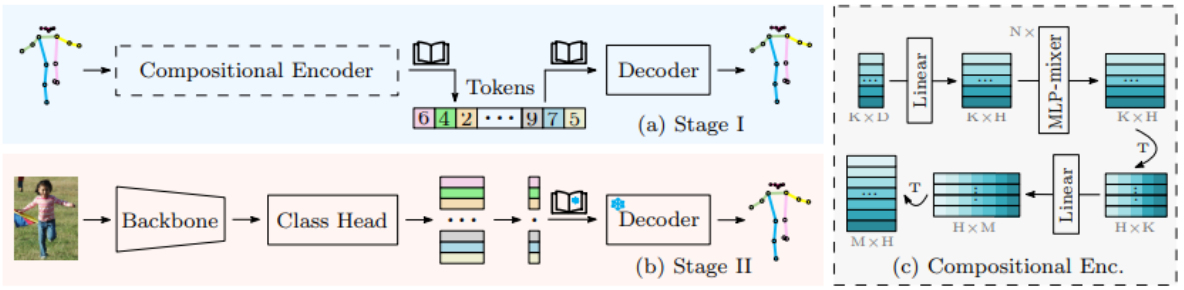
kouassi konan

Paper Summary

Title: Pose as Compositional Tokens (PCT),

Human pose estimation is a computer vision task that allows the computer to understand and recognize human pose on images and videos.
The paper presented here introduces a new representation for human pose called Pose as Compositional Tokens (PCT).
Unlike traditional methods that use coordinate vectors or heatmap embeddings in order to represent human pose, PCT utilizes a structured approach by representing a pose with M discrete tokens. This enables it to model the dependency between joints, and thus take them into account when detecting and representing the pose. This answers the problem addressed here, as in all the subjects of pose estimation, i.e. a reliable and more realistic representation of human pose.



PCT architecture

The PCT first converts the pose into a set of simplified tokens using a compositional encoder, then treats the process of understanding these tokens as a classification challenge, ultimately reconstructing the original pose from these classified tokens using a decoder.

The compositional encoder, codebook(“kind of dictionaries in which we can quantify elements, and these elements are represented by discrete indices), and decoder are jointly learned by minimizing a loss function.

In classification task (stage 2) The image features are extracted using a backbone network, and a classification head predicts token categories. The decoded tokens recover the final pose

PCT aims to learn tokens that represent meaningful pose substructures, balancing granularity and efficiency.
It avoids the extreme case where each marker corresponds to a single joint or an entire pose, thereby encouraging learning of the basic substructures used in all poses.
The codebook contains fewer entries than theoretically required, allowing the model to effectively learn larger structures.

According to the experiments, PCT achieves competitive or superior accuracy compared to state-of-the-art methods, particularly excelling in occluded joints. It excels in handling occlusion, requires no post-processing modules, and provides a good representation for 2D and 3D poses.

In the future authors suggest that It will be interesting to further reduce the ambiguities in pose estimation by exploring other cues under the discrete representation.

This paper was interesting to me cause i worked on Human pose estimation projects during my internships and when I started out, I used classic methods for estimating human poses, and often the algorithms used to represent poses required a high level of complexity in order to obtain realistic representations, and very often, in cases of ambiguity and occlusions, the representations were unrealistic. PCT now manages to produce more or less realistic representations.