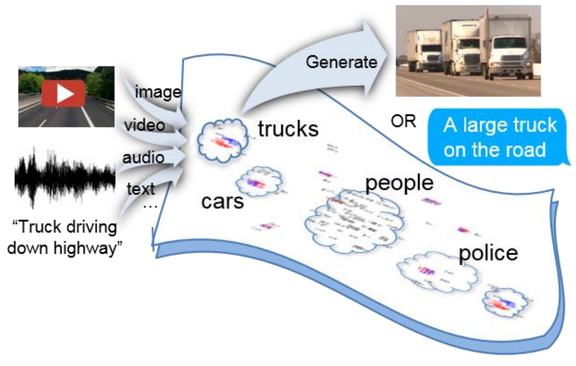
MULTIMODAL RECONSTRUCTION USING VECTOR REPRESENTATION

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

Introduction

- Recent work has demonstrated that neural embedding from multiple modalities can be utilized to focus the results of generative adversarial networks.
- However, little work has been done towards developing a procedure to combine vectors from different modalities for the purpose of reconstructing input.

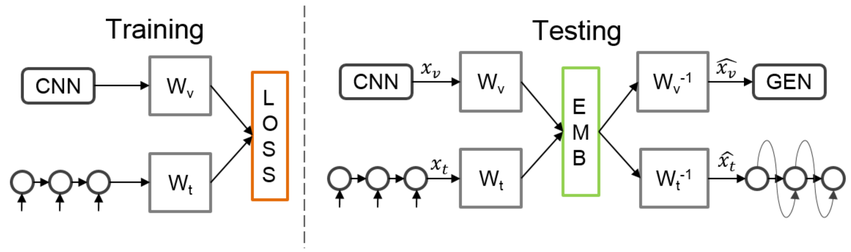
Common Vector Space



- In this paper, we propose learning a Common Vector Space (CVS) where similar inputs from different modalities cluster together.
- We develop a framework to analyze the extent of reconstruction and robustness offered by CVS.

[1] Oh Song, Hyun, et al. "Deep metric learning via lifted structured feature embedding." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016.

Architecture



- Training and testing modes for learning the common vector space.
- During training, the encoded input modalities are aligned through a loss function.

$$Loss = L_r + L_m$$

Reconstruction loss

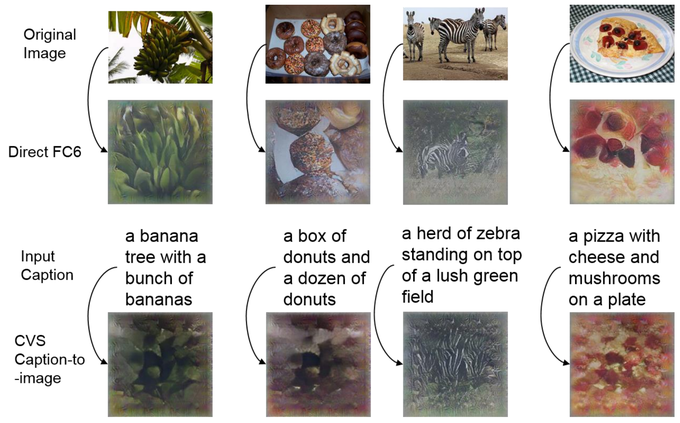
$$L_r = ||x^p - \hat{x^p}||_2$$

Metric loss between positive and negative pairs) [1]

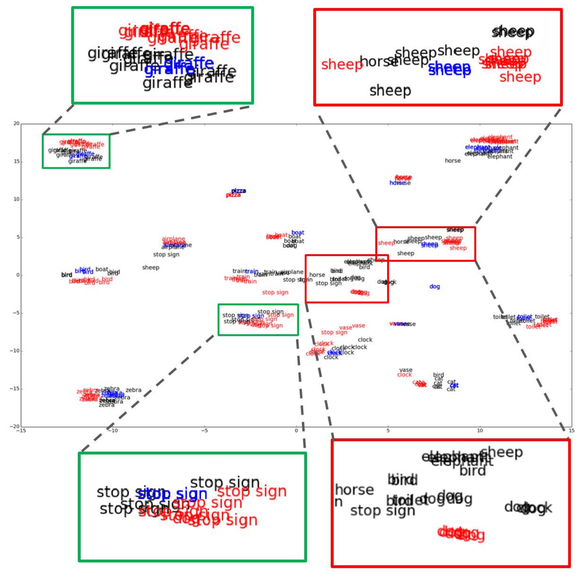
$$L_m = \frac{1}{2|\hat{P}|}\left(\lambda_{\hat{N}} \sum_{(i,j)\in\hat{P}} \left(\log\left(\sum_{(i,k)\in\hat{N}} \exp(\alpha - d_{i,k})\right) + \sum_{(j,k)\in\hat{N}} d_{j,k}\right) + \lambda_{\hat{P}} d_{i,j}\right)^2$$

- The learned weights are inverted in test phase to reconstruct the input modality.

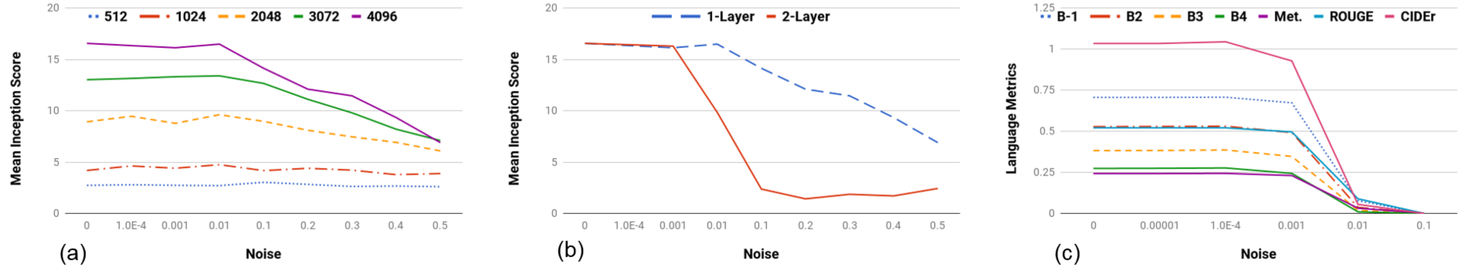
Results



Caption to image generation examples. Images generated through the direct FC6 vectors are the upper-bound on the quality of the generated images.



t-SNE visualization the common vector space on a validation set. Red, black and blue colors indicate captions, images and word categories, respectively.



Noise analysis of image and caption generator.