Bottled Brilliance: Gated Mixture of Experts for Biomedical Explainability

Sidharrth Nagappan
sn666@cam.ac.uk
University of Cambridge

Abstract

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim aeque doleamus animo, cum corpore dolemus, fieri tamen permagna accessio potest, si aliquod aeternum et infinitum impendere malum nobis opinemur. Quod idem licet transferre in voluptatem, ut postea variari voluptas distinguique possit.

Keywords Biomedical Explainability · CLIP · BioMedCLIP

1 Introduction

The adoption of deep learning in computational medicine, particularly in medical image classification, is often hampered by the notion of transparency. Black-box models provide predictions without clarifying the semantic features that drive these decisions, making them difficult to trust in critical environments. Concept Bottleneck Models (CBMs) emerged as an interpretable alternative, that incentivised models to route their decisions through a set of human-understandable concepts. However, these conventional CBMs relied on human annotations to mark those concepts in the first place. Language in a Bottle (LaBO) addressed these challenges by automating concept discovery using Large Language Models (LLMs). These concepts were then aligned to images using pre-trained vision-language models such as CLIP, allowing the formation of a concept bottleneck layer.

However, the end-to-end architecture depends on the richness of CLIP's neural representations, whose wide training-base lacks domain-specific grounding. While general visual representations (such as colour and shape) are learnt, they may overlook subtle morphological cues that are essential for deeply nuanced decisions. In contrast, domain-specific models such as BioMedCLIP, trained on scientific literature and imagery, possess specialist knowledge but may lack the broader visual diversity of CLIP.

In this mini-project, an extension of the LaBO framework that uses both CLIP and BioMedCLIP as complementary experts, is explored. Specifically, this is framed as a mixture-of-experts (MoE) problem, where CLIP is the generalist expert and BioMedCLIP is the specialist expert, and a learned gating network determines the relative contribution of each for every input image. The motivation is that different skin lesion may benefit from generalist knowledge (e.g. shape, colour patterns) or specialist biomedical cues (e.g. vascular structure, lesion-specific terminology) to varying degrees. A dynamic gating mechanism allows the model to adapatively

leverage either or both experts on a per-instance basis, improving flexibility, accuracy, and interpretability.

1. Does domain expertise improve interpretability and classification performance in biomedical image analysis?
2. Can a mixture-of-experts fuse information from these disparate concept space, and does this learned fusion outperform their uni-expert counterparts?

... *write what happened*

2 Related Work

3 Method

3.1 Biomedical Data

HAM10000 is a collection of 10,015 dermatoscopic images representing seven variations of skin lesions[1] that are compiled from various populations [1], and is commonly used as a benchmark dataset for medical vision encoders. We use the same training, validation and testing splits as the dataset providers.

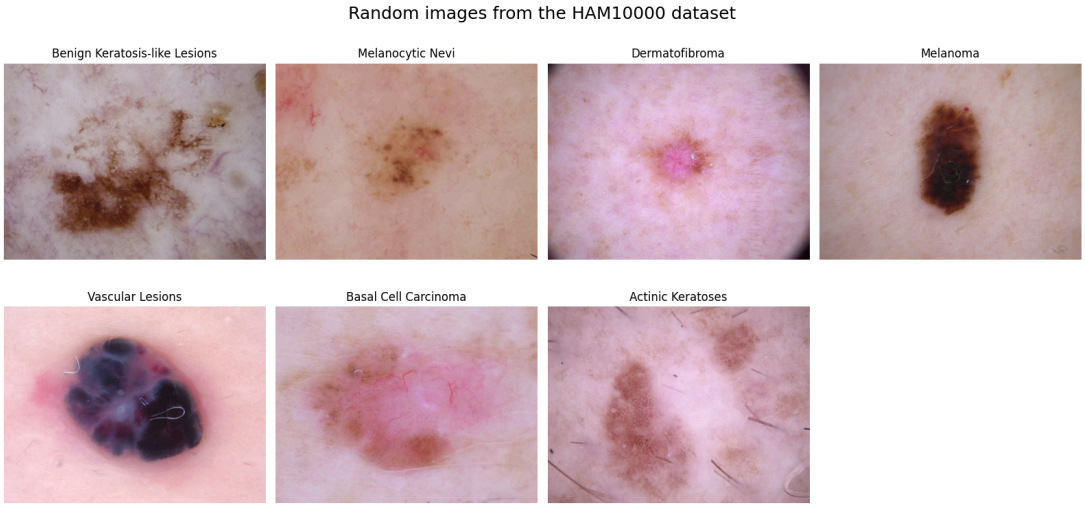

Figure 1: Sample Images from the HAM10000 dataset

3.2 A Representative Bottleneck

LaBO employed sentence parsing using a T5 to extract semantic concepts from LLM-generated sentences [2]. We conjecture that this approach is suboptimal, leads to information loss and the quality of the final model is dependent on the accuracy of the trained parsing model. Instead, we propose enforcing JSON structure via Pydantic in prompts we send to our LLM suite (LLAMA, DeepSeek, Meditron and OpenAI's 4o), directly extracting phrasal concepts without intermediate parsing [3, 4].

[1] melanoma, basal cell carcinoma, and benign keratosis-like lesions

Concepts generated for the generalist are augmented with the phrase: "You can be a bit technical."[2] Our enhanced prompt engineering outperforms the manual parsing algorithm, as shown in Table 2

Technical Prompt Generation

Background: Extract the concepts from the class to be used for dermatoscopic images. *You can be slightly technical when generating the concepts.*

Prompt: Describe the {*feature*} of the {*disease*} disease in HAM10000 that can be used as visual concepts for Skin Cancer classification.

3.3 Multi-Expert Submodular Optimisation

Submodular optimisation is used to select a discriminative set of concepts that maximise coverage of class semantics while minimizing redundancy. Specifically, we define a set function $f(S) = \alpha \cdot \text{coverage}(S) - \beta \cdot \text{redundancy}(S)$, and select the subset $S \subseteq C$ of concepts by approximately maximizing $f(S)$ via a greedy algorithm. As an improvement to the original algorithm:

1. We incorporate CLIP embeddings into the selection process to account for global similarity. This also implicitly makes sure that only textual concepts that are semantically understood by the VLM are part of the final selection. We modify this architecture for the mixture-of-experts scenario, doing expert-wise bottleneck maintenance.
2. Concepts are stemmed and filtered to remove those containing morphological variants of class name tokens, reducing concept leakage.

This mechanism proves particularly valuable for advanced biomedical terminology — such as *telangiectasia*, *ovoid* or *keratinization* — which are well-represented in BioMedCLIP's domain corpus but may not be meaningfully encoded by CLIP. By filtering concepts through this embedding-informed scoring process, we obtain a lean and discriminative concept set that adapts to each expert model. As seen in the example concept list, the generalist leans towards visual descriptors, while the specialist uses very specific terminology, whose visual context is implicitly encoded due to the training corpus[3]. It is unsurprising that common words such as "presence", "brown", "areas" and "pigmentation" are widely used in both corpora[4].

Global Specialist Concepts:	**Global Generalist Concepts:**
• Keratinization patterns	• Crusty texture
• Erythematous base	• Small diameter
• Focal Nodularity	• Pink
• Multilobular pattern	• Light brown

During MoE, we freeze the concept selection bottlenecks, and use those selected during their corresponding uni-expert training cycles. This allows for fair comparisons and ensures that the data distribution is the only independent factor.

[2] After several rounds of prompt engineering, this produced the best results.
[3] e.g. "Keratinization is defined as cytoplasmic events that take place"
[4] distributions computed using the nltk toolkits and the CLIP similarity scores

3.4 Mixture-of-Experts

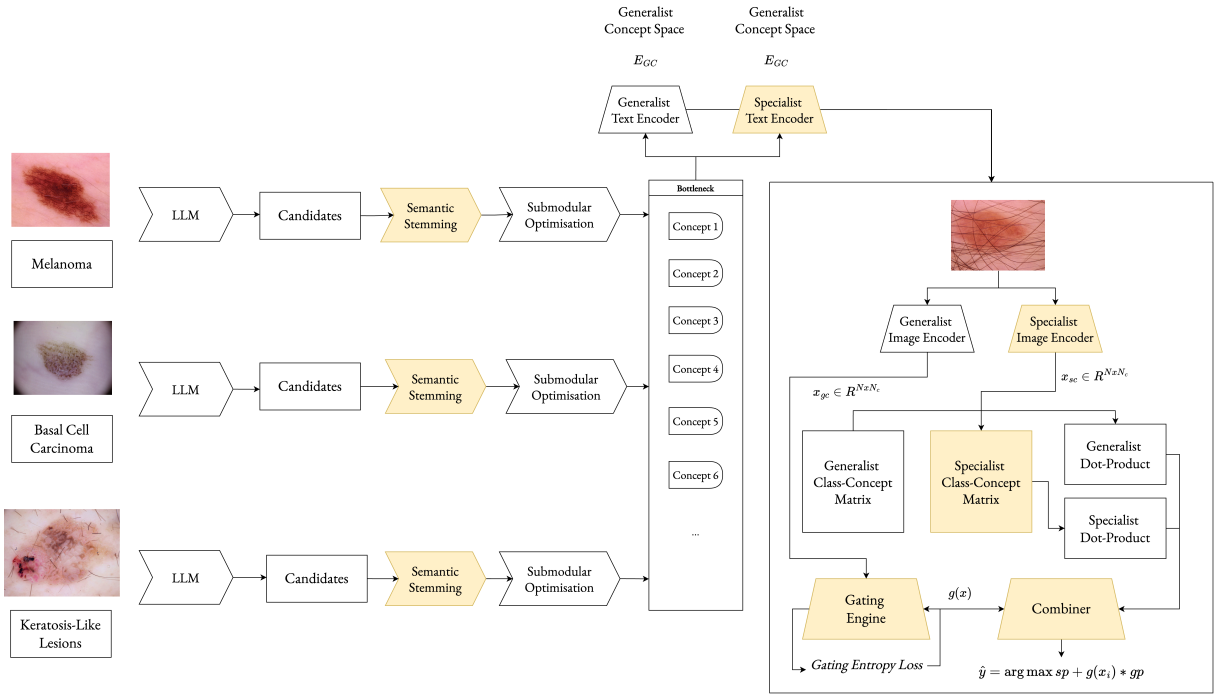

Figure 2: Bottled Brilliance Neural Architecture. Additions to LaBO are highlighted in yellow.

3.4.1 The Case for Expertise

As seen in the selection process, some lesions are distinguishable by general visual attributes like colour patterns and asymmetry, which CLIP catures well. However, we hypothesise that there may be useful biomedical descriptors, that lend a help when attempting to interpret decisions. For instance, a "multilobular pattern" has a distinct morphological shape, that when presented to an ordinary person, would appear obscure, but could highlight a clear analogy to a medical professional who would assess that analogy to build a deeper understanding of the diagnosis. The elucidated scenario comprises of a specialist audience, and a "common man" generalist. When representing this neurally, we find that it is reminiscent of the Mixture-of-Experts architecture.

3.4.2 The Experts

CLIP is the choice architecture in LaBO; it learns a transferable visual representation by contrastively training image and text encoders on 400 million (image, text) pairs. It's wide training corpus and general understanding of worldly knowledge makes it a suitable candidate for the generalist. BioMedCLIP is a multimodal biomedical foundation model, trained on PMC-15M, a dataset containing 15 million biomedical image-text pairs that are taken from scientific articles in PubMed Central (PMC). The corpus taxonomy includes dermatology photos, microscopy, histopathology and X-Rays.

To ensure fair comparison, we standardise the architecture by using ViT-B/16 for both experts, instead of the ViT-L/14 used in LaBO. While ViT-L/14 outperforms the base variant, large-scale BioMedCLIP models are not publicly available; however, scaling laws suggest that performance would improve proportionally across all variants [5].

3.4.3 Formulation

Our Gated Mixture-of-Experts approach combines similarity embeddings from both CLIP (E_C) and BioMedCLIP (E_B). Our approach uses precomputed image-to-concept dot products from each expert, and learns a concept-to-class association matrix for both. Formally, given an input image vector x_i, we pre-compute the generalist and specialist dot products based on their image and concept vectors:

$$\left\{D^g \in \mathbb{R}^{B\times m_g}, D^s \in \mathbb{R}^{B\times m_s}\right\} \tag{1}$$

where m_g and m_s denote the number of generalist and specialist concepts respectively. $A^g \in \mathbb{R}^{K\times m_g}$ and $A^s \in \mathbb{R}^{K\times m_s}$ are learnable association matrices that map concepts to class logits. Following the original paper, these associations are initialised with language model priors. Class-level predictions from each expert are computed as $S^g = D^g \times (A^g)^T$ and $S^s = D^s \times (A^s)^T$.

The gating network, tuned to inhibit over-parametrisation, is a two-layer neural network with a *LeakyReLU* activation and sigmoid output, defined as:

$$g(x_i) = \sigma(W_2(\text{LeakyReLU}(W_1(\text{LayerNorm}(x_i))))) \tag{2}$$

$g(x_i) \in [0, 1]$ dynamically determines the cross-expert weighting for each input:

$$S_i = g(x_i) \cdot S_i^s + (1 - g(x_i)) \cdot S_i^g \tag{3}$$

The Gated MoE model is trained by minimizing a total loss that consists of a classification loss (cross-entropy for single-label and binary cross-entropy for multi-label) and a gate entropy loss, that encourages the gating network to avoid overly deterministic decisions and completely depend on one of the experts (by collapsing $g(x_i) \to \{0, 1\}$). Additional regularisers are added to encourage prediction diversity (disincentivize model from collapsing similarity scores) and sparse concept-to-class activations[5]. In the results section, we share an ablation between different loss combinations.

$$\mathcal{L} = \text{CrossEntropy}(S, y) + \lambda_{\{\div\}} \cdot \left(-E_{\{i\}}\left[\text{Var}_{\{k\}}\left(S_{\{i,k\}}\right)\right]\right) + \lambda_{\{\text{L1}\}} \cdot (\|A^g\|_1 + \|A^s\|_1) \tag{4}$$

3.5 Experimental Infrastructure

All experiments were run on a single NVIDIA L40S GPU in the Department of Computer Science's GPU server, while Weights and Biases is used for experimental tracking [6]. We run all few-shot models for a maximum of 5000 epochs, while restricting fully supervised models to 1500 epochs[6]. To tackle the cold start issue for noisy gate parameters, we allow a 500 epoch warm start, where the only trainable parameters are the gate.

4 Results

4.1 Fully Supervised Baselines

We first evaluate the models in the complete models in a fully supervised setting. The original paper uses `ViT-L/14` in their architecture. However, for the sake of fair comparisons, we

[5] the original paper did not employ these losses in their final ablations, so we replicate those same decisions

[6] tuned to prevent overfitting where the training accuracy can quickly hit 100% due to under-parametrisation

Model Variant	Val. Acc. (%)	Val. Loss	Test Acc. (%)	Test Loss
ViT-B/16	79.1	0.5501	76.8	0.6126
BioMedCLIP	77.3	0.69656	75.03	0.7916
MoE	79.2	6.1445	**77.21**	0.8235

Table 1: Individual Model Performance Results

Under fully supervised conditions, standard CLIP still outperforms BioMedCLIP, likely because ample supervision allows for a sufficiently comprehensive representation—lessening the advantage of specialized domain expertise. The best performance is attained by the Mixture of Experts (MoE), outperforming both uni-expert counterparts by $\approx 2.85\%$. However, the elevated loss suggests that, while the model achieves strong accuracy, its occasional errors are highly confident misclassifications — potentially exacerbated by the gating mechanism's sharp routing decisions. When analysing the average gate distribution, we find that the gate's $g(x_i)$ begins to fluctuate before reaching a batch-wise average of 0.8 by the final training epoch. This shows increased dependence on the specialist in the majority of cases.

4.2 Few-Shot Learning

Architecture	Validation Accuracy					Test Accuracy				
	1-shot	2-shot	4-shot	8-shot	16-shot	1-shot	2-shot	4-shot	8-shot	16-shot
G (PC)	23.0	35.9	24.4	34.5	54.6	22.4	38.1	23.4	31.9	52.6
S	25.0	38.8	**49.4**	**63.0**	52.8	27.2	40.0	**48.8**	**61.7**	52.3
MoE	31.4	46.4	24.8	43.9	46.4	28.1	48.1	27.3	42.9	45.1
$\text{MoE}_{\text{entropy}}$	**48.2**	**49.4**	24.8	34.6	**54.6**	**45.8**	**50.2**	23.5	34.2	**52.8**

Table 2: Shot-by-Shot Results (in %) -

The specialist expert outperforms its counterparts by a substantial margin in ultra-low-shot settings (1-2 shots), showing that domain-specific knowledge provides strong performance boosts when there is minimal labelled data. However, under high-supervision and full-supervision, the specialist advantage diminishes, with the generalist outperforming it as soon as there is enough data to learn broad visual features and enrich the association matrix.

Interestingly, the Mixture-of-Experts excels at ultra-low-shot scenarios, where MoE partial insights are fused from both experts to provide surprisingly impressive performance ($> 70\%$) improvement. It must however be noted that the weights learned by gate may be suboptimal when there isn't sufficient supervision to enrich weighting decisions; in early epochs, the average gate weight $g(x_i)$ edges wavers $0.4 \leftrightarrow 0.6$ Under mid-range supervision, the results do not consistently favour MoE, since partial supervision is likely insufficient for the gate to converge on an optimal blend, adversely hurting performance instead.

A general observation is that the mixture-of-experts lacks sufficient supervision in few-shot settings to train a sufficiently rich gate, with the full benefits only visible during full supervision.

4.2.1 Intuition about Foundation Gates

One solution to this problem would be to train the gate on a tangentially similar task and port the weights, so the starting representation would have some intuition about which types

of images would be suitable for which expert. However, this would require clear cross-task alignment, and there is little formal proof to support this conjecture.

4.3 Expanding to NIH X-Rays

Linear Probe:

Val Accuracy: 89.883, Val std: 0.000 /home/sn666/.conda/envs/labo/lib/python3.9/site-packages/sklearn/metrics/*classification.py:1565: UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no true nor predicted samples. Use* `zero_division` *parameter to control this behavior.* warn_prf(average, modifier, f"{metric.capitalize()} is", len(result)) Test Accuracy = 88.101, Macro F1 = 0.092

4.4 Gating Fluctuations

The gate provides fascinating insights into the inner workings of the model

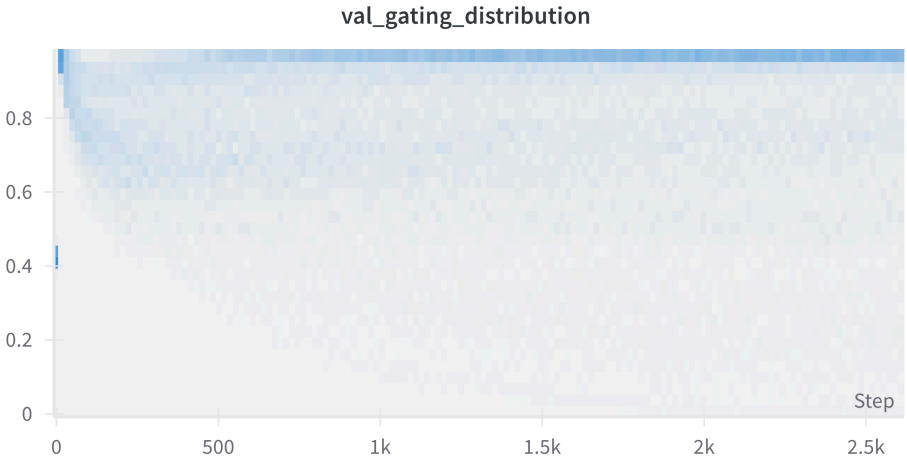

Figure 3: Validation Gating Distribution Under Full Supervision - There is clearly a collapse where the specialist expert is being almost completely relied on close to the end of training.

5 Conclusion and Limitations

- Interestingly, there is potential for leakage, since the concept's actual semantic definition can be overly synonymous in certain cases. However, since the specialist expert did not outperform the generalist under the fully supervised setting, the leakage isn't immediately obvious.

6 Notes

- Using ViT-B/16 to establish the baseline in this paper - because it outputs 512 dimensions, which is the same as MedCLIP and BioMedCLIP
- Motivation - biomedical explainability is important - do more specialised variants do a better job
- We can't directly assess the quality of the explanation, but we can implicitly assess them through the expressiveness of the concept alignment
- Hypothesis - combine generalist + specialist improve interpretability and concepts

- Try initialising association weights using `gen_init_weight_from_cls_name` – might be useful in few-shot scenario
- Some of the accuracies in the table were from the last epoch, not the best epoch - make sure to check

6.1 Research Questions

1. First, does separating concept spaces improve interpretability and classification performance?
2. Second, can learned fusion weights outperform naive averaging of similarity scores?
3. And third, does the specialist model contribute more on rare or complex conditions?

6.2 Methodology

- Run individual models - done
- Run BioMedCLIP with more specialised features
- Do hybrid gating between MedCLIP and BioMedCLIP - use different concept sets - only modify asso_opt.py file
- CLIP explainability - https://colab.research.google.com/github/hila-chefer/Transformer-MM-Explainability/blob/main/CLIP_explainability.ipynb#scrollTo=3ogYpvQAAH4s

If there's time:

- Linear probe NIH-XRay
- Apply best method from above

7 Results

Dataset	Concept Set	Variant	Shot	Val Acc	Val Loss	Test Acc	Test Loss
HAM10000	Generalist	ViT-B/16	All	0.791	0.55009	0.76915	0.61261
HAM10000	Generalist	ViT-L/14	All	0.792	0.61521	0.79900	0.61158
HAM10000	Generalist	BioMedCLIP	All	0.773	0.69656	0.75025	0.7916
HAM10000	Specialist	BioMedCLIP	All	0.7730	0.70034	0.73731	0.72475
HAM10000	MoE	ViT-B/16 + BioMedCLIP	All	sdf	dfs	0.7721	0.8235

Table 3: Individual Model Results

8

Bibliography

[1] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Scientific Data*, vol. 5, no. 1, Aug. 2018, doi: 10.1038/sdata.2018.161.

[2] C. Raffel *et al.*, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," *CoRR*, 2019, [Online]. Available: http://arxiv.org/abs/1910.10683

[3] H. Touvron *et al.*, "LLaMA: Open and Efficient Foundation Language Models." [Online]. Available: https://arxiv.org/abs/2302.13971

[4] DeepSeek-AI *et al.*, "DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning." [Online]. Available: https://arxiv.org/abs/2501.12948

[5] X. Zhai, A. Kolesnikov, N. Houlsby, and L. Beyer, "Scaling Vision Transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2022, pp. 12104–12113.

[6] L. Biewald, "Experiment Tracking with Weights and Biases." [Online]. Available: https://www.wandb.com/

APPENDIX A

A.1 Prompt Generation

A.2 Ablation Study on Mixture-of-Experts