Jia. A deep learning system for detecting diabetic retinopa-

thy across the disease spectrum. *Nature Communications*, 12(1):3242, 2021. 1

[3] R. Gargeya and T. Leng. Automated identification of diabetic retinopathy using deep learning. *Ophthalmology*, 124(7):962–969, 2017. 2

[4] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, R. Kim, R. Raman, P. C. Nelson, J. L. Mega, and D. R. Webster. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Journal of the American Medical Association*, 316(22):2402–2410, 2016. 1, 2

[5] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022. 2, 4

[6] J. Krause, V. Gulshan, E. Rahimy, P. Karth, K. Widner, G. S. Corrado, L. Peng, and D. R. Webster. Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. *Ophthalmology*, 125(8):1264–1272, 2018. 2

[7] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 8748–8763, 2021. 2

[8] Z. Teo, Y. Tham, M. Yu, M. L. Chee, T. H. Rim, N. Cheung, M. M. Bikbov, Y. X. Wang, Y. Tang, Y. Lu, I. Y. Wong, D. S. W. Ting, G. S. W. Tan, J. B. Jonas, C. Sabanayagam, T. Y. Wong, and C.-Y. Cheng. Global prevalence of diabetic retinopathy and projection of burden through 2045: Systematic review and meta-analysis. *Ophthalmology*, 128(11):1580–1591, 2021. 1

[9] X. Wang, Y. Lu, Y. Wang, and W.-B. Chen. Diabetic retinopathy stage classification using convolutional neural networks. pages 465–471, 07 2018. 3

[10] X. Wang, Y. Peng, L. Lu, Z. Lu, and R. M. Summers. Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9049–9058, 2020. 2

[11] C. P. Wilkinson, F. L. Ferris, R. E. Klein, P. P. Lee, C. D. Agardh, M. Davis, D. D. Dills, R. C. Kampik, E. M. Pararajasegaram, G. Verdaguer, and the Global Diabetic Retinopathy Project Group. Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. *Ophthalmology*, 110(9):1677–1682, 2003. 3

[12] E. B. Zaken, Y. Goldberg, and S. Ravfogel. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1–9, Dublin, Ireland, May 2022. Association for Computational Linguistics. 2

[13] P. Zhang, X. Wang, X. Wang, L. Lu, Z. Lu, and R. M. Summers. MedCLIP: Contrastive learning from unpaired medical images and text. *arXiv preprint arXiv:2210.10163*, 2022. 2

[14] S. Zhang, Y. Xu, N. Usuyama, H. Xu, J. Bagga, R. Tinn, S. Preston, R. Rao, M. Wei, N. Valluri, C. Wong, A. Tupini, Y. Wang, M. Mazzola, S. Shukla, L. Liden, J. Gao, A. Crabtree, B. Piening, C. Bifulco, M. P. Lungren, T. Naumann, S. Wang, and H. Poon. A multimodal biomedical foundation model trained from fifteen million image–text pairs. *NEJM AI*, 2(1), 2024. 3

[15] Z. Zhang, M. Cui, W. S. El-Sayed, T. Liang, M. Liu, W. Fang, and T. Ren. BiomedCLIP: A foundation model for multimodal medical imaging and text. *arXiv preprint arXiv:2303.00915*, 2023. 2