# TokenPose: Learning Keypoint Tokens for Human Pose Estimation笔记

- Paper: [TokenPose: Learning Keypoint Tokens for Human Pose Estimation]
- Code: [leeyegy/TokenPose]

## 1. Introduction

### 1.1 Why

- 关键点检测深度依赖各部位之间的视觉线索和结构约束等信息，**然而CNN方法缺乏显式地学习关键点之间约束关系的能力。**(Human pose estimation deeply relies on visual clues and anatomical constraints between parts to locate keypoints. Most existing CNN-based methods do well in visual representation, however, lacking in the ability to explicitly learn the constraint relationships between keypoints.)

### 1.2 What

- **TokenPose**(Token representation for human Pose estimation)：每一个关键点都被显式地编码为一个token，以同时从图像中学习**约束关系**和**外观信息。**(Each keypoint is explicitly embedded as a token to simultaneously learn constraint relationships and appearance cues from images.)
    - pure Transformer:
        - TokenPose-T*
    - hybrid Transformer:
        - TokenPose-S*: stem-net(CNN Backbobne)
        - TokenPose-B*: HRNet-W32-stage3(CNN Backbobne)
        - TokenPose-L*: HRNet-W48-stage3(CNN Backbobne)

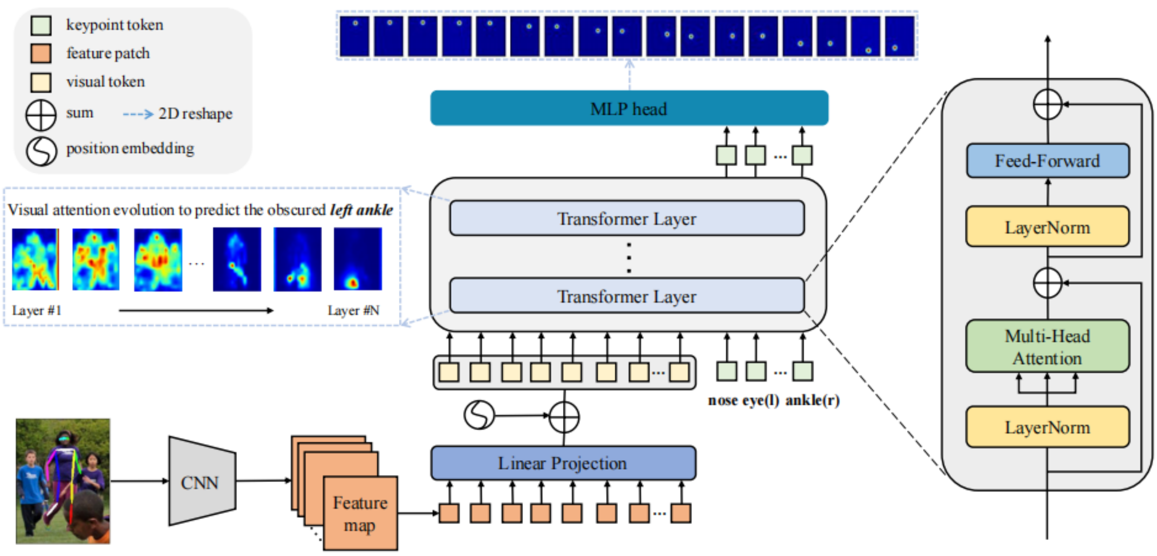
### 1.3 How

- TokenPose能从大量的数据中学习关键点之间的静态约束关系，这些信息被编码到关键点tokens中，关键点tokens向量可以通过向量相似度来记录关键点之间的关系。(It is worth noting that TokenPose learns the **statistic constraint relationships** between keypoints from large amounts of data. Such information is encoded into keypoint tokens that can record their relationships by **vector similarities**.)
- 在推理的过程中，TokenPose将关键点tokens与视觉tokens进行关联，视觉tokens对应的图像patches区域可能包含对应的target关键点。在某种程度上，这些关键点tokens起到的作用与解码相似，它将视觉信息从视觉tokens向量中解码出来，以获得最终的关键点预测结果。(During inference, TokenPose **associates keypoint tokens with those visual tokens** whose corresponding patches possibly contain the target keypoints. To some extent, such keypoint tokens work like decoders, which decode visual information from visual tokens to achieve the final predictions.)

### 1.4 Contributions

- 本文提出了用tokens向量来表征关键点实体。通过这种方法，**视觉线索**学习和**约束线索**学习可以被显式地合并在一个统一的框架下。(We propose to use **tokens** to represent each keypoint entity. In this way, **visual cue learning** and **constraint cue learning** are explicitly incorporated into a unified framework.)
- 本文研究了两种架构：CNN-Transformer混合架构、纯Transformer架构。本文的纯Transformer架构是二维人体关键点检测领域中第一个纯Transformer架构。(Both **hybrid** and **pure Transformer-based** architectures are explored in this work. To the best of our knowledge, our proposed TokenPose-T is the first pure Transformer-based model for 2D human pose estimation.)
- 在COCO和MPII两个数据集上，与CNN方法相比，在很少的参数和计算代价下，能够取得具有竞争力的先进性能。(We conduct experiments over two widely-used benchmark datasets: COCO keypoint detection dataset and MPII Human Pose dataset. TokenPose achieves competitive state-of-the-art performance with much fewer parameters and computation cost compared with existing CNN-based counterparts.)

## 2. Method




two different types of tokenizations:

- keypoint tokens: randomly initialized embeddings, each of which represents a specific type of keypoint
- visual tokens: flattened image patches

更新于2021-05-26