**MSc thesis project:**
**A human-in-the-loop framework for embedding human knowledge in CNNs via attention maps**

Visual explanation is used in deep learning to interpret the decisions of the CNNs. Broadly, these methods can be categorized as either requiring additional backpropagation or not. Response based methods do not require backpropagation and generate explanation and predictions simultaneously. Recently, [1] introduced attention branch network (ABN), a response-based visual explanation model by introducing a branch structure with an attention mechanism. In [2], they showed that human knowledge can be inserted into ABNs by modifying the attention maps and introducing a reconstruction loss in the attention branch. However, this method required a large amount of attention maps to be edited to achieve meaningful performance gain. [3] further improved the ABN models' performance by introducing multi-scale ABNs and showed that the human knowledge can also be inserted in form of bounding boxes (weaker annotations than object boundaries) and significant performance gain can be achieved from relatively few annotations. In this project we wish to explore strategies for maximizing model improvement by minimal user "correction" input. The strategies should be evaluated/compared on different classification datasets (natural scene benchmarking sets, microscopy images for mechanism of action recognition, microscopy/photography from Stora Enso).

**Core task:**

- Explore and compare strategies for prioritizing which images to annotate (correct) first for retraining and quantifying the achieved performance gains. The order of training samples to show the user can for example be sorted on the basis of wrong predictions first (confusion matrix off-diagonal peaks) or the amount of overall attention (since the models might focus on background). We can think of additional input parameters/training statistics from the user for this, for example, typical object size etc).

**Supporting/additional tasks:**

- Develop an end-to-end framework for human knowledge insertion in the CNNs consisting of: model training, user input, and retraining.
- Using SimSearch for finding and annotating repetitive mistakes from the models.
- Implementing and testing different retraining strategies for efficient performance.
- Evaluating the framework on different datasets.

**References**


1. Fukui, Hiroshi, et al. "Attention branch network: Learning of attention mechanism for visual explanation." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.
2. Mitsuhara, Masahiro, et al. "Embedding human knowledge into deep neural network via attention map." arXiv preprint arXiv:1905.03540 (2019).
3. Gupta, Ankit, and Ida-Maria Sintorn. "Towards Better Guided Attention and Human Knowledge Insertion in Deep Convolutional Neural Networks." arXiv preprint arXiv:2210.11177 (2022).


**Supervisors & subject reviewer[1]**


Ankit Gupta, Dept. IT (ankit.gupta@it.uu.se); Anindya Gupta, Stora Enso (anindya.gupta@storaenso.com); [1]Ida-Maria Sintorn, Dept. IT, (ida.sintorn@it.uu.se)