Proceedings of the workshop on People in Vision, Language, and the Mind @LREC2022 , pages 26–40
Marseille, 20 June 2022

Examining the Effects of Language-and-Vision Data Augmentation for Generation of Descriptions of Human Faces

Nikolai Ilinykh*, Rafal Černiavski†, Eva Elžbieta Sventickaitė†, Viktorija Buzaitė†, Simon Dobnik*
*Centre for Linguistic Theory and Studies in Probability (CLASP),
Department of Philosophy, Linguistics and Theory of Science (FLoV),
University of Gothenburg, Sweden
nikolai.ilinykh, simon.dobnik@gu.se
†Faculty of Languages, Department of Linguistics and Philology, Uppsala University, Sweden
rafal.cerniavski.2286, evaelzbieta.sventickaite.9060, viktorija.buzaite.1828@student.uu.se

Abstract

We investigate how different augmentation techniques on both textual and visual representations affect the performance of the face description generation model. Specifically, we provide the model with either original images, sketches of faces, facial composites or distorted images. In addition, on the language side, we experiment with different methods to augment the original dataset with paraphrased captions, which are semantically equivalent to the original ones, but differ in terms of their form. We also examine if augmenting the dataset with descriptions from a different domain (e.g., image captions of real-world images) has an effect on the performance of the models. We train models on different combinations of visual and linguistic features and perform both (i) automatic evaluation of generated captions and (ii) examination of how useful different visual features are for the task of facial feature classification. Our results show that although original images encode the best possible representation for the task, the model trained on sketches can still perform relatively well. We also observe that augmenting the dataset with descriptions from a different domain can boost performance of the model. We conclude that face description generation systems are more susceptible to language rather than vision data augmentation. Overall, we demonstrate that face caption generation models display a strong imbalance in the utilisation of language and vision modalities, indicating a lack of proper information fusion. We also describe ethical implications of our study and argue that future work on human face description generation should create better, more representative datasets.

Keywords: face description generation, data augmentation, feature manipulation

1. Introduction

Humans generally excel at recognising and defining everyday objects as well as human faces. However, *human face recognition* is a daily challenge to some. More than 2% of the population worldwide are affected by prosopagnosia (Corrow et al., 2016), the inability to distinguish individuals based on their facial features. As Lopatina et al. (2018) argue, the lack of the ability to recognise and describe a human face has underlying social importance, as impaired facial perception is a common indication of brain conditions, such as autism spectrum disorder. Therefore building automatic systems that can recognise human faces is essential to assist people with neurological conditions.

Although the task of face recognition has largely been solved, as state-of-the-art facial recognition models reach an accuracy of over 99% (Yan et al., 2019), the problem of **generating facial descriptions** has not received much attention. Prior research on facial cognition has shown that an attention-based model can generate captions for faces with a particular focus on emotions (Nezami et al., 2020). Similarly impressive is the performance of modern text-to-face models, which aim to generate realistic faces from short texts. Models such as the ones proposed by Nasir et al. (2019) or Sun et al. (2021) leverage powerful Generative Adversarial Networks (GANs) to produce pictures of faces that can be highly similar to natural images of faces. It has been also argued that generating facial descriptions with extra focus on words depicting emotions and sentiment is important to understand how different facial expressions can influence decision-making and inter-personal relations (Mathews et al., 2016). Nevertheless, despite these major achievements, grounding of facial features in language or generating captions of human faces remains arguably an open task, since the quality of generated descriptions remains questionable. One plausible reason is the lack of sufficient and representative data. Furthermore, features of the human face are relatively ambiguous; for instance, there is no conventional and objective measure to differentiate a small nose from a big one. Therefore we argue that it is necessary to examine both *models* and *different feature representations* for the task of automatic facial description generation, because the quality of generated texts directly affects not only the correctness of mentions of factual facial features (e.g., oval face, blond hair), but also how humans *socially* perceive and construct opinions about others in different situations and contexts (e.g., interpreting sentiment based on various facial clues).

In this paper we focus on the task of **facial description generation**. Specifically, we examine how data augmentation of either visual or linguistic representa-

tions affects performance of the face description generation model[1]. **First**, we investigate whether face captioning models demonstrate better performance when trained on various abstractions of original face images (sketches, composites, distortions). Individual features are much less pronounced in such abstractions since representations become more abstract and less specific. **Second**, we also investigate to what extent the facial captioning model can utilise visual representations and if an utterly grotesque abstraction (distorted images) affects the quality of generated captions. **Third**, we enrich the training set of facial descriptions with their counterparts, which are semantically equivalent, but differ in terms of the words and form. For example, for the description "this human has blond hair" we create the following paraphrase: "this human does not have brown hair". **Lastly**, we also evaluate the performance of statistical multi-label feature classification models trained on different visual features. With the latter, we study differences between visual abstractions and original images outside of the generation task. We conclude with a general discussion of the results and possible ethical implications of the study. We emphasise the importance of creating datasets of images of faces that would represent a more significant number of human groups and communities, while keeping in mind the right to the privacy of information.

2. Related Work

Visual Data Augmentation Data augmentation is the process of altering the dataset so as to increase the amount of data available for training. In terms of images, data augmentation usually involves rotating, flipping, resizing, and changing the colours of the images. Such a seemingly simple method often leads to considerable and consistent improvements in performance across a variety of models (Lim et al., 2019; Wang et al., 2019; Xie et al., 2020). More advanced methods of utilising data to improve the performance of models include noise reduction and image deformation. Noise reduction generally refers to the process of filtering out elements that appear to obstruct the view. For example, noise reduction is often used to eliminate Gaussian noise which can sometimes corrupt images that are being transformed (Mafi et al., 2019). Image deformation, on the other hand, is mostly used in sketch recognition and involves creating slightly changed versions of images. As formulated by Zheng et al. (2021), this method relies on learning temporal patterns in drawing a sketch and using them to deform the sketch. Having more sketches created through augmentations boosts performance sketch recognition models. Deformation can also be applied to images by performing domain adaptation (Wang et al., 2020). If the target domain involves abstraction, this method can be thought of as incorporating both noise reduction and image deformation, since the output of such a model is, for instance, a sketch which discards any non-essential information. It should be noted that, unlike pictures or images, sketches are often limited to just a few lines or strokes on white background. Thus, models are required to perform recognition from fewer features.

Language Data Augmentation Different methods are typically used to caption an image (Bernardi et al., 2017): from templates (Fang et al., 2015) to end-to-end systems (Kiros et al., 2014) with attention (Wang et al., 2016; Xu et al., 2015; Lu et al., 2017). More recently, a transformer architecture has been adopted for many multi-modal tasks including image captioning[2].
Augmenting datasets with additional captions incorporating certain linguistic variation has been shown to improve performance of captioning models. Zhang et al. (2015) replace words with synonyms based on the thesaurus from WordNet (Fellbaum, 2005), whereas Fadaee et al. (2017) propose an augmentation method for a machine translation model which targets rare words. Kobayashi (2018) implements contextual augmentation for convolutional and recurrent neural networks. In general, researchers use deletion, insertion, replacement or swap techniques at either character or word level to augment captions (Zhang et al., 2015).

3. Augmenting the Task Dataset

Motivation It is not immediately clear how to augment data for models that operate with multiple modalities. The key challenge is to change representations for both modalities in such a way that these changes are relatively comparable and have similar conceptual motivation behind. In general, data augmentation either adds or removes specific features. Such strategies allow for better understanding of how and what models learn. In terms of *visual augmentation*, we constructed different abstract representations (sketches) of images of faces. When generating sketches, we simultaneously *reduce* individual visual features (e.g., abstract sketches look much more similar to each other versus images of faces, which are more varied in terms of individual features) and bring abstract representations to the fore, *introducing* input representations to the captioning model which are more distilled (e.g., general facial features on sketches are more pronounced). In terms of *textual augmentation*, we *add* features by generating alternative descriptions of images, which introduce new vocabulary items to learn for the model. By generating such alternative texts, we also *exclude* direct correspondence of descriptions into images of faces and make grounding task for the model much harder, because generated descriptions of faces do not use the exact same words as the ground truth description. Overall, we believe that our augmentation methods introduce comparable conditions for both language

[1]Our work is an examination of whether vision-and-language model relies on biases in feature representations or learns spurious correlations (Agarwal et al., 2020).

[2]For an overview of many different architectures, we refer the reader to Bugliarello et al. (2021).

and vision, in which different types of information are either removed or added.

Face Description Dataset We use *CelebA-HQ* dataset (Karras et al., 2017) as our task dataset for training and testing facial description generation models. The dataset contains $30,000$ high-resolution images of human faces of celebrities with 10 natural language descriptions per each image. On average, each description is 15.53 tokens long, e.g., Figure 3. The dataset also provides binary annotations of 40 facial features. The size of the training set in all experiments is set to the first E entries from the dataset ($E = 10,000$).

Augmenting Vision Zhang et al. (2011) have shown that the human recognition rate of facial sketches is largely affected by the sketch quality and the level of detail. This finding indicates that image manipulation should be conducted very carefully: the face has to be still recognisable while its representation can become highly abstract, e.g. containing contours of some parts of faces. Therefore, we control the level of abstraction by generating *three* different sketches per image.

First, we run a simple auto-encoder architecture (Rumelhart et al., 1986) to transform images into sketches.[3] This is an unsupervised neural network which consists of an encoder that compresses data into vectors and passes them through multiple convolutional layers (Cun et al., 1990). Next, a decoder learns to reconstruct the original data as closely as possible from these vectors. Backpropagation is used to minimise the reconstruction loss. We train the model for 100 epochs with the Adam optimiser (Kingma and Ba, 2014). We further refer to this type of sketches as **Face-2-Sketch**.

Second, we follow Zhu et al. (2017) and implement a generative adversarial network for image-to-image translation task. This model is a combination of two networks, a generator and a discriminator, which use two unaligned sets of images, A and B, to identify their similarities and transform images from the first set of images to images of the second set. The model is using the cycle consistency loss expressed as follows:

$$\begin{aligned} \boldsymbol{L}(\mathbf{G},\mathbf{F},\mathbf{D}_A,\mathbf{D}_B) = \quad & \boldsymbol{L}_{GAN}(\mathbf{G},\mathbf{D}_A,\mathbf{A},\mathbf{B}) + \\ & \boldsymbol{L}_{GAN}(\mathbf{F},\mathbf{D}_B,\mathbf{B},\mathbf{A}) + , \\ & \lambda \boldsymbol{L}_{CYC}(\mathbf{G},\mathbf{F}) \end{aligned} \quad (1)$$

where $\mathbf{G}$ and $\mathbf{F}$ are mappings from image set $\mathbf{A}$ to $\mathbf{B}$ and vice versa, $\mathbf{D}_A$ and $\mathbf{D}_B$ are discriminators that are trained to differentiate between real and predicted images, and λ parameter controls the contribution of each loss for the final loss score. We achieve the best performance loss-wise with the GAN model after 5 epochs of training. We also train the model for 33 epochs in total to see the extent to which the model can over-fit and generate distorted, grotesque sketches. The resulting sketches might be highly dissimilar to the original images and we use them to investigate whether our facial description generator could still learn from highly unrecognizeable images. We thus use both models which we refer to as **GAN:Composite** and **GAN:Distorted** respectively. We set $\lambda = 10$, the learning rate $l = 0.0002$, batch size $b = 1$ and a weight decay $wd = 0.00001$ after each epoch.

Both **Face-2-Sketch** and **GAN** models were trained on the combination of three datasets: CUHK dataset (Wang and Tang, 2009a) consisting of 188 face-sketch pairs, AR dataset (Martinez and Benavente, 1998) with 123 photo-sketch pairs, and CUHK Face Sketch FERET Database (CUFSF) (Wang and Tang, 2009b; Zhang et al., 2011) of 1,194 sketches, for which we additionally obtained the FERET (Phillips et al., 1998) database with pictures of 1,194 people. We resize the pictures and sketches to 200×250. In addition, since FERET dataset contains pictures from various angles, we manually cleaned the dataset, leaving only one profile picture per person. Examples of the images received with different visual augmentation methods are shown in Figure 1, check Figure 4 for more examples.

Augmenting Language Kafle et al. (2017) use two methods for data augmentation for Visual Question Answering on real-world images: (i) template-based generation of texts based on rich object annotations of images and (ii) LSTMs (Hochreiter and Schmidhuber, 1997) to generate texts that resemble structure of the original texts. While images in captioning datasets such as MSCOCO (Lin et al., 2014) include a large variety of objects, images of faces are much more rigid in terms of observable parts: nose, mouth, etc. Parts of the face can differ on the level of attributes (shape, flatness, openness, for example) and a simple method to augment our dataset with more descriptions of each face is to *generate new sentences by changing verbs, adjectives and adverbs* which typically depict attributes.

In our search for the most suitable method for language augmentation we decided to examine an existing tool, the `nlpaug`[4] library. This library allows us to try a variety of existing language models and use different word embedding representations extracted by feeding captions to such models as word2vec(Mikolov et al., 2013), GloVe(Pennington et al., 2014), FastText(Mikolov et al., 2018), BERT(Devlin et al., 2019), DistilBERT(Sanh et al., 2019), and RoBERTA(Zhuang et al., 2021). Based on the similarity of extracted embeddings, `nlpaug` either (i) substitutes words in captions with synonyms or (ii) inserts additional words inside captions. In addition, this library allows us to use WordNet hierarchies from the nltk library (Bird et al., 2009) in order to manipulate with the original descriptions, replacing words with either synonyms or antonyms. Examples of the captions obtained with dif-

[3] We adapt the code from `https://www.kaggle.com/theblackmamba31/photo-to-sketch-using-autoencoder/notebook`.

[4] `https://github.com/makcedward/nlpaug`

Figure 1: Example of the image from the task dataset. We show original, composite, sketch-based and distorted images in order from the most left one to the right.
Ground truth description: This person is attractive, and young and has bags under eyes, wavy hair, arched eyebrows, and mouth slightly open.
Feature annotations: Arched_Eyebrows, Attractive, Bags_Under_Eyes, Blond_Hair, Heavy_Makeup, High_Cheekbones, Mouth_Slightly_Open, No_Beard, Smiling, Wavy_Hair, Wearing_Earrings, Wearing_Lipstick, Wearing_Necklace, Young.
Augmented description: This person is not unattractive, and not old and doesn't have flat under eyes, straight hair, straight eyebrows, and mouth completely closed.

ferent `nlpaug` methods and the ground truth facial description are shown in Table 2. During manual examination of resulting descriptions, we noticed that the augmented captions were in most cases incorrect: they neither followed the proper English grammar nor they referred to the features present or absent in the picture.[5] To make sure that augmented captions do not contradict with images, we developed *a rule-based algorithm* that replaces all verbs, adjectives, and adverbs (a set of 28 word types in total) with antonyms. The list of words to be replaced was designed manually by two authors of the paper. Different replacements were agreed through discussion; the whole list is shown in Table 3. First, we carefully selected antonyms from thesauri and dictionaries so as to ensure that the antonyms refer to what is considered *the opposite of facial features*, e.g. round face $\rightarrow$ square face; blond hair $\rightarrow$ black hair. Next, we negated each antonym, e.g. square face $\rightarrow$ not square face; black hair $\rightarrow$ not black hair. Example of the resulting description and ground truth text are shown in Table 2, in which we can see that both augmented caption and ground truth description correspond to each other due to the mention of the same feature but in a different way, e.g. "is attractive" and "is not unattractive", "young" and "not old". Note that the experiments that we report in this paper were conducted *only* with captions generated with our rule-based algorithm and list of the words to be replaced; we did not use any `nlpgaug`-based methods for our experiments due to bad quality of augmented captions.
We note that replacing verbs, adjectives and adverbs with their antonyms or negated counterparts *guaranteed* that the negative captions were still semantically correct, since the generated captions addressed the features that the faces lacked rather than possessed, e.g. the person has wavy hair $\rightarrow$ the person does not have straight hair. We believe it is important to see whether the models will be able to pick up an important linguistic cue - negation - and tailor its output accordingly (Niu and Bansal, 2018). Our method of data augmentation also enforces the model to learn to reason with language (e.g., wavy hair is not straight hair), which could potentially improve the quality of feature grounding between language and vision. At the same time, the vocabulary of the model is increased because of the introduction of antonyms which make language modality more prominent as a feature. Note that the combination of antonyms and negated relations ("with" $\rightarrow$ "without") creates ambiguity and therefore such descriptions are harder to learn: "hat" can be identified by visual features but it is unclear what visual features "without a hat" can be identified with. Overall, due to language augmentation the task becomes much harder and we expect that this will be reflected in the performance.

Manual Evaluation of Augmentation Examples from each of the four sets of images can be seen in Figure 4. The quality of the sketches differs greatly across all models: sketches of white women, who constituted nearly half of the dataset, were most accurate, whereas sketches of other people were more distorted overall. We believe that to the naked eye, **GAN:Composite** were the most successful in terms of condensing the facial features. When it comes to the **Face2Sketch** subset, the quality was considerably poorer. Nevertheless, some of the features are still visible. Images in the **GAN:Distorted** subset appeared to have additional noise that partially masks some of the facial features. Examples of augmented captions are shown in Table 2 describing features of the original image in Figure 1.

[5]Interestingly, our manual examination also indirectly evaluated augmentation methods introduced in `nlpaug`, showing that these methods have many flaws.

4. Facial Description Generation

Model Our face description generator is a simple CNN-LSTM encoder-decoder network with attention (Xu et al., 2015)[6]. The model is trained with cross-entropy loss as well as doubly stochastic regularisation. We pick the best checkpoint based on the BLEU score (Papineni et al., 2002) on the validation set and do early stopping. We train the model for 20 epochs and set the batch size to $b = 5$, learning rate to $lr = 1e - 4$ for the encoder and $lr = 4e - 4$ for the decoder, dropout to $d = 0.5$ and gradient clipping $gc = 5$.

Training and Evaluation The description generation model is trained on either original images (**Baseline**) or on one of the three types of visual manipulations (**GAN:Composite**, **GAN:Distorted**, **Face-2-Sketch**). We add `start` and `end` to the captions and pad shorter descriptions. We use 5 captions per image for training. As the vocabulary of the descriptions is rather limited, we manipulate training data by augmenting the captions with a mix of original and generated descriptions with a ratio of 3:2 (3 original and 2 augmented captions) and name this condition **Aug-Anton 3:2**. We also replace all five descriptions per image with augmented ones for **Aug-Anton 5**. In addition, we augment training data by injecting the model with a small portion of the caption from *Flickr8k* dataset (Hodosh et al., 2013): we add 12.5% in training and validation sets which corresponds to 1000 and 125 of image-caption pairs respectively (both images and captions were added to the our task dataset). The latter model is referred to as **Aug-Caption**. The vocabulary expanded to 100 when data with antonyms was produced and to 470 when a variant with image captions is used. We evaluate the models on three types of data: (i) the original images, (ii) the composites produced by **GAN:Composite**, and (iii) distorted images from **GAN:Distorted**. By running our models on different evaluation sets we aim to measure whether the distillation of features has a desirable effect on captions.

Results We report BLEU-1 (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005) and ROUGE (Lin, 2004) scores for generated captions. Table 1 shows the results of automatic evaluation of generated face descriptions and Figure 2 shows examples of descriptions generated with different vision and language augmentation methods. Red-coloured values indicate best models among those, which were trained with original, composites, sketches or distorted images, e.g. visual augmentation. Blue-coloured values depict best models among those which were trained on a dataset in which we augmented only the textual side (original images were used for training). We note that in our evaluation of linguistically augmented models we compared generated texts against their non-augmented counterparts. For example, while the **Aug-Anton 3:2** model has been trained on both "tall" and "not short" in respective captions, it has been evaluated only against the one that has "tall" in it. With this harsh evaluation we aimed to see whether models learn more distinct representations for target words ("tall") when trying to contrast them with their negated antonyms.

[6]We use the code from `https://github.com/sgrvinod/a-PyTorch-Tutorial-to-Image-Captioning`.

METEOR	**1. img**	**2. cmp**	**3. dst**
A. Baseline	**72.87**	60.27	60.35
B. GAN:Composite	59.47	**72.76**	66.95
C. Face-2-Sketch	**72.87**	61.86	61.36
D. GAN:Distorted	57.17	64.22	**70.93**
E. Aug-Caption	69.98	39.06	46.03
F. Aug-Anton 3:2	**72.51**	**62.34**	**61.29**
G. Aug-Anton 5	41.02	32.71	33.35
BLEU-1	**1. img**	**2. cmp**	**3. dst**
A. Baseline	**48.12**	30.41	29.18
B. GAN:Composite	26.84	**43.76**	33.71
C. Face-2-Sketch	39.91	24.22	25.39
D. GAN:Distorted	27.75	36.29	**43.69**
E. Aug-Caption	**49.71**	12.94	17.79
F. Aug-Anton 3:2	39.09	**30.65**	**32.41**
G. Aug-Anton 5	13.84	7.10	8.71
ROUGE	**1. img**	**2. cmp**	**3. dst**
A. Baseline	**64.36**	53.13	54.41
B. GAN:Composite	54.36	**62.07**	57.67
C. Face-2-Sketch	59.58	50.11	51.19
D. GAN:Distorted	53.27	62.07	**62.65**
E. Aug-Caption	**65.81**	44.41	48.03
F. Aug-Anton 3:2	59.46	**54.31**	**54.08**
G. Aug-Anton 5	42.33	35.52	35.76

Table 1: Automatic evaluation of generated facial descriptions. We report results for three NLG metrics: METEOR, BLEU-1 and ROUGE split into three tables. In each table each row depicts a type of the (non-)augmented data that the model has been trained on. The first set of models include those which were trained on either original dataset (*Baseline*) or visual augmentations (captions were kept untouched). The second set of models below a dashed line shows models trained with different language augmentations but with original images. The columns show the type of data each model has been evaluated on: *img* stands for original images, *cmp* and *dst* are for composites and distorted images respectively.

We first analyse the performance of the models which were trained with *visual* augmentations (B - D). The **Baseline** model, which is trained on original images, performs best when tested on original images, which is expected. Notably, **Face-2-Sketch** that is trained on facial sketches achieves the same METEOR score and is also the second best in terms of BLEU and ROUGE scores when tested on original images. This indicates that our model either (i) cannot fully use original visual representations and this is why its performance is close to the model trained on sketches or (ii) the model is actually able to sufficiently learn from sketches of faces.

When tested on composite and distorted images, the best models are the ones that were trained on the corresponding visual augmentations. As expected the **Baseline** model suffers the most when tested on non-original images. Interestingly, the **Face-2-Sketch** model shows one of the worst performances when tested on composites and distorted images, while it is on par with the baseline when tested on original images. The result implies that only a particular level of abstraction of faces is exploited by the model to generate better descriptions: a simple auto-encoder, although producing very abstract representations, outperforms the generative adversarial network which likely generates sketches with high contrast, high level of details and high distortions as shown by the examples in Figure 4. We conclude that it is important to consider the network type and abstractness of its output when performing visual augmentation of multi-modal datasets.

The bottom parts of the table below the dashed lines show performance of the models augmented with different *linguistic* representations (E - G). For two out of three metrics, the model that has been jointly trained on both facial descriptions and image captions (**Aug-Caption**) performs best when tested on original images. Partial augmentation with descriptions with the same meaning but different form (**Aug-Anton 3:2**) leads to the second-best performance with the exception of the METEOR metric where this model performs best. This can be attributed to the fact that METEOR is designed specifically for better synonym matching and linking of paraphrased sentences and therefore its high score indirectly reflects that our method of mixing original descriptions with paraphrased descriptions (training for **Aug-Anton 3:2**) is helpful for the model. In contrast, using only augmented descriptions results in a drop in performance, possibly because the model is not able to learn grounding of descriptions in visual features. The model is required to perform extra reasoning to ground augmented descriptions since they correspond to a variety of visual features. It has been shown that METEOR generally correlates better with human judgements unlike BLEU or ROUGE (Elliott and Keller, 2014) which means augmenting facial descriptions with our simple method can generate more human-like descriptions. When tested on *composite* and *distorted* images, **Aug-Anton 3:2** performs best across all metrics. Interestingly, in terms of BLEU, **Aug-Caption** and **Aug-Anton 5** show a much lower performance than **Aug-Anton 3:2** when tested on both composite and distorted images. It is possible that when visual features are very different from what the model has been trained on (trained on original images, but tested on composites and distorted), the model starts relying on fine-grained differences in linguistic augmented descriptions which also introduce contrast in form but not in meaning. At the same time, training the model on augmented descriptions only (**Aug-Anton 5**) results in a very low performance in terms of BLEU (7.10 and 8.71 for composites and distorted respectively), because the model does not have access to a suitable representation in either of the modalities. Also, the fact that models F and G were evaluated against untouched captions might lead to generally lower metric results compared to model E.

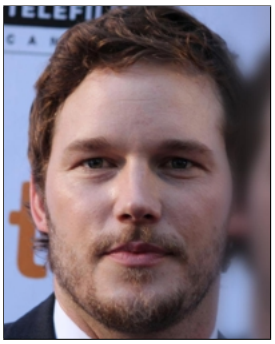
Overall, note that **Aug-Caption** has shown a relatively good performance in terms of all testing conditions for METEOR and ROUGE. When we test this model (model E in Table 1) on original images, straightforward replacement of words (models F and G) does not bring better learning, but using captions from a different domain does. This is because captions from a different domain introduce a larger variety of syntactic structures and semantic relations between words in text. In comparison, our manual linguistic augmentation does not change either syntax or semantics of descriptions - it simply introduces new words into the vocabulary. At the same time, the model which learns to discriminate between descriptions which are identical in terms of their meaning but different in terms of their form (model F) achieves higher scores across multiple conditions and metrics. Therefore we conclude that augmenting language has a positive effect on the model's performance when (i) there is a strong form-based contrasting signal from descriptions like in **Aug-Anton 3:2** model, and (ii) the data is infused with descriptions from a similar multi-modal domain (**Aug-Caption**), e.g. image captioning. We also believe that future work should examine the extent of how much does the face description generation model benefit from being trained on captions from different multi-modal domains and tasks.

5. Multi-Label Feature Classification

In addition to caption generation we also evaluate the augmented visual datasets on another task, facial feature classification.

Model We train two statistical classifiers: Random Forest and k-Nearest Neighbours. We use the annotations of K features for every image provided by the authors of the dataset. Each classifier takes a feature vector of the image as its input $v_n \in \mathbf{R}^{1\times D}$, where $D = 2048$, and learns to predict one of the K feature annotations, $K = 40$. Examples of the feature annotation are shown in Figure 4. Note that most of these features could overlap with the vocabulary of the image captioning model, but some of them are also more abstract, e.g. *5_o_Clock_Shadow*. *Blurry*. We aim to examine the effect of different visual representations on the performance of the classification model.

Training and evaluation We use a randomly selected sample of 9,000 images as a training set and another 1,000 as the test set to train and evaluate all models on the *CelebA-HQ* dataset. We use loss as the objective function and other standard parameters with the scikit-learn API (Pedregosa et al., 2011). The performance of the multi-label linear classification mod-



Original description:
The person has big lips, sideburns, goatee, mustache, and brown hair. He is wearing necktie.

Evaluated on original images:
Baseline: the man has sideburns and wears necktie
GAN:Composite: this man has big lips and black hair and is wearing hat
GAN:Distorted: this person has bags under eyes and is wearing lipstick
Face-2-Sketch: the man has bags under eyes and big nose
Aug-Caption: the person is young and has big nose and bags under eyes
Aug-Anton 3:2: this person has bags under eyes and big nose
Aug-Anton 5: this man differ old and refuse bags under eyes and little nose

Evaluated on composites:
Baseline: this man has big nose and big lips
GAN:Composite: this person has bags under eyes and big nose and is wearing necktie
GAN:Distorted: this woman has big nose and is wearing lipstick and hat
Face-2-Sketch: the man has big nose and bags under eyes
Aug-Caption: the person is chubby and has goatee and big nose
Aug-Anton 3:2: the person has bags under eyes and big lips
Aug-Anton 5: the person differ old and refuse pale skin and white hair

Evaluated on distorted images:
Baseline: the woman has big lips and wears lipstick and earrings
GAN:Composite: this person has big lips and is wearing hat
GAN:Composite: this person has bags under eyes big nose and sideburns
Face-2-Sketch: the person has big lips and wears lipstick
Aug-Caption: the person has gray hair and big nose and is wearing necklace
Aug-Anton 3:2: the person has mouth slightly open and big lips *Aug-Anton 5:* the person differ smiling and refuse mouth slightly closed bags under eyes and low cheekbones

Figure 2: Example of an image with the original description and texts generated by our models described in Table 1.

els was evaluated with reference to both the micro and macro averages of precision, recall, and F-score. We gave equal weight to precision and recall in calculating the F-score.

Results The results are shown in Figure 5. In terms of the F1-score, we do not observe any noticeable differences between performances of different features across both micro- and macro-averaged results. The same holds for the results on recall metric. Most notably, both k-NN and Random Forest model have the highest macro-average precision and recall on **Face-2-Sketch** features, which, we argue, is the least informative of the facial features. As can be seen from the graphs, macro-averaging is generally in a lower range than micro-averaging, demonstrating that model's performance on the non-majority classes is worse than on the majority classes. This result reflects that the model can mostly predict some of the most frequent facial features, which are often represented in the dataset (such as *female* and *attractive*), yet fail to predict rare features, such as *goatee* and *receding_hairline*. We leave a deeper investigation of the effect that the dataset imbalance has on the performance of the model on the feature classification task for future work.

Overall, visual features seem to be very similar with each other since using them interchangeably with each other does not affect the results on the feature classification task. High similarity of different visual features can also be one of the reasons why different models for visual augmentations (A-D in Table 1) do not differ so much from each other in terms of different evaluation metrics. In comparison, language augmentation methods (E-G) can affect performance of the model to a larger extent, e.g. **Aug-Anton 5** decreasing the overall performance to BLEU of 13.84 on the original images. Therefore we argue that the model is much more sensitive to language augmentation possibly because visual representations are very similar to each other and are not distinctive enough as the results on feature classification task demonstrate. This indirectly supports the idea that multi-modal architectures strongly learn to benefit from the language modality instead of visual representations whatsoever because language is more informative and mostly harder to predict (Frank et al., 2021).

6. General Discussion

Automatic evaluation of facial descriptions has revealed that, first and foremost, the original photos are generally better suited for the task. This may be due to the fact that both **Baseline** and **Aug-Caption** receive fully-coloured images as input, whereas other conditions such as **GAN:Composite**, **GAN:Distorted** and **Face-2-Sketch** are trained on gray-scale generated images distributed over 3 colour channels. This indicates that using sketches or other abstract representations of faces does not necessarily improve the quality of generated descriptions. In addition, as can be seen from Figure4, the images obtained with **Face-2-Sketch** are the most abstract ones. Nonetheless, automatic evaluation metrics for models trained on this data are higher compared to **GAN:Composite** and **GAN:Distorted**. As such, the reason for these differences could be that during the pooling process the features are meshed in such a manner that the high-quality of images may not

be necessary for rather adequate performance. Finally, introducing only antonyms without mixing them with original descriptions results in incorrect or impossible grounding of visual features with descriptions which are not encoded by those visual features. It is important to introduce both correct descriptions and their augmented versions so that the model learns from both texts, which are semantically equivalent but differ in terms of their form.

6.1. Ethical Implications

The current study touches upon ethical implications of representation ability in data used for computer vision and natural language processing tasks. We note that our task dataset, *CelebA-HQ*, is over-represented with high-quality images of humans of specific race, gender and ethnicity. This potentially leads to considerable bias in models, since the models are predominantly exposed to a very limited groups since most datasets include Caucasian and Asian people. Ensuring that a bigger number of groups are represented in the dataset is costly and difficult. Alternative ways of debiasing and exposing models to more diverse set of images of faces are highly needed. At the same time, what matters is the correctness and fairness of face descriptions: they should depict only concrete face features without any subjective, sensitive or offensive descriptions.

The language augmentation approach proposed in this paper is an attempt at exposing the models to features that are not present in the dataset and thus compensate for the lack of representation of images through linguistic knowledge. For instance, it is challenging or even counterproductive to generate synthetic faces with various features that are not represented in the dataset. On the other hand, generating augmented descriptions with semantically similar words is a relatively simple yet effective way towards exposing the model to features that are not present in the data, yet possible. We acknowledge that our simple approach is without a doubt insufficient for ensuring a better coverage of different groups of people, as human features, unlike synonymity-antonymity, are non-binary: the colour of hair can be blond, black, brunette, whichever other colour, or there could be no hair at all. Nevertheless, we believe that future work should focus on the language augmentation method of face description datasets with the emphasis on creating semantically correct, but also diverse descriptions.

7. Conclusion and Future Work

In this project, we aimed to investigate the effects of visual and linguistic augmentation as means of improving automatic generation of facial descriptions. In particular, we operated with different levels of visual abstractions and paraphrases of descriptions and tracked how these modifications alter the generated texts. We also investigated how different visual representations affected the feature classification with linear models.

Our results show that original images are generally more useful for the facial description generation task. However, different feature manipulation have a different effect on the resulting texts: augmenting linguistic representations in a contrasting way (keeping original descriptions and adding artificially created ones) has a larger effect on model's learning ability unlike augmenting data from the vision side. For the latter, using auto-encoded sketch-like features of faces is generally more preferable rather than using facial composites, possibly due to the level of abstractness of sketches. Also, we have shown that linguistic augmentation of the dataset with captions from a different domain could lead to better face descriptions.

In terms of future work, we suggest the following experiments: in terms of visual augmentation, first, manipulate the model in such a manner that it could accommodate training on different types of visual data in parallel. One approach may be to experiment with different combinations of sets of images, composites, and distorted pictures through dense layers and examine how it would affect the captions. Furthermore, the images could be manipulated to limit one or more colour channels at a time, thus, more information could be extracted on how the colouring of the images affects the training and, in turn, the attention and quality of the generated captions. In terms of language augmentation, we propose to run the experiments in parallel with data in multiple languages to assess whether features that are mapped to certain tokens in different languages are the same, e.g. feature grounding task.

8. Acknowledgements

NI and SD were supported by a grant from the Swedish Research Council (VR project 2014-39) for the establishment of the Centre for Linguistic Theory and Studies in Probability (CLASP) at the University of Gothenburg.

9. Bibliographical References

Agarwal, V., Shetty, R., and Fritz, M. (2020). Towards causal vqa: Revealing and reducing spurious correlations by invariant and covariant semantic editing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9690–9698, June.

Atliha, V. and Šešok, D. (2020). Text Augmentation Using BERT for Image Captioning. *Applied Sciences*, 10:5978.

Banerjee, S. and Lavie, A. (2005). Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Bernardi, R., Cakici, R., Elliott, D., Erdem, A., Erdem, E., Ikizler-Cinbis, N., Keller, F., Muscat, A., and Plank, B. (2017). Automatic description generation

from images: A survey of models, datasets, and evaluation measures (extended abstract). In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 4970–4974.

Bird, S., Klein, E., and Loper, E. (2009). *Natural language processing with Python*. O'Reilly.

Bugliarello, E., Cotterell, R., Okazaki, N., and Elliott, D. (2021). Multimodal pretraining unmasked: A meta-analysis and a unified framework of vision-and-language BERTs. *Transactions of the Association for Computational Linguistics*, 9:978–994.

Cornia, M., Stefanini, M., Baraldi, L., and Cucchiara, R. (2020). Meshed-memory transformer for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10578–10587.

Corrow, S. L., Dalrymple, K. A., and Barton, J. J. (2016). Prosopagnosia: current perspectives. *Eye and brain*, 8:165.

Cun, Y. L., Boser, B., Denker, J. S., Howard, R. E., Habbard, W., Jackel, L. D., and Henderson, D., (1990). *Handwritten Digit Recognition with a Back-Propagation Network*, page 396–404. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Duchaine, B. (2011). Developmental prosopagnosia: Cognitive, neural, and developmental investigations. In Andy Calder, et al., editors, *Oxford Handbook of Face Perception*, pages 821–838. Oxford University Press.

Elliott, D. and Keller, F. (2014). Comparing automatic evaluation measures for image description. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 452–457, Baltimore, Maryland, June. Association for Computational Linguistics.

Fadaee, M., Bisazza, A., and Monz, C. (2017). Data Augmentation for Low-Resource Neural Machine Translation. *ArXiv*, abs/1705.00440.

Fang, H., Gupta, S., Iandola, F., Srivastava, R. K., Deng, L., Dollár, P., Gao, J., He, X., Mitchell, M., Platt, J. C., et al. (2015). From captions to visual concepts and back. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1473–1482.

Fellbaum, C. (2005). Wordnet and wordnets. In Keith Brown, editor, *Encyclopedia of Language and Linguistics*, pages 665–670, Oxford. Elsevier.

Frank, S., Bugliarello, E., and Elliott, D. (2021). Vision-and-language or vision-for-language? on cross-modal influence in multimodal transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9847–9857, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Kafle, K., Yousefhussien, M., and Kanan, C. (2017). Data augmentation for visual question answering. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 198–202, Santiago de Compostela, Spain, September. Association for Computational Linguistics.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015.

Kiros, R., Salakhutdinov, R., and Zemel, R. (2014). Multimodal neural language models. In *International conference on machine learning*, pages 595–603. PMLR.

Kobayashi, S. (2018). Contextual Augmentation: Data Augmentation by Words with Paradigmatic Relations. *ArXiv*, abs/1805.06201.

Li, J., Yu, X., Peng, C., and Wang, N. (2017). Adaptive representation-based face sketch-photo synthesis. *Neurocomputing*, 269:152–159.

Li, X., Yin, X., Li, C., Hu, X., Zhang, P., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., Choi, Y., and Gao, J. (2020). Oscar: Object-semantics aligned pre-training for vision-language tasks. *ECCV 2020*.

Lim, S., Kim, I., Kim, T., Kim, C., and Kim, S. (2019). Fast autoaugment. In *NeurIPS*.

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., and Dollár, P. (2014). Microsoft coco: Common objects in context. cite arxiv:1405.0312Comment: 1) updated annotation pipeline description and figures; 2) added new section describing datasets splits; 3) updated author list.

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Lopatina, O. L., Komleva, Y. K., Gorina, Y. V., Higashida, H., and Salmina, A. B. (2018). Neurobiological aspects of face recognition: The role of oxytocin. *Frontiers in Behavioral Neuroscience*, 12.

Lu, J., Xiong, C., Parikh, D., and Socher, R. (2017).

Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 375–383.

Ma, E. (2019). NLP Augmentation. https://github.com/makcedward/nlpaug.

Mafi, M., Martin, H., Cabrerizo, M., Andrian, J., Barreto, A., and Adjouadi, M. (2019). A comprehensive survey on impulse and gaussian denoising filters for digital images. *Signal Processing*, 157:236–260.

Mathews, A., Xie, L., and He, X. (2016). Senticap: Generating image descriptions with sentiments. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, page 3574–3580. AAAI Press.

Mikolov, T., Chen, K., Corrado, G. S., and Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. In *ICLR*.

Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., and Joulin, A. (2018). Advances in Pre-Training Distributed Word Representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Nasir, O. R., Jha, S. K., Grover, M. S., Yu, Y., Kumar, A., and Shah, R. R. (2019). Text2facegan: Face generation from fine grained textual descriptions. In *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)*, pages 58–67. IEEE.

Nezami, O. M., Dras, M., Wan, S., and Paris, C. (2020). Image captioning using facial expression and attention. *Journal of Artificial Intelligence Research*, 68:661–689.

Nie, W., Narodytska, N., and Patel, A. (2019). RelGAN: Relational generative adversarial networks for text generation. In *International Conference on Learning Representations*.

Niu, T. and Bansal, M. (2018). Adversarial oversensitivity and over-stability strategies for dialogue models. In *CoNLL*.

Panetta, K., Samani, A., Yuan, X., Wan, Q., Agaian, S. S., Rajeev, S., Kamath, S., Rajendran, R., Rao, S. P., Kaszowska, A., and Taylor, H. A. (2020). A Comprehensive Database for Benchmarking Imaging Systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42:509–520.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Peng, C., Gao, X., Wang, N., Tao, D., Li, X., and Li, J. (2016). Multiple Representations-Based Face Sketch–Photo Synthesis. *IEEE Transactions on Neural Networks and Learning Systems*, 27:2201–2215.

Pennington, J., Socher, R., and Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. In *EMNLP*.

Perov, I., Gao, D., Chervoniy, N., Liu, K., Marangonda, S., Umé, C., Dpfks, M., Facenheim, C. S., RP, L., Jiang, J., et al. (2020). Deepfacelab: A simple, flexible and extensible face swapping framework. *arXiv preprint arXiv: 2005.05535*.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J., (1986). *Learning Internal Representations by Error Propagation*, page 318–362. MIT Press, Cambridge, MA, USA.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Sun, J., Li, Q., Wang, W., Zhao, J., and Sun, Z. (2021). Multi-caption text-to-face synthesis: Dataset and algorithm. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2290–2298, New York, NY, USA. Association for Computing Machinery.

Wang, C., Yang, H., Bartz, C., and Meinel, C. (2016). Image captioning with deep bidirectional lstms. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 988–997.

Wang, N., Gao, X., and Li, J. (2018). Random sampling for fast face sketch synthesis. *Pattern Recognit.*, 76:215–227.

Wang, Y., Pan, X., Song, S., Zhang, H., Huang, G., and Wu, C. (2019). Implicit semantic data augmentation for deep networks. *Advances in Neural Information Processing Systems*, 32.

Wang, Q., Meng, F., and Breckon, T. P. (2020). Data augmentation with norm-vae for unsupervised domain adaptation. *arXiv preprint arXiv:2012.00848*.

Xia, W., Yang, Y., Xue, J., and Wu, B. (2021). Tedigan: Text-Guided Diverse Face Image Generation and Manipulation. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2256–2265.

Xie, Q., Dai, Z., Hovy, E., Luong, T., and Le, Q. (2020). Unsupervised data augmentation for consistency training. *Advances in Neural Information Processing Systems*, 33:6256–6268.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., and Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR.

Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., and He, X. (2018). Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1316–1324.

Yan, M., Zhao, M., Xu, Z., Zhang, Q., Wang, G., and Su, Z. (2019). Vargfacenet: An efficient variable group convolutional neural network for lightweight face recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J. G., Salakhutdinov, R., and Le, Q. V. (2019). XLNet: Generalized Autoregressive Pretraining for Language Understanding. In *NeurIPS*.

Yu, J., Xu, X., Gao, F., Shi, S., Wang, M., Tao, D., and Huang, Q. (2020). Toward Realistic Face Photo-Sketch Synthesis via Composition-Aided GANs. *IEEE Transactions on Cybernetics*, PP:1–13, 03.

Zhang, Y., Ellyson, S., Zone, A., Gangam, P., Sullins, J., McCullough, C., Canavan, S., and Yin, L. (2011). Recognizing face sketches by a large number of human subjects: A perception-based study for facial distinctiveness. In *2011 IEEE International Conference on Automatic Face and Gesture Recognition and Workshops, FG 2011*, pages 707–712, 04.

Zhang, X., Zhao, J. J., and LeCun, Y. (2015). Character-level Convolutional Networks for Text Classification. *ArXiv*, abs/1509.01626.

Zhang, P., Li, X., Hu, X., Yang, J., Zhang, L., Wang, L., Choi, Y., and Gao, J. (2021). Vinvl: Making visual representations matter in vision-language models. *CVPR 2021*.

Zheng, Y., Yao, H., Sun, X., Zhang, S., Zhao, S., and Porikli, F. (2021). Sketch-specific data augmentation for freehand sketch recognition. *Neurocomputing*, 456:528–539.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*.

Zhuang, L., Wayne, L., Ya, S., and Jun, Z. (2021). A Robustly Optimized BERT Pre-training Approach with Post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China, August. Chinese Information Processing Society of China.

10. Language Resource References

Hodosh, M., Young, P., and Hockenmaier, J. (2013). Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics (Extended Abstract). *J. Artif. Intell. Res.*, 47:853–899.

Tero Karras and Timo Aila and Samuli Laine and Jaakko Lehtinen. (2017). *Progressive Growing of GANs for Improved Quality, Stability, and Variation*.

Martinez, A. and Benavente, R. (1998). *The AR Face Database: CVC Technical Report, 24*. Department of Computer Science, Universitat Autònoma de Barcelona, January.

Phillips, P. Jonathon and Wechsler, Harry and Huang, Jeffrey and Rauss, Patrick J. (1998). *The FERET database and evaluation procedure for face-recognition algorithms.*

Wang, Xiaogang and Tang, Xiaoou. (2009a). *Face Photo-Sketch Synthesis and Recognition*.

Xiaogang Wang and Xiaoou Tang. (2009b). *Face Photo-Sketch Synthesis and Recognition*.

Zhang, W., Wang, X., and Tang, X. (2011). Coupled information-theoretic encoding for face photo-sketch recognition. In *CVPR 2011*, pages 513–520.

Appendix

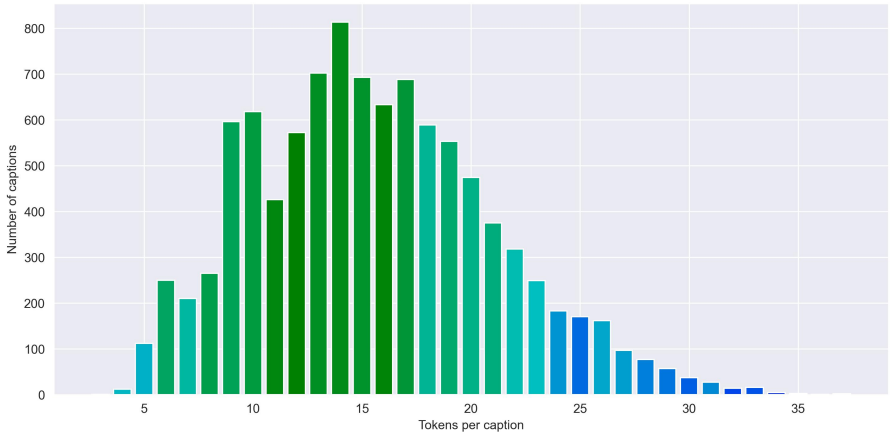


Figure 3: Distribution of captions based on their length in the CelebA-HQ dataset. The horizontal axis depicts token count per caption, the vertical axis represents caption count.

(a) Original caption: This man has double chin, bags under eyes, high cheekbones, mustache, big nose, goatee, and eyeglasses and wears hat. He is chubby. Feature annotations: Bags_Under_Eyes, Big_Lips, Big_Nose, Chubby, Double_Chin, Eyeglasses, Goatee, High_Cheekbones, Male, Mouth_Slightly_Open, Mustache, Smiling, Wearing_Hat.

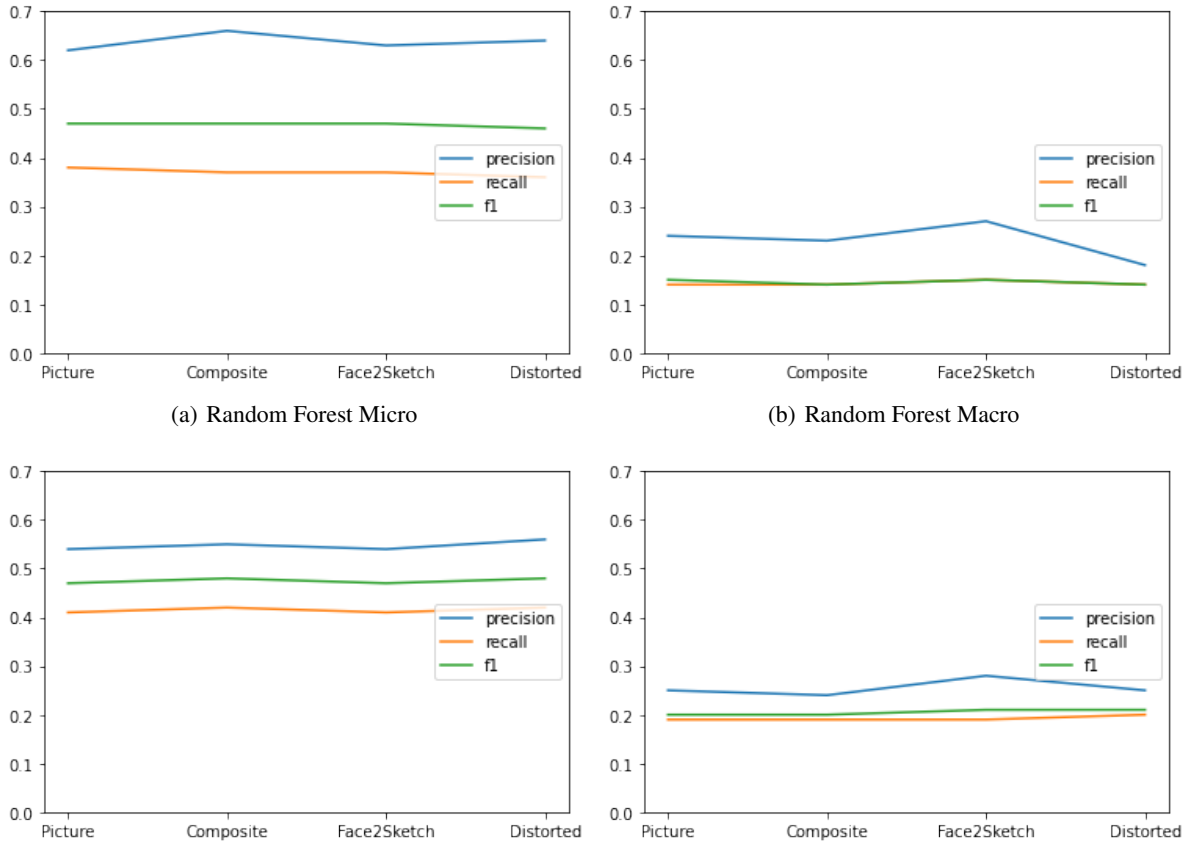
(b) Original caption: She is young and has mouth slightly open. Feature annotations: Mouth_Slightly_Open, No_Beard, Wearing_Necklace, Young

(c) Original caption: This person has mustache, big nose, and receding hairline. He is bald and wears necktie. He has beard.Feature annotations: Bags_Under_Eyes, Bald, Big_Nose, Chubby, Double_Chin, Male, Mouth_Slightly_Open, Mustache, Receding_Hairline, Smiling, Wearing_Necktie

Figure 4: Original picture in grey-scale versus the generated sketches. The images och each person display (from left to right): Original photo, Composite, Face2Sketch and Distorted models.

Original sentence: This person is attractive, and young and has bags under eyes, wavy hair, arched eyebrows, and mouth slightly open.
word2vec: *substitution* This person is desirable, and spurn_materialism and has poly_bags ##under before_igniting_gunpowder, corkscrew_curls hair, arched eyebrows, and mouth minimally Pat_Barberot_Orchestra. *insertion* Massachusetts This person Mauer is UNH.N attractive, and young Indrajit and has Arun bags under eyes, Jack wavy Assa hair, arched eyebrows, and JUSTIN mouth slightly open.
GloVe: *substitution* This person is attractive, and both years has bags even eyes, wavy glasses, symmetrical eyebrows, taken tributary changed open. *insertion* This pask person aparece is attractive, and cnni young and has heberle bags under eyes, handson wavy friele hair, arched eyebrows, and mouth 102,500 slightly open.
fasttext: *substitution* Moreover person is attractive, and young and has bags beside eyes, wavy strawberry-blonde, bow-shaped question, thereafter mouth slowly locked. *insertion* Trinitresque This person LLU is attractive, and –Boston young and Finesilver has RoW bags under eyes, wavy Daksha hair, Jakar arched eyebrows, and Masturbator mouth slightly open.
BERT: *substitution* the man is attractive, and young and dark amber under eyes, wavy hair, arched eyebrows, but face tinted pink. *insertion* sometimes this person is attractive, short and so young and also has bags hiding under eyes, wavy silver hair, highly arched eyebrows, throat and mouth slightly open.
DistilBERT: *substitution* prehistoric lizard appeared attractive, appears young and has orange under thighs, red hair, arched ears, and mouth slightly open. *insertion* but this female person is attractive, and young and young has bags under blue eyes, wavy auburn hair, extremely arched eyebrows, and whose mouth slightly exposed open.
RoBERTA: *substitution* This female is attractive, and young and has bags under eyes, wavy hair, arched eyebrows, y mouth slightly open. *insertion* This person is attractive, fresh and also young and has bags under eyes, wavy hair, arched eyebrows, and mouth is slightly open.
WordNet (synonyms): *substitution* This person comprise attractive, and young and has bags under eyes, wavy hair, arched eyebrows, and mouth slightly open.
WordNet (antonyms): *substitution* This person differ repulsive, and old and lack bags under eyes, wavy hair, arched eyebrows, and mouth slightly unreceptive.
Manual (antonyms): This person is not unattractive, and not old and doesn't have flat under eyes, straight hair, straight eyebrows, and mouth completely closed.

Table 2: Examples of caption augmentation with different methods available in the `nlpaug` tool. The image that these descriptions were produced for is the first image from Figure 4. For each model we show the results of both word-level substitution and insertion. Our best method is based on manual replacement of antonyms and is shown last.



(a) Random Forest Micro

(b) Random Forest Macro

(c) k-NN Micro

(d) k-NN Macro

Figure 5: Results of multi-label feature classification with different visual representations across multiple types of classifiers. The legend in each graph shows our evaluation metrics: precision, recall, F1-score. X axis corresponds to four different vision augmentation conditions, where **Picture** stands for original images. We set the maximum values on the y axis to 0.7 for visualisation purposes. We also report results for both Macro- and Micro-Averaged results per evaluation metric.

arched	straight
attractive	unattractive
bald	hairy
big	small
black	white
blond	dark
bushy	thin
chubby	skinny
double	single
grey	colourful
has	doesn't have
heavy	light
high	low
is	isn't
narrow	wide
no	any
open	closed
oval	square
pale	glowing
pointy	blunt
receding	widow's peak
rosy	pale
slightly	completely
smiling	frowning
straight	wavy
wavy	straight
wears	doesn't wear
young	old

Table 3: Each of the verbs, adjectives and adverbs found in original facial descriptions (left column) has been replaced with an antonym (right column) in our caption augmentation experiment. Note that some antonyms ("widow's peak") are much harder to ground into visual features.