COST EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

SHORT TERM SCIENTIFIC MISSION (STSM) – SCIENTIFIC REPORT

The STSM applicant submits this report for approval to the STSM coordinator

Action number: IC1307
STSM title: Producing facial images from text descriptions
STSM start and end date: 05/02/2018 to 16/02/2018
Grantee name: Blaz Meden

PURPOSE OF THE STSM/

(max.500 words)

The purpose of the STSM was to bridge the gap between natural language processing and computer vision technologies. We decided to explore the possibilities of producing photo-realistic facial images using recently introduced deep generative neural networks by using natural language descriptions as an input.

Generative neural networks and generative adversarial networks are promising methods for new content generation and learning the essential latent statistics about large image-based datasets. To establish the connection with the vision and language we intended to research if such generative models can gain the capability of generating new content based on natural language inputs. To accomplish this task we combined technologies from natural language processing field (NLP) and deep convolutional neural networks (CNNs), especially recent generative neural networks (GNNs).

To demonstrate the merits of our approach, we applied our developed models to the modality of faces. The concrete example here is the problem of how to estimate one's looks from the attributes. This could be useful for determining suspect identities from verbal descriptions in various areas of criminal investigations or forensics for example. Various attributes or metadata could be used to describe visual characteristics of a human face as accurately as possible. These include the facial shape (geometry), displayed facial expression or expression action units (if present), hair style and color, presence or absence of accessories of any kind (glasses, rings, moustache, beard, etc.), gender, age (in a form of estimated age or simply classification to major age groups), ethnicity, femininity or masculinity factors or even the parameter of attractiveness. More advanced options would be to describe parts of a face separately. These detailed descriptions could include shape information for each face part. Additional way of adding more information to the description is to include the pose parameters.

The input to the entire developed system is a natural language description of the attributes that should be visually recognized at a produced facial image that would be generated as an output.

Funded by the Horizon 2020 Framework Programme
of the European Union

DESCRIPTION OF WORK CARRIED OUT DURING THE STSMS

(max.500 words)

Our work consisted of several steps, which are described below.

STEP 1. DATASET ANNOTATION

We updated existing Radbound Faces Dataset [1] with new labels from our chosen subset of visual attributes. The dataset consists of 57 adult subjects, each of them depicting 8 facial expressions, totalling at 456 frontal face images available to be used in training. Full dataset also consists of non-frontal images and kid subjects, but we limited ourselves on frontalized subset of adult images, to be able to use facial hair attribute during augmentation.

Each annotation is represented as bag of words. We provided such descriptive annotation for all 57 adult subjects. We limited ourselves to small subset of describable attributes that our system would support initially. These include shape of the head (as 5 categories: square, round, oval and oblong), hair color (as 3 categories: bright, brown, black), presence of wearing glasses (present or not), presence of facial hair (applied only to male subjects, present or not) and gender.

STEP 2. NETWORK TRAINING AND DATA AUGMENTATION

We prepared implementation for our generative part of the system, including structured inputs, dense layers and deconvolutional layers. The implementation is based on the previous work [2] of Michael D. Flynn (https://github.com/zo7/deconvfaces). We modified the architecture to suit our objective by defining additional inputs, based on our attribute selection.

We performed live image augmentation during training process (by detecting key facial landmarks using the approach from [3] and then adding accessories with probabilistic threshold landmark positions. Randomly selected glasses were added with the probability of 0.5. Facial hair was added only if the subject in the input batch was male (with the probability of 0.5), otherwise not. Example of detected landmarks and augmentation procedure is is illustrated in Fig. 1. A subset of final augmented facial images is displayed in Fig. 2.

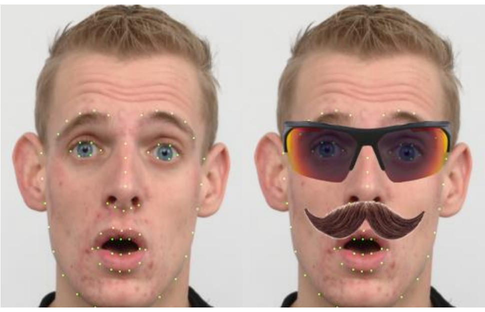

Fig. 1. Detected landmarks (left) and blended accessories after the detection (right).

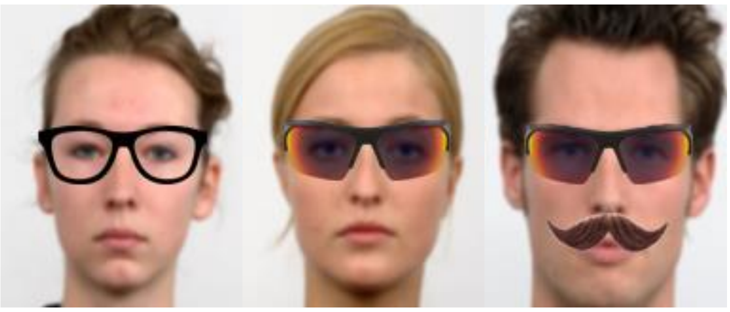

Fig. 2. Examples of image augmentation: added reading glasses (left), with sunglasses covering eye region (middle), with sunglasses and facial hair (right).

We obtained best visual results with generative model trained with 5 deconvolutional layers over 1000 epochs with Adam optimizer (using default parameters). Batch size was set to 32, weights were initialized from random uniform distribution using default parameters. Loss function was defined as Mean Square Logarithmic Error (MSLE). We used Peak Signal-to-Noise Ratio (PSNR) as an accuracy metric to monitor image quality during training.

STEP 3. PLATFORM ASSEMBLY

We assembled module for rapid keyword extraction, based on Rapid Keyword Extraction (RAKE) algorithm [4], which formed a basis for our NLP processing module. The input to the module are natural language sequences, which contain keywords about subject's appearance. The output is represented as extracted keywords. These are mapped to the inputs of the trained GNN as visual attributes, matching the attribute format used during GNN training.

STEP 4. QUALITATIVE ANALYSIS

We performed qualitative visual analysis of the obtained results. The discussion is available at the section with main obtained results.

DESCRIPTION OF THE MAIN RESULTS OBTAINED

(max. 500 words)

We obtained quantitative results and analyzed them visually to estimate the visual quality of generated examples.

Fig. 3. displays some randomly generated examples (produced by our model), which represent visually adequate results (generated faces are complete, without visible artefacts, and mostly match the visual attributes, that were passed into the system). In the first row of Fig 3. (left) we display generated image with enabled following attributes: male, black glasses, black eyes, round face. Next image in the first row of Fig 3 used similar attributes, but without present glasses and with face shape changed to rectangular category. Second row of Fig. 3. displays two more visual examples, where glasses were properly placed over the face, which was also visually good looking.

Fig. 4. displays some of the randomly generated images that resulted in a poor quality reconstruction of the input description (here most of the problems are caused due to the fact, that some of the combinations were not present in the training data, i.e. missing regions or imperfect facial features). We suspect that we could remedy these failure cases by using larger dataset and provide wider arrangement of visual descriptors into the training set. For some combinations (for example, using sunglasses as the input attribute) we can see, that the network did not converge, and therefore the visual results are only partially developed (Fig. 4.).

The poor quality examples are opening new points for further research and searching for improvements of the training process and network model architecture. We think that these preliminary results are promising, since they were obtained in such short period and they could be furtherly improved with additional research on the topic of generative networks.

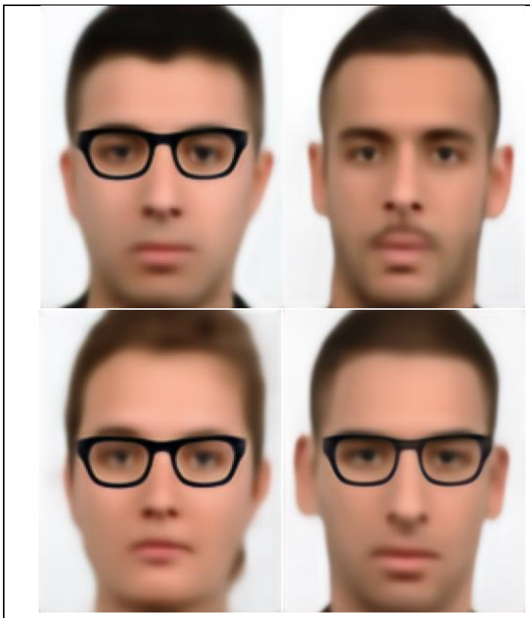

Fig. 3. Visually good looking generated examples where the presence or absence of the accessories is well defined, as well as mostly completed facial region with no visible major artefacts.

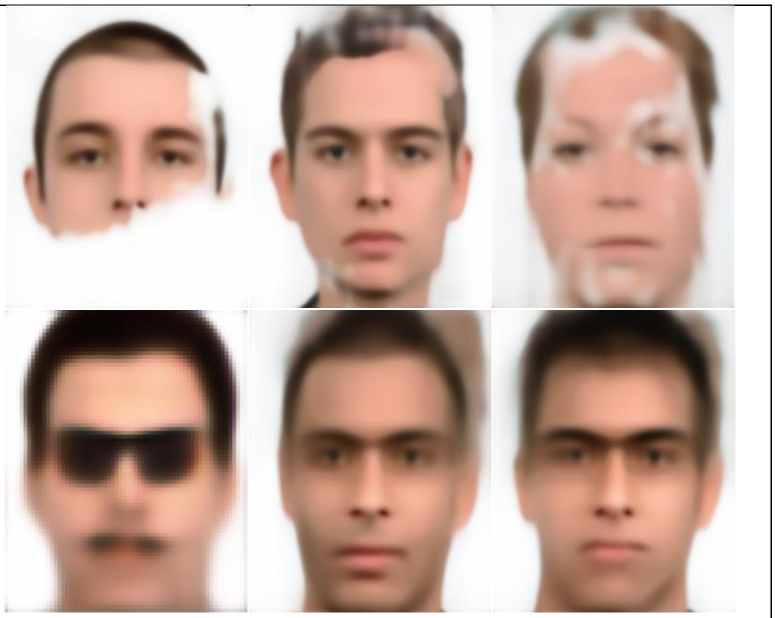

Fig. 4. Examples of poorly generated examples. First row -- visible missing regions (missing entire part of the face, missing hairline, poor placing of skin and hair). Second row – showing regions with blurry transitions (especially poorly defined hairline in the middle and right). We can also observe the inability of the network to converge in some cases (left).

FUTURE COLLABORATIONS (if applicable)

(max.500 words)

During our visit and work with prof. Carlos M. Travieso-González, we discussed and agreed on, that future collaboration is an open and acknowledged option. Below we describe two areas that are in our common research interest.

First option is to continue working on selected topic from this STSM. This would address our proposed improvements from this report, including improvement of the generator part in sense that it will be able to output more clear looking facial images with broader support for visual attributes. New architectures and deep learning paradigms could be used to upgrade existing generative neural network (such as adversarial training paradigm, attention mechanisms and semantic awareness) to effectively achieve these proposed improvements. Additional data has to be gathered and annotated with proper labels in order to supply newly developed models with large enough dataset for proper training. Another key remark that we want to expose as possible improvement is to perform end-to-end training of our system. In order to do that, it is required to replace current NLP processing module with RNN-or-LSTM-like reccurent neural architecture [5] and attach it to the generative model to be able to perform training of the entire pipeline together. This means, however, that previously defined visual attributes (that we established by hand) won't be applicable anymore as all information about the encoding of the description words and their visual semantic attributes would be hidden in a latent space of features inside the joined neural model (consisting of NLP processing network joined with generative network for outputing images). This is an appealing option of using two neural architectures together (to bridge the gap between natural language and image data in an elegant way), but on the other hand this means that, we cannot really understand, what the network is actually learning within its latent feature space. This is why we initially developed two separate modules, to have more control over the training process and hence, obtain better understanding of training a generative model to produce facial images from meaningful visual attributes, established by hand. Another problem of end-to-end training is the problem of how to properly define the criteria of optimization (i.e. loss function) that will be sufficient for achieving stable convergence of the model during training. More research is needed in this part, where it‘s not clear, whether conditioning with additional labels or other prior knowledge is required within the process of model training.

Besides the ideas performed during our STSM, we are sharing other common grounds that we could address by working together. Particularly, we are interested in developing new models, based on deep learning, to solve problems from different areas (one concrete example is to build a deep classifier that would recognize the species of plants from bark image dataset, which was gathered during the research work on local flora diversity). Here, it is important that we develop the models and then evaluate them on an existing evaluation protocol, which will compare our findings with the recognition performances of the existing descriptor-like approaches.

References:

[1] Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D.H.J., Hawk, S.T., & van Knippenberg, A. (2010). Presentation and validation of the Radboud Faces Database. Cognition & Emotion, 24(8), 1377—1388. DOI: 10.1080/02699930903485076

[2] M. Flynn. (2017) Generating faces with deconvolution networks. [Online]. Available: https://zo7.github.io/blog/2016/09/25/generating-faces.html

[3] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, 2014, pp. 1867-1874., DOI: 10.1109/CVPR.2014.241

[4] S. Rose, D. Engel, N. Cramer, and W. Cowley, Automatic Keyword Extraction from Individual Documents. John Wiley & Sons, Ltd, 2010, pp. 1–20. [Online]. Available: http://dx.doi.org/10.1002/9780470689646.ch1

[5] K. Hwang, M. Lee, and W. Sung, “Online keyword spotting with a character-level recurrent neural network,” CoRR, vol. abs/1512.08903, 2015. [Online]. Available: http://arxiv.org/abs/1512.08903