ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence
and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6.
Available from http://www.i6doc.com/en/.

Transferring Style in Motion Capture Sequences with Adversarial Learning

Qi Wang[12], Mickael Chen[3], Thierry Artières[12], Ludovic Denoyer[3] *†

1- Ecole Centrale de Marseille
2-LIF, AixMarseille Université - CNRS UMR 7279
3- Laboratoire d'Informatique de Paris 6, Université Pierre et Marie Curie

Abstract. We focus on style transfer for sequential data in a supervised setting. Assuming sequential data include both content and style information we want to learn models able to transform a sequence into another one with the same content information but with the style of another one, from a training dataset where content and style labels are available. Following works on image generation and edition with adversarial learning we explore the design of neural network architectures for the task of sequence edition that we apply to motion capture sequences.

1 Introduction

Synthesizing realistic motion capture data is a key issue in the animation domain. Statistical generative models have been proposed for designing synthesis systems that generate rather realistic animation by learning statistical models from large corpora of motion capture data [1, 2]. A related task is motion editing when one wants to generate a new sequence from an existing one by transforming it in some way, e.g. in the style of another sequence. Motion edition has been studied in the last decade by researchers in animation and graphics fields with dedicated models [3, 4, 5]. Recently neural networks learned within an adversarial learning framework has become a key strategy for learning accurate generative models for complex data [6]. Such models have achieved impressive results, mainly on images and videos. A number of extensions of the seminal work by Goodfellow et al. have been proposed in particular for image edition [7, 8] and more generally for the disentanglement of content and style in images [9, 10, 11] and videos [12]. Following this trend most recent works that have been done on motion capture data, for synthesis as well as for edition [13, 14], have focused on neural networks, building on their well known capability of automatically learning relevant representation that has led to recent achievements in computer vision, natural language processing and speech recognition. We propose here to build on such previous works to design neural network architectures allowing to perform sequence edition and we evaluate their potential on motion capture data performed under emotion on a well known dataset of the field [15].

*Qi Wang's Ph.D. thesis is funded by China Scholarship Council.
†Part of this work has been funded by the French ANR project Deep In France.

2 Background and related works

Our proposal is based on adversarial learning and in particular on **adversarial autoencoders** [16]. An adversarial autoencoder consists in an autoencoder (composed of an encoder $\mathbf{e}$ and of a decoder $\mathbf{d}$) and a discriminator $\mathbf{D}_a$ that are jointly learned. The autoencoder ($\mathbf{e}$, $\mathbf{d}$) is learned to encode input data x into a representation (or encoding) space $\mathbf{e}(x)$ then to reconstruct the input from this representation, i.e. making that $\mathbf{d}(\mathbf{e}(x)) \approx x$. Besides the encoder is also learned to fool the discriminator $\mathbf{D}_a$ which aims to distinguish between noise samples z, drawn from a chosen prior noise distribution p_n, usually a Gaussian distribution, and the encodings of true data $\mathbf{e}(x)$. This learning makes the encoder map the input data into a representation space with the noise prior distribution so that after learning, the decoder may be used as a generative model by first sampling z from the prior distribution p_n then by computing $\mathbf{d}(z)$.

Sequential Adversarial Autoencoder (SAAE) [14] are an adversarial extension of sequence to sequence autoencoders (SA), a particular case of sequence to sequence models (S2S) [17]. An SA is composed of an encoder and of a decoder that are implemented as recurrent neural networks (RNNs) exploiting RNNs' ability to transform a variable lengthed sequence into a fixed sized vector. It aims at reconstructing an input sequence at its output while compressing the input sequence in a low dimensional representation space, the encoding space. A SAAE consists in the association of a SA and of a discriminator D_a which aims at discriminating between random vectors following a prior distribution (e.g. a Gaussian distribution p_n) and the encodings of training sequences [14]. As for adversarial autoencoders once an SAAE has been learned, its decoder part (a RNN) may be used as a generative model. SAAE have been shown promising results for synthesizing from scratch new realistic motion capture sequences.

3 Adversarial Learning for Style Transferring between motion capture sequences

Following works on image editing with adversarial learning we build on SAAE for designing a system able to transfer style from a sequence to another one. Note that although we focus on motion capture data from now on, we believe our work is generic enough for dealing with other signal and/or sequential data. We consider now that the training set consists of a number of sequences that are performed under a particular style, where the number of styles is finite. Actually we focus on motion capture sequences that correspond to various activities (e.g. walk, run) performed under emotion (e.g. pride, fear). We consider the setting where the style information is available at training time.

We first describe a NN architecture, Model1, that is dedicated to transform a given sequence by changing its style (emotion), where the style is considered a categorical variable. The architecture of the model is shown in left part of Figure 1. It includes two main components. The first one is an Sequential Adversarial Autoencoder (SAAE) as described above where the decoder is slightly different from the one used in SAAE since it takes as input, every time step, a one hot encoding of the input sequence's style/emotion label. The rationale behind this is, after learning, to enable generating new sequences conditioned on a chosen emotion. To achieve this the emotion information should be as less present in the input sequence encoding as possible. This motivates the second main component of the model. We add a style discriminator, D_s, that is learned to recover the style information of the input sequence from its encoding representation. Following the idea in [18], the generator is learned so as to fool this discriminator by back-propagating the reverse of the style discriminator gradient in the encoder. Once such a model has been learned one may exploit the decoder to generate new sequences conditioned on a chosen emotion label. The model is learned by optimizing the following loss:

$$\min_{\mathbf{e},\mathbf{d}} \max_{D} \quad \mathbf{E}_{x,s\sim p_d}\left[\delta(\mathbf{x},\mathbf{d}(\mathbf{e}(\mathbf{x}),s))\right] + \mathbf{E}_{x,s\sim p_d}\left[log\left(1-D_a(\mathbf{e}(x))\right)\right] \\ +\mathbf{E}_{z\sim p_n}\left[logD_a(z)\right] - \mathbf{E}_{x,s\sim p_d}\left[H_s\left(D_s(\mathbf{e}(x))\right)\right]$$

where δ is a distance between sequences (Euclidean distance in our case), p_d stands for the empirical distribution of data (pairs of a sequence and of its style label, (x,s)), and H_s stands for the cross entropy criterion for the style discriminator D_s.

In Model1 discrete style label is used for controlling the style of generated sequence. However, discrete label is not enough to encoder enough variations and subtleties of style. Thus we propose Model2 illustrated in the right part of Figure 1 to learn continuous embedding of style. In Model2 the encoder produces two encodings, one for the core information of the input sequence, called content, (e.g. the activity which is performed) while the other one encodes the style information (e.g. the emotion). Both encodings are input to the decoder. As in SAAE a discriminator enforces the content encoding to obey a given prior distribution. Two style discriminators are added to the model. A first one takes as input the content encoding, it is used in the same way as above. Its parameters are learned to recover the style information from the content encoding but the reverse of its gradient is back-propagated in the encoder to make content encoding free from the style information. A second style discriminator takes as input the style encoding. It is learned to recover the style label and its gradient is back-propagated in the encoder as is, so that the encoder should learn to actually include the style information in the style encoding. We do not detail the loss for lack of place. Once such a model has been learned one may transfer the style of a sequence x_2 to another sequence x_1 as follows (Figure 2). Both sequences are processed by the encoder, yielding a pair of latent codes for each of the two sequences x_i, (c_i, s_i). Then the decoder is used to process the pair of latent codes (c_1, s_2) composed of the content latent code of sequence x_1 and the style latent code of sequence x_2.

4 Experiments

We performed experiments on the Emilya Dataset [15]. It includes motion capture sequences for 8 activities performed under 8 emotions, by 12 actors. We

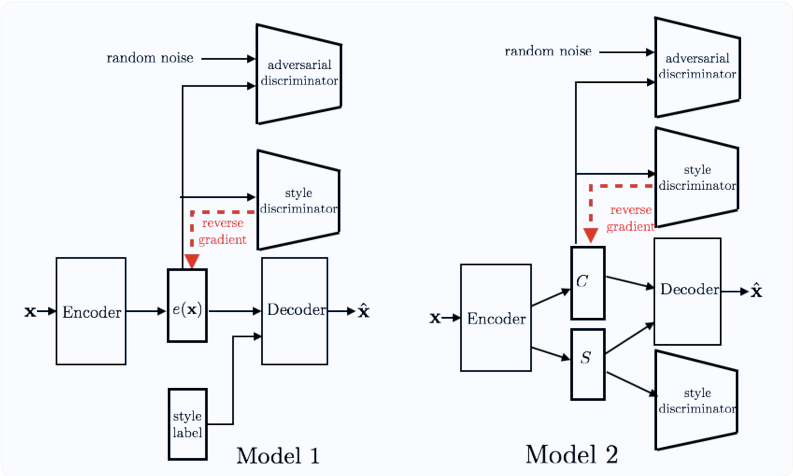

Fig. 1: Architectures of Model1 and Model2.

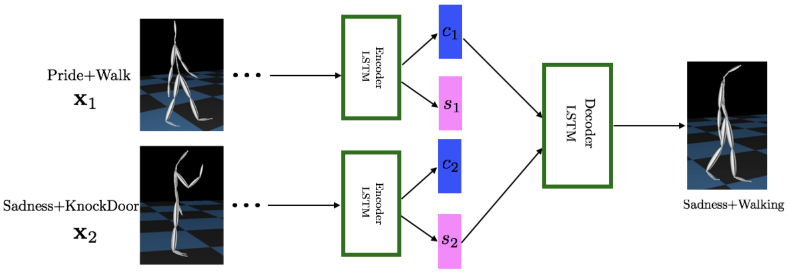

Fig. 2: Transferring style between sequences.

split sequences in windows of 200 frames (about 1.7s length) to focus on short dynamics. We split the dataset across $(activity, emotion, actor)$ labels to make sure they are distributed evenly in training, validation and test set. The training/validation/test sets include 78 982/21 614/38 637 sequences of two hundreds 70-dimensional frames. We compared our models to SAAE and to Variational Autoencoders for sequences (SVAE), which do not exploit the style information. Note that to generate sequences with non adversarially learned SVAE we first estimate, after training a Gaussian distribution of the latent codes computed from training sequences. Then we use this distribution and the decoder part as we explained for generating with SAAE.

We provide below objective results to measure the behaviour of our models. Some animations also can be found here [1]. We first evaluate the quality of the learned generative models by estimating the *likelihood of the test data*. To compute such a likelihood we follow [6] and use a Gaussian Parzen Estimator. We fit the Gaussian Parzen estimator on generated sequences to get an estimated

[1] `https://drive.google.com/open?id=1Nalyw9nW5Qd9dvWFXxkx0zl0c82a2e0n`

PDF of the generative model. To exploit the idea on sequences we consider fixed length generated sequences that we reshape as vectors. Then we randomly select 10 000 test sequences and we compute the mean log-likelihood of these under the estimator. Table 1a compares the results for SAAE, SVAE and the proposed models (with a 5 dimensional style encoding in Model2). As shown here our adversarially learned model handling the style information reach a significantly higher likelihood than both SAAE and SVAE. Next we report in Table 1b statistics that provide insights on the diversity of the sequences generated by a particular model, their quality, and their completeness (meaning generate data cover the whole variety of true sequences). Statistics are computed as follows for one particular generative model. We generated a set of 60 000 sequences. For each generated sequence we computed its minimum distance to a true sequence from the validation and test set ($\approx$ 60 000 sequences). We report this average minimum distance as G2T criterion (Generated to True). We compute similarly T2G (True to Generated). For Model1 and Model2,compute the same distance metric between transformed sequences and true sequences.

We provide results obtained with SAAE and SVAE as a reference. One sees that our models, relying on a disentanglement of content and style, allow generating more realistic generated sequences (lower G2T) and that all modes of the true distribution are well covered (low T2G), than models that do not take into account the style.

Models	Likelihood
SAAE	1730 ± 11
SVAE	1719 ± 19
Model1	1796 ± 11
Model2-2dim	1809±11
Model2-3dim	1815±11
Model2-5dim	1808± 10

(a)

Models	*G2T*	*T2G*
SAAE	1.11	0.926
SVAE	1.08	0.90
Model1	0.9069	0.835
Model2-5dim	0.8274	0.766

(b)

Table 1: (a). Likelihood estimation on test set (b).Distance statistics

Models	**True Sequences**	**M1**	**M2-2dim**	**M2-3dim**	**M2-5dim**
Accuracy	82%	45.33%	41.58%	48.02%	55.52%

Table 2: Emotion classification accuracy on test sequences and on sequences generated by Model1 and Model2 (with style encoding sizes of 2 to 5).

Finally Table 2 reports accuracy of an emotion classifier operating on sequences (with a similar architecture as the encoder in our models) that has been learned on training sequences and that is evaluated on style transformed sequences. Although the achieved accuracy is significantly lower than the performance achieved on true sequences from the test set, it may be seen that sequences generated with Model2 allow reaching up to 55% accuracy demonstrating the ability of our framework to indeed transfer style between sequences.

5 Conclusion

We presented a neural network architecture able to perform sequence edition tasks by transfer style from a sequence to another. We demonstrated the ability of this transfer on motion capture data. We plan to investigate deeper the potential of our framework on other kind of sequential data.

References

[1] S. Levine, J.M. Wang, A. Harauxand Z. Popović, and V. Koltun. Continuous character control with low-dimensional embeddings. *ACM Transactions on Graphics*, 31(4):1–10, 2012.

[2] Gregor Hofer and Hiroshi Shimodaira. Automatic head motion prediction from speech data. In *INTERSPEECH*, pages 722–725, 2007.

[3] S. Xia, C. Wang, J. Chai, and J. Hodgins. Realtime style transfer for unlabeled heterogeneous human motion. *ACM Transactions on Graphics (TOG)*, 34(4):119, 2015.

[4] K Grochow, S L Martin, A Hertzmann, and Z Popovic. Style-based inverse kinematics. *Acm Transactions on Graphics*, 23(3):522–531, 2004.

[5] M Ersin Yumer and Niloy J Mitra. Spectral style transfer for human motion between independent actions. *ACM Transactions on Graphics (TOG)*, 35(4):137, 2016.

[6] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative Adversarial Nets. *NIPS 27*, 2014.

[7] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *CoRR*, abs/1606.03657, 2016.

[8] Guim Perarnau, Joost van de Weijer, Bogdan Raducanu, and Jose M. Álvarez. Invertible conditional gans for image editing. *CoRR*, abs/1611.06355, 2016.

[9] M. Mathieu, J. Jake Zhao, P. Sprechmann, A. Ramesh, and Y. LeCun. Disentangling factors of variation in deep representations using adversarial training. *CoRR*, abs/1611.03383, 2016.

[10] M. Chen and L. Denoyer. Multi-view generative adversarial networks. *CoRR*, abs/1611.02019, 2016.

[11] M. Chen, L. Denoyer, and T. Artieres. Multi-view data generation without view supervision. *CoRR*, abs/1711.00305, 2017.

[12] Emily Denton and Vighnesh Birodkar. Unsupervised learning of disentangled representations from video. *CoRR*, abs/1705.10915, 2017.

[13] Daniel Holden, Jun Saito, and Taku Komura. A deep learning framework for character motion synthesis and editing. *ACM Transactions on Graphics*, 35(4):1–11, 2016.

[14] Qi Wang and Thierry Artieres. Motion capture synthesis with adversarial learning. In *International Conference on Intelligent Virtual Agents*, pages 467–470. Springer, 2017.

[15] Nesrine Fourati and Catherine Pelachaud. Emilya: Emotional body expression in daily actions database. In *LREC*, pages 3486–3493, 2014.

[16] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, and Ian J. Goodfellow. Adversarial autoencoders. *CoRR*, abs/1511.05644, 2015.

[17] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to Sequence Learning with Neural Networks. *Nips*, pages 3104–3112, 2014.

[18] Y. Ganin and V. Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*. JMLR Workshop and Conference Proceedings, 2015.