Motion Prediction

Deepak Gopinath (dgopinath@), Jesse Smith (hjessmith@), Sasha Sheng (sash@)

1. Motivation:
    a. Human motion prediction is the problem of forecasting future body poses given observed pose sequence. The problem requires encoding both spatial and temporal aspects of human motion, and generating sequences conditioned on that.
    b. The problem is at the intersection of graphics and computer vision, with applications spanning human-computer interaction (think robots and self driving cars), motion synthesis (to automatically animate characters), and virtual and augmented reality. It is a particularly challenging sequence modelling task owing to the stochastic nature of human motion, leading to a large space of possible pose predictions.
2. Approaches:
    a. The problem is formulated as a sequence modeling task -- the input is 120 poses (frames) i.e., 2s of motion at 60Hz, and the output is 24 poses (400ms). Motion data consists of two parts -- **skeleton** and time series of **poses**. The structure of a character is defined by the "skeleton" with hierarchical arrangement of joints (like hip, back, neck, left arm etc.). A **pose** is defined as a list of 3D angles that describe the orientation of all joints in a particular frame.
    b. Solutions for the problem have been inspired by techniques in language and vision. The problem was first proposed in [1], which used RNNs to read and generate one frame at a time. Future works used seq2seq models [2], CNNs [3], and even reinforcement learning (as imitation learning) [3] to solve the task.
    c. The current state of the art model architecture, Spatio-Temporal Transformers [4], uses separate self-attention modules to encode spatial and temporal features, to extract pose representations.
    d. How is the problem approached today, and what are the limits of current practice?
    e. Simple baselines for the task could be using RNN, seq2seq RNN, and Transformer. These baselines are readily available on the open sourced fairmotion library.
    f. Potential approaches students can start with, without major change from the baselines could be Temporal Convolutional Networks [5], and seq2seq Transformers [6].
3. Metrics:
    a. Dataset:
        i. The recently released AMASS dataset unifies several motion capture datasets, and is the largest publicly available dataset with 42 hours of mocap data. It can be downloaded from the

website https://amass.is.tue.mpg.de/, after registration. It is free to access the dataset for academic use.
    ii. For this task, the dataset splits for train, test and validation have been defined in the work "Structured Prediction Helps 3D Human Motion Modelling" Aksan et al. We shall be using the same for consistency. The fairmotion library contains filename-split mapping.
b. The error metrics are designed to calculate the deviation between predicted pose and target pose at every frame:
    i. Mean Angle Error (MAE): Euclidean distance between target and predicted Euler joint angles
    ii. Positional Error: Euclidean distance between predicted and target values of 3D position of each joint of the body.
c. State-of-the-art:
    i. The current state of the art model architecture is the Spatio-Temporal Transformer (https://arxiv.org/abs/2004.08692). The implementation of the model has not been open sourced yet.

4. Scope:
a. What are possible directions that students could explore? E.g. In terms of modeling, data, efficient computation.
    i. Students can explore novel model architectures that can more effectively encode temporal data and/or spatial data. They would aim to improve benchmark metrics like MAE and Positional Error on the AMASS dataset.
    ii. The task involves generation of short number of frames (400ms). Students can work on designing models that can generate motion over a much longer time frame. This would be directly useful for application of motion prediction models for character animation. Refer to papers on motion synthesis for more details.
    iii. They can propose better metrics that can quantify the naturalness or physical plausibility of the generated motion.
b. How much work in each direction would justify a good grade?
    i. Students would be expected to write a report based on the literature survey they conduct, as part of understanding the task.
    ii. They would be expected to propose concrete, reasonable approaches that can easily beat baseline models. They would not be expected to beat SOTA performance. Grades would not depend on final performance of the model, but on the novelty of ideas, and effort put into the project.
c. What is the ideal size for a team tackling this project?
    i. The ideas size of a team would be 2-3, considering the focus of the project is going to be on modelling. The project would involve literature survey and brainstorming too.

5. Resources:
    a. The AMASS dataset can be used to train models and benchmark performance. Use the dataset splits and pre/post-processing in `fairmotion` to maintain consistency.
    b. Baseline models can be trained on a single GPU in 3-4 hours. Training a full SOTA model takes 12 hours on single GPU. Prototyping and development can be done on CPU or GPU.
    c. Started code with data loading, preprocessing, training baselines, and evaluation can be found at https://github.com/facebookresearch/fairmotion

References:

[1] "*Recurrent Network Models for Human Dynamics"*, Fragkiadaki et al
[2] "*On human motion prediction using recurrent neural networks"*, Martinez et al
[3] "*Convolutional Sequence to Sequence Model for Human Dynamics"*, Zhang et al
[4] *"Imitation Learning for Human Pose Prediction"*, Wang et al
[5] *"Temporal convolutional networks for action segmentation and detection"*, Lea et al
[6] *"Attention is all you need",* Vaswani et al.