ExaM: Unsupervised Concept-Based Representation Learning to Better Explain Models in Vision Tasks

Supplementary Material

1. Visual Results

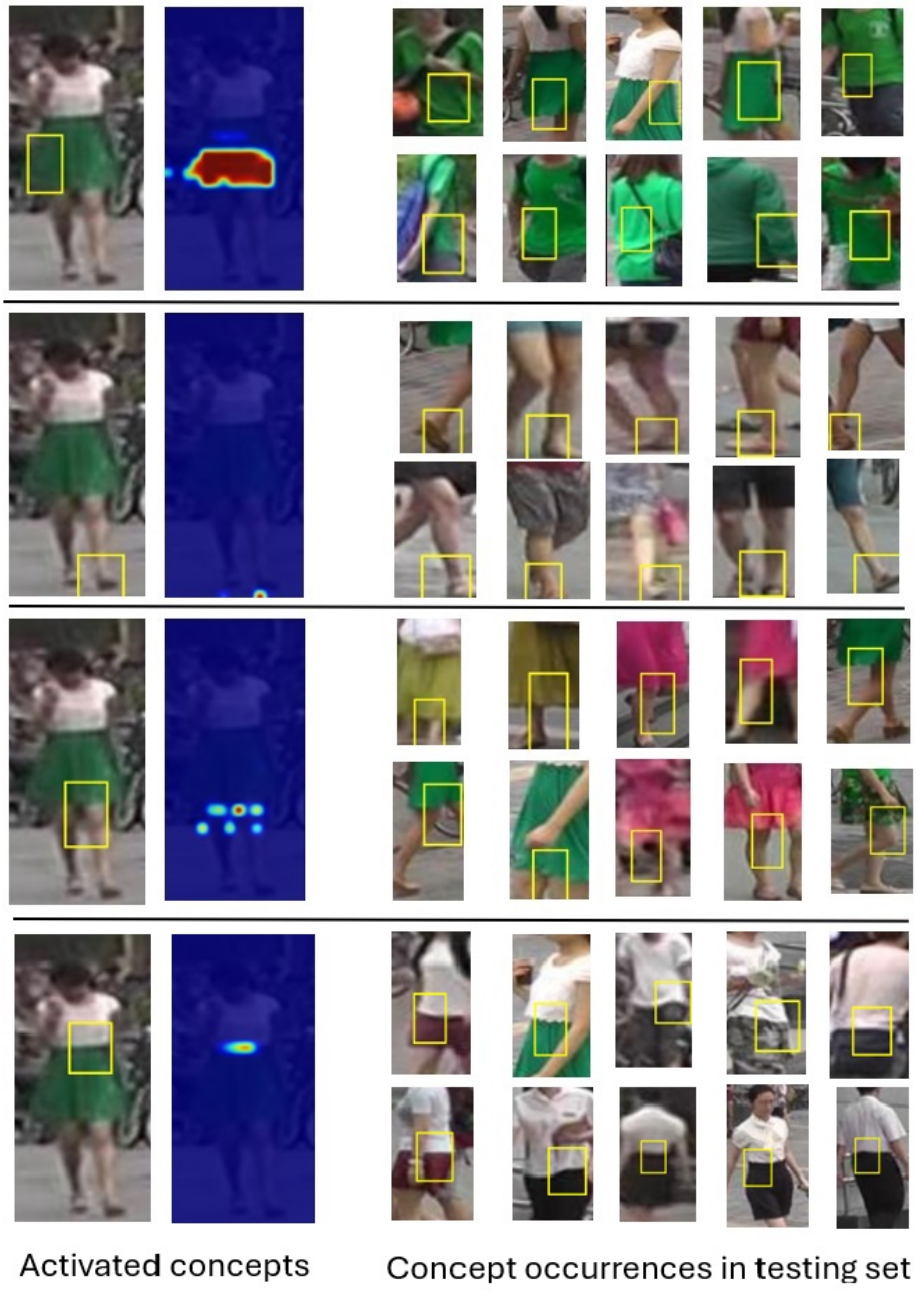

Figure 1. Example of concepts extracted for a single identity occurrence. Each row displays the primary activation zone for the concept within the image, along with the corresponding activation map and a set of identities from the testing set where the concept has also been activated.

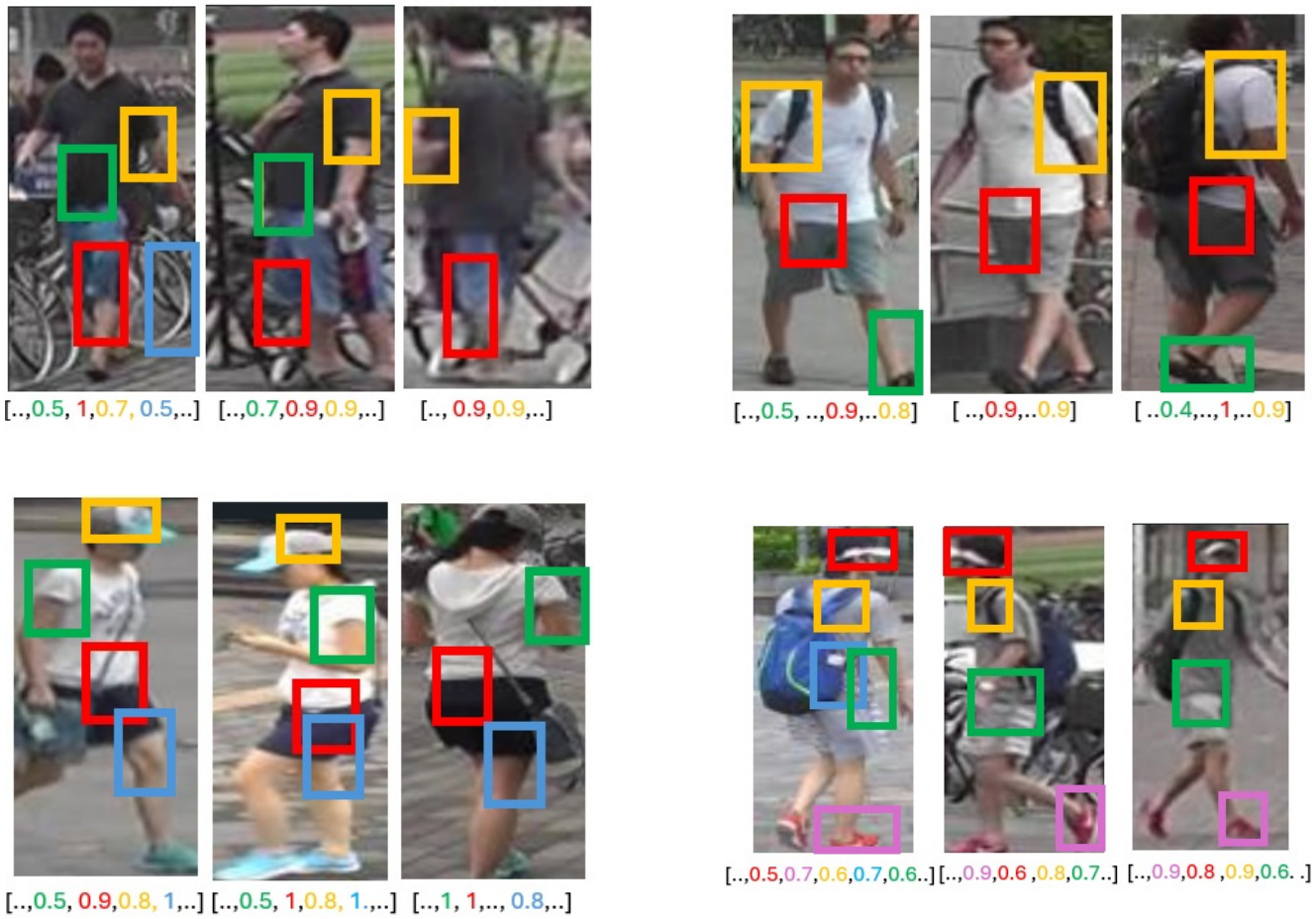

Figure 2. Example of the concepts detected for a person ReID task on Market-1501. Each block of three images represents the same identity. We observe that concepts are shared across different images of the same person, although some might not be activated due to appearance differences, e.g., orientation (front or back view).

Figure 3. More concepts from the unsupervised concept discovery on the test set (Market1501). Each row is associated with a concept and represents the top six images from the test set where the concept activation value is more than a fixed threshold of 0.5.

2. Model Implementation

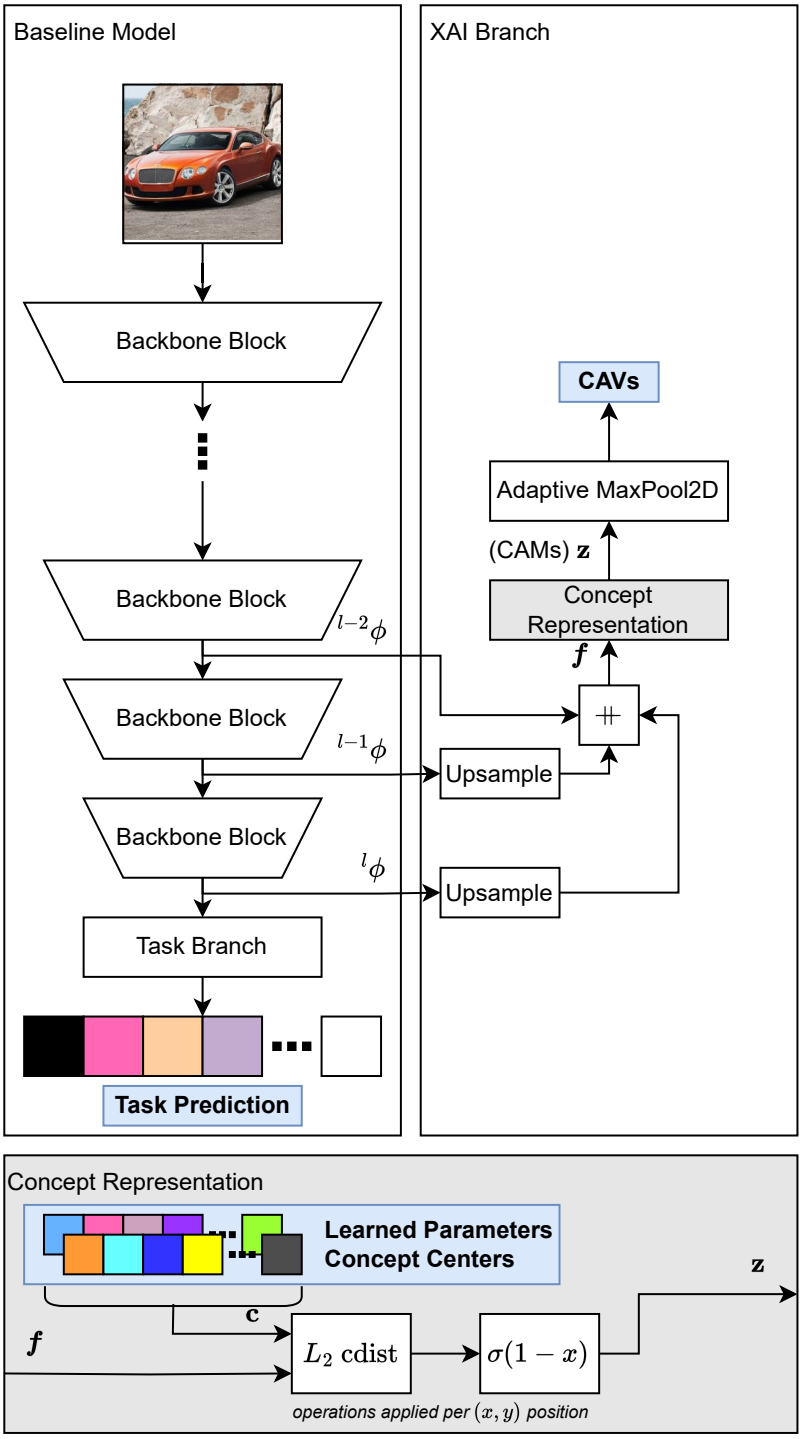

Figure 4. Architecture of the ExaM method. As described, the ExaM method architecture relies on adding a separate interpretability branch (named *XAI branch*) which outputs the Concept Activation Vectors (CAVs) and Concept Activation Maps (CAMs). Therefore, the task branch can retain the full baseline task performance, while the CAVs and Concept Activation Maps are used to interpret the prediction using patch-based concepts learned during training.

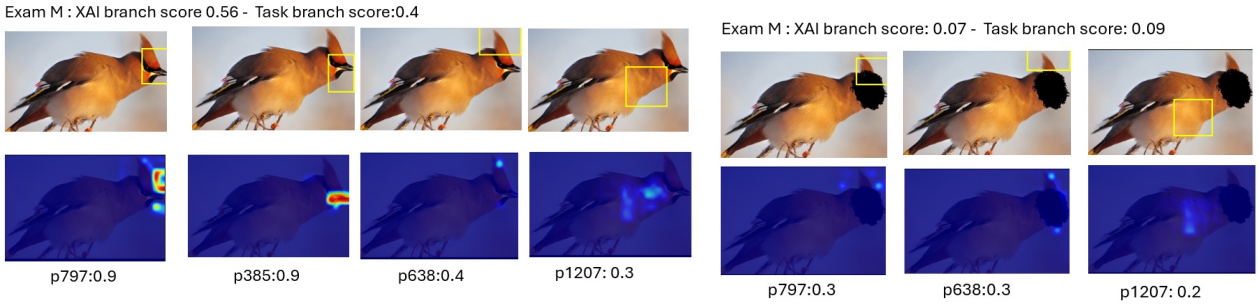

Figure 5. ExaM M concepts extracted for images with varying levels of alteration. Each column shows the primary activation zone within the image, along with the corresponding activation map for a specific element p of the CAV and its associated score. The XAI branch score and the concept activations aligns with the main task branch score when important visual information is removed.

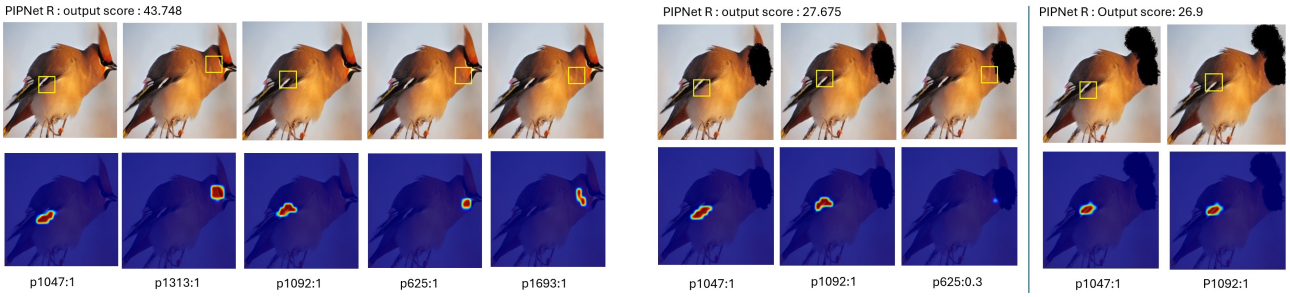

Figure 6. PIP-Net R concepts extracted for images with varying levels of alteration. Each column shows the primary activation zone within the image, together with the corresponding activation map for a specific element p of the CAV and its associated score.