Multimodal Alignment using Variational Inference and Wasserstein Distance

Sheikh Shams Azam

3rd Ph.D. Student

Elmore Family School of ECE

Motivation

Motivation

Image Text Alignment

Fox jumps over dog

Image and text come from the same source or define some structure together but cannot be combined naively. So, we want to develop an architecture to learn a joint and "meaningful" representation.

Document Understanding

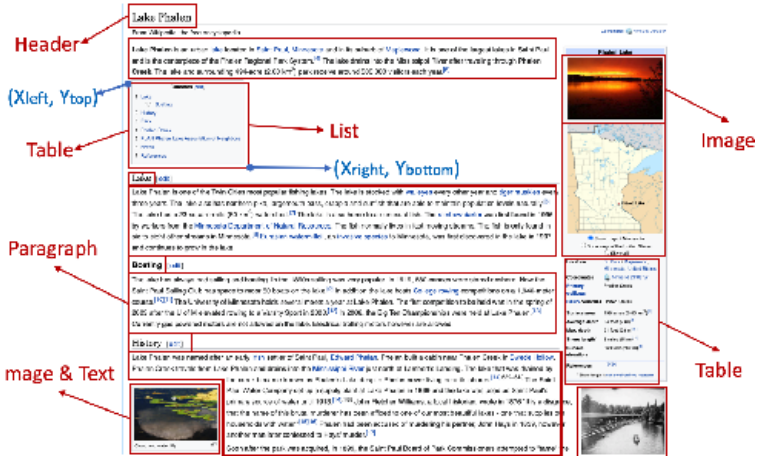
Motivation

Image Text Alignment

Fox jumps over dog

Image and text come from the same source or define some structure together but cannot be combined naively. So, we want to develop an architecture to learn a joint and "meaningful" representation.

Document Understanding

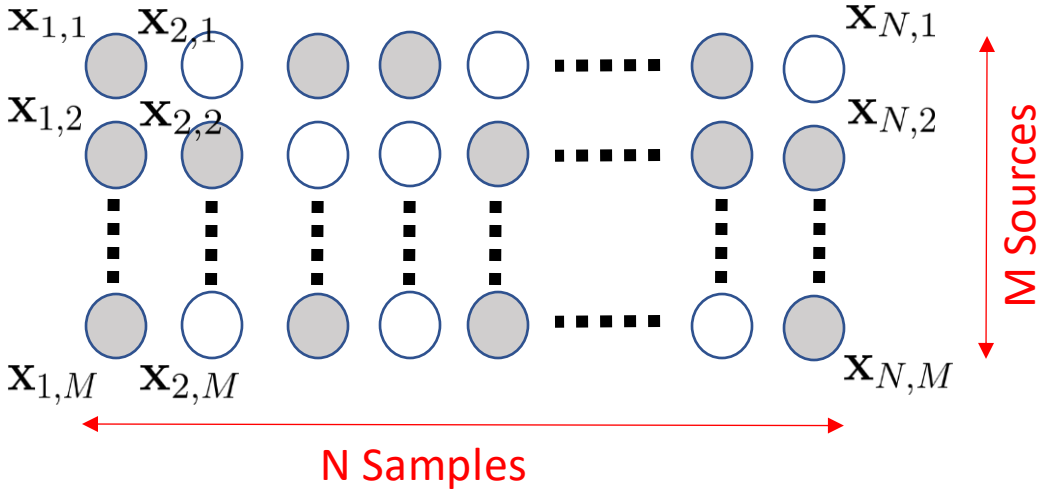


Data is not perfect

unobserved

observed

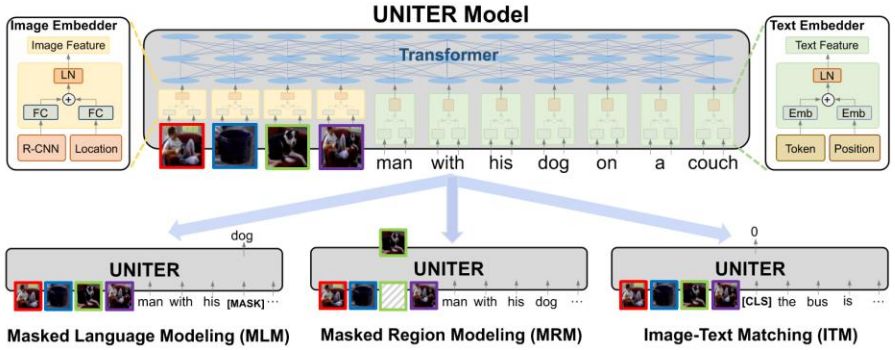
$\mathbf{x}_{n,m}$ m-th mode of n-th sample



Related Works

Related Works

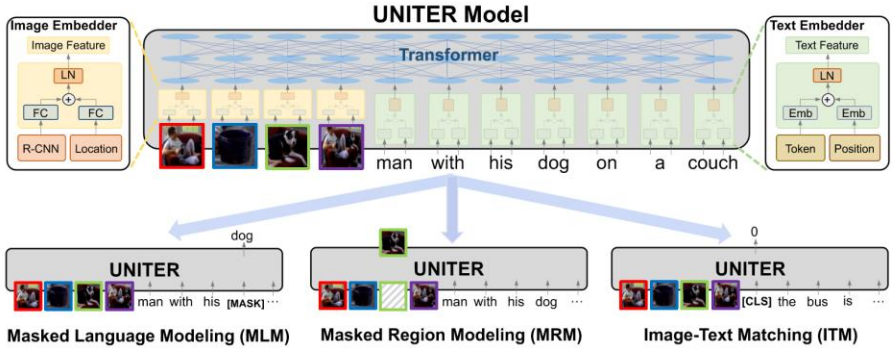
Implicit Alignment Techniques



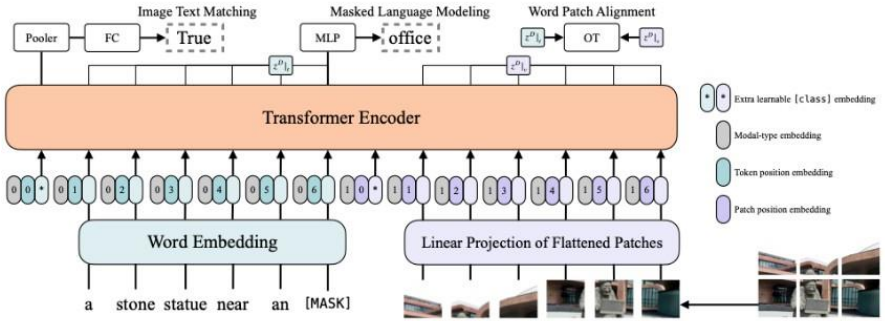
UNITER by Chen et al. ECCV, 2020.

Related Works

Implicit Alignment Techniques



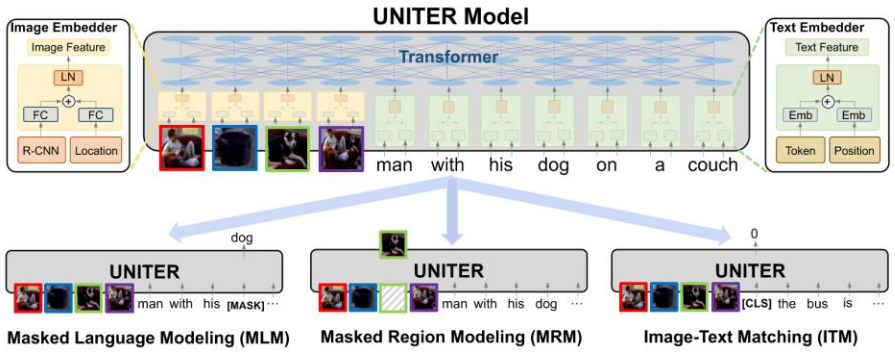
UNITER by Chen et al. ECCV, 2020.



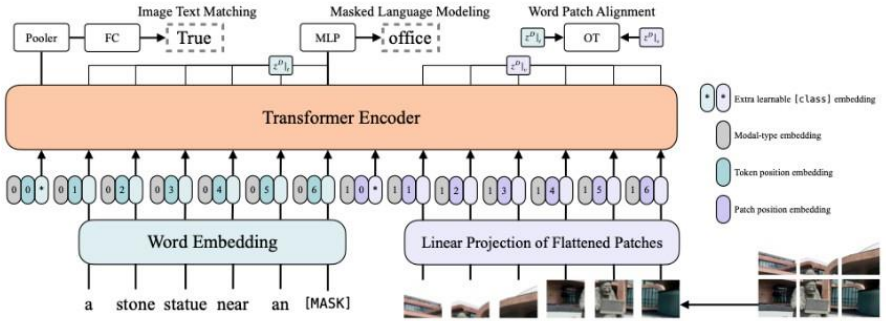
VILT by Kim et al. ICML, 2021.

Related Works

Implicit Alignment Techniques

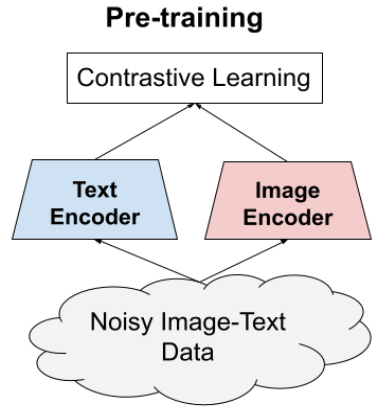


UNITER by Chen et al. ECCV, 2020.



VILT by Kim et al. ICML, 2021.

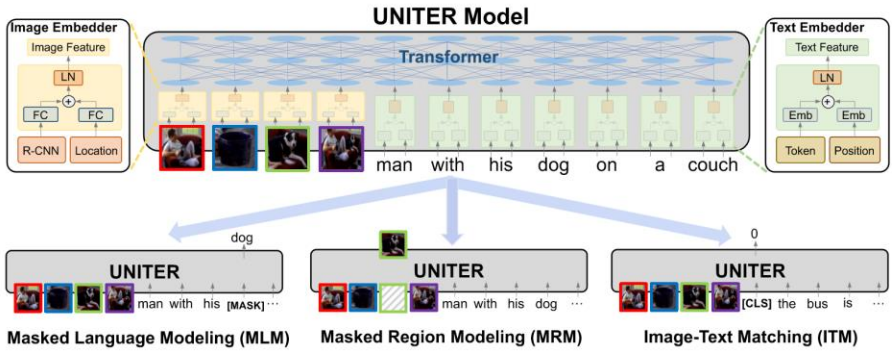
Explicit Alignment Techniques



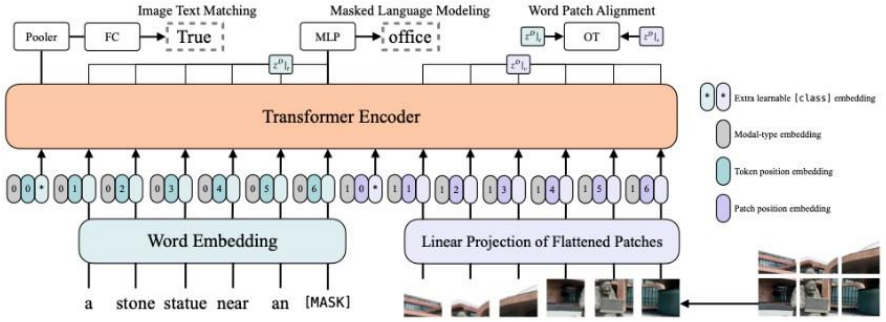
ALIGN by Jia et al. ICML, 2021.

Related Works

Implicit Alignment Techniques

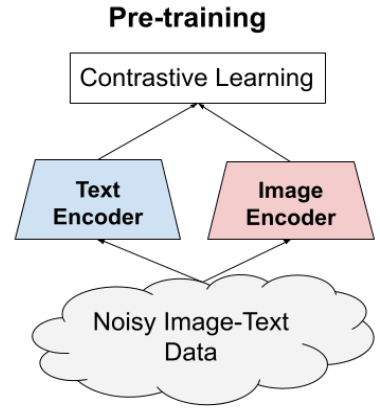


UNITER by Chen et al. ECCV, 2020.

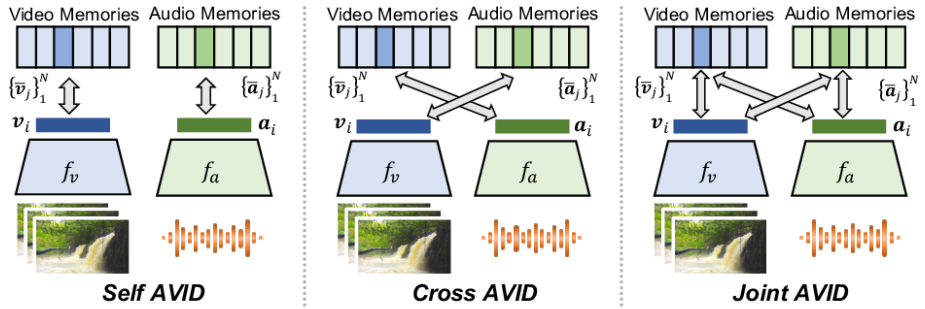


VILT by Kim et al. ICML, 2021.

Explicit Alignment Techniques



ALIGN by Jia et al. ICML, 2021.



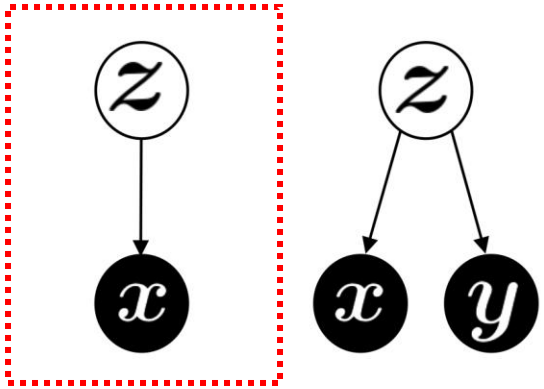
AVID by Morgado et al. CVPR, 2021.

Some Drawbacks

- Models are very large.
- Training takes several weeks.
- No direct comparison of models.
- Indirect comparison through downstream finetuning tasks.
- Not theoretically grounded. Self-supervised tasks are ad-hoc.

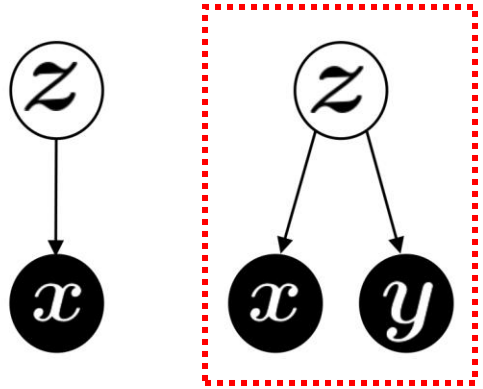
VAE for Multimodal Alignment

VAE for Multimodal Learning



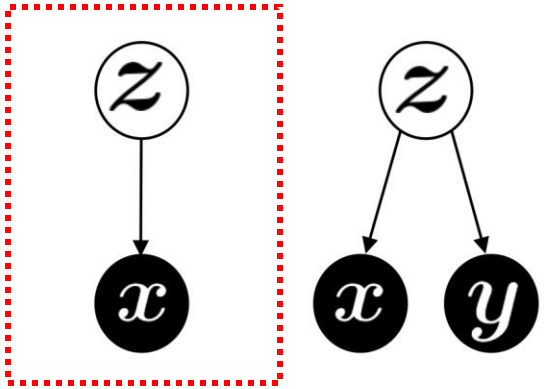
- Original VAE was proposed for unimodal density estimation.

VAE for Multimodal Learning



- Original VAE was proposed for unimodal density estimation.
- We extend unimodal VAE (Kingma & Welling, 2014) to a multimodal setting.

VAE for Multimodal Learning



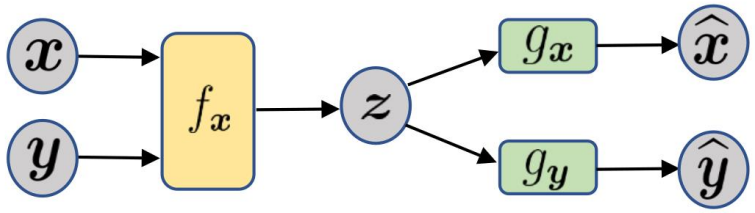
- Original VAE was proposed for unimodal density estimation.
- We extend unimodal VAE (Kingma & Welling, 2014) to a multimodal setting.
- We also want to ensure that we can do inference (inferring z) with only one modality (x or y or both).

Implementation

Implementation at: https://github.com/shams-sam/MultiModal

Implementation

Naïve Extension of VAE to Multimodal VAE

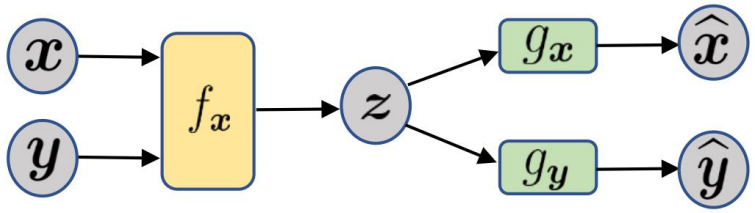


- Encoder decoder are neural networks.
- Not straightforward to do inference under a single modality (partially observed).

Implementation at: https://github.com/shams-sam/MultiModal

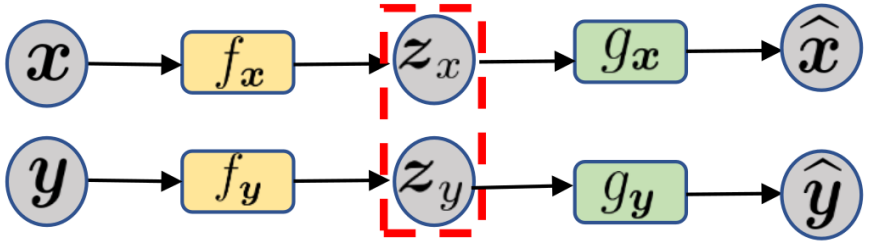
Implementation

Naïve Extension of VAE to Multimodal VAE



- Encoder decoder are neural networks.
- Not straightforward to do inference under a single modality (partially observed).

Multimodal VAE with Factorized Encoders

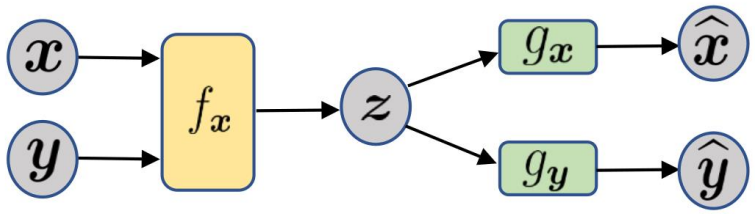


Consensus among distributions

Implementation at: https://github.com/shams-sam/MultiModal

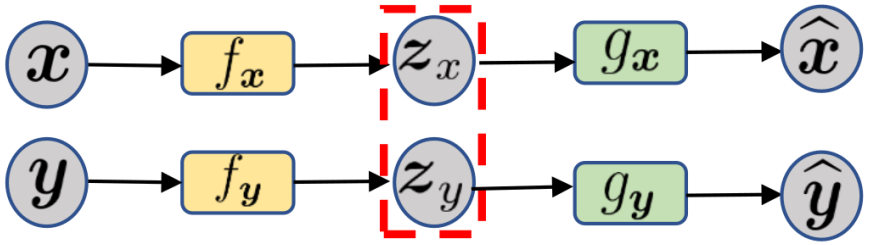
Implementation

Naïve Extension of VAE to Multimodal VAE



- Encoder decoder are neural networks.
- Not straightforward to do inference under a single modality (partially observed).

Multimodal VAE with Factorized Encoders



Consensus among distributions

Wasserstein Distance between 2 Gaussians

$$d^2 = \|\mu_1 - \mu_2\| + \mathrm{Tr}(\Sigma_1 + \Sigma_2 - 2(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2})$$

Implementation at: https://github.com/shams-sam/MultiModal

Results

Results

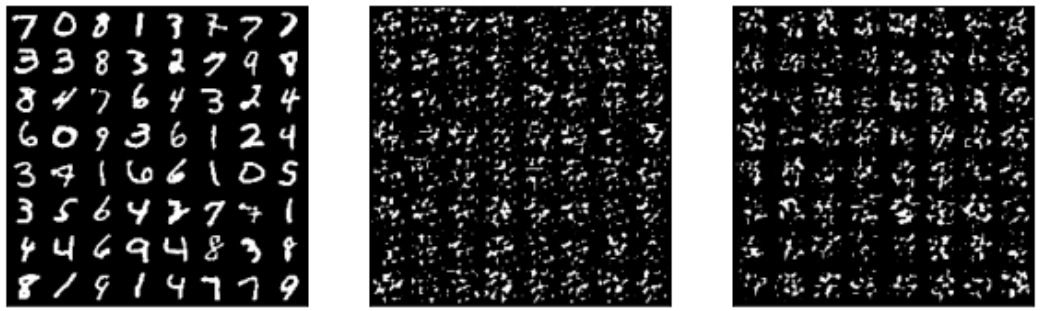
Proof of concept using MNIST

- **Image input is mode 1**
- **Corresponding label is mode 2**
- **Objective is to learn a joint distribution of the two**
- **Evaluation by inspecting the quality of images reconstructed.**

Results

- (a) reconstruction when all inputs are observed.
- (b) reconstruction when only image input is observed.
- (c) reconstruction when only label input is observed.

Baselines



(a) (b) (c)

Samples from MVAE-POE (Multimodal VAE with Product of Experts)

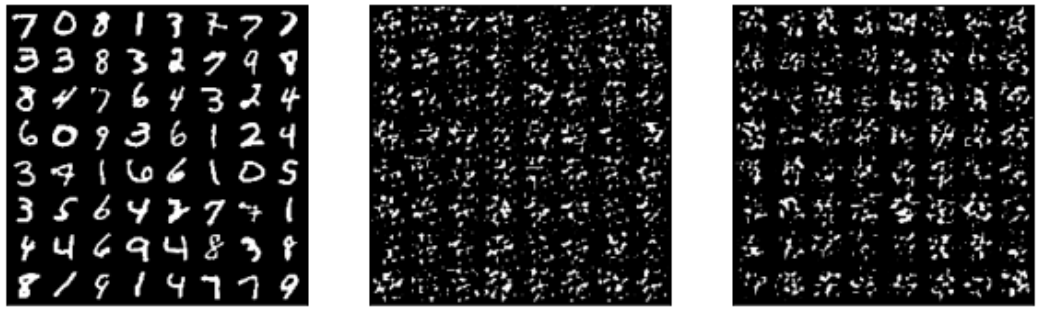
(a) (b) (c)

Samples from MVAE (our architecture without Wasserstein Loss)

Results

- (a) reconstruction when all inputs are observed.
- (b) reconstruction when only image input is observed.
- (c) reconstruction when only label input is observed.

Baselines



(a) (b) (c)

Samples from MVAE-POE (Multimodal VAE with Product of Experts)

(a) (b) (c)

Samples from MVAE (our architecture without Wasserstein Loss)

(a) (b) (c)

Samples from Multimodal Wasserstein VAE

Future Works

Future Works

More than 2 modalities

Wasserstein distance applies to two distributions only. Pairwise application of W-distance to multiple distributions would add combinatorically more terms in the optimization. Studies along Barycenter problem in Optimal transport might be helpful in this case.

- lots of red light and reflections from them tinting the vehicles
- a person on bicycle riding down a street in a busy city.
- a bicyclist rides down a city street at dusk.

Future Works

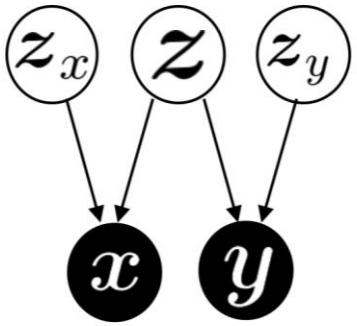
More than 2 modalities

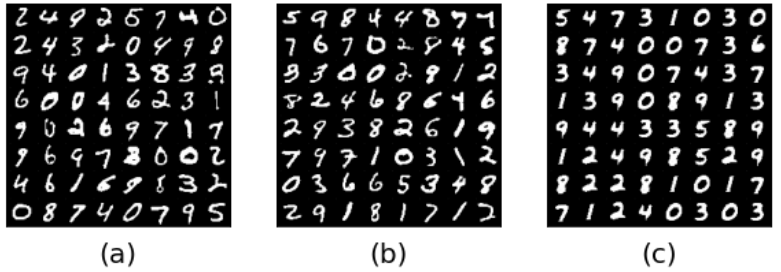
Wasserstein distance applies to two distributions only. Pairwise application of W-distance to multiple distributions would add combinatorically more terms in the optimization. Studies along Barycenter problem in Optimal transport might be helpful in this case.

- lots of red light and reflections from them tinting the vehicles
- a person on bicycle riding down a street in a busy city.
- a bicyclist rides down a city street at dusk.

Factorizing the Latent Code

Observe some correlation between the factors learnt by the Factorized Multimodal Wasserstein VAE. Can we learn factors that are minimally entangled (or correlated). This might be helpful in generating data with fine-grain control.





Samples from Multimodal Wasserstein VAE

Future Works

More than 2 modalities

Wasserstein distance applies to two distributions only. Pairwise application of W-distance to multiple distributions would add combinatorically more terms in the optimization. Studies along Barycenter problem in Optimal transport might be helpful in this case.

- lots of red light and reflections from them tinting the vehicles
- a person on bicycle riding down a street in a busy city.
- a bicyclist rides down a city street at dusk.

Disentangling the Factors

Observe some correlation between the factors learnt by the Factorized Multimodal Wasserstein VAE. Can we learn factors that are minimally entangled (or correlated). This might be helpful in generating data with fine-grain control.

Continuous modeling for Sequences

Continuous modeling of sequence modality with non-sequence modality, e.g., learn a fine-grained alignment of a subset of the caption with part of the image.