Paper Summary

This paper demonstrates scalable and dynamic ideas for a wholesome human-artificial intelligence (AI) interacting system, especially aiding in differential diagnosis in the field of Medicine. In differential diagnosis, the medical practitioner (user) proposes a hypothesis and some alternate hypotheses which are then systematically eliminated to arrive at the final diagnosis. So, in such scenarios it is helpful to analyse patient information with similar ailments and understand their underlying contributors. The main issue with current AI models is that the results returned as a result of query (Content-Based Image Retrieval Systems (CBIR)) may be algorithmically similar but not clinically aka semantic Gap. This decreases the trust of the user in uses such a system in decision making. The paper explores some refinement techniques incorporated in the models as per user feedback and their utility and introduces a model which returns Similar Images like Yours (SMILY).

Table 1 Summary of the analysis on the proposed refinement tools

Refinement Tools	***Evaluation Method***	***Tool Management and User Experience***
Refine by Region *- Physical cropping the image to relevant area*	The user is asked to either rate the relevance of the image result or crop the image (extra caution needs to be exercised to take care of the aspect ratio of the input to the model). After refinement 88% results were desirable.	This is the only technique that needed users to work with the image. Reasonable iteration length per user.
Refine by example *- Sometimes the system inherently provides good example . Reinforce this behavior by updating the query embeddings to the embeddings averaged from the search results*	More quantitative in nature -- refinement by emphasis.	Slower iterations. Initially users were excited since it is so intuitive but eventually they couldn't assess if it indeed was an improvement, hence slower iterations and confounding user logs *"You think in pictures first then words." - User*
Refine by Concept *– Sometimes you don't want the query to return similar images instead some augmented versions to aid in proving/disproving your hypothesis . Manual sliders are provided to adjust the concept the user wants to emphasise.*	Refinement done by one pathologist and tested by giving query results to another pathologist to see if he can identify the concept prominently without ambiguity. Not obvious improvement in this case.*	Very Useful. Users requested support for more concepts. Requires some human labeled data although not a lot. System design makes overall creation of concepts dynamic and scalable

The paper discusses some interesting decision making and coping techniques with black box ML as follows:

- Reinforces the current hypothesis's likelihood
- Iterative processes allow room for ample reflection on the correctness of the thought process, thus helping generate new ideas
- Feature segmentation one by one -- helps in better focus
- Narrowing the semantic gap between the system and domain expert by refinement
- Understand the sensitivities of the ML model -- helpful in distinguishing human errors from machine-errors

Overall such refinement techniques show a lot of promise in instilling the user's trust and minimising decision making errors in such applications. Most importantly, this work is an evidence of how AI can augment domain experts and not replace them. The authors suggest such ideas can be scalable and be effectively used other application areas as well.

Things I liked about the paper

This paper was my introduction to the idea of ***Concept Activation Vectors*** which I find very appealing. This aligns with the branch of AI dealing with *model-interpretability*. And for such critical use cases where domain expertise is qualitative, a better insight into the features governing the input embedding by the deep-learning models is crucial.

Areas that can use improvement

*The evaluation of refinement by example and concept seemed to overlap quite a bit (which is acceptable) but the results from refinement by concept were not thoroughly reasoned. There wasn't any uniform consensus in the results from this refinement (whether such a feedback works for all users or only a specific pool of users) which could have used some more discussion by the authors. It would have been interesting to understand the features where the medical community does not have any unanimous agreement on.