Name: Mukund Dhar
UCF ID: 5499369

Report on **"HUMAN MOTION DIFFUSION MODEL"**

The paper introduces Motion Diffusion Model (MDM), a carefully modified diffusion-based generative model the domain of human motion. The architecture is transformer-based and lightweight and hence, fits the temporal and non-spatial information better. The concept of geometric losses is applied in the diffusion architecture to improve generation of human motion especially without jitter. This is possible since the in the model MDM predicts the sample instead of the noise in each diffusion step. The framework can use different forms of conditioning enabling it to perform three tasks: text-to-motion, action-to-motion, and unconditioned generation. The trade-off between photorealism and diversity is handled by training the model classifier-free and with sampling both conditionally and unconditionally from the same model. The MDM framework is able to generate state-of-the-art results in the tasks of motion-to action as well as action-to-motion in terms of FID, Diversity, and Multimodality. The semantic editing of specific body parts is also demonstrated by setting a motion prefix and suffix by the model. In conclusion, the Motion Diffusion model offers a motion generation framework that produces state-of-the-art results in several of the tasks as well as well thought-out domain knowledge.

Strengths: MDM presents a unique classifier-free diffusion model, featuring a transformer-encoder, that predicts the signal instead of the noise allowing it to utilize geometric losses as well. It can be applied to various conditioning to generate results on several motion generation tasks: text-to-motion, action-to-motion, and unconditioned generation. This approach is not sensitive to the type of architecture used and gives state-of-the-art results in many of the conditioned-generation tasks. The model also provides editing applications either conditionally or unconditionally.

Weaknesses: The paper points to one of the difficult motion generation tasks being unconstrained synthesis. The lack of labeled data makes this a challenging task and hence it is not able to beat the state-of-the-art. The model also consumes a large amount of time during the inference. It takes about 1000 forward passes for a single result.

Questions:

- If the classifier-free approach of the model gives such good results, it hints to make questions on whether we are utilizing the potential of using the data information completely.
- What are real-world applications of using this model that takes more time to train than others?

Possible ideas:

Since it uses a transformer, availability of more amounts of labeled data can only result in better training. The model can also think of motion editing before inference.