Gated Mixture of Experts

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 First keyword · Second keyword · etc.

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 specilist 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.

... write what happened

2 Related Work

3 Method

3.1 Biomedical Dataset

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

3.2 Concept Generation

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.

3.3 The Experts

3.4 Mixture-of-Experts

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 obtain the generalist and specialist dot products:

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

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). Additional regularizers can optionally be added to encourage prediction diversity (disincentivize model from collapsing

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

similarity scores) and sparse concept-to-class activations; the original paper did not employ these losses in their final ablations, so we replicate those same decisions.

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

3.5 Experimental Infrastructure

All experiments were run on a single NVIDIA L40S GPU in the Department of Computer Science's GPU server. We run all few-shot models for a maximum of 5000 epochs, while restricting fully supervised models to 1500 epochs[2]

4 Results

5 Conclusion and Limitations

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

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

- 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 1: Individual Model Results

8 Introduction

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, augeri amplificarique non possit. At.

9 Heading: first level

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magnam aliquam quaerat.

9.1 Heading: second level

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magnam aliquam quaerat.

9.1.1 Heading: third level

Paragraph Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magnam aliquam quaerat.

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magnam aliquam quaerat.

10 Math

Inline: Let a, b, and c be the side lengths of right-angled triangle. Then, we know that: $a^2 + b^2 = c^2$

Block without numbering:

$$\sum_{k=1}^{n} k = \frac{n(n+1)}{2}$$

Block with numbering:

As shown in Equation 5.

$$\sum_{k=1}^{n} k = \frac{n(n+1)}{2} \tag{5}$$

More information:

- https://typst.app/docs/reference/math/equation/

11 Citation

You can use citations by using the `#cite` function with the key for the reference and adding a bibliography. Typst supports BibLateX and Hayagriva.

```
#bibliography("bibliography.bib")
```

Single citation [3]. Multiple citations [3, 4]. In text Vaswani A, Shazeer NM, Parmar N, et al [3]

More information:

- https://typst.app/docs/reference/meta/bibliography/
- https://typst.app/docs/reference/meta/cite/

12 Figures and Tables

header 1	header 2
cell 1	cell 2
cell 3	cell 4

Table 2: Lorem ipsum dolor sit amet.

Figure 1: Lorem ipsum dolor sit amet, consectetur adipiscing.

More information

- https://typst.app/docs/reference/meta/figure/
- https://typst.app/docs/reference/layout/table/

13 Referencing

Figure 1 Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do., Table 2.

More information:

- https://typst.app/docs/reference/meta/ref/

14 Lists

Unordered list

- Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do.
- Lorem ipsum dolor sit amet, consectetur adipiscing elit.

Numbered list

1. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do.
2. Lorem ipsum dolor sit amet, consectetur adipiscing elit.
3. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor.

More information:

- https://typst.app/docs/reference/layout/enum/
- https://typst.app/docs/reference/meta/cite/

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] A. Vaswani *et al.*, "Attention is All you Need," in *NIPS*, 2017.

[4] G. E. Hinton, O. Vinyals, and J. Dean, "Distilling the Knowledge in a Neural Network," *ArXiv*, 2015.

APPENDIX A

A.1 Appendix section

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, augeri amplificarique non possit. At etiam Athenis, ut e patre audiebam facete et urbane Stoicos irridente, statua est in quo a nobis philosophia defensa et collaudata est, cum id, quod maxime placeat, facere possimus, omnis voluptas assumenda est, omnis dolor repellendus. Temporibus autem quibusdam et.