SEMANTICALLY INVARIANT TEXT-TO-IMAGE GENERATION

Shagan Sah, Dheeraj Peri, Ameya Shringi, Chi Zhang, Miguel Dominguez, Andreas Savakis, Raymond Ptucha
Rochester Institute of Technology, New York, USA
Contact- sxs4337@rit.edu

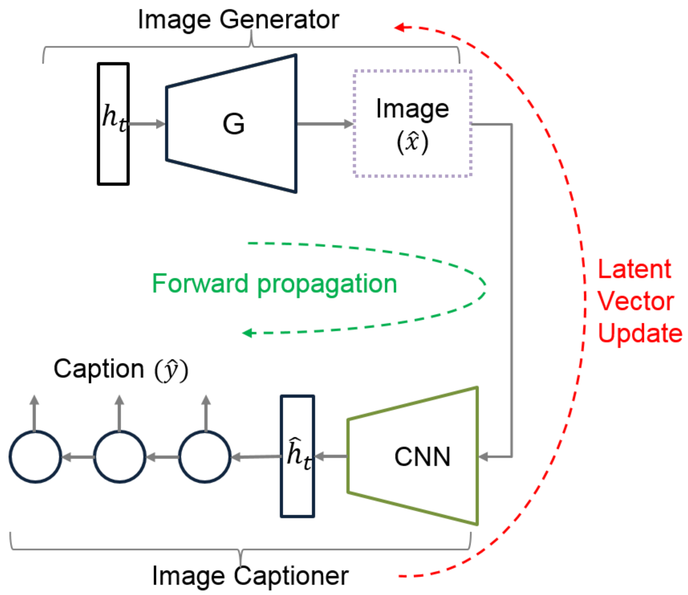
Common Vector Space

- An ambitious goal for machine learning in the vision and language domain is to be able to represent different modalities of data that have the same meaning with a common latent representation.
- A sufficiently powerful model should be able to store similar concepts in a similar representation or produce any of these realizations from the same latent space.
- Successfully mapping visual and textual modalities in and out of this latent space would significantly impact the broad task of information retrieval.
- We propose a cross-domain model capable of converting between text and image.

Text-to-Image Generation

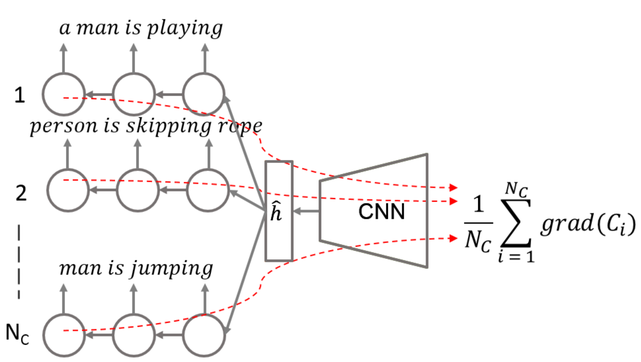
- Recent work in image generation has shown significant improvements in image quality when text is used as a prior.
- We propose two improvements to the text conditioned image generation-
  - A n-gram metric based cost function is introduced that generalizes the caption with respect to the image.
  - Multiple semantically similar sentences are shown to help in generating better images.

Architecture

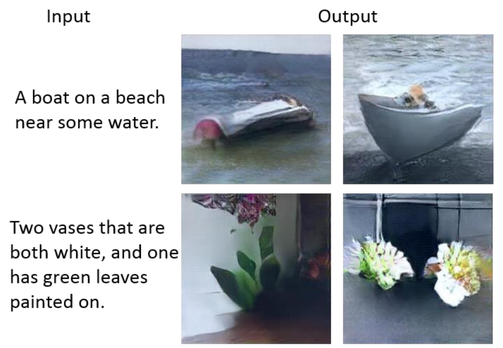


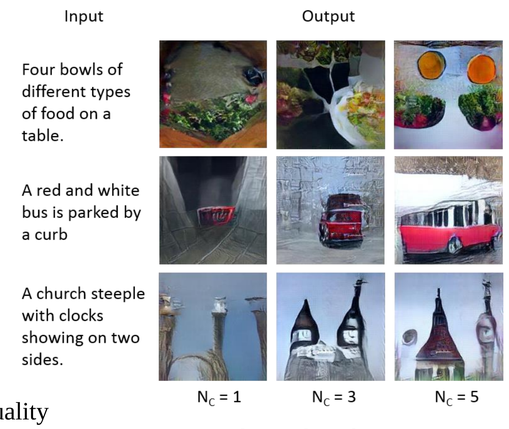
- The model consists of two pre-trained modules -- an image generator (G) that inputs a latent representation h_t and generates an image $\hat{x}$; and an image captioner that inputs an image $\hat{x}$ and generates a caption $\hat{y}$.
- To update the latent vector h_t, cross-entropy between the generated caption $\hat{y}$ and a ground truth caption y is used while the weights for the generator and CNN are fixed.

Condition on Multiple Captions



Results





N_C is the number of captions

Evaluation of the generated image quality using the inception and detection scores.

Method	Inception	Detection
Baseline (FC-6)	5.77 ± 0.96	0.762
PPGN [5]	6.71 ± 0.45	0.717
MMVR (B-1)	7.22 ± 0.81	0.713
MMVR ($N_c = 5$)	$\mathbf{8.30 \pm 0.78}$	**1.004**

N_C	Inception	Detection
1	7.22 ± 0.81	0.713
3	8.04 ± 0.57	0.915
5	$\mathbf{8.30 \pm 0.78}$	**1.005**

Detection Metric for Evaluation

Pre-trained YOLO object detector tested on synthetically generated images.

[5] Nguyen, Anh, et al. "Plug & Play Generative Networks: Conditional Iterative Generation of Images in Latent Space." *CVPR*. Vol. 2. No. 5. 2017.