# 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
[†]Faculty of Languages, Department of Linguistics and Philology,
Uppsala University, Sweden

nikolai.ilinykh, simon.dobnik@gu.se[*]
rafal.cerniavski.2286, evaelzbieta.sventickaite.9060,
viktorija.buzaite.1828@student.uu.se[†]

P-VLAM 2022, co-located with LREC 2022

# Why?

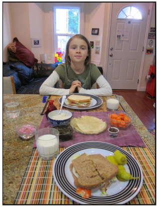
- Face recognition and description is central to social interaction and it has an impact on decision-making and inter-personal relations

# Why?

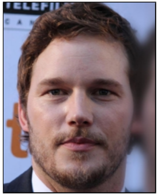
- Face recognition and description is central to social interaction and it has an impact on decision-making and inter-personal relations
- Building face description systems is beneficial for humans with prosopagnosia

# Why?

- Face recognition and description is central to social interaction and it has an impact on decision-making and inter-personal relations
- Building face description systems is beneficial for humans with prosopagnosia
- Face description generation involves subjective language and requires a fine-grained understanding of specific parts of images (in blue)



**COCO:**
A girl is sitting at a table set with sandwiches and milk.



**CelebA-HQ:**
The person has big lips, sideburns, goatee, mustache, and brown hair. He is wearing necktie.

# In this study we . . .

- Examine the fit of a standard image description generation model for face description generation

# In this study we . . .

- Examine the fit of a standard image description generation model for face description generation
- Analyse the impact of using abstract visual representations for face description generation

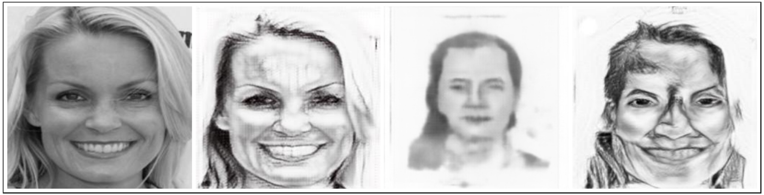
# In this study we . . .

- Examine the fit of a standard image description generation model for face description generation
- Analyse the impact of using abstract visual representations for face description generation
- Test the effects of text augmentation on the quality of generated descriptions

# In this study we . . .

- Examine the fit of a standard image description generation model for face description generation
- Analyse the impact of using abstract visual representations for face description generation
- Test the effects of text augmentation on the quality of generated descriptions
- Evaluate the quality of visual abstractions for facial feature classification

# Visual augmentation



- We train a description generation model on different visual abstractions
- From left to right: **original, composite, sketch, distorted**
- Sketches are generated by applying an auto-encoder on original images
- Composites are generated by a GAN trained for 5 epochs; distorted images are generated by the same GAN but after 33 epochs

# Linguistic augmentation

**Original:**
This person is attractive, and young and has bags under eyes, wavy hair, arched eyebrows, and mouth slightly open.

**Augmented:**
This person is not unattractive, and not old and doesn't have flat under eyes, straight hair, straight eyebrows, and mouth completely closed.

- We also train a model with original images but different/augmented descriptions
- We replace verbs, adjectives and adverbs with their antonyms and negate them
- The idea is to produce a description that is semantically close to the original text but different in terms of their form
- We also augment original dataset with a subset of the Flickr8k dataset, bringing semantic knowledge from a different multi-modal domain

# Generation: training details

- A simple CNN-LSTM model with attention
- The best model is chosen based on BLEU on the validation set after 20 epochs
- Greedy decoding

# Generation: training details

- A simple CNN-LSTM model with attention
- The best model is chosen based on BLEU on the validation set after 20 epochs
- Greedy decoding
- The models we train:
  - Visual augmentation:
  - **Baseline**: original captions and images
  - **GAN:Composite**: original captions and composite images (after 5 epochs)
  - **GAN:Distorted**: original captions and composite images (after 33 epochs)
  - **Face-2-Sketch**: original captions and sketches

  ---

  - Linguistic augmentation:
  - **Aug-Anton 3:2**: original images with 3 original and 2 augmented captions each
  - **Aug:Anton 5**: original images with 5 augmented captions each
  - **Aug-Caption**: original dataset plus a subset of Flickr8k

# Results

| 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 |

# Results

| **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 |

# Results

| **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 |

# Results

| **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 |

# Results

| **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 |

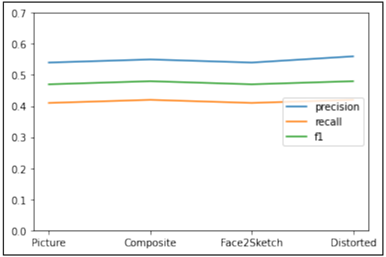
# Feature classification: training details

- We use visual representations (original, composites, sketches and distorted) and train a classifier to predict image features based on feature annotations
- Each classifier takes an image and learns to predict one of 40 features

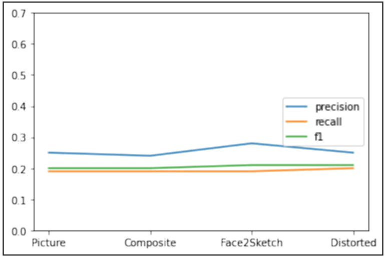
# Feature classification: training details

- We use visual representations (original, composites, sketches and distorted) and train a classifier to predict image features based on feature annotations
- Each classifier takes an image and learns to predict one of 40 features
- The models we train:
  - Random forest and k-nearest neighbours
  - Randomly select 9000 images as a training set and 1000 as the test set
  - We evaluate in terms of micro and macro averages of precision, recall and F-score

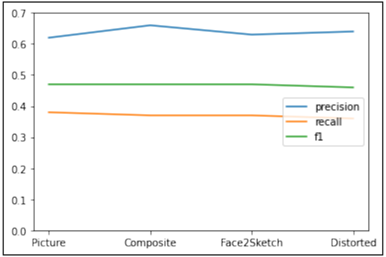
# Feature classification: results



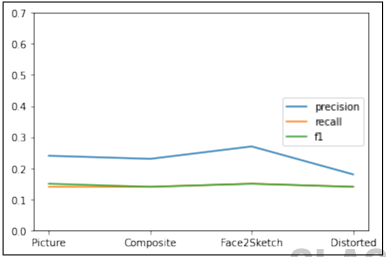
knn, micro



knn, macro



random forest, micro



random forest, macro

# Conclusions and future work

- In this study we examined the effects of visual and linguistic data augmentation on generation of descriptions of faces
- Our results indicate that original images are generally more useful. However, it is still possible to "distill" original images to such level of abstractions (e.g., sketches), in which the model still performs relatively well
- The model benefits from training on captioning data from a different domain

# Conclusions and future work

- In this study we examined the effects of visual and linguistic data augmentation on generation of descriptions of faces
- Our results indicate that original images are generally more useful. However, it is still possible to "distill" original images to such level of abstractions (e.g., sketches), in which the model still performs relatively well
- The model benefits from training on captioning data from a different domain
- Future work should focus on
    - improving the dataset to increase the variety of faces and descriptions to include all social groups
    - examining the possibility of representing images in terms of hierarchy of abstractions
    - extending the task to different languages

# Conclusions and future work

- In this study we examined the effects of visual and linguistic data augmentation on generation of descriptions of faces
- Our results indicate that original images are generally more useful. However, it is still possible to "distill" original images to such level of abstractions (e.g., sketches), in which the model still performs relatively well
- The model benefits from training on captioning data from a different domain
- Future work should focus on
    - improving the dataset to increase the variety of faces and descriptions to include all social groups
    - examining the possibility of representing images in terms of hierarchy of abstractions
    - extending the task to different languages

**Thank you for your attention!**