Pose Transformers: Human Motion Prediction with Non-Autoregressive Transformers

Angel Martínez-González, Michael Villamizar and Jean-Marc Odobez

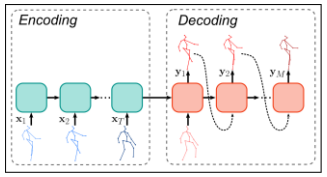
idiap RESEARCH INSTITUTE

EPFL

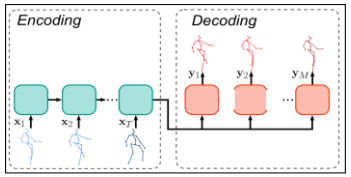
Motivations

- Sequence-to-sequence motion prediction
- AR decoding: potential error accumulation
- AR decoding: computational costly

AR decoding



Parallel decoding



Contributions

- Parallel decoding: efficient inference
- Non-autoregressive Transformer architecture
- Single model: activity and motion prediction

Related Work

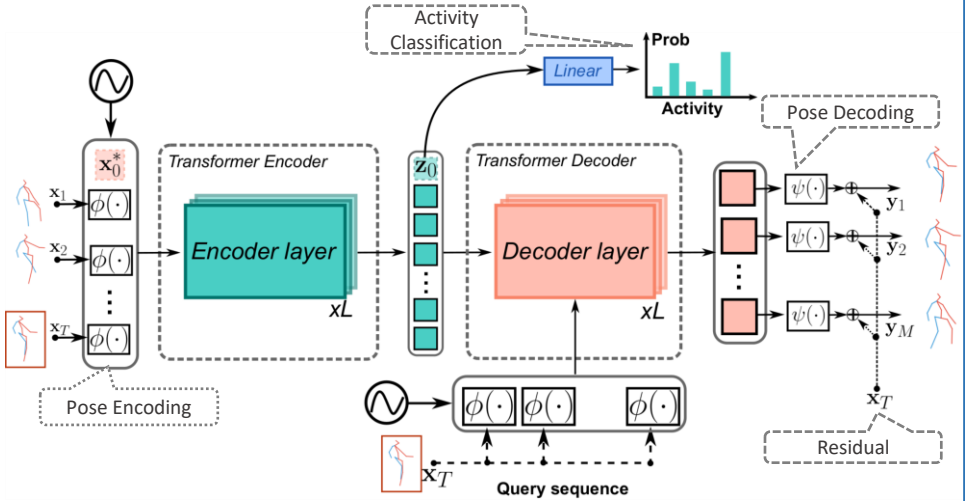
[1] Martinez et al, On Human motion prediction with recurrent neural networks, CVPR 17
[2] Carion et al, Detection Transformers, ECCV 2020
[3] Dosovitskiy et al, Vision Transformers, ICLR 2021
[4] Mao et al, History repeats itself, ECCV 2020
[5] Gu et al, Non-Autoregreesive Machine Translation, ICRL 2018

Acknowledgements

MuMMER mummer-project.eu

CONACYT Consejo Nacional de Ciencia y Tecnología

Method

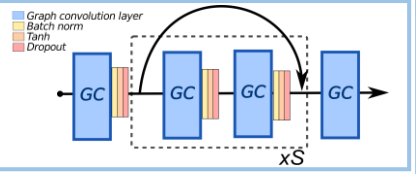


- Query sequence: repeat $\boldsymbol{x}_T$
- Residual and parallel motion decoding
- Activity token – encode activity from motion
- Speed – **Non-AR: 149.2 SPS**; AR: 8.9 SPS

Pose Encoding & Decoding

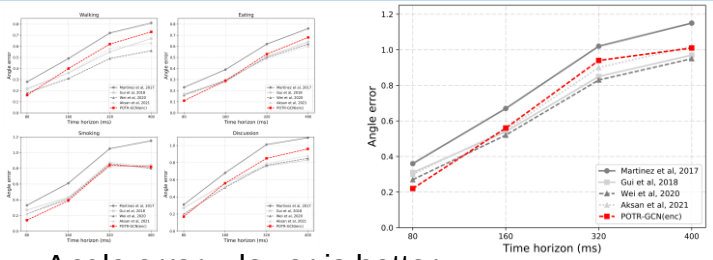
Investigated architectures

1. ϕ and ψ are linear layers
2. Graph Convolutional Network



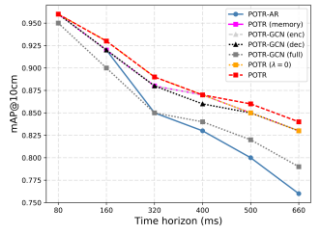
Code & Models
https://github.com/idiap/potr

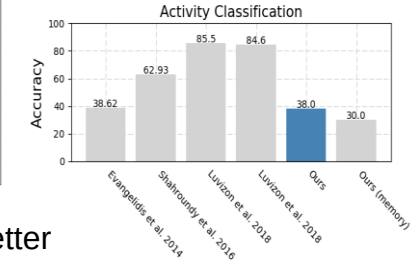
H36M – Motion



- Angle error – lower is better
- Best in shorter horizons

NTURGB – Motion & Activity





- mAP – higher is better
- Skeleton-based classification – low accuracy

Attention Visualization