Experimental Setup: Datasets and Features

- Datasets in Human Multimodal Language
  - Multimodal Personal Trait Recognition
  - Multimodal Sentiment Analysis
  - Multimodal Emotion Recognition
  - Recorded Dyadic Dialogues (EMOCAP)

- Multimodal Features
  - Language: pre-trained Glove word embeddings
  - Visual: facial action units from Facet
  - Acoustic: MFCCs from COVAREP
  - Align by PVFA

Multimodal Representation Learning

Multimodal Factorization Model (MFM)

- Bayesian Network

\[
\begin{align*}
&X_{1:M} \sim \text{multimodal data from } M \text{ modalities, } Y \text{ labels} \\
&Y_{1:M} \sim \text{generated multimodal data, } Y \text{ generated labels} \\
&Z_{d} \sim \text{modality-specific latent variables, } F \text{ factors}
\end{align*}
\]

Generation: Factorization over Joint Distribution

\[
\begin{align*}
&\mathcal{Q}(Z_{1:M} | Y) = \mathcal{Q}(Z_{d} | X_{1:M}) \mathcal{Q}(Z_{a} | Y_{1:M})
\end{align*}
\]

Inference: Joint-Distribution Wasserstein Distance

\[
\begin{align*}
W_{1}(p_{y|y}, p_{y}) &= \inf_{\Pi_{y|y} \in \Pi} \mathbb{E}_{\Pi_{y|y}} \left[ \phi_{y|y}(y) \right] + \mathbb{E}_{\Pi_{y|y}} \left[ \phi_{y}(y) \right]
\end{align*}
\]

where \( \phi_{y|y} \) is the prior over \( Z = \{Z_{d}, Z_{a}(1:M)\} \), and \( \phi_{y} \) is the aggregated posterior of the proposed approximate inference distribution \( \hat{p}(Z|X_{1:M}, Y) \).

Relaxed Generative-Discriminative Objective

- Mean-field assumption:

\[
\begin{align*}
&\mathbb{E}_{p_{y}(y)} \left[ \Psi \left( \phi_{y}(y) \right) \right] \\
&\text{Objective:} \\
&\sum_{y} \mathbb{E}_{p_{y}(y)} \left[ \Psi \left( \phi_{y}(y) - \phi_{y}(\hat{y}) \right) \right] + \lambda \mathbb{E}_{p_{y}(y)} \left[ \Psi \left( \phi_{y}(\hat{y}) \right) \right]
\end{align*}
\]

MFM achieves strong performance on 6 multimodal time-series datasets

MFM can be applied on any multimodal fusion encoder