Data augmentation through space linearization

Grigorios G. Chrysos[1], **Yannis Panagakis**[1,2], **Stefanos Zafeiriou**[1]
[1] Department of Computing, Imperial College London, UK
[2] Department of Computer Science, Middlesex University London, UK
{g.chrysos, i.panagakis, s.zafeiriou}@imperial.ac.uk

Abstract

The state-of-the-art learning-based methods in machine learning tasks require a vast amount of samples to be trained. Such methods have hundreds of millions of parameters, which translates to more parameters than training samples to train them on. A standard approach to ameliorate the lack of labelled training samples is data augmentation (e.g. cropping, affine transformations). The two major categories of data augmentation methods are i) simple transformations (e.g. affine transformations, pixel intensity adaptations), ii) tailored to the task in hand (e.g. 3D model based). Instead, we propose to learn local transformations to synthesize new images that include the same scene/object with small adaptations. To that end, we introduce a three-stage approach that finds a low-dimensional (approximately) linear space. We learn a forward transformation from the image to the latent space, we perform a linear operation in this space and learn the inverse transformation (from latent to the image space) to synthesize a new image. We illustrate how this three-stage approach can be used to build powerful adaptive models for deformable facial tracking.

1 Introduction

In the era of deep learning, in which the scale of the dataset has a vast effect on the performance, different data augmentation methods are commonplace during training. These label-preserving transformations help to i) avoid over-fitting, ii) provide enough training samples for learning the millions of parameters (He et al. [2016]). We argue that the most commonly used augmentation methods are either simple (rotation, zoom) or task-specific (tailored for a handful of tasks) and propose instead a method that performs powerful local transformations to the image.

The data augmentation methods employed can be divided into i) image-level transformations, ii) model-based transformations. The first category includes affine transformations or pixel-level adjustments/noise, e.g. color perturbation (Krizhevsky et al. [2012]). Even though such methods can be applied in a variety of computer vision tasks, they consist a limited group of transformations in which partial invariance has been achieved in the past. The latter category, i.e. model-based transformations, e.g. the 3D profiling of Zhu et al. [2016] or the novel-view synthesis from 3D models (Rematas et al. [2014], Chandra et al. [2016]), typically require a very accurate 3D object model (Zhu et al. [2016], Tran et al. [2017]) or synthetic data (Rematas et al. [2014]). Such 3D models are available for only a small number of classes, while the realistic generation from 3D models/synthetic data is still an open problem (Bousmalis et al. [2017]), hence these model-based transformations are available for a subset of the computer vision tasks.

We argue that local transformations can be employed as a third category of augmentations. The code idea assumes the existence of an (approximately) linear low-dimensional space, which can describe the high dimensional nonlinear image space. We form a three-stage approach that learns the transformations from the image space to this latent space and enables us to perform a linear operation in the low-dimensional space. Specifically, i) we learn a nonlinear transformation that maps the image

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

to this low-dimensional space, then ii) we perform a linear transformation that maps the original latent representation to the representation of a slightly transformed image (e.g. a pair of successive frames in a video). As a third step, we learn the inverse transformation, i.e. mapping from the low-dimensional space to the transformed image, hence a linear operation in the latent space results in a nonlinear change in the image space. Both the forward (from image to latent space) and the inverse transformations are approximated by neural networks, specifically an Adversarial Autoencoder and a Generative Adversarial Network. Even though a generic model for augmenting all classes could be theoretically learned, we introduce object-specific transformations. We experimentally demonstrate how our proposed method can be used for augmenting the images of human faces[1].

2 Method

We want to find a low-dimensional and (approximately) linear space that can represent the object-specific attributes that our images contain. We learn both the transformation from the image space to this latent space, as well as the inverse transformation. Then, the latent representation $\boldsymbol{d}^{(t)}$ (corresponding to an arbitrary image $\boldsymbol{i}^{(t)}$) can be linearly transformed into a different representation $\boldsymbol{d}^{(t+x)}$, which we can transform back to the image space as $\boldsymbol{i}^{(t+x)}$. The newly synthesized image $\boldsymbol{i}^{(t+x)}$ differs from $\boldsymbol{i}^{(t)}$ with some local nonlinear transformation, i.e. the same object and scene are included, hence the original image has been augmented with a local transformation. We describe below the three-stage approach that achieves this augmentation.

2.1 Stage I: From image to latent representation

We follow an unsupervised learning approach for extracting the image representations. We utilize an Adversarial Autoencoder (Makhzani et al. [2015]). An Adversarial Autoencoder is composed of i) a generator (Autoencoder), ii) a discriminator. The discriminator aims in discerning the synthesized samples (outputs of generator) from the true data distribution, while the generator's task is to generate samples that resemble the true distribution's samples. The generator accepts an image $\boldsymbol{i}^{(t)}$, encodes it to $\boldsymbol{d}^{(t)}$ (we assume that the latent vector $\boldsymbol{d}^{(t)}$ lies in the latent space we want to find) and then decodes it to $\hat{\boldsymbol{i}^{(t)}}$.

2.2 Stage II: Linear transformation in the latent space

Our goal is to create a realistic synthetic image (which is a nonlinear transformation of image $\boldsymbol{i}^{(t)}$) by performing a linear transformation in the latent representation of $\boldsymbol{i}^{(t)}$. A simple but very powerful way to perform this change is to learn a linear regression model in the latent space. Assuming we have N pairs of correlated images available[2], i.e. $\boldsymbol{I} = \{(\boldsymbol{i}_k^{(t_k)}, \boldsymbol{i}_k^{(t_k+x_k)})\}$, we obtain (from Stage I) $\boldsymbol{\mathcal{D}} = \{(\boldsymbol{d}_k^{(t_k)}, \boldsymbol{d}_k^{(t_k+x_k)})\}$. This is formally expressed as:

$$\boldsymbol{d}^{(t_k+x_k)} = \boldsymbol{A} \cdot [\boldsymbol{d}^{(t_k)}; 1] + \boldsymbol{\epsilon} \tag{1}$$

where $\boldsymbol{\epsilon}$ is the noise; we compute $\boldsymbol{A}$ (closed-form solution) from real video samples.

2.3 Stage III: From latent to image representation

The last step consists in learning the transformation from the latent representation $\hat{\boldsymbol{d}}^{(t+x)}$ to the image representation $\boldsymbol{i}^{(t+x)}$. We utilize GAN (Goodfellow et al. [2014]) as they have demonstrated results of high visual quality (Ledig et al. [2017], Pathak et al. [2016]). The input to the GAN is the (transformed) latent representation $\hat{\boldsymbol{d}}^{(t+x)}$ along with noise, while the expected output is a (nonlinearly) transformed version of $\boldsymbol{i}^{(t)}$, i.e. an image $\hat{\boldsymbol{i}^{(t+x)}}$.

[1] The facial space is highly nonlinear, while such a method would have significant application in a host of face-related tasks.

[2] During the learning process those can be images from successive frames of a sequence of frames.

2.4 Prediction

Once the learning is completed, the structure for prediction is greatly simplified, see Fig. 1. Specifically, the image $\boldsymbol{i}^{(t)}$ is encoded (only the pre-trained encoder is used); the resulting representation $\boldsymbol{d}^{(t)}$ is multiplied by $\boldsymbol{A}$ to obtain $\hat{\boldsymbol{d}}^{(t+x)}$, while the last step consists of feeding $\hat{\boldsymbol{d}}^{(t+x)}$ to the GAN to synthesize a new image $\hat{\boldsymbol{i}}^{(t+x)}$.

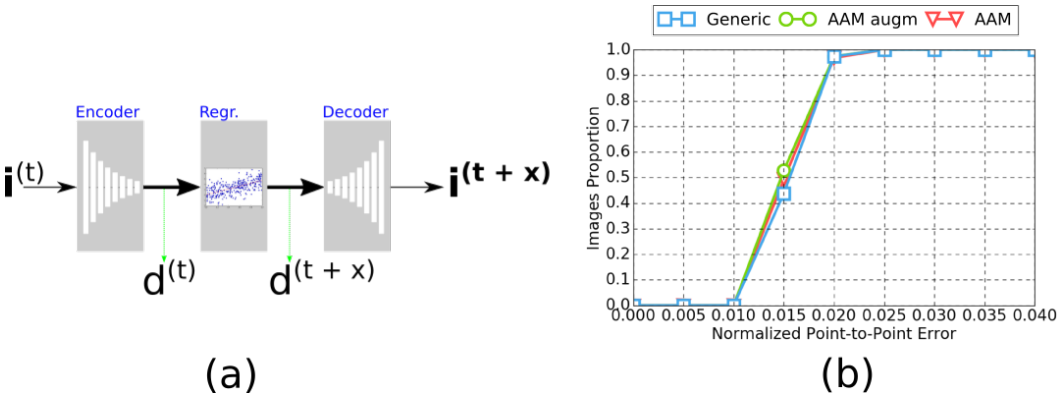

Figure 1: (Preferably viewed in color) (a) Prediction procedure from an image $\boldsymbol{i}^{(t)}$, (b) Cumulative Error plot comparing the adaptive model learned with and without augmentation (Sec. 3.1).

3 Experiments

To quantitatively validate our model, we have used our proposed method to augment the samples used for learning adaptive deformable models for tracking. Let us first provide the implementation details along with information about the datasets used.

The networks of Sec. 2.1 and Sec. 2.3 share the same architecture, i.e. the encoder/decoder in both cases is composed by 8 layers followed by batch normalization (Ioffe and Szegedy [2015]); both discriminators consist of 5 layers, while the latent dimension is fixed to 1024 for all cases.

The databases of Guo et al. [2016] (Stage I) and Chrysos and Zafeiriou [2017a] (Stages II and III) are utilized for the learning parts, while the database of Shen et al. [2015] for the experimental validation.

3.1 Augmented adaptive deformable models

Even though our data augmentation technique is not confined to geometric tasks, we indicate its application in building adaptive deformable models (Sánchez-Lozano et al. [2016], Chrysos and Zafeiriou [2017b]). Adaptive deformable models typically rely on very few samples to learn the statistics of the specific video. As already demonstrated in Chrysos and Zafeiriou [2017b] adaptive deformable models can surpass the performance of meticulously trained generic methods, hence if we can introduce additional invariances to the adaptive models, we could improve the performance of deformable tracking methods.

The experiment was conducted in a video of 300VW that was not used in any learning part. The video is comprised of more than 900 frames; the generic tracking was achieved by the state-of-the-art landmark localization of Yang et al. [2017]. Using K-Means (with clusters $K = 10$) in the extracted shapes, we have sampled one shape from each cluster and used those shapes and corresponding appearances to learn an Active Appearance Model (AAM) of Cootes et al. [2001]. Two different alternatives were considered: i) a model with the original shapes, ii) a model trained with our data augmentation method. The latter one included the original shapes/appearances along with their augmentations, i.e. each image was augmented and the new shape was extracted with Yang et al. [2017]. The model with the augmented shapes included at most two times the original number of shapes. These AAM models were then used to refine the shapes of the generic method. In Fig. 1 the CED curves comparing i) the original fitting results (denoted 'Generic'), ii) the AAM with the cluster shapes (denoted 'AAM'), iii) the AAM with the augmented shapes (denoted 'AAM augm') are

visualized. Even though very few frames were selected, both adaptive methods improve the results of the meticulously trained generic method. In addition, our augmented samples improve the AAM bases, hence result in an improved fitting.

4 Conclusion

In this work, we have introduced a three-stage approach for locating a latent space where a linear change in the image representation (in that space) results in nonlinear changes in the image space. The three proposed stages include a forward transformation (from the image to the representation space), a linear transformation in that space, an inverse transformation (from the transformed representation to the image space). We have experimentally validated that our model can be used for augmentation in facial images and illustrated how this can be used for augmenting the samples used for constructing adaptive deformable models.

Acknowledgement

The work of G. Chrysos has been funded by an Imperial College DTA. The work of Y. Panagakis is partially supported by the EPSRC project EP/N007743/1 (FACER2VM). The work of S. Zafeiriou has been partially funded by the FiDiPro program of Tekes (project number: 1849/31/2015), as well as the EPSRC project EP/N007743/1 (FACER2VM).

References

K. Bousmalis et al. Unsupervised pixel-level domain adaptation with generative adversarial networks. *CVPR*, 2017.

S. Chandra, G. Chrysos, and I. Kokkinos. Surface based object detection in rgbd images. In *BMVC*, 2016.

GG Chrysos and S. Zafeiriou. Deep face deblurring. *CVPR Workshops*, 2017a.

GG Chrysos and S. Zafeiriou. Pd^{2}t: Person-specific detection, deformable tracking. *T-PAMI*, 2017b.

TF Cootes, GJ Edwards, and CJ Taylor. Active appearance models. *T-PAMI*, 23(6):681–685, 2001.

I. Goodfellow et al. Generative adversarial nets. In *NIPS*, 2014.

Y. Guo et al. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *ECCV*, pages 87–102, 2016.

Kaiming He et al. Deep residual learning for image recognition. In *CVPR*. IEEE, 2016.

S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015.

A. Krizhevsky et al. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012.

C. Ledig et al. Photo-realistic single image super-resolution using a generative adversarial network. *CVPR*, 2017.

A. Makhzani et al. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015.

D. Pathak et al. Context encoders: Feature learning by inpainting. In *CVPR*, pages 2536–2544, 2016.

K. Rematas et al. Image-based synthesis and re-synthesis of viewpoints guided by 3d models. In *CVPR*, 2014.

E. Sánchez-Lozano et al. Cascaded continuous regression for real-time incremental face tracking. In *ECCV*, 2016.

J. Shen et al. The first facial landmark tracking in-the-wild challenge: Benchmark and results. In *ICCV Workshops*, 2015.

Luan Tran et al. Disentangled representation learning gan for pose-invariant face recognition. In *CVPR*, 2017.

J. Yang et al. Stacked hourglass network for robust facial landmark localisation. In *CVPR Workshops*, 2017.

X. Zhu, Z. Lei, Z. Liu, H. Shi, and S Z Li. Face alignment across large poses: A 3d solution. In *CVPR*, 2016.