Targeted Image Data Augmentation Increases Basic Skills Captioning Robustness

Valentin Barriere[2,3*], Felipe del Rio[1,3*], Andres Carvallo de Ferari[3*], Carlos Aspillaga[1,3], Eugenio Herrera-Berg[3], Cristian B. Calderon[3]

[1] Universidad Católica de Chile [2] Universidad de Chile [3] CENIA

Motivation

- Humans develop all kinds of **cognitive abilities** that allows us to interact with the world in **countless different contexts**.
- ANNs often **struggle in generalizing to out-of-context** examples**.**
- Datasets only incorporate **partial information** regarding the potential **correlational structure of the world**.

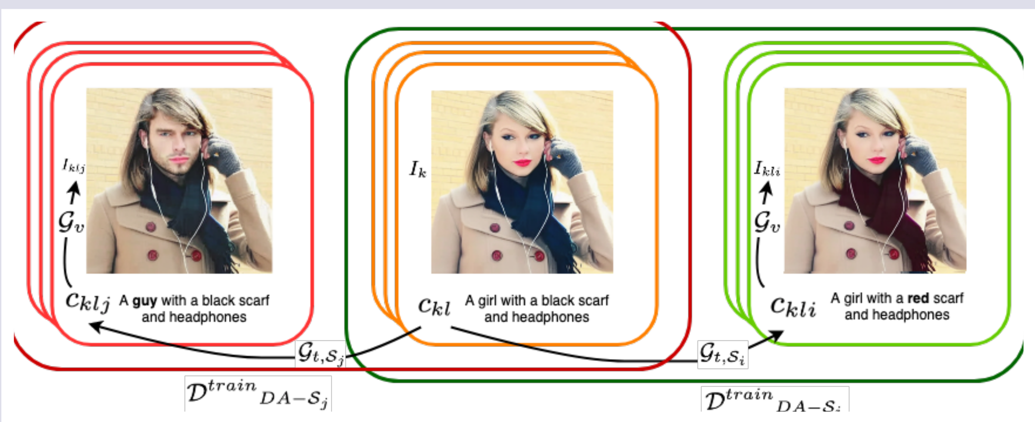

Fig. 1. Images generated using TIDA using different skills.

Targeted Image-editing Data Augmentation

TIDA, a two-step method for **generating new data examples** for a **particular skill:**

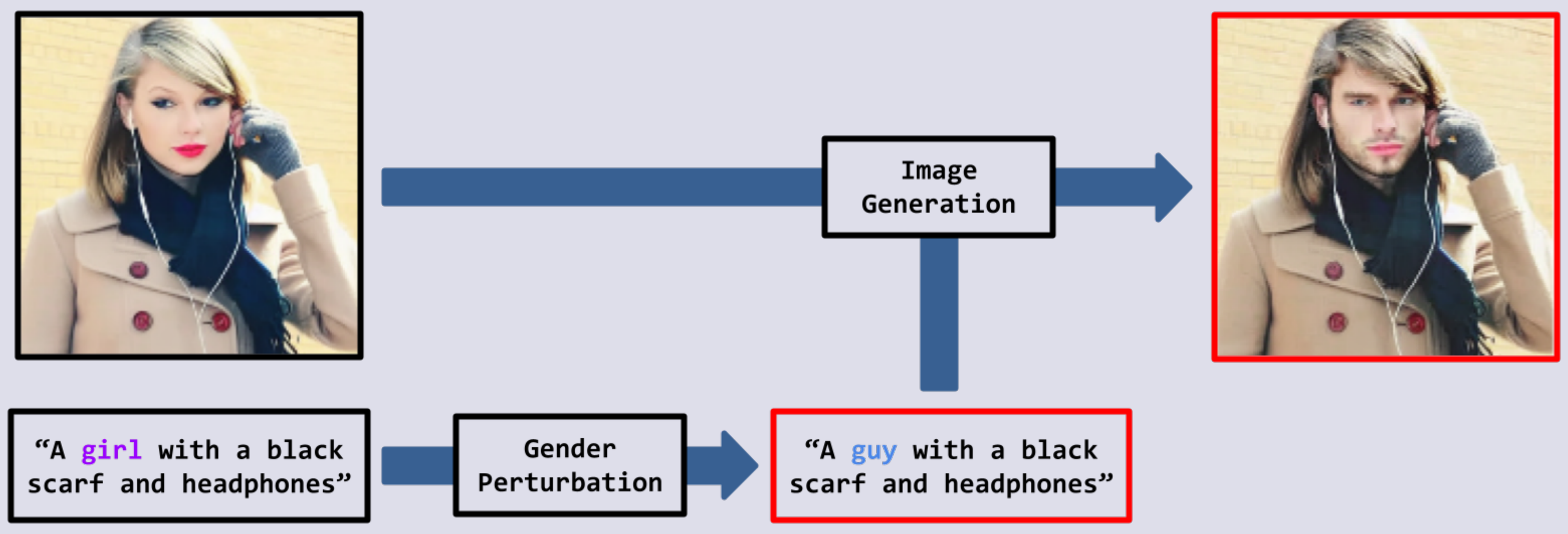

Fig. 2. Targeted Data Augmentation process.

1. Skill-related Retrieval

Create a binary classifier to **detect the presence of a skill** in a caption.

Select samples from both the train and test datasets to create a **subset for each skill**.

2. Targeted Data Augmentation

For each image in a skill subset, we create **new captions** for it, using text generators functions that **perturb its original caption**.

With these **new captions** we use **text-to-image generator** to create a new image.

Methodology

- Image Captioning task using the **Flickr30k** dataset.
- Three basic human skills:
  - **Color** recognition.
  - **Counting** capability.
  - **Gender** detection.
- **Baseline:** Model trained with dataset augmented by **generating images from random captions** of the dataset

Implementation Details

- **Stable Diffusion** [1] to generate images.
- **BLIP** [2] model for Image Captioning.

Overall Results

Performance of the BLIP model trained using the data generated by TIDA as measured by different image captioning metrics.

Test / Train	#DA	BLEU@1-4 $\mathcal{D}^{test}_{clr}$	$\mathcal{D}^{test}_{ctg}$	$\mathcal{D}^{test}_{gdr}$	$\mathcal{D}^{test}$	RefCLIPScore $\mathcal{D}^{test}_{clr}$	$\mathcal{D}^{test}_{ctg}$	$\mathcal{D}^{test}_{gdr}$	$\mathcal{D}^{test}$
$\mathcal{D}^{tr}$	0	51.8	44.0	49.9	49.7	79.9	79.3	79.8	80.3
$\mathcal{D}^{tr}_{RAND}$	60k	51.3	44.1	49.2	49.6	80.0	79.5	79.7	80.2
$\mathcal{D}^{tr}_{COLOR}$	20k	*51.7*	44.0	*49.3*	49.5	79.8	79.4	79.6	80.1
$\mathcal{D}^{tr}_{COUNT}$	20k	*51.7*	*44.4*	49.2	49.7	79.9	*79.5*	79.7	80.2
$\mathcal{D}^{tr}_{GENDER}$	20k	51.2	43.4	48.5	48.8	*80.0*	79.2	*79.9*	80.3
$\mathcal{D}^{tr}_{ALL}$	60k	**51.8**	**44.9**	**50.1**	**50.5**	**80.1**	**79.7**	**80.1**	**80.5**

Table 1. TIDA Performance with different metrics.

- Overall **best scores on each test set** are obtained with the model that uses the **combination** of the **three types of data-augmentation** techniques.
- Counterintuitively, skill-related TIDA are **not achieving the best scores in their respective test sets.**

Use of Skill-Related Words

- Investigate specific semantic words and evaluate the propensity of the model to use those words in the right context.
- Measure the inclusion of skill-related words in the captions, as compared to the ground-truth.

Findings

- The model use skill-associated words more often when the caption should contain one and less when it should not.
- Sometimes, as for gender, precision is decreased.

Skill / Train	Color F1	Counting F1	Gender F1
$\mathcal{D}^{tr}$	66.7	69.4	**74.1**
$\mathcal{D}^{tr}_{RAND}$	67.0	**75.5**	73.4
$\mathcal{D}^{tr}_{COLOR}$	68.4	69.2	72.4
$\mathcal{D}^{tr}_{COUNT}$	68.1	71.0	73.2
$\mathcal{D}^{tr}_{GENDER}$	66.1	72.3	72.4
$\mathcal{D}^{tr}_{ALL}$	**68.6**	73.4	**74.1**

Table 2. Skill-related inclusion.

Conclusions

- TIDA allows for gains regarding classical metrics.
- TIDA helps the image captioning model to use those words more efficiently

References

[1] Rombach, R. et al. High-resolution image synthesis with latent diffusion models. CVPR, 2022.

[2] Li, J. et al. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. ICML, 2022.

Acknowledgements

Research partly funded by National Center for Artificial Intelligence CENIA, FB210017, BASAL, ANID

Check our paper!