Enhancing Human-Robot Interaction through Multi-Human Motion Forecasting

Mohammad Samin Yasar and Tariq Iqbal

msy9an@virginia.edu , tiqbal@virginia.edu

University of Virginia | School of Engineering & Applied Science

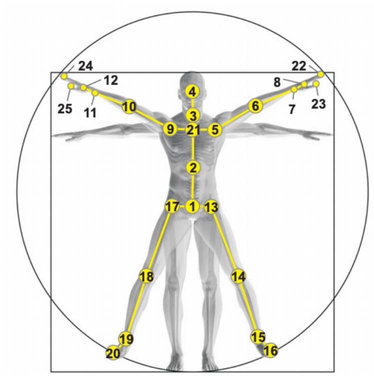
Collaborative Robotics Lab

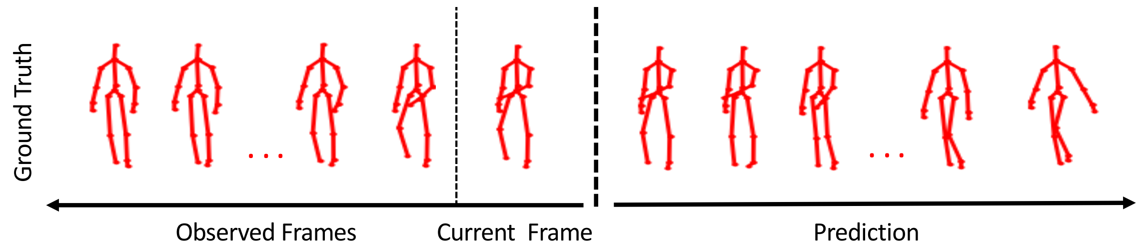

CAUTION
FOGGED DISINFECTANT
BUILD WITH ROBOTS
KUKA

Credit: KUKA AG

What is needed for robots to collaborate with humans?

Anticipate Human Intent and Motion





- Treat anticipation as a sequence learning problem
- Model spatial correlation between joints within a frame
- Model temporal correlation across joints over a horizon
- Highly stochastic over long-term, with high variations across frame

Existing Work

- Deterministic: Learns a point estimate over the motion data
  Seq2Seq[1], Seq2Seq-SPL[2]
  - Cannot model uncertainties in human motion
- Probabilistic: Learns a distribution over the motion data
  HP-GAN[3], VAE[4]
  - Separate objective function for learning a distribution
  - Requires careful hyper-parameters selection and annealing

[1]J. Martinez, M. J. Black, and J. Romero, "On human motion prediction using recurrent neural networks," in IEEE CVPR, 2017.
[2]E. Aksan, M. Kaufmann, and O. Hilliges, "Structured prediction helps 3d human motion modelling," in IEEE ICCV, 2019
[3]E. Barsoum, J. Kender, and Z. Liu, "Hp-gan: Probabilistic 3d human motion prediction via gan," in IEEE CVPRW, 2018
[4]S. Toyer, A. Cherian, T. Han, and S. Gould, "Human pose forecasting via deep markov models," in International DICTA, 2017.

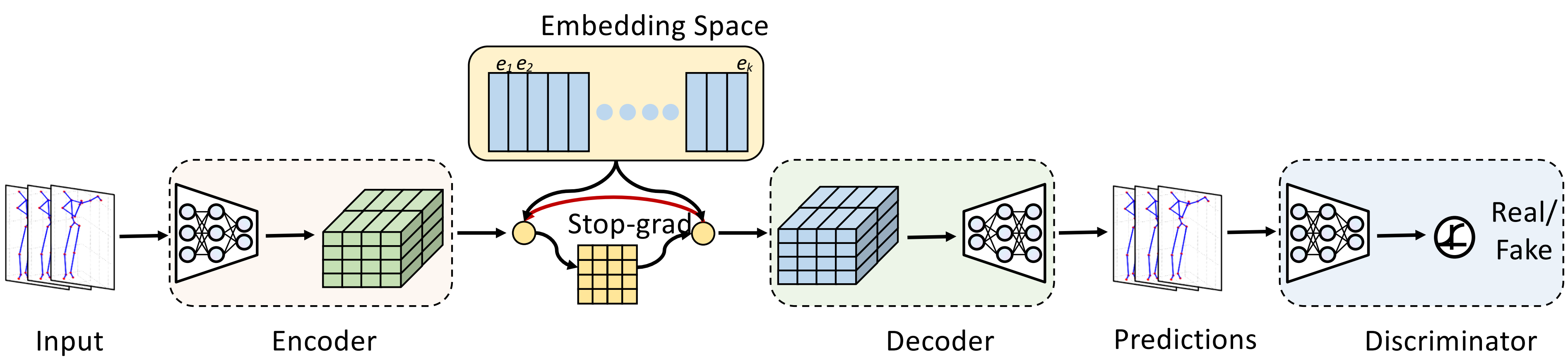
Research Gap

- Learning a robust representation of the past motion.
- Improving temporal and spatial correlation in the motion prediction.
- Leveraging the appropriate objective function.

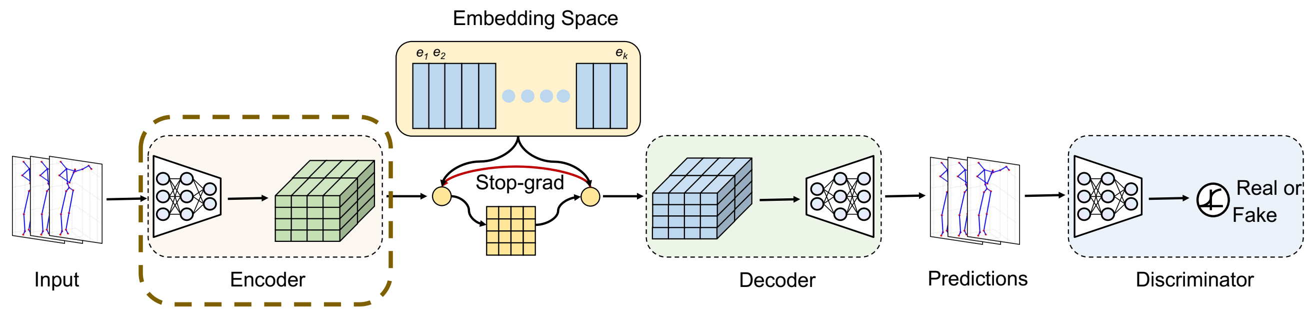
Contributions

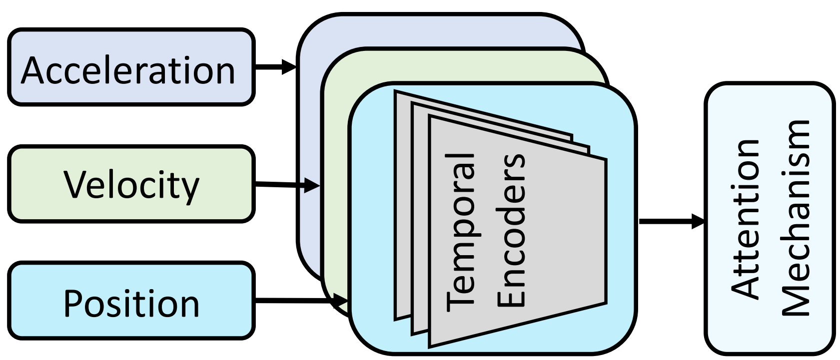
- In this work, we propose VADER, a novel sequence learning framework that
  - Learns a robust representation over the observed human motion,
  - Uses the expressive powers of codebooks to learn discrete representations over the observed motion data,
  - Is not restricted by any static priors,
  - Explicitly models interaction in multiple humans via a lightweight attention mechanism.
- VADER outperformed previous state-of-the-art approaches across three difference scenarios: single-agent, multiple-agent and human-robot collaboration over short and long-term horizons.

Unified Architecture of VADER



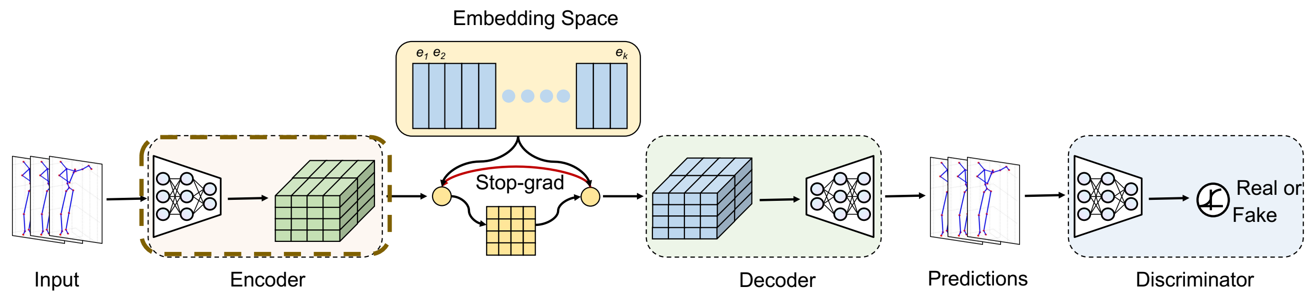
- Our framework augments the encoder-decoder framework with codebook learning and distribution matching.
- We use adversarial training to improve the temporal and spatial coherency by penalizing predictions that deviates from the ground-truth distribution.

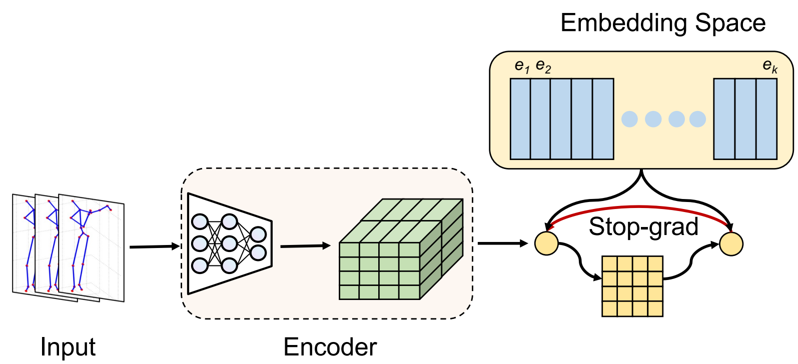




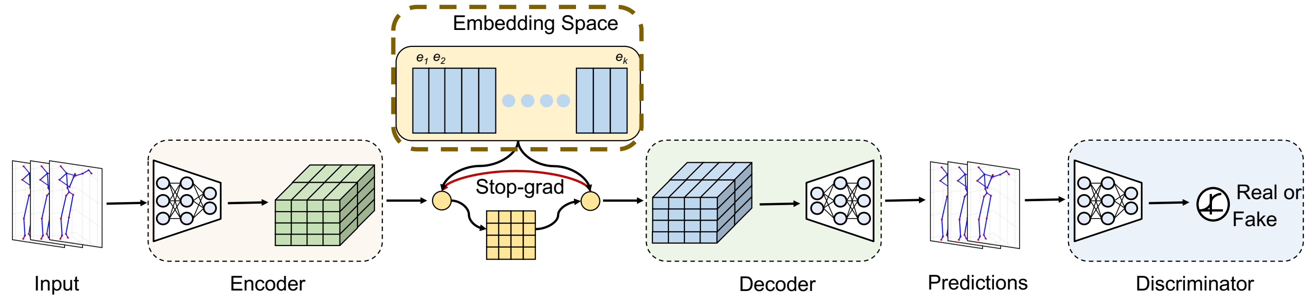
To obtain a robust representation over observed trajectory:

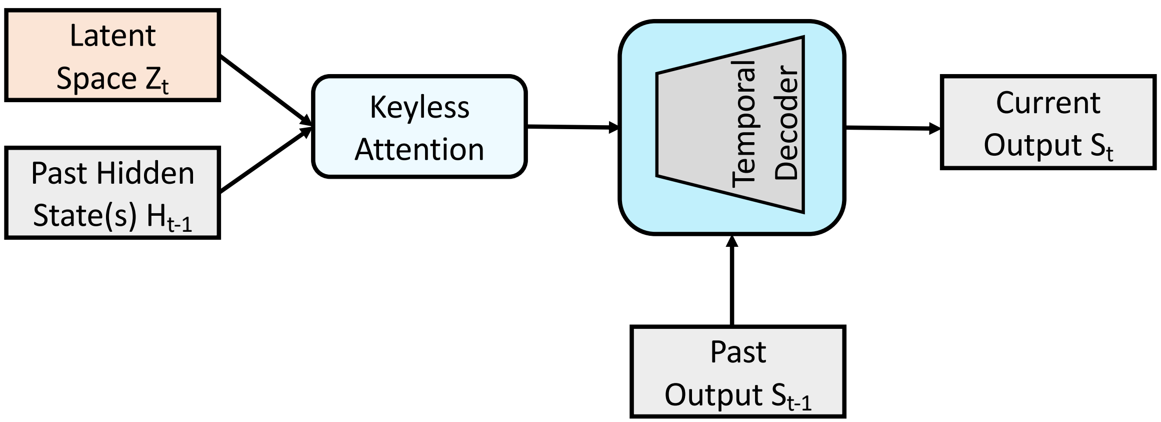
- We extract velocity and acceleration features from position, and explicitly model all three representations.
- The representations are passed to an attention module, that learns a robust characterization over past observations.



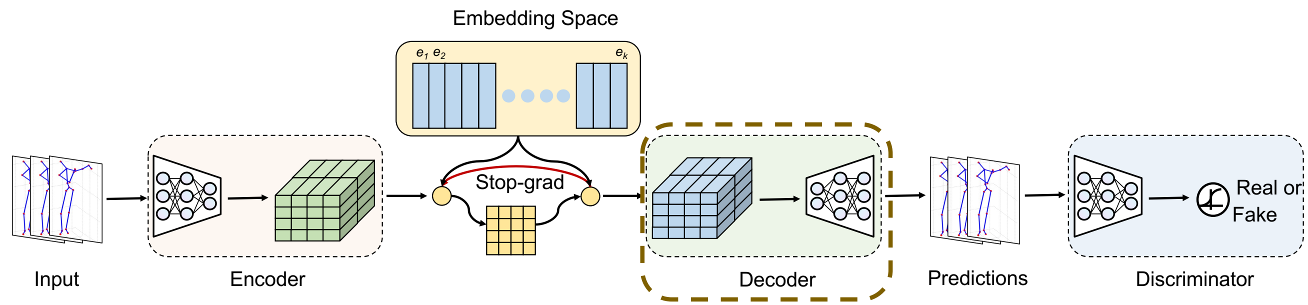


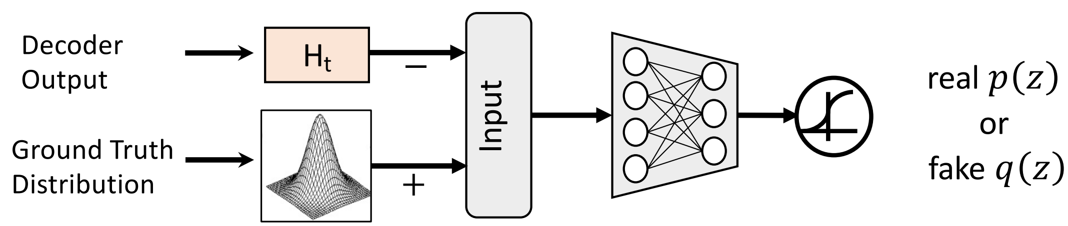
- We propose the use of a codebook for calculating the latent space, using Vector Quantization.
- The output of the encoder is used to calculate the discrete latent space using the nearest neighbor lookup from the shared embedding space.





- The decoder learns to condition its output on the previous hidden state(s) and the latent representation, that summarize past frames.
- This provides performance gains, particularly over long-term horizons.





- We use adversarial training to align the prediction with the ground truth.
- The adversarial loss complements the Reconstruction Loss by penalizing predictions that deviate from the ground-truth distribution.

Quantitative Evaluation

- We evaluated the performance of our framework on two widely used human-activity datasets, one social interaction dataset and on data collected from human-robot collaboration (HRC) experiments:
  - **UTD-MHAD** for single-agent motion prediction
  - **NTU-RGBD+D 60** dataset for multi-agent motion prediction
  - **CMU Panoptic** dataset for multi-agent motion prediction
  - **KTH-HRC** dataset on human-robot collaboration experiments
- Our evaluation metric is the **Mean Squared Error** between the ground-truth and the predicted poses at each timestep.

Results: **Single-agent** motion prediction (UTD-MHAD)

Approaches	Frames					
	2	4	8	10	13	15
Zero-Velocity	11.31	27.91	68.79	89.09	116.95	133.05
Seq2seq	8.90	19.09	39.03	47.45	57.84	63.30
Seq2seq-SPL	8.17	17.63	36.86	45.02	55.20	60.72
Scalable	**6.39**	14.33	31.63	39.12	48.57	53.74
VADER	6.61	**14.22**	**29.82**	**36.23**	**43.83**	**47.81**

- Our method **outperformed** state-of-the-art models on majority of the evaluated benchmarks, suggesting improved representation learning and sequence modeling.

Results: **Multi-agent** motion prediction (NTU RGB+D 60 dataset)

Approaches	Frames					
	2	4	8	10	13	15
Joint Learning	9.68	15.84	29.88	37.52	49.55	57.93
Joint Learning + Social	9.71	15.97	30.36	38.25	50.70	59.39
Scalable	9.66	15.66	29.05	36.16	47.20	54.84
VADER	**9.65**	**15.48**	**28.57**	**35.64**	**46.71**	**54.39**

- For multi-agent motion prediction, our method **outperformed** state-of-the-art models on all evaluated benchmarks, suggesting that the attention mechanism at the decoder can best represent the inter-agent dynamics among all the agents.

Results: **Multi-agent** motion prediction (CMU Panoptic dataset)

Approaches	Frames					
	2	4	8	10	13	15
Joint Learning	1.334	2.29	4.15	5.09	6.55	7.56
Joint Learning + Social	1.396	2.39	4.35	5.35	6.87	7.90
Scalable	1.327	2.22	3.94	4.79	6.07	6.94
VADER	**1.321**	**2.19**	**3.84**	**4.66**	**5.89**	**6.75**

- For multi-agent motion prediction, our method **outperformed** state-of-the-art models on all evaluated benchmarks, suggesting that the attention mechanism at the decoder can best represent the inter-agent dynamics among all the agents.

Results: **Human-Robot** Collaboration (KTH-HRC)

Approaches	Frames					
	5	10	20	30	35	40
Zero-Velocity	0.11	0.34	1.18	2.38	3.07	3.81
Seq2seq	0.18	0.55	1.67	3.11	3.91	4.74
Seq2seq-SPL	0.17	0.42	1.20	2.33	2.98	3.66
Scalable	**0.06**	**0.20**	0.72	1.61	2.21	2.91
VADER	**0.06**	**0.20**	**0.69**	**1.55**	**2.15**	**2.88**

- Our method **outperformed** state-of-the-art models on all evaluated benchmarks, suggesting improved representation learning and sequence modeling.

Results: Ablation Study

Approaches	Frames					
	2	4	8	10	13	15
VADER w. TCN encoder-decoder	9.68	19.71	37.27	44.02	52.35	57.21
VADER with TCN encoder	7.85	16.29	33.49	40.68	49.47	54.27
VADER w/o GAN objective	8.08	16.76	33.57	40.39	48.54	52.87
VADER w/o attention	10.19	22.21	45.26	54.78	66.85	73.92
VADER	**6.61**	**14.22**	**29.82**	**36.23**	**43.83**	**47.81**

Summary

- We proposed VADER, a novel sequence-learning approach that seeks to overcome some of the longstanding challenges of motion prediction.
- In VADER, we proposed the use of vector quantization to learn a discrete latent space, with no restrictions of a static prior
- Next, we proposed using the discriminator loss to compliment the MSE objective to improve the accuracy of motion prediction.
- Finally, to account for the interdependence of human motion, we incorporated a lightweight attention mechanism to condition predictions on other humans

VADER: Vector-Quantized Generative Adversarial Network for Motion Prediction

Mohammad Samin Yasar and Tariq Iqbal

msy9an@virginia.edu , tiqbal@virginia.edu

University of Virginia | School of Engineering & Applied Science

Collaborative Robotics Lab