Responses to the questions and comments of Andrea Manno-Kovács, PhD

I would like to express my sincere gratitude to Andrea Manno-Kovács, PhD for her comments and suggestions on my dissertation, as well as for accepting the request to serve as a reviewer. I have made every effort to respond to the questions and remarks with the utmost care and accuracy.

Questions and Comments on Chapter 2

Q1 *How robust is the proposed STN in case of bigger rotations or distortions?*

A1 Regarding larger rotations and distortions, the current STN architecture has natural limitations: extreme geometric transformations can cause the localization network to stall, and fall into local minima, leading to such alignment issues that can be observed in Figure 1. However, for the operational range required by our dataset, the provided transformation remains robust enough to support the downstream task.

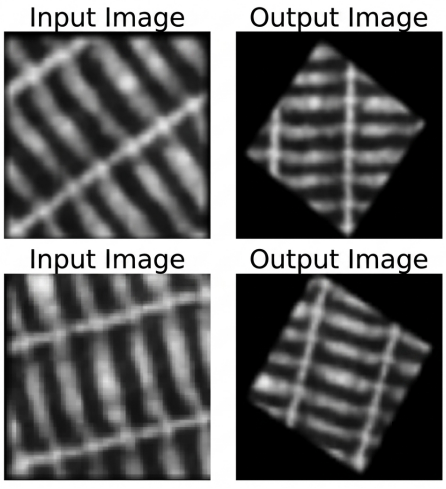

Figure 1: Sample images from the Texture 1 dataset before and after rotation using the Spatial Transformer Network (STN). Although the primary objective was to align the pattern's dominant lines vertically — a goal the STN generally successfully achieved — minor alignment discrepancies can still be noted.

Questions and Comments on Chapter 3

Q2 *Is the proposed method scalable for other models, like Vision Transformers?*

A2 In a ViT-based architecture, the network generates embedding vectors (e.g., via the final hidden states or the [CLS] token). My proposed approach can be directly applied on top of these embeddings as a classification head.

Q3 *Can deep randomized neural networks replace backpropagation techniques in other deep learning models?*

A3 While deep randomized networks may not completely replace backpropagation for feature learning in large architectures, they serve as a highly efficient replacement in the classification and adaptation stages of models like FCNNs and CNNs [2] (see Section C.5.2 in the Appendix of the Dissertation). The primary advantage of replacing backpropagation with these kind of models is gaining computational speed. This method particularly excels in two specific scenarios:

- Deployed on edge devices where computationally expensive backpropagation-based retraining is not feasible.
- In environments where the data changes — such as in few-shot learning scenarios or the sudden appearance of novel classes — backpropagation-based fine-tuning is often too slow and hardware-intensive.

In this scenarios a randomized neural network head can integrate new classes rapidly by recomputing the weights.

Questions and Comments on Chapter 4

Q4 *The author responded earlier that the created OGYEIv2 database is public and available to the general public. It would have been useful to add the link of the database to the dissertation as well.*

A4 The link to the OGYEIv2 dataset is, in fact, included in the dissertation. For reference, it is provided in Table 1.1 (Chapter 1) and detailed in Chapter 4, Subsection 4.2.

Q5 *Why was the EfficientNetV2-S network used as backbone? Were there any experiments or other considerations behind this choice?*

A5 The choice of opting for the EfficientNetV2-S architecture as the backbone was motivated by both state-of-the-art (SOTA) performance considerations and empirical optimization during the development phase. EfficientNetV2 was heavily optimized by Google to serve as a highly effective, strong general-purpose backbone, making it well-suited for feature extraction in downstream tasks. While the baseline method [1] utilized a more modest architecture, my early-stage experiments focused on upgrading the network that generates embeddings.
I initially experimented with a simpler EfficientNet-B0 ($\sim$5.3M parameters), but later moved to the EfficientNetV2 family to leverage its superior feature extraction capabilities and faster training speeds.The selection of the Small (S) variant ($\sim$21.4M parameters) represents a delicate trade-off between model capacity and computational efficiency.
My experiments indicated that opting for larger configurations (such as the Medium or Large models) increased the computational overhead without yielding significant improvements in accuracy. Therefore, EfficientNetV2-S provided the optimal balance for my task.

Q6 *How sensitive is the proposed model to incomplete or inaccurate textual description information?*

A6 The model was not explicitly benchmarked against incomplete or inaccurate textual descriptions, as the official patient information leaflets in our dataset provided standardized and accurate visual descriptions.
However, there were cases where the official documentation was unavailable — specifically for dietary supplements. In these scenarios, we manually created descriptions following the structural template and terminology of official leaflets to maintain consistency.
It is safe to say that incomplete or inaccurate text acts similarly to heavy data augmentation: misleading information disrupts the alignment in the embedding space, making the embeddings noisy and degrading classification performance. Furthermore, if descriptions are incomplete (e.g., missing essential attributes like color or shape), it reduces the separation between negative pairs in the embedding space. This shrinks the margin, making it harder for the loss function to enforce a clear decision boundary.

In extreme cases where text contains zero usable information, the framework can simply fall back to the baseline visual-only loss function.

References

[1] Suiyi Ling et al. “Few-shot pill recognition”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 9789–9798.

[2] Amr M Nagy and László Czúni. “Classification and fast few-shot learning of steel surface defects with randomized network”. In: *Applied Sciences* 12.8 (2022), p. 3967.