Interpreting mechanisms of prediction for skin cancer diagnosis using multi-task learning

1. Rebuttal

We thank all the reviewers for their comments and suggestions, to which we will respond in this document. *Acronyms*: **CIM**: changes in the manuscript; DL: deep learning; SF: sharing fraction metric.

A similar idea [...] in Taghanaki et al. [2]. We thank the reviewer for suggesting this paper; it will be discussed when introducing the concept of learnable gates. However, we note that the objective of [2] is different as it deals with reducing the memory impact of skip connections instead of multi-task learning. **CIM:** Beginning of Section 4.1.

The proposed method does not [...] attribute the gap [...] ImageNet-pretraining. We thank the reviewers for the comments regarding the performance gap. An apples-to-apples comparison would have been better suited as also noted by Reviewer #7. One way to test whether our method would benefit from additional data could be to concatenate the metadata features to the last layer of the diagnosis task (DIAG) and classify the new feature vector with a few fully connected layers. Setting up these new experiments properly will require some time and results will be available after the camera-ready submission to the workshop.

While the authors motivate [...] for 'DL practitioners' than for dermatologists. We agree with the reviewer that at this stage the proposed method is not yet ready to be of direct interest to dermatologists' practise and that the information derived from the method can be of more interest to DL practitioners. However, we believe that understanding the associations learned by the model would ultimately benefit the clinicians by providing insights on what the model has learned from the data.

A similar argument [...] deductions from the results presented in Figure 4. From Figure 4 we note that the diagnosis task (DIAG) requires more feature maps shared from the other tasks, i.e. higher SF. While our method uses the 7 criteria as tasks, it is not "aware" of the existing relation among those and the melanoma diagnosis in the 7-point checklist rule. In our best-performing experiment DIAG shows higher SF scores with the major criteria. Similarly, we note that in the 7-point rule the major criteria are weighed more (2 points) than the minor ones (1 point).

[...] ask expert dermatologists for their feedback [...]. We will contact our dermatologist collaborators to discuss the implications of our work in this paper.

Figure S1 - S4: why [...]. A possible reason for the behaviour not being monotonic might be that during training many tasks are competing for parameter space, thus making it harder to train the model.

Refrain from capitalizations [...] Multiple grammatical [...]. We thank the reviewer for the comment and will edit the manuscript accordingly. **CIM:** Mid-sentence capitalizations have been removed. Multiple grammar errors and typos have been edited.

The overall writing [...]. We thank the reviewer for the comment and have rephrased the indicated text. **CIM:** Section 7: rephrasing of the text.

The impact [...] apples-to-apples comparison would certainly clarify the impact of the proposal. As also noted by Reviewer #6 there is a gap in the performance with [1]. We appreciate the suggestion that making an apples-to-apples comparison would help in highlighting the impact of the current work and will conduct the necessary experiments as described in an earlier point.

A revision of the formatting and data of the reference is mandatory [...]. We thank the reviewer for the comment and have double-checked all the references in the work.

Although this work focuses [...] performance differences in diagnosis accuracy, recall, and precision. A similar comparison has been carried out through the "gates-off" experiment: in this experiment the gates regarding tasks other than the t-th task have been all set to a constant value of 0, thus rendering the model a black-box. Compared to our best-performing experiment "standard", we report a 10% increase in average accuracy across the 7-point criteria and a slight increase in the diagnosis task performance.

It would be better to present figure 3 [...]. We thank the reviewer for the comment and have moved figure 3 at the beginning of the section as suggested. **CIM:** Figure 3 (now Figure 1) has been moved to the beginning of Section 4 (with in-text reference). As a consequence, figures 1 to 3 have been re-numbered.

It would be interesting to see [...] how it affects the performance. We thank the reviewer for the suggestion and we will conduct the proposed experiments to evaluate how this impacts the performance of the model.

Other changes: 1) Added acknowledgements paragraph; 2) Minor changes in text and captions to fit the 8 page limit.

References

[1] Jeremy Kawahara, Sara Daneshvar, Giuseppe Argenziano, et al. Seven-Point Checklist and Skin Lesion Classification Using Multitask Multimodal Neural Nets. *IEEE Journal of Biomedical and Health Informatics*, 23(2):538–546, Mar. 2019. 1

[2] Saeid Asgari Taghanaki, Aicha Bentaieb, Anmol Sharma, S. Kevin Zhou, Yefeng Zheng, et al. Select, Attend, and Transfer: Light, Learnable Skip Connections. In *Machine Learning in Medical Imaging*, pages 417–425. Springer, Cham, Oct. 2019. 1