**Hyperparameter investigations:**

**There is**

**one distribution to model variant parameter s, and**

**one distribution to model invariant parameter z and**

**one distribution to model output.**

* Datasets

1. Benchmark: discriminative output, noise, samples 1, 10 or 10, 10. Result: 1.0.
2. CVPR (SOTA result: 0.1): discriminative output, no noise, samples 1, 10 result:0.2, or 10, 10. Result: 0.3.

* Optimization

1. Epoch: 1000
2. Learning rate
3. Training in multiple steps with pretraining/initialization or end to end

* Model

1. Number of samples of different distributions

Equally set for all distributions or not.

How many samples to use for expectations?

1. Output prediction approach: CVPR: No noise

Discriminative vs generative

Stochastic (with noise or small number of samples for expectation) vs deterministic

1. Latent dimensions: CVPR set to 2.
2. GMM number of clusters (for s): CVPR set to 3.
3. Architecture models of encoders and decoders LSTM vs MLP.
4. Loss: adding terms to loss (contrastive loss)
5. Coupling layer for both s and z priors.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Experiment** | **Output Model** | **Learning rate** | **Sample dist Y** | **Sample Dist Z,X,S** | **Contrastive Loss** | **Coupling Layer S Prior** | **Coupling Layer Z Prior** | **Encoder S** | **Encoder Z** | **End to End** | **Epoch** | **Result ADE** |
| E1 | Point( without log variance terms) | No (1e-3) | 10 | 10 | no | yes | yes | mlp | mlp | yes | 1000 | 0.3 |
| E2 | Point( without log variance terms) | Yes, One cycle | 10 | 10 | no | yes | yes | mlp | mlp | yes | 500 | 0.32 |
| E3 | Point( without log variance terms) | No (1e-3) | 1 | 10 | no | yes | yes | mlp | mlp | yes | 400 | 0.34 |
| E4 | Point( with log variance terms) | No (1e-3) | 10 | 10 | no | yes | yes | mlp | mlp | yes | 1000 | 0.31 |
| E7 | Point( with log variance terms) | No (1e-3) | 10 | 10 | yes | yes | yes | mlp | mlp | yes | 1000 |  |
| E6 | Point( with log variance terms) | No (1e-3) | 1 | 10 | yes | yes | yes | mlp | mlp | yes | 1000 |  |
| E5 (without fully connected before decoders) | Point( with log variance terms) | No (1e-3) | 10 | 10 | yes | yes | yes | mlp | mlp | yes | 1000 | 0.32 |
| E8 (without fully connected before decoders) | Point( with log variance terms) | No (1e-3) | 1 | 10 | no | yes | yes | mlp | mlp | yes | 1000 | 0.3 |
| E9 (without fully connected before decoders) | Point( with log variance terms) | No (1e-3) | 1 | 10 | yes | yes | yes | mlp | mlp | yes | 1000 | 0.23 |
| E17 | Point  FC+Decoder  Dimensions:  d\_CouplingLayer=32  d\_FC\_before\_Dec=32  d\_FC\_after\_Enc=32,32  d\_latent=8 | No (1e-3)  One cylce | 10 | 10 | no | yes | yes | mlp | mlp | yes | 1000 | 0.30 |
| E18 | Point  FC+Decoder  Dimensions:  d\_CouplingLayer=32  d\_FC\_before\_Dec=32  d\_FC\_after\_Enc=32,32  d\_latent=8 | No (1e-3)  One cylce | 10 | 10 | yes | yes | yes | mlp | mlp | yes | 1000 | 0.30 |

Dataset ETH-UCY

**Note**: Initialization doesn’t affect the result significantly.

Component’s of GMM of S

Num sample in presence of the model’s noise.

Robustness experiments with different seeds.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Experiment** | **Output Model** | **Learning rate** | **Sample dist Y** | **Sample Dist Z,X,S** | **Contrastive Loss** | **Coupling Layer S Prior** | **Coupling Layer Z Prior** | **Encoder S** | **Encoder Z** | **End to End** | **Epoch** | **Result ADE** |
| E1 | Point  Dimensions:  d\_CouplingLayer=8  d\_FC\_after\_Enc=8,8  d\_latent=2 | No (1e-3) | 1 | 10 | no | yes | yes | lstm | lstm | yes | 1300 | 0.82 |
| E2 | Point  FC+Decoder  Dimensions:  d\_CouplingLayer=32  d\_FC\_before\_Dec=32  d\_FC\_after\_Enc=32,32  d\_latent=8 | No (1e-3) | 1 | 10 | no | yes | yes | lstm | lstm | yes | 500 | 0.76 |
| E3 | Point  Model Noise  FC+Decoder  Dimensions:  d\_CouplingLayer=32  d\_FC\_before\_Dec=32  d\_FC\_after\_Enc=32,32  d\_latent=8 | No (5e-3) | 10 | 10 | no | yes | yes | lstm | lstm | yes | 1000 | 0. |
| E10 | Point  FC+Decoder  Seed=1  Bk=20  Domain shifts  Dimensions:  d\_CouplingLayer=32  d\_FC\_before\_Dec=32  d\_FC\_after\_Enc=32,32  d\_latent=8  Domain shifts | No (1e-3)  One cylce | 1 | 10 | no | yes | yes | lstm | lstm | yes | 1000 | 0. |
| E11 | Point  FC+Decoder  Seed=2  Bk-20  Domain shifts  Dimensions:  d\_CouplingLayer=32  d\_FC\_before\_Dec=32  d\_FC\_after\_Enc=32,32  d\_latent=8 | No (1e-3)  One cylce | 1 | 10 | no | yes | yes | lstm | lstm | yes | 1000 | 0. |
| E12 | Point  FC+Decoder  Seed=3  Bk=20  Domain shifts  Dimensions:  d\_CouplingLayer=32  d\_FC\_before\_Dec=32  d\_FC\_after\_Enc=32,32  d\_latent=8 | No (1e-3)  One cylce | 1 | 10 | no | yes | yes | lstm | lstm | yes | 1000 | 0. |
| E13 | Point  FC+Decoder  Seed=4  Bk=20  Domain shifts  Dimensions:  d\_CouplingLayer=32  d\_FC\_before\_Dec=32  d\_FC\_after\_Enc=32,32  d\_latent=8 | No (1e-3)  One cylce | 1 | 10 | no | yes | yes | lstm | lstm | yes | 1000 | 0. |
| E14 | Point  FC+Decoder  Seed=5  Bk=20  Domain shifts  Dimensions:  d\_CouplingLayer=32  d\_FC\_before\_Dec=32  d\_FC\_after\_Enc=32,32  d\_latent=8 | No (1e-3)  One cylce | 1 | 10 | no | yes | yes | lstm | lstm | yes | 1000 | 0. |
| E15 | Point  FC+Decoder  Seed=1  Bk=1  With model noise  Dimensions:  d\_CouplingLayer=32  d\_FC\_before\_Dec=32  d\_FC\_after\_Enc=32,32  d\_latent=8 | No (1e-3)  One cylce | 10 | 10 | no | yes | yes | lstm | lstm | yes | 1000 | 0. |
| E16 | Point  FC+Decoder  Seed=1  Bk=1  Domain shift  With noise  Dimensions:  d\_CouplingLayer=32  d\_FC\_before\_Dec=32  d\_FC\_after\_Enc=32,32  d\_latent=8 | No (1e-3)  One cylce | 10 | 10 | no | yes | yes | lstm | lstm | yes | 1000 | 0.90 |
| E17 (Zahra) | Point  FC+Decoder  Seed=1  Bk=1  Dimensions:  d\_CouplingLayer=32  d\_FC\_before\_Dec=32  d\_FC\_after\_Enc=32,32  d\_latent=8 | No (1e-3)  One cylce | 10 | 10 | no | yes | yes | lstm | lstm | yes | 1000 | 0.878 |
|  |  |  |  |  |  |  |  |  |  |  |  |  |