Name: Aliya Amirzhanova
Course: Scientific Computing (M.Sc.)
Student number: 3525113
Date: August 30, 2018

Seminar: Explainable Machine Learning SS2018

Deep Feature Interpolation for Image Content Changes

published by P. Upchurch, J. Gardner, G. Pleiss, R. Pleiss, N. Snavely, K. Bala and K. Weinberger

Aliya Amirzhanova

Image content changes is one of the most appealing and challenging area in Computer Vision and graphics. This field has been getting a high attention because of its intriguing visual results and possibility of performing semantic transformations. Following this tendency, this seminar report presents a research paper for image content changes by Upchurch, Gardner, Pleiss, Pleiss, Snavely, Bala, and Weinberger. In their work [12] authors presented a new technique - Deep Feature Interpolation (DFI), which according to Upchurch et al. can be considered as a baseline for image transformations.

Contents

1 Introduction

Nowadays the image content change is a captivating field in Computer Vision. It is attractive due its alluring performance of being capable to reproduce images with a desired attribute of any choice. Today there are many research are being conducted in this area. One of the recent works in this area is the approach of Upchurch et al. [12]. They focused on a Deep Feature Interpolation technique, which is considered to produce an output with a desired different attribute in a straightforward manner. Hence, this seminar paper discusses the scientific paper by Upchurch et al. [12], giving a detailed view on their used method and evaluating it via different perspectives showing advantages of the used technique and disadvantages compared to the other methods.

1.1 Outline

This paper is structured in the following way. The next subsection describes motivation to conduct a research in a field of the image transformation and later gives an overview on a related work in this area. The following section shows a detailed view of the DFI method. Later, a short introduction of the used dataset in the experiment of Upchurch et al. is provided. Then acquired results with the evaluation on those results are illustrated. The last section concludes the report summarizing insights on the article's strengths, weaknesses, and impact.

1.2 Motivation

As it has been stated previously, today the image transformation is a popular topic of interest. Thus, the consecutive question arises, "Why are image content transformations inspirational?" Generating new image or transforming old ones are viewed as the new image formation or an artistic work most conceivably with the help of Neural Networks with its own features of creativity. A new image creation has been also found its interpretation in the work "The Poetics of Space" of Gaston Bacherald, where he states, "when the image is new, the world is new", meaning that the world is progressing forward. Moreover, there a parallel can be drawn with an appearance and development of new methods and techniques, which are considered as drivers of the progress at some degree. Furthermore, semantic transformation has always been on a high level of curiosity for people. These transformations are able answering the most popular question "what if...?". For example, "what if my hair will be blond", "what if I wear glasses" or "what if I look like a man" and etc. In addition, consider an online shopping scenario, where a

user is looking for the shoes and has found a pair that mostly suits him, but he would like them to be a little taller, or wider, or in a different color. Thus, the study of Upchurch et al. [12] make it possible to do such kind of transformation. As for image inpainting, it is an important research field in image restoration that can be used to retouch damaged images and videos, remove text, and conceal errors in videos. Image inpainting has a very high application value and has recently received an increasing attention. It is a challenging task, as reconstruction of large regions requires connoisseurship and precision. Hence, the work of Upchurch et al. deals also with image inpainting, presenting DFI on missing regions.

2 Related work

Today there exists many related works concerning the current topic. These works can be divided into several groups regarding used techniques. One of them is generative methods. As prime examples might be considered approaches of Larsen et al. [8] and Radford et al. [11]. Larsen et al. introduced "...an autoencoder that leverages learned representations to better measure similarities in data space" [8]. What is appealing here is that authors combined a variational autoencoder with a generative adversarial network, making it possible to outperform VAEs in terms of visual representation. Radford et al. presented deep convolutional generative adversarial networks (DCGANs), which showed ability of "...easy manipulation of many semantic qualities of generated samples..." [11] alongside with the interpolation capabilities of GAN. Another approach from this group can be considered the work of Brock et al. [1]. They also used a hybridization of the VAE and GAN, which called Introspective Adversarial Network, by manipulating latent space variables for editing the image content. Despite that the method achieves competitive results, it is limited in the image resolution. Another approach of Lample et al. [7] is designed to learn a disentangled latent space using "Fader Networks", where an explicit control on some attributes of interest is achieved. There are many works on GAN or its extensions such as InfoGAN [2] for transforming effects. In contrast the method of Upchurch et al. [12] can achieve similar results on a discriminatively trained feature space. These methods of using GAN are the most similar to the DFI approach; however, those methods basically rely on specially trained generative autoencoders. Moreover, those generative methods primely deal with generating new images instead of inpainting or reconstructing the images.

Image editing can also be done by minimizing the witness function of the Maximum Mean Discrepancy as in the approach of Gardner et al. [3]. MMD is a statistic test which measures the difference between source and target probability distributions. This general method produces believable results and is versatile in tasks as can be applied to edit the contents of faces, cities and nature scenes. The main disadvantage of this work is time and memory consumption, which grows linearly.

In addition, there are several alternative ways to make image content changes such as a live puppetry system using a real-time 3D scanner [13] , the cross-dissolve technique [6] or facial reenactment system using matching techniques [4]. Compared to those methods, the work of Upchurch et al. [12] is less complicated and produces believable results.

In case of image inpainting recent works using Convolutional Neural Networks (CNN) has performed a capability of generating meaningful content for missing regions [10]. However, most of them require complex architectures with many parameters, a huge amount of dataset and long training times [10], while DFI method resolves this issue.

A totally different approach has been offered by Isola et al. [5]. Their method's idea is to use PatchGAN, mainly GAN with focus on the local patches' structure, which decreases the effect of downsampling. However, their method does not work well with non-texture data, as it fails to transfer semantic features [9].

3 Deep feature interpolation

This section describes the method in detail. Preceding to the introduction of algorithm, a few words on interpolation will be given.

3.1 Interpolation

Interpolation is a smooth transition between two known locations. Making interpolation in an image space is not effective as it produces ghosting and artifacts, while in the latent space it is smooth and convincing. The interpolation is done by walking along the connecting shortest line as in the example (fig. 1):

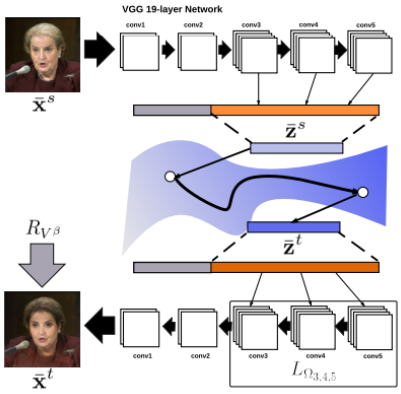


Figure 1: An example of deep manifold traversal in latent space [3]

3.2 DFI algorithm

The method itself works in the following way. Provided with x – a test image, which should be changed in a believable form with respect to a given attribute, assume that there is also access to the $S^t = \{x_1^t, ..., x_n^t\}$ – a set of target images with the desired attribute (men with facial hair) and $S^s = \{x_1^s, ..., x_m^s\}$ – a set of source images without the attribute (men without hair). Thus, also provided with a pre-trained VGG network it is possible to obtain a new representation of an image, which is denoted as: $x -> \phi(x)^*$, where $\phi(x)$ – a vector consisting of concatenated activations of the convnet when applied to the image x (deep feature representation of x) [12]. Then the algorithm itself can be described with following high-level steps [12]:

1. Map images in the target and source sets into deep feature representation through the pre-trained convnet. Obtain ϕ^t and ϕ^s.

2. Compute the mean feature values for each set of images. Compute mean $\bar{\phi}^t$ and $\bar{\phi}^s$, define their difference as the attribute vector w: $w = \bar{\phi}^t - \bar{\phi}^s$.

3. Map the test image x to a point $\phi(x)$ and move along the vector w: $\phi(x) + \alpha w$.

4. Reconstruct the transformed output image z by solving the reverse mapping into pixel space: $\phi(z) = \phi(x) + \alpha w$. The usage of pre-trained models provide the simplicity of architecture and less time consumption.

Therefore, the transformed output image z can be acquired by discriminatively iterating N times:

$$z = argmin_z \frac{1}{2} \|(\phi(x) + \alpha w) - \phi(z)\|_2^2 + \lambda_{V^\beta} R_{V^\beta}(z), \quad (1)$$

where R_{V^β} is the Total Variation regularizer, which is responsible for smooth transitions between neighboring pixels [12].

4 Experiments and discussions

This section will introduce used dataset and several results on that data obtained by the authors, and following by thorough discussion.

4.1 Dataset

The data used by authors for performing and testing DFI method is:

- Labeled Faces in the Wild (LFW) dataset with 13, 143 images of faces and containing predicted annotations for 73 different attributes. Testing was conducted using 38 images and 6 different attributes.

- A collection from CelebA, MegaFace, Helen and Google image search images (overall 100,000) for performing DFI on high-resolution images.

- A shoes subset of UT Zappos50k with 29, 771 images.

All the data images were aligned using DLIB tool, cropped and resized.

4.2 Results

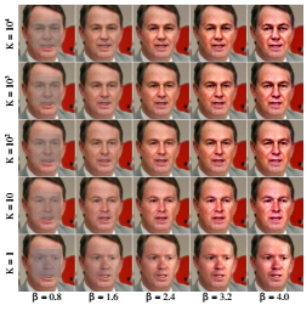
In case of face attribute changes the authors compared their results with AEGAN output (fig. 2). As it can be seen from the image, DFI produces believable and, at some degree, natural results. Also, compared to AEGAN, DFI's output are superior in quality in case of glasses change and preserving the race at some level. It is expalainable due to the special consideration of the target and source sets (sets are restricted to the K nearest neighbors).

Figure 2: Experimental results of DFI versus AEGAN [12].

As a result of filling missing regions, DFI manages to output visually reliable results. In case of shoes inpainting, it is clearly indicated (fig. 3) that DFI is able to follow not only the color of the shoe, but also its fashion style.

Figure 3: Inpainting on shoes [12].

Fig.4 demonstrates the change of the parameter β (strength of transformation) and K (size of source/target sets). Low β produces ghosting, and too large one – unnatural outputs; K controls variety, too much variety – unnatural, lack of variety creates artifacts [12].



(a) An inpainting case.

(b) A semantic transformation case.

Figure 4: Varying the free parameters [12].

A really difficult task for DFI method is to edit the image when almost the half of the region is missed (fig. 5). The output of the method obtained by the authors are implausable and absolutely unnatural.

Figure 5: An example of DFI output failing to fulfill inpainting [12].

5 Summary and conclusion

To sum up the authors implemented the DFI method for semantic transformations and inpainting purposes.

On one hand, surely, the DFI algorithm is able to provide competitive results, either it is related with face attribute changes or with image inpainting. The algortihm works relatively fast, compared to heavy neural network architectures, which require a lot of time for training. In addition, the simplicity of the DFI method based on linear interpolation in the deep feature space makes it plausable in usage, compared to GAN methods' structure. Also, this approach is versatile in different attribute changes.

On the other hand, despite that the method provides results of a high perceptual quality, it has several limitations. Firstly, the most important requirement for performing DFI is image alignment. All used images have to be aligned and resized. Secondly, it is arduous for DFI to achieve proper outcome when the challenging task involves inpainting half of the region of the images. Moreover, this approach is costly, and reliance on a

pre-trained model makes it unanticipated and unforeseen to other attributes or factors, which are not considered during the pre-training [7].

One of the potential further development for this method might be incorporation of real-time transfer techniques [12] to overcome unforseen factors due to pre-trained models and also to increase the time performance.

However, despite all the constraints, the method proposed by Upchurch et al. might be used as a baseline approach for high-resolution image transformation, especially if all the recommendations for future work are fulfilled.

References

1. A. Brock, T. Lim, J. Ritchie, and N. Weston. “Neural Photo Editing with Introspective Adversarial Networks”. *arXiv:1609.07093v3 [cs.LG]*, 2017.
2. X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. “InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets”. *arXiv:1606.03657v1 [cs.LG]*, 2016.
3. J. R. Gardner, P. Upchurch, M. J. Kusner, Y. Li, K. Q. Weinberger, K. Bala, and J. E. Hopcroft. “Deep Manifold Traversal: Changing Labels with Convolutional Features”. *arXiv:1511.06421 [cs.LG]*, 2016.
4. P. Garrido, L.Valgaerts, O.Rehmsen, and T.Thormahlen. “Automatic Face Reenactment”. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4217–4224.
5. P. Isola, J. yan Zhu, T. Zhou, and A. A. Efros. “Image-toImage Translation with Conditional Adversarial Networks”. *arXiv:1611.07004*, 2016.
6. I. Kemelmacher-Shlizerman, E. Shechtman, R. Garg, and S. M. Seitz. “Exploring Photobios”. *ACM Transactions on Graphics (TOG)* 30, 2011.
7. G. Lample, N. Zeghidour, N. Usunier, A. Bordes, L. Denoyer, and M. Ranzato. “Fader Networks: Manipulating Images by Sliding Attributes”. *arXiv:1706.00409v2 [cs.CV]*, 2018.
8. A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther. “Autoencoding beyond pixels using a learned similarity metric”. *arXiv:1512.09300v2 [cs.LG]*, 2016.
9. C. Li and M. Wand. “Precomputed Real-Time Texture Synthesis with Markovian Generative Adversarial Networks”. *arXiv:1604.04382v1 [cs.CV]*, 2016.
10. N. van Noord and E. Postma. “Light-weight pixel context encoders for image inpainting”. *arXiv:1801.05585v1 [cs.CV]*, 2018.
11. A. Radford, L. Metz, and S. Chintala. “Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks”. *arXiv:1511.06434 [cs.LG]*, 2016.
12. P. Upchurch, J. Gardner, G. Pleiss, R. Pleiss, N. Snavely, K. Bala, and K. Weinberger. *Deep Feature Interpolation for Image Content Changes*. 2017.
13. T. Weise, H. Li, L. V. Gool, and M. Pauly. “Face/Off: Live Facial Puppetry”. *Eurographics/ ACM SIGGRAPH Symposium on Computer Animation*, 2009.