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?

- Building face description systems is beneficial for humans with prosopagnosia.

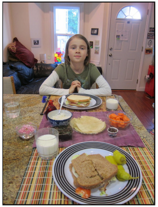
Why?

GÖTEBORGS UNIVERSITET

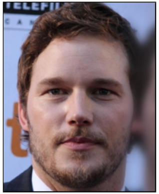
- Building face description systems is beneficial for humans with prosopagnosia.
- Face recognition and description is central to social interaction and it has an impact on decision-making and inter-personal relations.

Why?

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



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



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

In this study we . . .

- Examine the fit between the task of face description generation and standard image description generation model.
- Analyse the impact of different feature representations on the task of face description generation.
- Evaluate the quality of visual abstractions for facial feature classification.

Visual augmentation

GÖTEBORGS UNIVERSITET

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

Linguistic augmentation

GÖTEBORGS UNIVERSITET

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.

- Next, we train the model with original images, but different 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 its form (augmentation on the textual side).
- We also augment original dataset with the subset of Flickr8k dataset, bringing knowledge from a different multi-modal domain.

Generation: training details

- A simple CNN-LSTM with attention.
- The best model is chosen based on BLEU on the validation set after 20 epochs.
- Decoding: greedy.

Generation: training details

- A simple CNN-LSTM with attention.
- The best model is chosen based on BLEU on the validation set after 20 epochs.
- Decoding: greedy.
- 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

GÖTEBORGS UNIVERSITET

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

CLASP centre for linguistic theory and studies in probability

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 out of 40 features.

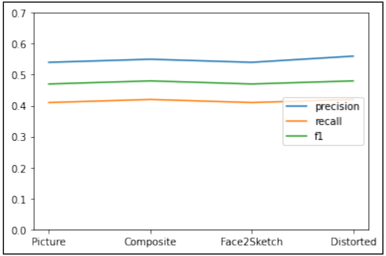
Feature classification: training details

GÖTEBORGS UNIVERSITET

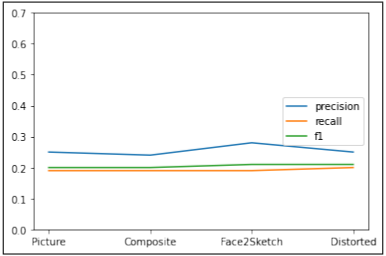
- We use visual representations (original, composites, sketches and distorted) and train a classifier to predict image features based on feature annotations.
- Each classifier takes an image and learns to predict one out of 40 features.
- The models we train:
  - Random forest and k-Nearest neighbours
  - 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.

CLASP centre for linguistic theory and studies in probability

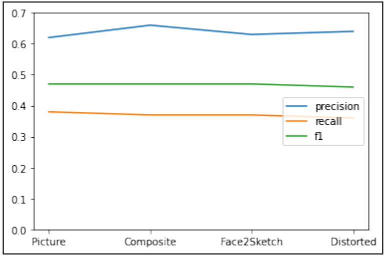
Feature classification: results



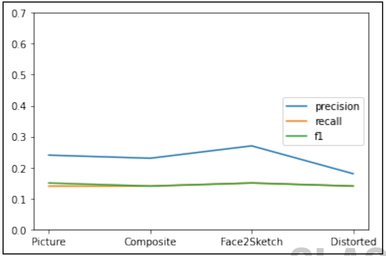
knn, micro



knn, macro



random forest, micro



random forest, macro

Conclusions and future work

GÖTEBORGS UNIVERSITET

- In this study we examined the effects of visual and linguistic data augmentation
- 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 would still perform 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
- 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 would still perform relatively well
- The model benefits from training on captioning data from a different domain.
- Future work should focus on
  - improving the dataset to include more races, genders and ethnicities
  - examining the possibility of representing images in terms of hierarchy of abstractions
  - extending the task including descriptions in different languages

Conclusions and future work

GÖTEBORGS UNIVERSITET

- In this study we examined the effects of visual and linguistic data augmentation
- 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 would still perform relatively well
- The model benefits from training on captioning data from a different domain.
- Future work should focus on
  - improving the dataset to include more races, genders and ethnicities
  - examining the possibility of representing images in terms of hierarchy of abstractions
  - extending the task including descriptions in different languages

Thank you for your attention!

CLASP centre for linguistic theory and studies in probability