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. 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.

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.

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

\[
\text{Loss} = L_r + L_m
\]

Reconstruction loss
\[
L_r = \| x^p - \hat{x}^p \|_2
\]

Metric loss between positive and negative pairs [1]
\[
L_m = \frac{1}{2 |P|} \sum_{(i,j) \in P} \left( \log \left( \sum_{(i,k) \in N} \exp(\alpha - d_{i,k}) \right) + \sum_{(j,k) \in N} d_{j,k} + \lambda_p d_{i,j} \right)^2
\]

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

\[\text{Noise analysis of image and caption generator.}\]