There is Noise on Your Face, but there is no Nose: An Exploration of Data Augmentation Methods for Captioning Facial Features

Rafal Černiavski
Uppsala University
{rafal.cerniavski.2286@student.uu.se}

Eva Elžbieta Sventickaitė
Uppsala University
{e.e.sventickaite@gmail.com}

Viktorija Buzaitė
Uppsala University
{viktorija.buzaite.1828@student.uu.se}

Abstract

In this research, we investigated the effects of data distortion and augmentation on the quality of the descriptions of facial features. To do so, we used *CelebA-HQ* and *Flickr8k* datasets. We trained models for three distinct tasks: CycleGAN as well as autoencoder model for composite generation and distortion, WordNet-based antonym substitution model for caption augmentation, and a CNN-RNN encoder-decoder with attention for caption generation. The generated captions were evaluated based on feature extraction as well as linguistic accuracy. The produced composites were also used to train linear classifiers to see whether abstraction could aid linear models in accurate feature recognition. The results show that by generating composites not only the noise is reduced, but also essential information for learning is lost. As such, the best performance was achieved with the introduction of noisy data. It was the only approach that improved face captioning models, as image augmentation, as well as caption augmentation, led to results comparable to the baseline in terms of linguistic accuracy.

1 Introduction

Image captioning is the task of generating a logically and grammatically correct description of an image in natural language. Due to the cross-disciplinary nature of the task, it has drawn attention from both Computer Vision (CV) and Natural Language Processing (NLP) communities. With combined efforts from the two communities, the results by image captioning models on various benchmarks have arguably achieved human-like quality (Li et al., 2020; Zhang et al., 2021). Yet it could be argued these models achieve such accuracy due to the generalized nature of the training and assessment data. The images used for training these models contain numerous distinct categories of objects, which could be seamlessly defined based on their most prominent and recurring features.

Generally, humans are very adaptive at recognizing and defining distinct everyday objects as well as faces, however, many have to tackle face recognition as an everyday challenge. More than 2% of the population worldwide are affected by prosopagnosia (Corrow et al., 2016), inability to distinguish individuals based on their facial features, which suggests that differentiating and defining faces is a task demanding more complex and fine-grained processing as compared to general image processing.

In the hopes of finding more efficient ways for facial description generation, we have decided to investigate the effects of data augmentation and distortion methods drawn both from CV and NLP fields. In this project, the following experiments were implemented:

1. We generated facial composites from celebrity images and evaluated the caption generation model's ability to describe the face of a person in an image as compared to the generated sketch.

2. Augmented the facial composites and evaluated the effects of the distortion on the caption generation by comparing the results with the previous experiment.

3. Assessed whether image augmentation has a more distinct effect as compared to caption augmentation or noise introduction on the precision of the caption generation both in regards to feature extraction and linguistic accuracy.

4. Evaluated the effects of noise reduction and image distortion on linear models for facial feature classification.

We hypothesized that by generating facial composites the most prominent features from the image are condensed. Subsequently, any further processing or modeling using ML techniques would be streamlined as the noise in the picture is limited. In other words, we believed that we could achieve better performance of captioning and improve linear feature recognition models by training them on facial composites, for they emphasize facial features. Lastly, we hoped that we could train the model to produce more fine-grained captions by feeding it with augmented captions.

This work first introduces some background knowledge, then data and tools are presented. They are followed by descriptions of the methods used in the experiments. Finally, results are presented and discussed.

2 Background

The experiments in this project combine knowledge amassed from both Computer Vision and Language Technology fields. This section provides essential information on image and caption generation as well as augmentation.

2.1 Image Augmentation, Noise Reduction, and Distortion

Data augmentation is the process of altering the dataset at hand so as to increase the amount of data available for training. In the case of images, the process usually involves rotating, flipping, resizing, and changing the colors of images. Such a seemingly simple method often leads to considerable and consistent improvements on a variety of models (Lim et al., 2019; Wang et al., 2019; Xie et al., 2020).

More advanced methodologies of utilizing data to improve the performance of a model include noise reduction or 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 an image that is being transferred (Mafi et al., 2019). Image deformation, on the other hand, is mostly used in sketch recognition and involves creating a slightly changed version of an image. As formulated by Zheng et al. (2021), the method relies on learning the temporal patterns of drawing a sketch and using them to deform the sketch. As a result, having more sketches to learn from, the augmentation method is effective in boosting the performance of models on sketch recognition. Deformation can also be performed on images and pictures by performing domain adaptation (Wang et al., 2020). If the target domain involves abstraction, the method can be said to also incorporate both noise reduction and image deformation, since the output of such 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, the models are required to perform recognition from considerably fewer data.

2.2 Caption Augmentation

Big language models require a lot of varied data to produce satisfactory results. Although the datasets used in this study have a relatively huge amount of labels, the descriptions are rather uniform with a limited vocabulary and expressions. To tackle this problem, the image caption augmentation method is used to produce more encompassing descriptions as a means to better describe a picture (Atliha and Šešok, 2020).

Caption augmentation has produced positive results in previous research. Zhang et al. (2015) replaced words with their synonyms based on thesaurus from WordNet (Fellbaum, 2005), whereas Fadaee et al. (2017) proposed an augmentation method for a machine translation model which targeted rare words. In addition, Kobayashi (2018) implemented contextual augmentation for convolutional and recurrent neural networks. Usually, for word augmentation task, the deletion, insertion, replacement, or swap techniques are employed on either character or word level (Zhang et al., 2015).

More recently, BERT (Devlin et al., 2019) has been proven to be a state-of-the-art model in many NLP tasks. Its success lies behind the transformer architecture implemented with attention mechanism, with the help of encoder and decoder. Since the encoder mechanism is able to process the whole sequence at once and not sequentially like older models, BERT is able to learn and generalize based on all captions at once, which renders it a powerful tool for image captioning and text augmentation.

2.3 Image Captioning

Image captioning can be done by applying one of the three methods: a classical approach of using templates (Fang et al., 2015), a time-proven implementation of encoder-decoder architecture (Kiros et al., 2014) or its improved versions in regards to attention (Wang et al., 2016; Xu et al., 2015; Lu et al., 2017), as well as the newest approach, transformers or its variants (Dosovitskiy et al., 2020; Cornia et al., 2020).

The encoder-decoder model could be considered an old image captioning approach, as it was adapted from neural machine translation for the task in 2014. Nevertheless, its capabilities are near state-of-the-art when considering caption generation for static images. With such wide-ranging success, applying enhanced encoder-decoder architectures for image captioning is still prevalent today. Subsequently, this work was inspired by Xu et al. (2015), where the authors introduced the encoder-decoder model with attention. The model is composed of a Convolutional Neural Network (CNN) as the encoder and a Recurrent Neural Network (RNN) with attention as the decoder. As the images have background and foreground with different levels of details, the attention mechanism not only allowed the model to better process the image and link features to specific semantic space, but also enabled researchers to visualize the process of model learning by creating attention maps.

2.4 Grounding of Facial Features

The ability to recognize and describe a human face is a complex task that involves multiple brain regions. As argued by Lopatina et al. (2018), it has underlying social importance, as impaired facial perception is a common indication of brain conditions, such as autism spectrum disorder. Grounding of the facial features is therefore an important task in pursuit of a human-like cognition for artificial intelligence.

Prior research in the field of facial cognition has achieved impressive results. Nezami et al. (2020) trained an Attention-based model to produce captions for images with a special focus on the emotions expressed by the human faces in the dataset. In addition to scoring close to state of the art, their model was able to generate exceptionally fine-grained captions that correctly identified the emotion. As impressive is the performance of the modern Text-to-Face models the goal of which is to generate a realistic face from a brief description. Models such as the ones proposed by Nasir et al. (2019) and Sun et al. (2021) leverage powerful Generative Adversarial Networks (GANs) to produce pictures of faces that can be hardly distinguished from real pictures. Nevertheless, despite the major achievements in the field, grounding of facial features or captioning of the human face arguably remains an open task.

3 Experimental Setup

In this research project, we explore the effects of image and caption augmentation strategies for grounding of facial features. More specifically, we sought to find out whether reducing or increasing the amount of noise in the pictures and their sketches could lead to improvements in the models' ability to ground the features. We examine the groundedness of facial features by the models' ability to produce accurate and informative captions as well as linear classification of features. The following sections describe the data and the models we used to augment the data and subsequently improve the quality of generated captions.

3.1 Data

We used the *CelebA-HQ* dataset (Karras et al., 2017) in all of our experiments. The dataset contains 30,000 high-resolution images of human faces (celebrities) with ten respective descriptions, uniquely created for each image. The dataset additionally contains binary annotations of 40 facial features. We decided to use the first 10,000 entries from the dataset in all of our experiments in order to minimize the resources needed to train the models.

We additionally used a small portion of *Flickr8k* [1] data as well as three supplementary datasets of generated face-composites described in section 4.1.

3.2 Models

The models used in our research can be classified into the following categories:

(i) **Image Augmentation and Composite Generation**

We used two divergent models to generate the facial composites. Firstly, we implemented

[1] https://www.kaggle.com/adityajn105/flickr8k

a simple autoencoder-based architecture to transform pictures to sketches[2]. We trained the model for 100 epochs with the Adam optimizer. We further refer to the model as *Face2Sketch*. As for the second model, we implemented a CycleGAN-based Zhu et al. (2017) domain adaptation models[3]. We attempted training the model for 33 epochs in total; however, we observed that the best and most sketch-like results were produced with just five epochs of training. With 33 epochs of training, the model appeared to drastically overfit and therefore producing distorted, grotesque sketches. We nevertheless were interested to learn whether the models could still learn facial features from distorted composites. We thus used the weights of both 5 and 33 epochs to produce composites of celebrities. We further refer to the sketches generated by the model trained for 5 epochs as *Composites*, as the sketches predicted by the model trained for 33 epochs as *Distorted*.

(ii) **Caption Generation**
For caption generation a CNN-RNN encoder-decoder with attention was applied (Xu et al., 2015)[4]. The original model was trained on 120 epochs, yet in this experiment, we trained all the models on 20 epochs, which has proven to be efficient as the models mostly converged on 15-18th epoch. The batch size was also lessened, from 32 to 5 as to optimize the training. In order to ensure result comparability we have chosen not to experiment with parameterization of the model, therefore, all the other parameters were used as in the original, namely, learning rate for the encoder 1e-4 and 4e-4 for the decoder, dropout 0.5, and gradient clipping at 5.

(iii) **Caption augmentation**
For caption augmentation experiment, various models and methods, built-in in python library *nlpaug*[5] created by Ma (2019) were tried. The one that best served our purpose was an *AntonAug* augmenter which substitutes the word with its antonym, provided in WordNet. The model follows a simple architecture by substituting verbs, adjectives, and adverbs to words of opposite meaning. It was a suitable match for the dataset in this study which is based on labeled features present on human faces. We assumed that the additional information on which features are lacking would make the data more varied because it is very rare that a dataset contains labels to what is *not* in the given picture. For the maximum benefit, we set the augmentation probability to 1 in the model architecture in order to convert all captions to negative features.

(iv) **Linear Classification**
Lastly, we believe that linear classification models would be most susceptible to data augmentation and distortion. Therefore, to verify the hypothesis, we trained two linear multi-label classifiers, namely Random Forest and k-Nearest Neighbours. The tensors of grayscale images were mapped against the encoded binary feature annotations of the *CelebA-HQ* dataset. 9,000 randomly sampled data points were used to train the models, while the remaining 1,000 were used as the test set.

4 Methods

The following sections provide detailed information and the step we took in carrying out each of the five components of the study.

4.1 Composite Generation and Distortion

We trained the models described in (i) to automatically generate facial composites. We used the following three datasets to train the models: CUHK dataset with 188 face-sketch pairs (Wang and Tang, 2009a), AR dataset with 123 photo-sketch pairs (Martinez and Benavente, 1998), and CUHK Face Sketch FERET Database (CUFSF) dataset of 1,194 sketches (Wang and Tang, 2009b; Zhang et al., 2011), for which we additionally obtained the FERET database with pictures of 1,194 people (Phillips et al., 1998). We resized the pictures and sketches to 200x250. In addition, since The FERET dataset contained pictures from various angles, we manually cleaned the dataset, leaving only one profile picture per person.

[2] https://www.kaggle.com/theblackmamba31/photo-to-sketch-using-autoencoder/notebook

[3] https://github.com/leehomyc/cyclegan-1

[4] https://github.com/sgrvinod/a-PyTorch-Tutorial-to-Image-Captioning

[5] https://github.com/makcedward/nlpaug

4.2 Caption Augmentation

For this task we implemented previously mentioned python library *nlpaug* by Ma (2019). At first, we tested word-level augmentation models by substituting existing words to synonyms or inserting additional words to caption labels. For this task we used implemented 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), RoBERTA (Zhuang et al., 2021) and simple architecture WordNet from nltk library (Bird et al., 2009) models. These various architectures were used in order to test which model produced the most desirable output to enlarge our caption vocabulary. Nevertheless, since no tested model produced coherent captions which would enrich our labels, we did not incorporate them.

Instead, we resorted to enlarging our dataset by substituting words with their respective antonyms in order to test models' ability to categorize facial features which are not present. The caption generation models were trained on two settings with augmented captions. Firstly, for each image in the training set, we kept three labels describing the existing features and added two augmented labels with the missing features. Secondly, we trained each model on five labels per image which solely contained negative (missing) features. We refer to this model as *Caption2Anton*.

4.3 Caption Generation

For evaluating the effects of image augmentation and distillation as well as the consequences of introducing noise into the data, image captioning was achieved by using the (a) original images, (b) facial composites generated by *Composites*, sketches extracted from *Face2Sketch* and (c) distorted composites generated by *Distorted* (d). For the aforementioned training, only the original captions from *CelebA-HQ* dataset were used. In the preprocessing steps, start and end indicators were added and the shorter descriptions were padded. For all training, 5 captions per image were used.

As previously mentioned, although informative, the descriptions in the dataset consist of only 69 unique tokens, which correspond to the possible facial and other features. As the vocabulary of the descriptions is rather limited, we manipulated the training data by augmenting the captions with *AntonAug2:3* (described in previous sections), and trained the caption generating model on the original photos with a mix of new and original captions with a ratio of 2:3 as well as *AntonAug5*, where all 5 captions were generated by the *AntonAug* model. Also, the training data was changed by introducing general image-caption noise by injecting the model with a small portion of *Flickr8k* data (12.5% in training and validation sets, which stands for 1000 and 125 of image-caption pairs, respectively). The latter model is reffered to as *Noisy*. The vocabulary expanded to 100 when data with antonyms was produced and to 470 when a noisy variant of the model was created.

4.4 Caption Evaluation

The caption evaluation consisted of two parts: assessing the linguistic accuracy of the caption as well as the feature representation in the caption. The linguistic similarity was assessed by establishing METEOR score[6], as introduced by Banerjee and Lavie (2005), whereas the feature representation was examined with our in-house precision algorithm, described below.

As the input into the algorithm, both the predictions and bigrams were stripped of any verbs, punctuation, stopwords, start and end indicators, or similar. The algorithm measured the precision of feature bi-grams in the produced caption against the gold standard. If the caption was identical to the gold standard in regards to features presented, a full score was given. Otherwise, the generated caption was compared to all 10 variants in the gold standard, its per-bigram precision was extracted and stored. The highest generated gold-caption pair score was kept and the overall score was averaged throughout all captions.

4.5 Feature Classification

The performance of the multi-label linear classification models was evaluated with reference to both the micro and macro averages of precision, recall, and f-score. Precision is an evaluation measure that allows evaluating how many of the predicted labels are correct, while recall serves to measure how many of the correct labels were predicted. F-score is a harmonic mean of the two. We gave equal weight to precision and recall in calculating the f-score. It is important to note that we implemented two models to ensure that the observed performance when trained on a certain type

[6] https://github.com/zembrodt/pymeteor

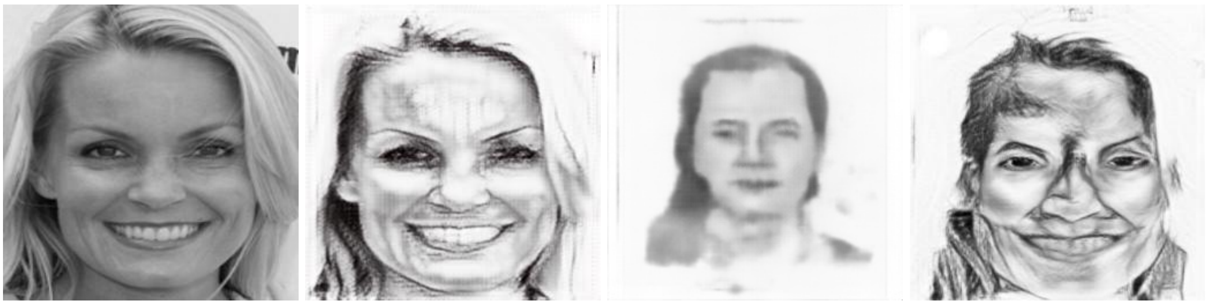

(a) Picture (b) Composite (c) Face2Sketch (d) Distorted

Figure 1: Original picture in greyscale versus the generated sketches

of data would not be specific to the particular linear model. Therefore, in evaluating the models, we focus on the differences in the performance of the models on the subsets rather than comparing the models.

5 Results

In this section, the performance of the models and their outputted scores will be discussed and evaluated in detail. Relevant examples of generated captions can be found in the Appendix.

5.1 Image and Caption Augmentation and Distortion

Examples from each of the four sets of images can be seen in Figure 1. Overall, we believe that the *Composites* were the most successful data in terms of eliminating noise yet not discarding the prominent features, as a human would likely still be able to recognize the person in the image should they know them. When it comes to the *Face2Sketch* subset, the quality is considerably poorer; nevertheless, some of the features are still visible. Lastly, the images in the *Distorted* subset appear to have additional noise in the image that partially masks some of the facial features.

With regards to caption augmentation, various models of different architecture have been tried in order to enrich the caption vocabulary. Mostly substitution and insertion methods with synonyms were used. As is apparent in Table 1 which describes the features of Figure 1, in many cases captions contain rather contextually unrelated words to human appearance. Nevertheless, models with transformer architecture performed significantly better, creating rather coherent descriptions. Although, the vocabulary did not increase significantly. Due to this reason, it was decided to only use substituted antonyms to be able to train the models in this study on descriptions which are lacking.

5.2 Feature extraction evaluation

As presented in Table 3, the results of feature evaluation show that, contrary to our belief, the ability to extract features in a human-like manner was more in-tune with the baseline model as compared to reduced noise data models, such as *Composites* and *Distorted.* Furthermore, the results mostly follow a straightforward trend: the model performs best on the original data it was trained on, which can be seen from baseline, *Composites* and *Distorted* model performances. However, it is important to note the two distinct outliers: first, the *Face2Sketch* model was more precise when tested on original images as compared to composites or distorted images, and, second, the *Noisy* model outperformed the baseline, yet its accuracy was the lowest when tested on composites and distorted images.

Furthermore, from the analysis of attention visualization, when comparing *AntonAug* and baseline, an example provided in Figure 2, it is evident that the model maps the same features as the original model, yet with antonymous meaning (e.g. ”old”, ”lacks blonde hair”)[7]. In other words, it did not indicate missing features but swapped feature words with their antonyms.

5.3 Linguistic Evaluation

The evaluation with *METEOR* score highlighted some of the trends found in feature extraction evaluation, yet also provided additional differing insights. Similar to the feature extraction evaluation, *METEOR* score is highest when the

[7] The original can be found in the Appendix, Figure 4

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.

Table 1: Caption augmentation results

model is tested on the training data for baseline, *Composites* and *Distorted*. However, the main difference can be seen in the performance of *Face2Sketch*, Noisy, as well as *AntonAug*. Although *Face2Sketch* performed rather poorly in regards to feature extraction when tested on composites and distorted, the linguistic accuracy scores much closer to the other models. Whereas the score for Noisy provides most information: even though the feature extraction score was highest for this model, it seems that introducing noise into the data distorts the model's grammatical capabilities. Finally, the comparison of *AntonAug5 AntonAug3:2* suggests that partial noise in the linguistic data does not affect the model much. It is yet unclear what threshold of augmentation could be reached for the model not to be affected by the antonyms.

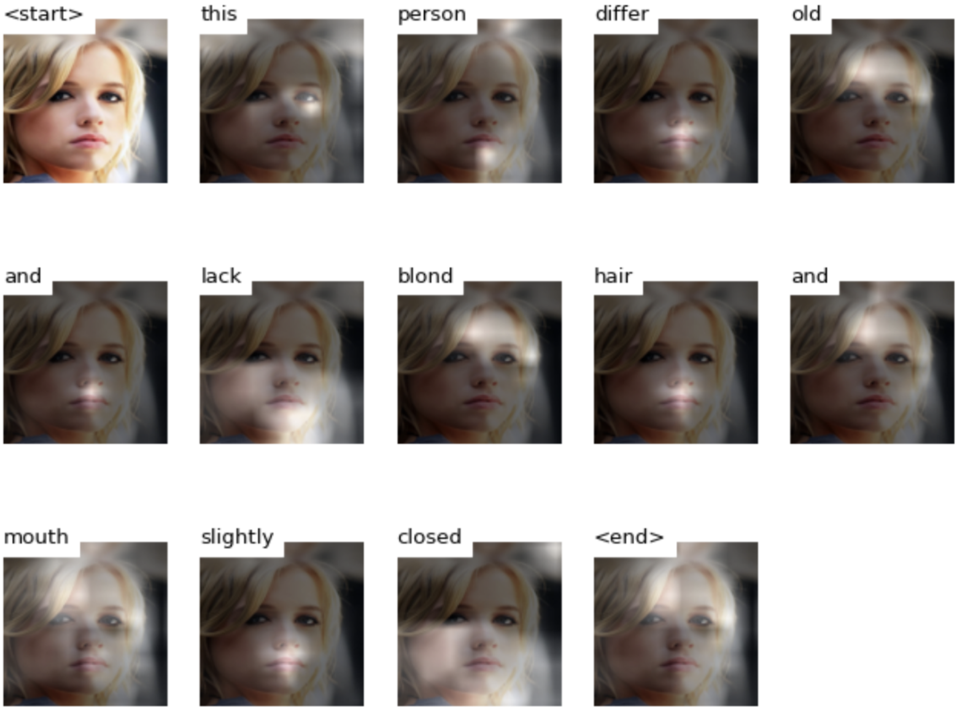


Figure 2: The visualisation of attention of AntonAug5 as evaluated on the original images. The lighter areas represent which areas are more significant when processing the sequence.

	1. img	2. cmp.	3. dist.
A. baseline	0.729	0.603	0.604
B. Composites	0.595	0.728	0.669
C. Face2Sketch	0.729	0.619	0.614
D. Distorted	0.572	0.642	0.709
E. Noisy	0.699	0.391	0.460
F. AntonAug5	0.410	0.327	0.333
G. AntonAug3:2	0.725	0.623	0.613

Table 2: The results of linguistic accuracy using METEOR score. The rows depict on which data the model was trained on, the columns present what data it was tested on. 1. img stands for images, 2. cmp for composites and 3. dist. for distorted.

	1. img.	2. cmp.	3. dist.
A. baseline	33.87	19.86	19.74
B. Composites	17.27	29.75	21.24
C. Face2sketch	28.9	14.15	14.6
D. Distorted	14.03	22.03	27.77
E. Noisy	35.29	12.38	14.15
F. AntonAug5	17.79	10.74	11.19
G. AntonAug3:2	27.52	18.53	20.52

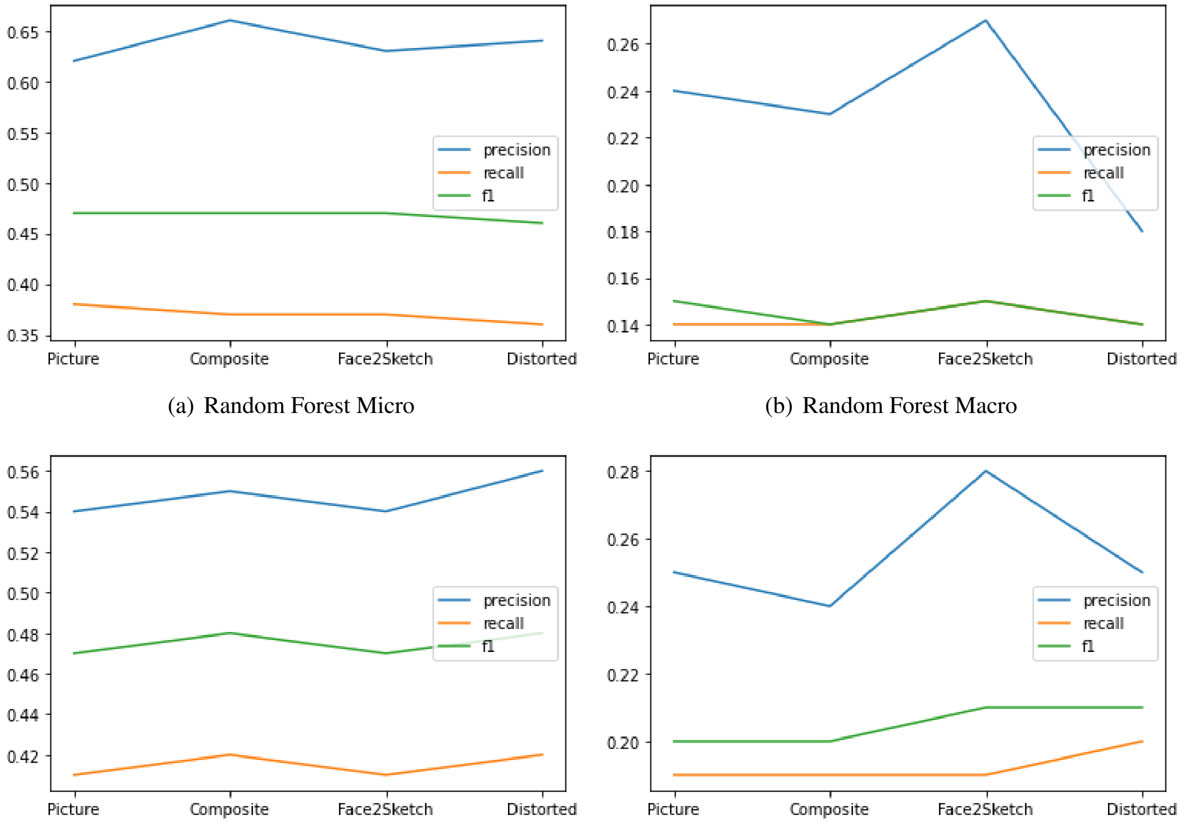
Table 3: Feature extraction results. The rows depict on which data the model was trained on, the columns present what data it was tested on. 1. img stands for images, 2. cmp for composites and 3. dist. for distorted. Values are depicted in %.

5.4 Feature Classification

The performance of linear classifiers is shown in Figure 3. Overall, the performance of the linear classifiers is rather stable across the subsets. Most notably, both models have the highest macro average precision and recall on Face2Sketch, which, from our qualitative analysis, was the least indicative of the facial features. Given that the differences in the classification of facial features were rather insignificant, it is plausible that the performance differs mostly due to variance and the differences are likely within the range of a standard deviation. Moreover, as can be observed from the micro average precision and recall scores, the models mostly learned to predict some of the most frequent facial features.

6 Discussion

The caption evaluation has revealed that, first and foremost, in regards to feature extraction, the original photos are better suited for the task of facial features grounding. This may be due to the fact that both baseline and *Noisy* receive fully-coloured



(a) Random Forest Micro

(b) Random Forest Macro

(c) k-NN Micro

(d) k-NN Macro

Figure 3: Performance of linear classifiers on the original picture versus Composite, Face2Sketch, and Distorted.

images as input, whereas *Composites*, *Distorted* and *Face2Sketch* are trained on grayscale generated images distributed over 3 color channels. This indicates that lessening the noise does not improve the learning process. In addition, as can be seen from Figure 1, the *Face2Sketch* images are furthest from human-interpretable format, nonetheless, the feature extraction score is higher compared to *Composites* or *Distorted*. As such, it could be argued that during the pooling process the features are meshed in such a manner that the high-quality of the images may not be necessary for rather adequate performance. However, when the linguistic accuracy of captions is considered, the introduction of overall noise hurts the performance the most, yet caption augmentation can score surprisingly high. Finally, introducing only antonyms without the mix with original captions results in mapping features to antonyms, not finding ”lacking” features, as hypothesized. This shows, that the model is able to discern the correct features even if the noise added to the captions seems to be unrelated to facial features.

Overall, data augmentation methods used in this research lead to satisfactory results. We believe that both the subsets of images and captions produced emphasized divergent aspects and features. For instance, the Composite subset decreased the amount of noise in pictures, whereas the antonymy-based captions offered an alternative way of grounding the facial features through feature that the faces lack rather than the features they exhibit. Nevertheless, both qualitative and quantitative analysis shows that the models were rather mislead by both the augmented images and sketches, indicating that we did not manage to enhance its facial cognition abilities.

6.1 Future work

For future work, we suggest the following directions: in regards to image augmentation effects, first, manipulate the model in such a manner that it could accommodate training on different 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 color channels at a time, thus, more information could be extracted on how the coloring of the images affects the training and, in turn, the attention and quality of the generated captions. Whereas for linguistic augmentation outcomes, we propose running the experiments on parallel multiple language data at a time and assessing whether features are mapped to certain tokens in different languages the same. An additional task would be masking words in the training data and assessing how the model adapts to the lack of data.

7 Conclusion

In this project, we sought to investigate the effects of image and caption augmentation as means of improving facial feature grounding. More specifically, we attempted both reducing and increasing the noise in both the images and captions and tracked how it altered the produced captions. We also investigated how the noise in images affected the feature recognition of linear models.

The results of our research show that for facial feature recognition and extraction, the reduction of noise only hurts the performance. In addition, by providing noisy data the model becomes better suited for capturing features, yet the performance decreases in regards to linguistic capabilities. As such, the results suggest that the best approach may be using high-quality rich image data for feature extraction with clear, but linguistically limited descriptions and then applying linguistic post-processing with context-aware language embeddings on captions to enrich them with more diverse syntactic structures.

References

Viktar Atliha and Dmitrij Šešok. Text Augmentation Using BERT for Image Captioning. *Applied Sciences*, 10:5978, 2020.

Satanjeev Banerjee and Alon Lavie. 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, 2005.

Steven Bird, Ewan Klein, and Edward Loper. Natural Language Processing with Python. 2009.

Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10578–10587, 2020.

Sherryse L Corrow, Kirsten A Dalrymple, and Jason JS Barton. Prosopagnosia: current perspectives. *Eye and brain*, 8:165, 2016.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, 2019.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. Data Augmentation for Low-Resource Neural Machine Translation. *ArXiv*, abs/1705.00440, 2017.

Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. From captions to visual concepts and back. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1473–1482, 2015.

Christiane Fellbaum. Wordnet and wordnets. In Keith Brown, editor, *Encyclopedia of Language and Linguistics*, pages 665–670, Oxford, 2005. Elsevier. URL http://wordnet.princeton.edu/.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *CoRR*, abs/1710.10196, 2017. URL http://arxiv.org/abs/1710.10196.

Ryan Kiros, Ruslan Salakhutdinov, and Rich Zemel. Multimodal neural language models. In *International conference on machine learning*, pages 595–603. PMLR, 2014.

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

Xiujun Li, Xi Yin, Chunyuan Li, Xiaowei Hu, Pengchuan Zhang, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. *ECCV 2020*, 2020.

Sungbin Lim, Ildoo Kim, Taesup Kim, Chiheon Kim, and Sungwoon Kim. Fast autoaugment. In *NeurIPS*, 2019.

Olga L. Lopatina, Yulia K. Komleva, Yana V. Gorina, Haruhiro Higashida, and Alla B. Salmina. Neurobiological aspects of face recognition: The role of oxytocin. *Frontiers in Behavioral Neuroscience*, 12, 2018. ISSN 1662-5153. doi: 10.3389/fnbeh.2018.00195. URL https://www.frontiersin.org/article/10.3389/fnbeh.2018.00195.

Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 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, 2017.

Edward Ma. NLP Augmentation. https://github.com/makcedward/nlpaug, 2019.

Mehdi Mafi, Harold Martin, Mercedes Cabrerizo, Jean Andrian, Armando Barreto, and Malek Adjouadi. A comprehensive survey on impulse and gaussian denoising filters for digital images. *Signal Processing*, 157: 236–260, 2019. ISSN 0165-1684. doi: https://doi.org/10.1016/j.sigpro.2018.12.006. URL https://www.sciencedirect.com/science/article/pii/S0165168418303979.

A.M. Martinez and R. Benavente. The ar face database. 06 1998.

Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. In *ICLR*, 2013.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. Advances in Pre-Training Distributed Word Representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA).

Osaid Rehman Nasir, Shailesh Kumar Jha, Manraj Singh Grover, Yi Yu, Ajit Kumar, and Rajiv Ratn Shah. Text2facegan: Face generation from fine grained textual descriptions. In *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)*, pages 58–67. IEEE, 2019.

Omid Mohamad Nezami, Mark Dras, Stephen Wan, and Cecile Paris. Image captioning using facial expression and attention. *Journal of Artificial Intelligence Research*, 68:661–689, 2020.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. GloVe: Global Vectors for Word Representation. In *EMNLP*, 2014.

P. Jonathon Phillips, Harry Wechsler, Jeffrey Huang, and Patrick J. Rauss. The feret database and evaluation procedure for face-recognition algorithms. *Image Vis. Comput.*, 16(5):295–306, 1998. URL http://dblp.uni-trier.de/db/journals/ivc/ivc16.html#PhillipsWHR98.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108, 2019.

Jianxin Sun, Qi Li, Weining Wang, Jian Zhao, and Zhenan Sun. *Multi-Caption Text-to-Face Synthesis: Dataset and Algorithm*, page 2290–2298. Association for Computing Machinery, New York, NY, USA, 2021. ISBN 9781450386517. URL https://doi.org/10.1145/3474085.3475391.

Cheng Wang, Haojin Yang, Christian Bartz, and Christoph Meinel. Image captioning with deep bidirectional lstms. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 988–997, 2016.

Qian Wang, Fanlin Meng, and Toby P Breckon. Data augmentation with norm-vae for unsupervised domain adaptation. *arXiv preprint arXiv:2012.00848*, 2020.

Xiaogang Wang and Xiaoou Tang. Face photo-sketch synthesis and recognition. *IEEE Transactions on Pattern Analysis and Machine In-*

telligence, 31(11):1955–1967, 2009a. doi: 10.1109/TPAMI.2008.222.

Xiaogang Wang and Xiaoou Tang. Face photo-sketch synthesis and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31:1955–1967, 2009b.

Yulin Wang, Xuran Pan, Shiji Song, Hong Zhang, Gao Huang, and Cheng Wu. Implicit semantic data augmentation for deep networks. *Advances in Neural Information Processing Systems*, 32, 2019.

Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. *Advances in Neural Information Processing Systems*, 33: 6256–6268, 2020.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Making visual representations matter in vision-language models. *CVPR 2021*, 2021.

Wayne Zhang, Xiaogang Wang, and Xiaoou Tang. Coupled information-theoretic encoding for face photo-sketch recognition. pages 513–520, 06 2011. doi: 10.1109/CVPR.2011.5995324.

Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. Character-level Convolutional Networks for Text Classification. *ArXiv*, abs/1509.01626, 2015.

Ying Zheng, Hongxun Yao, Xiaoshuai Sun, Shengping Zhang, Sicheng Zhao, and Fatih Porikli. Sketch-specific data augmentation for freehand sketch recognition. *Neurocomputing*, 456:528–539, 2021.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 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 2021. Chinese Information Processing Society of China.

Appendix

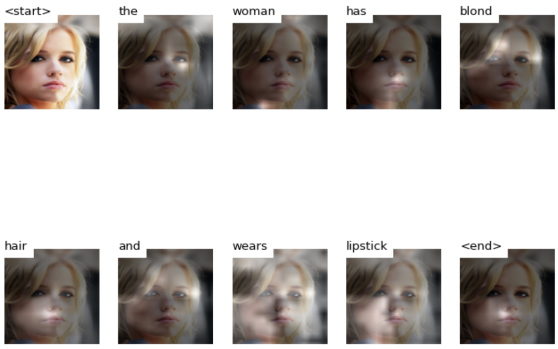
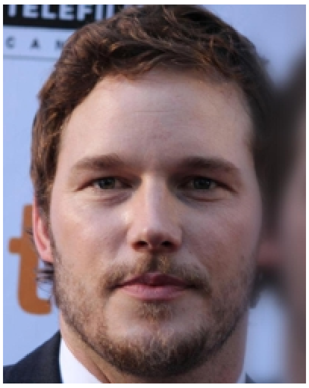


Figure 4: The visualisation of attention of baseline as evaluated on the original images. The lighter areas represent which areas are more significant when processing the sequence.

7.1 Caption examples

Below we provide some of the examples of captions as generated by the models described in Section 3.

Image id: 9737



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

- On original images:
  - *baseline:* the man has sideburns and wears necktie
  - *Composite:* this man has big lips and black hair and is wearing hat
  - *Distorted:*this person has bags under eyes and is wearing lipstick
  - *Face2Sketch:*the man has bags under eyes and big nose
  - *Noisy:*the person is young and has big nose and bags under eyes
  - *AntonAug5:* this man differ old and refuse bags under eyes and little nose
  - *AntonAug3:2:*this person has bags under eyes and big nose
- Tested on composites:
  - *baseline:* this man has big nose and big lips
  - *Composite:* this person has bags under eyes and big nose and is wearing necktie
  - *Distorted:* this woman has big nose and is wearing lipstick and hat
  - *Face2Sketch:* the man has big nose and bags under eyes
  - *Noisy:* the person is chubby and has goatee and big nose
  - *AntonAug5:* the person differ old and refuse pale skin and white hair
  - *AntonAug3:2:* the person has bags under eyes and big lips
- Tested on distorted images:
  - *baseline:* the woman has big lips and wears lipstick and earrings
  - *Composite:* this person has big lips and is wearing hat
  - *Distorted:* this person has bags under eyes big nose and sideburns
  - *Face2Sketch:* the person has big lips and wears lipstick
  - *Noisy:* the person has gray hair and big nose and is wearing necklace
  - *AntonAug5:* the person differ smiling and refuse mouth slightly closed bags under eyes and low cheekbones
  - *AntonAug3:2:* the person has mouth slightly open and big lips