Dataset Creation for Visual Entailment using Generative AI

Rob Reijtenbach, Suzan Verberne, Gijs Wijnholds. Presenter: Mehrnoosh Sadrzadeh

Universiteit Leiden
The Netherlands

Introduction

- **Natural language inference (NLI)** is a classification problem for pairs of two texts:

  a premise and a hypothesis

- The pair is labeled as entailment, neutral, or contradiction

- In **visual entailment** (VE) tasks, the premise is substituted by an image; the hypothesis is still in text
- Example:

\+

- *Two woman are holding packages.*
- *The sisters are hugging goodbye while holding to go packages after just eating lunch.*
- *The men are fighting outside a deli.*

=

- *Entailment*
- *Neutral*
- *Contradiction*

Premise *Hypothesis* *Answer*

https://github.com/necla-ml/SNLI-VE

Our motivation and goals

- Dataset creation for visual entailment is costly
- We evaluate the use of **generative AI for VE dataset creation**
- This allows cheaper and easier dataset creation.

- We introduce a synthetic version of the SNLI-VE dataset called Synthetic-NLI-VE
- We evaluate the quality of models trained on the synthetic dataset on real data and compare to models trained on real data

Data

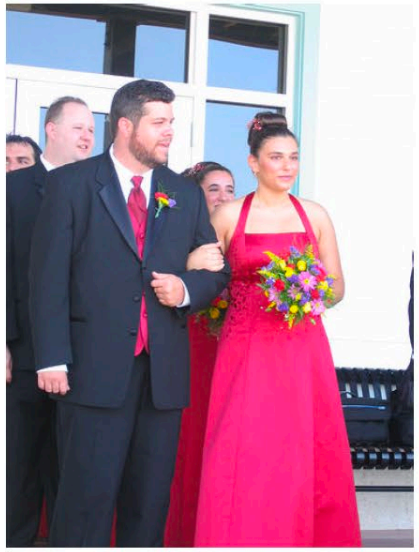
- **SNLI-VE** (Xie et al., 2019)
  - combining the SNLI dataset with the Flickr30k dataset
  - 31,783 photos of everyday activities with 5 different captions

- **SICK-VTE** (Iokawa et al., 2024)
  - a visual entailment version of (a subset of) the SICK dataset with a multilingual component
  - For 488 unique images there are 2,899 sentence pairs, with 1,930 Entailment and 969 Contradiction.

Xie et al. 2019. Visual entailment: A novel task for fine-grained image understanding.
Iokawa et al. 2024. Multilingual visual-textual entailment benchmark with diverse linguistic phenomena.

Example

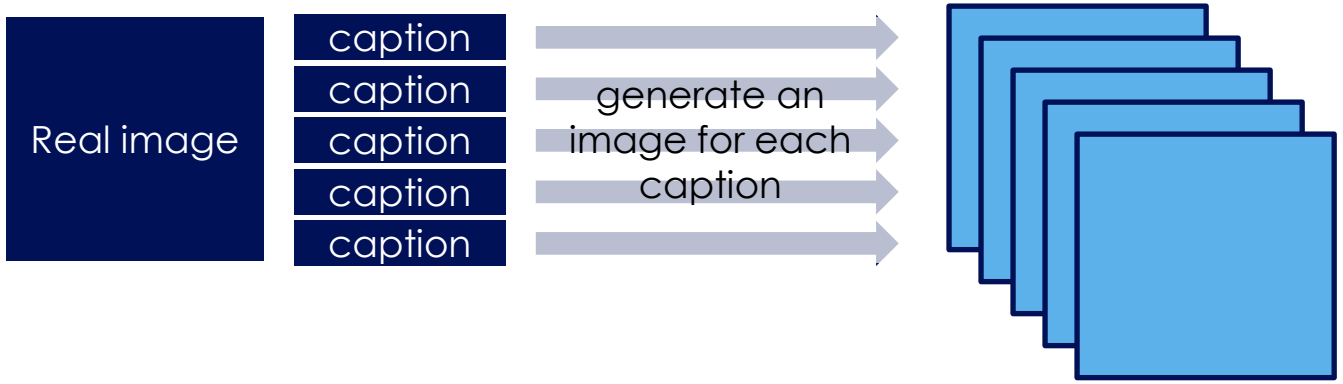
- Example of a photo from the SNLI dataset with 5 different captions



- A bearded man, and a girl in a red dress are getting married.
- A wedding party walks out of a building.
- The group of people are assembling for a wedding.
- A man and woman dressed for a wedding function.
- A woman holds a man's arm at a formal event.

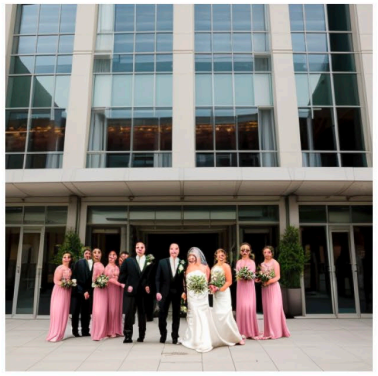
Methods (1)

- **Image generation**:
  - Use the caption (premise) text from SNLI as input prompts in a generative model
  - creating an image for every premise caption.
  - This results in a dataset similar to SNLI-VE, but with a unique image for every premise.



Methods (2)

- **Image generation**:
  - Use the caption (premise) text from SNLI as input prompts in a generative model
  - creating an image for every premise caption.
  - This results in a dataset similar to SNLI-VE, but with a unique image for every premise.



A wedding party walks out of a building.

The group of people are assembling for a wedding.

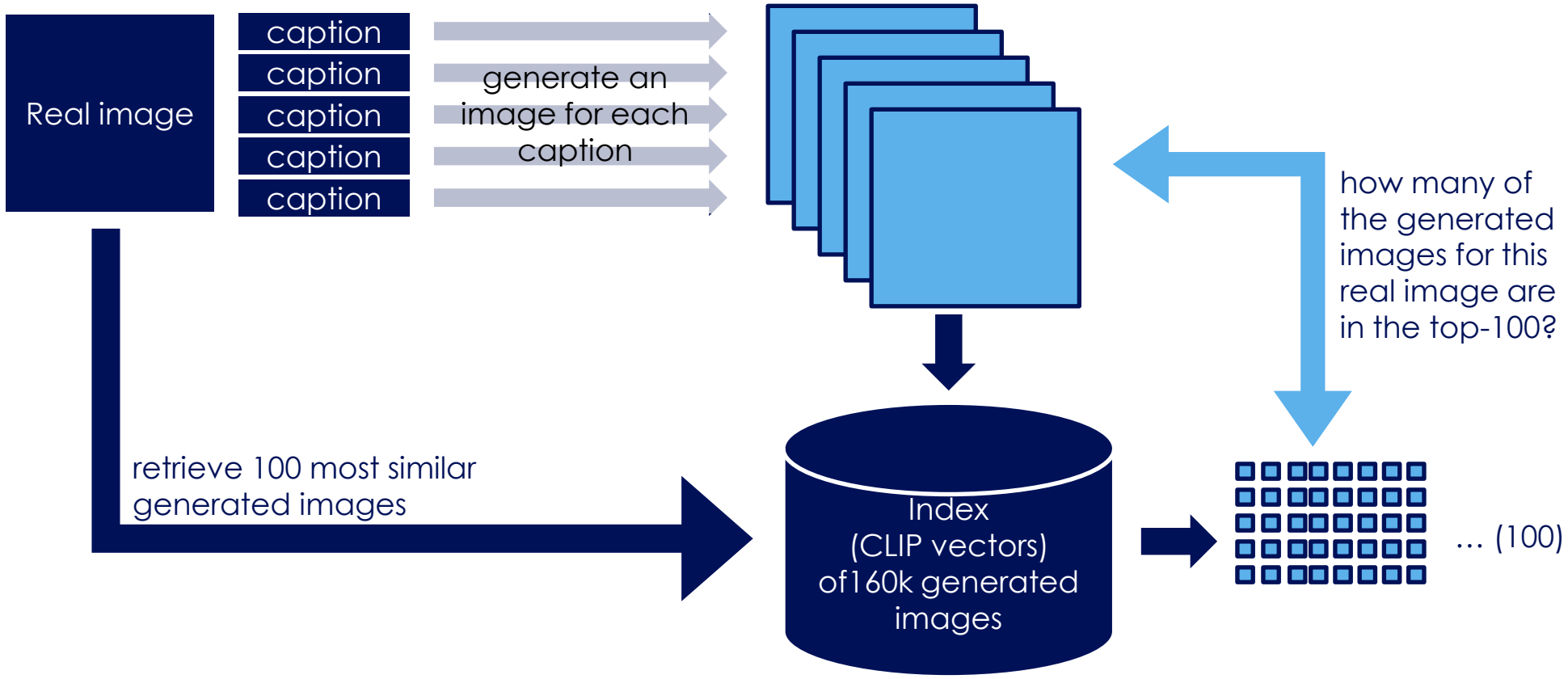
A man and woman dressed for a wedding function.

- **Model**: Stability AI's Stable Diffusion, checkpoint: Realistic Vision v51

https://github.com/Stability-AI/generative-models

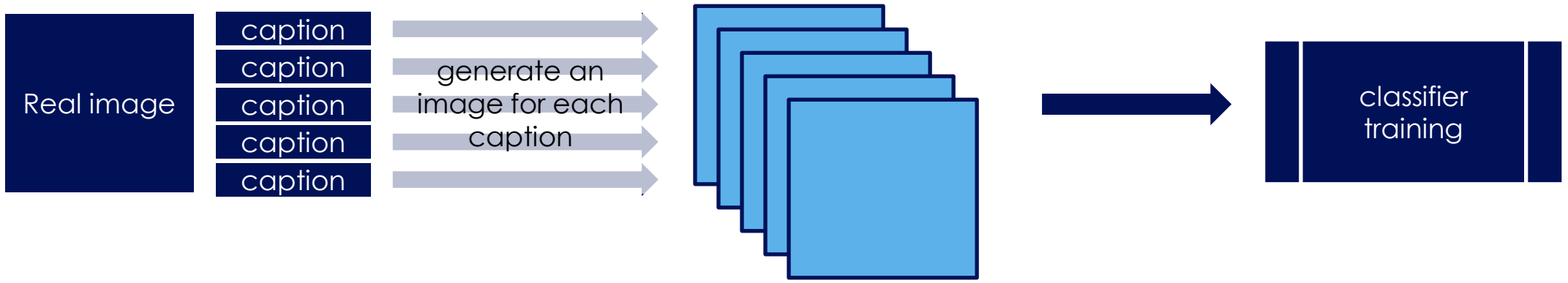
Evaluation (1)

- **Intrinsic evaluation**: Ranked similarity
  - For a given original image, rank the generated images from the full synthetic dataset
  - The 5 generated images corresponding to the original image should be in the top-100



Evaluation (2)

- **Extrinsic evaluation**: classification
  - Using the synthetic images to train a classifier
  - Model: CLIP for visual entailment (Song et al. 2022)
  - Evaluation on real data, comparison to a model trained on real data

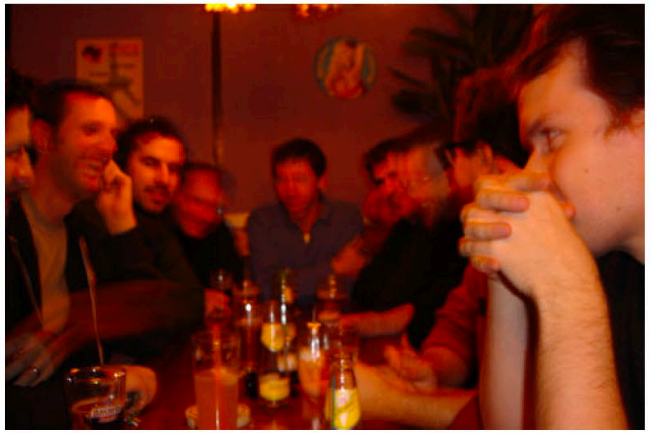


Haoyu Song et al. (2022) CLIP models are few-shot learners: Empirical studies on VQA and visual entailment

Results: intrinsic evaluation

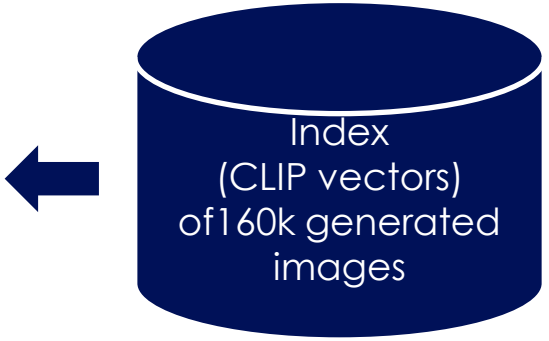
- Of the 100 most similar synthetic images to each real image, on average only **1.6 were generated from one of that real image's captions**
- This is because out of ~160k generated images many other images are similar to the real image

Example: this was counted as a negative because the generated image (b) was not generated for one of the captions of this original image (a). But they are clearly similar.



(a) Original

(b) Generated



Results: extrinsic evaluation (in-domain)

- Accuracies of both models on both test sets (original and generated) of SNLI-VE:

Train set	Test: Original	Test: Generated
Original	70.3%	71.1%
Generated	68.9%	73.2%

- The model trained on original images performs better on the generated test set than it does on the original test set.
  - This could suggest that the generated test set is "easier" to classify.
- The model trained on generated data and tested on original data has a somewhat lower performance in this experiment, **but the difference is small**.

Extrinsic evaluation (out-domain)

- Accuracies of both models on the SICK-VTE (original and synthetic) datasets:

Train set	Test: Original	Test: Synthetic
Original	50.7%	51.4%
Generated	47.2%	47.6%

- Performance is **relatively poor**, given a majority baseline of 66%.
  - This result is in line with the findings of Talman and Chatzikyriakidis (2019), who found similar issues when transfering models trained on the SNLI dataset to the SICK dataset.
- Again, the model trained on generated data performs **slightly worse** compared to the model trained on original data.

Conclusions

Synthetic dataset generation for visual entailment is a viable option

The dataset proved to have similar utility in a classification task compared to the original, human-created data

But: Cross-domain evaluation was relatively poor

Future work could investigate:

- further optimization of the generative model
- generating more than one image per caption to increase diversity in the data
- the use of multiple classification models for evaluation

Thank you!

https://huggingface.co/datasets/robreijtenbach/Synthetic-NLI-VE

Universiteit
Leiden
The Netherlands