Motion Capture Synthesis with Adversarial Learning

Qi Wang, Thierry Artières

Motion Capture Synthesis with Adversarial Learning*

Qi Wang[1,2] ✉ and Thierry Artières[1,2]

[1] Ecole Centrale Marseille, France
[2] LIF, Université d'Aix Marseille and CNRS, France

Abstract. We propose a new statistical modeling approach that we call Sequential Adversarial Auto-encoder (SAAE) for learning a synthesis model for motion sequences. This model exploits the adversarial idea that has been popularized in the machine learning field for learning accurate generative models. We further propose a conditional variant of this model that takes as input an additional information such as the activity which is performed in a sequence, or the emotion with which it is performed, and which allows to perform synthesis in context.

1 Introduction

Synthesizing realistic motion capture data is a key issue in the animation domain. A number of statistical models have been proposed for designing such synthesis systems with the expectation of producing highly realistic animation by learning these statistical models from large corpora of motion capture data [2,6]. These works rely on markovian models such as Hidden Markov models (HMMs) or Conditional Random Fields (CRFs). Such models rely on an assumption on the shape of the probability distribution of the data, which is usually strong and may lead to bad performance. We propose a new generative framework that we call Sequential Adversarial Auto-encoder (SAAE). It builds on recent advances in adversarial learning and on sequence autoencoders [9,4]. One may interest of adversarial learning is that it does not make any assumption on the distribution of the data. We propose here a specifi model for dealing with sequences and with motion capture data.

Moreover in the animation area, one needs to capture multiple variant styles of motions for controlling each animation character, a key feature is the high level control a designer may have on a synthesized animation. Few motion editing techniques have been proposed, such as inverse kinematics, style transfering, etc [7,5,8]. With this goal in mind we propose a variant of our framework enabling taking into account in the learning stage as well as during the synthesis stage such side information (e.g. activity, emotion, age, gender etc.)

* We are very grateful to Catherine Pélachaud for fruitful discussion and for access to and help with the Emilya dataset

2 Sequential Adversarial AutoEncoders (SAAE)

SAAE is built on the basis of two advanced ideas: seq2seq models[10], adversarial autoencoder [9]. Seq2seq has been proposed for machine translation tasks [10] and it is well adapted when there are complex and eventually long term dependencies or forward references in the sequence data. It consists of an encoder and a decoder and both of them are built on RNN. Given a sequence $\mathbf{x} = x_1, ..., x_T$, the encoder encodes $\mathbf{x}$ into a latent vector: $z = Enc(\mathbf{x})$ which is a fixed dimensional vector (usually low dimensional). Then, the decoder aims to reconstruct the input sequence from the latent vector z. It is expressed as $\hat{\mathbf{x}} = Dec(Enc(\mathbf{x}))$. The model is trained by minimising the reconstruction error on the training set, $\mathbf{E}_{x\sim p_{data}}\left[\Delta(\mathbf{x}, Dec(Enc(\mathbf{x})))\right]$ where Δ is a distance between a sequence $\mathbf{x}$ and its reconstruction by the autoencoder $\hat{\mathbf{x}} = Dec(Enc(\mathbf{x}))$.

In the generation process, we wish we can generate a new motion sequence from a given latent vector z which includes all the information of the motion to be generated. For achieving this, we need to know the distribution of the latent vector z. Therefore, we exploit adversarial learning to enforce the distribution of the latent codes z to satisfying a prior distribution $p(z)$. Then in the generation stage, a z can be sampled from the prior distribution $p(z)$ and fed into the decoder to obtain a new motion sequence. The adversarial learning part introduces a discriminator, D, into the modeling whose training samples come from two sources: On the one hand, the outputs of the encoder $Enc(\mathbf{x})$, for real data $\mathbf{x}$. On the other hand, noise vectors sampled from a prior distribution $p(z)$. The discriminator is trained to distinguish which source an input comes from while the encoder aims at increasing the probability that the discriminator makes a mistake, i.e. fooling the discriminator. It is a two player game. The whole training process of SAAE can be expressed as follows (1) and its structure is illustrated in Fig.1.

$$\min_{Enc,Dec}\max_{D}\Big\{\mathbf{E}_{x\sim p_{data}}\left[\Delta(\mathbf{x}, Dec(Enc(\mathbf{x})))\right] + \mathbf{E}_{x\sim p_{data}}\left[log\left(D(Enc(x))\right)\right] + \mathbf{E}_{z\sim p(z)}\left[log(1-d(z))\right]\Big\} \tag{1}$$

where the $D(\cdot)$ represent the output of the discriminator.

In order to exploit the contextual information (e.g. activity, emotion...), one may add contextual information as input to both the encoder and the decoder in the training process. This implicitly induces the model to learn the dependency between the motion and the contextual information. In any case after learning a SAAE, the decoder may be used to generate a sequence as explained above.

3 Experiments

We performed experiments with the Emilya Dataset [3]. It includes motion capture sequences performed by 12 actors corresponding to 8 activities performed

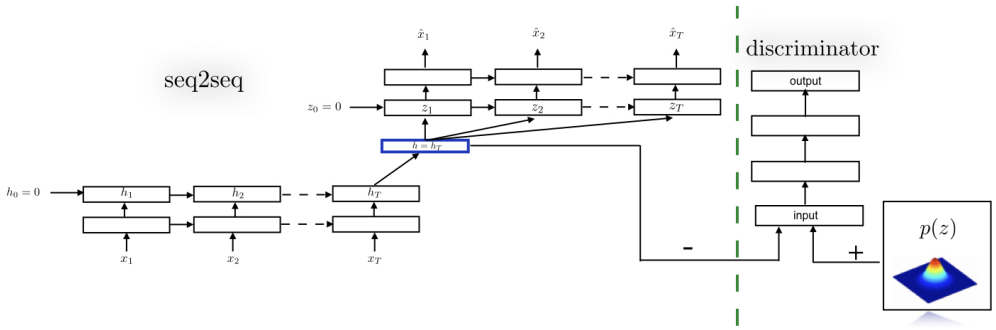

Fig. 1: Illustration of the Sequential Adversarial AutoEncoder (unfolded in time)

under 8 emotions. Each actor recorded each of 8 activity under each of the 8 emotions. All data are captured by 120Hz rate. Data are segmented, i.e. one sequence corresponds to a single activity performed by an actor under one emotion.

We first compare the standard seq2seq and SAAE with respect to their performance as a generative model.We follow [4,1] and use Gaussian Parzen Estimator. For each of these two models, we use its generated sequences to fit a Gaussian Parzen estimator which is the estimated PDF represented by this model. Then we randomly select 10 000 test sequences and we compute the mean log-likelihood of these under each of these two estimators. Note that to get a generative model from the standard seq2seq model we first estimate the distribution of latent vectors assuming a Gaussian distribution. The results are reported in Table 1 and show that SAAE outperforms Seq2Seq as a generative model on the test set.

Models	EmilyaDataset
Seq2Seq	1704.85 $\pm$ 5.117
SAAE	**1722.53 $\pm$ 3.344**

Table 1: Likelihood of the training data by a generative model gained from traditional (Seq2Seq) and from an adversarial (SAAE) learning.

Figure 2 shows examples of generated motion sequences from these two conditional SAAE. These qualitative results show that conditional SAAE can learn the variations corresponding to the contextual label and generate plausible motions matching the specified activity or emotion. More animation examples are available at `https://drive.google.com/drive/folders/0B8-1q1MIO1iJalhPcmxvZmpiMFk`.

4 Conclusion

We proposed to mix two recent ideas, adversarial learning and sequence to sequence models for proposing new models for the synthesis of motion capture

Fig. 2: Examples of generated motion from activity-conditioned ("Lift","Simple Walk") and from emotion-conditioned SAAE ("Sadness","Pride")

animations, with a conditional variant enabling synthesizing new sequences with a high level of monitoring.

References

1. Denton, E., Chintala, S., Szlam, A., Fergus, R.: Deep Generative Image Models using a Laplacian Pyramid of Adversarial Networks. Arxiv pp. 1–10 (2015)
2. Ding, Y., Prepin, K., Huang, J., Pelachaud, C., Artières, T.: Laughter animation synthesis. In: AAMAS, 2014
3. Fourati, N., Pelachaud, C.: Emilya: Emotional body expression in daily actions database. In: LREC. pp. 3486–3493 (2014)
4. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative Adversarial Nets. Advances in Neural Information Processing Systems 27 pp. 2672–2680 (2014)
5. Grochow, K., Martin, S.L., Hertzmann, A., Popovic, Z.: Style-based inverse kinematics. Acm Transactions on Graphics 23(3), 522–531 (2004)
6. Hofer, G., Shimodaira, H.: Automatic head motion prediction from speech data. In: INTERSPEECH. pp. 722–725 (2007)
7. Holden, D., Saito, J., Komura, T.: A deep learning framework for character motion synthesis and editing. ACM Transactions on Graphics 35(4), 1–11 (2016)
8. Huang, J., Wang, Q., Fratarcangeli, M., Yan, K., Pelachaud, C.: Multi-variate gaussian-based inverse kinematics. In: Computer Graphics Forum (2017)
9. Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I.: Adversarial Autoencoders. arXiv pp. 1–10 (2015), `http://arxiv.org/abs/1511.05644`
10. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to Sequence Learning with Neural Networks. Nips pp. 3104–3112 (2014)