# Variational Auto-Encoder for Multimodal Representation Learning

**Sheikh Shams Azam** [1]

## Abstract

In this work we develop a variational multimodal autoencoder to learn multimodal representation. We start by evaluating the common characteristics of existing multimodal architectures that heavily rely on transformers and then design an alternative variational inference based architecture derived from a probabilistic graphical model perspective. Particularly, the variational autoencoder for unimodal datasets is extended to multimodal use-cases by utilizing Wasserstein distance for aligning modalities. We evaluate the ulility of the Wasserstein distance optimization by comparing our model to appropriate baseline. Next, we introduce an inductive bias of factorized priors in the latent space with an aim to relax the alignment restrictions on the latent encodings from different modalities. One of the main advantages of this work is the use of tractable models for encoder, theoretical probabilistic interpretation and robust inference for data when all the modalities aren't observed. We compare our variational autoencoder architecture with prior works and present the comparisons in the experiment section. Finally, we discuss the shortcomings of the current model and the future extensions to the work.

## 1. Introduction

Understanding the latent structure for observed real-world datasets is one of the foundational motivations of generative models and probabilistic graphical models in general. Latent variable density estimation is thus a fundamental problem in generative modeling and therefore implicit and explicit density models such as Gaussian Mixture Models (GMMs) (Reynolds, 2009), Variational Autoencoders (VAE) (Kingma & Welling, 2013), Invertible Normalizing Flows (NFs) (Rezende & Mohamed, 2015), and Generative Adversarial Networks (GANS) (Goodfellow et al., 2014) are widely been studied in field. However, most of the studies involving density estimation focuses on unimodal observations, i.e., there is a single stream of data observed, e.g., images in MS-COCO (Lin et al., 2014). However data observed in real world is multimodal more generally, i.e., same source produces multiple streams of observed data, e.g., a particular scene in MS-COCO can be defined using an image, or one of the several captions, or using the targets that are present in the image. Thus, multimodal representation learning can be seen as a generalized generative modeling wherein we try to model the joint distribution (instead of marginal distribution of a single mode) of various modalities assumed to be generated from the same source. There are several real-world use-cases wherein we observe multimodal data and wish to learn the joint distribution. This may include applications in (i) medicine where we have several patient information that we wish to model to learn an overall representation of patient health, or (ii) a general case wherein we wish to model the image and its corresponding caption using a joint encoded representation.

Since the introduction of transformers in the pioneering paper "Attention is all you need" (Vaswani et al., 2017), a majority of multimodal frameworks revolve around attention-based architectures that use self-supervised learning via auxiliary tasks such as mask prediction, mode alignment, etc. However, transformer based models have several drawbacks such as: (i) extremely large model size, (ii) self-supervised tasks are designed in a heuristic way and are not strictly theoretical, (iii) there are no direct evaluation methods, i.e., it is not possible to evaluate the model merely by observing the performance on the self-supervised task, rather it is done by comparing the performance on downstream task during fine-tuning, and (iv) transformers learn information rich models and less focus lies on learning a rich encoding, i.e., the encodings from transformers are best utilized when using the same transformer for fine-tuning and not the encoded data for transfer learning.

In order to circumvent these drawbacks, we analyze the multimodal representations learning through a different perspective. We shift our focus towards probabilistic graphical models, specifically variational autoencoders (VAE) (Kingma & Welling, 2013) and formulate an extension of the VAE architecture to the multimodal setting. Probabilistic models, specifically VAEs offer several advantages: (i) lightweight architecture, (ii) the unsupervised task is based on variational inference that stems from a concrete theoretical formulation and theoretically guarantees maximization of likelihood, (iii) models can be directly compared to each other without depending on downstream performance during

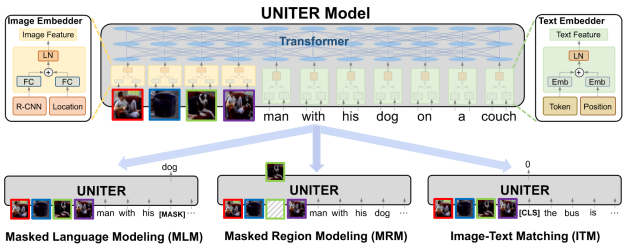


*Figure 1.* An overview of UNITER (Chen et al., 2020) architecture (figure adapted from the original work). Observe that the image feature extractor is a region proposal network (specifically R-CNN) augmented with location features which are required by the transformer to learn positional dependencies of image patches.

finetuning, and (iii) the latent representation is more expressive as the models used for encoding and decoding have limited size making it suitable for dimensionality reduction and/or transfer learning. We further explore architectural design that allows us to do the inference under partially observed input modalities or factorized latent representation to model common and independent factors among modalities as we will present in Section 3.

In Section 2 we review the contributions from prior works in the domain of multimodal learning. Two of the works use transformer-based models for implicit interaction of modalities, while another two give intuition about how to directly model interaction in multimodal settings by encouraging explicit alignment of distributions. In Section 3 we briefly dicuss our formulation, followed by presentation of results and comparisons in Section 4. Finally, Section 5 presents the conclusions and Section 6 discusses the future works.

## 2. Selected Papers Review

In this section, we review 4 of the prior works (2 using transformer-based pre-training and 2 using contrastive learning based pre-training) to draw inspiration from for the design of our model in Sec. 3.

### 2.1. UNITER by Chen et al. (2020)

**Summary.** UNiversal Image-TExt Representation Learning (UNITER) by Chen et al. (2020) is one of the initial works motivated by the BERT-style unsupervised training of transformers using pre-training tasks. One of the key differences that UNITER has with respect to (w.r.t.) existing transformer models is the ability to process multiple modalities, i.e., visual and language inputs and learning a joint representation by designing several multimodal pre-training tasks that promote inter-modal interaction to capture information from all input modalities. The four key pre-training tasks used to train UNITER are: Masked Language Modeling (MLM), Masked Region Modeling (MLM), Image-Text Matching (ITM), and Word-Region Alignement (WRA). The architecture of UNITER can be seen in Fig. 1. The ablation studies show the positive impact of conditional masking as well as WRA with OT on the downstream performances.

**Contributions.** There are 3 key contributions of this work: (i) the authors introduce a conditional multimodal language modeling framework. Unlike joint random masking wherein both the text token and corresponding image patch is masked, conditional masking masks an image/text but conditions it on the corresponding fully observed text/image. (ii) The authors also propose the use of Optimal Transport (OT) for the WRA task as an explicit alignment metric. It is argued that usage of OT as a metric for WRA encourages the learnt representations to have a fine-grained alignment among its word and image region, when compared to other contemporary methods which rely on implicit alignment through other tasks, and (iii) the UNITER architecture sets new state-of-the-art performance by a considerable margin over six of the chosen nine vision+language (V+L) tasks, which includes: Visual Question Answering, Image-Text Retrieval, Referring Expression Comprehension, Visual Commonsense Reasoning, Visual Entailment, and Natural Language for Visual Reasoning (NLVR).

**Weaknesses.** The weakness of this work are mainly in the model evaluation and design choices, specifically the image feature extractor, R-CNN (Girshick et al., 2014) used. Regarding model evaluation, while the authors claim that WRA using OT is one of their main contributions, the benefit it gives over ITM (which also enforces alignment) is not convincingly proven in the experiments section. Table 2 of the Chen et al. (2020) shows marginal improvement using WRA with OT only on a subset of the tasks considered. Regarding image feature extractor, the authors choose a region proposal network, R-CNN for image feature extraction. This choice can lead to 2 major limitations: (i) most R-CNN models are very large and affect the speed of UNITER architecture, (ii) although large, the expressive power of a pre-trained R-CNN optimized to generate bounding boxes is upper bounded by the labelled dataset it was trained on. So the R-CNN (static during training) cannot fully benefit from the unsupervised training on unlabelled datasets during UNITER training.

### 2.2. ViLT by Kim et al. (2021)

**Summary.** Vision-and-Language Transformer (ViLT) by Kim et al. (2021) aims to replace the pretrained region proposal networks used in UNITER-like architectures. Specifically, the authors use a trainable linear projection of flattened patches as input to the transformer encoder as shown in Fig. 2. Authors believe that a light linear projection trained joinly with the overall transformer model would not

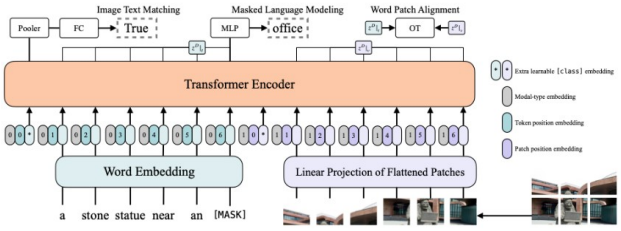


*Figure 2.* An overview of ViLT (Kim et al., 2021) architecture (figure adapted from the original work). The architectures uses a light trainable linear projection instead of heavy R-CNN used in UNITER for image feature extraction. The pre-training tasks are similar to UNITER but ablation studies are more comprehensive.

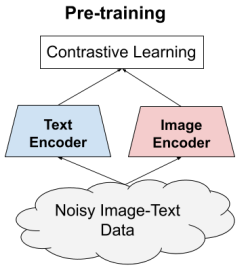


*Figure 3.* An overview of ALIGN (Jia et al., 2021) architecture (figure adapted from the original work). Noisy image-text data in practice has over a billion data points. Text encoder is a BERT-large model and image encoder is an EfficientNet. Unlike most of the contemporaries, ALIGN pretrains using contrastive learning.

only be more efficient in terms of speed but also have more expressive power than a static pre-trained R-CNN as used in UNITER. The ViLT architecture is monolithic, thus removing any error propagation among disjoint models that use pre-trained image embedder. The benefit of a light linear projection instead of R-CNN is evident: authors observe a boost of upto 60× in running time of ViLT over other multimodal architectures (execution time of 15ms for ViLT vs 900ms for UNITER). ViLT architecture uses a light embedder for both the text and visual inputs which are then fed to a transformer encoder for multimodal interaction enforced by several pre-training tasks as shown in Fig. 2.

**Contributions.** The key contribution of ViLT include: (i) the authors propose one of the simplest architectures for multimodal training which can be trained end-to-end and gives significant boost both in runtime speed and parameter reduction. (ii) The authors achieve competent or equivalent performance while reducing the overall size of model and increase runtime speed. (iii) ViLT also introduces a new pre-training tasks called whole word masking instead of using the conventional byte-pair encoded word masking which was originally introduced by Devlin et al. (2019) in BERT.

**Weaknesses.** One of the main limitation of ViLT can be attributed to its performance. While it is not trivial that ViLT has a substantially smaller memory footprint, it comes at the cost of performance of the model. ViLT consistently underperforms other multimodal architectures on several tasks. It is hard to comment on the exact cause of this drawback as the authors do not study the effect different embedders apart from linear projection used in the proposed model. Also, the ablation study shows that the adapted MRM task does not contribute towards downstream performance. This is a unintuitive finding, but the authors do not have any discussion about what might be the reason for such an observation.

### 2.3. ALIGN by Jia et al. (2021)

**Summary.** A Large-scale ImaGe and Noisy-text (ALIGN) model by Jia et al. (2021) tries to overcome another one of the shortcomings of UNITER-like models. As discussed in Sec. 2.1, one of the shortcomings of using a region-proposal network as image feature extractor is the limitation of its expressive power which is determined by the quality of labelled dataset it is trained on. The authors of this work propose a methodology that tries to overcome this issue by designing a learning framework which can be trained to state-of-the-art performance using a noisy labelled training dataset. The process of collecting this noisy dataset is simpler than building a non-trivially curated dataset used to train any of the region proposal networks like R-CNN or YOLO (Redmon et al., 2016). Also, another key difference among ALIGN and other works in studied in Sec. 2.1 & 2.2, is that ALIGN does not depend on a transformer based architecture. Instead, it utilizes contrastive learning (learns to keep similar samples together by pushing dissimilar ones far apart) to align the outputs from two separate encoders encoding image and text each as shown in Fig. 3. Moreover the authors also show that data collection for ALIGN is much simpler because of its ability to work with noisy labelling. The authors curate a dataset with over a billion noisy image-text pair using techniques for partial matching of concepts and filtering stopwords in the text, which is then used for ALIGN pre-training. Learnt representations are claimed to be as good as ones learnt using transformer style pre-training tasks.

**Contributions.** The key contributions of this work include: (i) the authors question the efficiency of networks that rely heavily on curated training dataset, and propose a method of using noisy text-labelled datasets instead. (ii) This work also questions the de-facto approach of transformer based pre-training utilized in multimodal modeling and proposes a contrastive learning architecture as an alter-

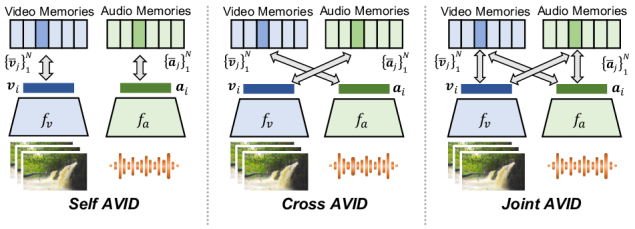


*Figure 4.* An overview of AVID (Morgado et al., 2021) (figure adapted from the original work). The authors study Self (inter-modal NCE) AVID, Cross (cross-modal NCE) AVID, and Joint (Self+Cross) AVID constrastive learning paradigms, and find that Cross-AVID is has the best effect on downstream performance.

native. (iii) The utility of ALIGN is realized by showing its performance on zero-shot image classification as well as setting new state-of-the-art (SOTA) performance on Flickr30k and MSCOCO image-text retrieval benchmarks (even compared to sophisticated transformer models). (iv) Authors also build a large noisy label image-text dataset which can be leveraged for similar training tasks in the future.

**Weaknesses.** The main drawbacks of this work are the missing discussions on reproducibility and speed of training. While the authors mention that one of motivations for ALIGN is the usage of noisy labels for multimodal training, the discussion on the speed of training over the large dataset (over a billion image-text pairs) is missing. Since the pre-training of ALIGN is done on such a large dataset and individual encoders, BERT-large and EfficientNet (Tan & Le, 2019), have millions of parameters, it is hard to ascertain whether the performance gains observed are due to the large dataset and large models or the design choices of the model. Without a comparison study over standard sized dataset among ALIGN and UNITER, no conclusive statements can be made about the effectiveness of contrastive learning pre-training over transformers based pre-training. In terms of nomenclature, the authors claim a contrastive learning approach not dependent on attention based pre-training, but the architecture internally uses BERT (Devlin et al., 2019) for text embedding, which is an transformer based model. The authors say that the model is trained on 1024 TPUv3 cores, with a batch-size of 16384 for 1.2 million steps. Replicating the results of such a large model or building pre-training on top of ALIGN is almost impossible for most researchers due to inaccessibility of such hardware requirements.

### 2.4. AVID by Morgado et al. (2021)

**Summary.** Audio-Visual Instance Discrimination (AVID) by Morgado et al. (2021) studies the video representation learning by leveraging cross-modal agreement among video and audio modalities. The work gives a comprehensive insight into the effect of inter-modal interactions in a multimodal representation learning framework. Unlike most of the other works discussed so far, Morgado et al. (2021) work with audio and video modalities instead of image and text. Nonetheless, the conclusions drawn are pertinent to image and text learning as well. The main task of interest in this work is to learn a representation of a video by using as much input signal as possible from the data stream, i.e., utilizing both video and audio modalities. While the common task used for such a dataset is "in-sync" prediction (given a video clip and audio clip, predict if they are aligned or not), the authors instead use a contrastive learning framework which uses Noise-Contrastive Estimates (NCE) (Gutmann & Hyvärinen, 2010) with negative sampling. Authors argue that a conventional "in-sync" prediction task does not utilize the cross-domain knowledge fully. The authors study effect of NCE for cross-modal, inter-modal and joint-modal (sum of inter and cross-modal) settings as summarized in Fig. 4 and develop AVID.

**Contributions.** The main contributions of this work include: (i) a comprehensive study of different settings under which multimodal contrastive training can be performed and drawing clear conclusion about the utility of cross-modal interactions during NCE calculation. (ii) Authors also analyze the shortcomings of Self-AVID and propose techniques such as cross-modal agreement (CMA) that further improves the performance Cross-AVID. Specifically, CMA uses a within-mode positive discrimination (wMPD) term to calibrate within-mode similarities among positive samples. This is necessary because, it is observed that Cross-AVID, while good at aligning cross-modal instances, fails to capture similarities within a mode. Also, authors find that NCE can suffer from noisy negative sampling and propose measures to overcome which further helps improve the performance.

**Weaknesses.** One of the main weakness of the work lies in the fact that while the modeling and conclusions drawn are more generic and apply to several types of multimodal learning settings (image-text, audio-video, etc.), the authors limit themselves to only audio-video settings during evaluation. The authors begin by saying wMPD alleviates the issues observed in Self-AVID and use it to improve performance in Cross-AVID learning, however the discussion on effect of wMPD on Joint-AVID is missing, and experiments evaluating those would be insightful in understanding the overall impact of wMPD.

## 3. Multimodal Wasserstein-VAE

### 3.1. Motivation

As mentioned in the Sec. 1, one of the main objectives of this work is to come up with a lightweight architecture for

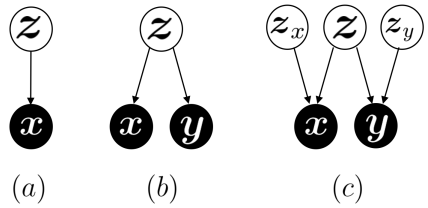


*Figure 5.* Graphical model for (a) unimodal latent variable model for VAE, and it's (b) extension to a multimodal latent variable model for M-VAE. While, in (a) VAE tries to approximate the latent variable $\boldsymbol{z}$ for observed variable $\boldsymbol{x}$, in (b), M-VAE tries to find the latent variable $\boldsymbol{z}$ for both observed variables $\boldsymbol{x}$ and $\boldsymbol{y}$. (c) shows the factorized model of the latent representation wherein there must be an overlap among information in $\boldsymbol{z}$ but each modality might also have independent factors given by $\boldsymbol{z}_x$ and $\boldsymbol{z}_y$.

multimodal representation learning. The works discussed in Sec. 2 focus on solving multimodal representation learning via self-supervised objectives using transformer-based architectures or contrastive loss for intermodal interaction. However, the models utilized for these architectures are generally very large with several million parameters and take very long to train. Given the limited GPU available to us, we take a step back to see if we can come up with a lightweight architecture. We start by evaluating if existing work in unimodal unsupervised learning, specifically variational autoencoder (VAE) (Kingma & Welling, 2013) can be extended to multimodal setting, i.e., a multimodal variational autoencoder (M-VAE). The motivations for using this methodology are:

- *Lightweight Architecture:* It is possible to ensure small model size and faster training time in VAE.
- *Theoretically Motivated:* The underlying logic for multimodal VAE is grounded in the concepts of probabilistic graphical models, variational inference. The optimization technique is based on expectation maximization and thus has stict theoretical guarantees in terms of optimizing the complete log-likelihood.
- *Direct Comparison:* Different methodologies based on explicit generative models can be directly compared using the likelihood scores. We can also do a qualitative comparison of encoded representation by observing samples generated through reconstruction.

### 3.2. Formulation

We start with the analyzing the unimodal VAE. Intuitively the graphical model in Fig. 5(a) implies that we observe data $\boldsymbol{x}$ and wish to find a latent variable $\boldsymbol{z}$ that serves as the reason or cause for generation of corresponding $\boldsymbol{x}$. Using the principles of probabilistic graphical models, the joint distribution of the random variables (r.v.) in the graph is given by $p(\boldsymbol{x}, \boldsymbol{z}) = p(\boldsymbol{z})p(\boldsymbol{x}|\boldsymbol{z})$. Consequently, the likelihood function of observed variable $\boldsymbol{x}$ parameterized by $\boldsymbol{\theta}$ can be expressed as: $p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{\boldsymbol{z}} p(\boldsymbol{z})p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})$. Extending this to the multimodal setting, we get the graphical model in Fig. 5(b), where the latent variable $\boldsymbol{z}$ is the common cause for both the observed modes $\boldsymbol{x}$ and $\boldsymbol{y}$ of the data. The joint distribution of variables $\boldsymbol{x}, \boldsymbol{y}$ and $\boldsymbol{z}$ is given by $p(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) = p(\boldsymbol{z})p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{y}|\boldsymbol{z})$ and the corresponding joint likelihood function is given by

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{y}) = \sum_{\boldsymbol{z}} p(\boldsymbol{z})p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})p_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{z}). \tag{1}$$

Similar to unimodal VAE, for multimodal VAE (M-VAE) the final objective is to maximize the likelihood function over observed data given by,

$$\begin{aligned} p_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{y}) &= \prod_i p_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}) \\ &= \prod_i \sum_{\boldsymbol{z}} p(\boldsymbol{z})p_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}|\boldsymbol{z})p_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{z}), \end{aligned} \tag{2}$$

where $i \in \{1, \cdots, N\}$ indexes over the data samples $\left\{(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)})\right\}_{i=1}^{N}$, and $N$ is the number of data samples observed. The probabilities in (2) are multiplied because data is assumed to be i.i.d distributed.

A good candidate for the prior $p(\boldsymbol{z})$ would be the conditional distribution $p(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{y})$ obtained using the Bayes rule,

$$p(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{y}) = \frac{p(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})}{p(\boldsymbol{x}, \boldsymbol{y})} \overset{(i)}{=} \frac{p(\boldsymbol{z})p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{y}|\boldsymbol{z})}{p(\boldsymbol{x}, \boldsymbol{y})}. \tag{3}$$

where $(i)$ follows from the graphical model in Fig. 5(b).

While we can evaluate $p(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{y})$, for a given $z$ sample, it is not clear how to sample $\boldsymbol{z}$ from the distribution. Hence, we use the variational inference to find a distribution $q_{\boldsymbol{\phi}}(\boldsymbol{z})$ parameterized by $\boldsymbol{\phi}$ with the properties: (i) easy to sample from $q_{\boldsymbol{\phi}}(\boldsymbol{z})$, and (ii) approximates the original $p(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{y})$ well, i.e., $q_{\boldsymbol{\phi}}(\boldsymbol{z}) \approx p(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{y})$. We tackle (i) by choosing Gaussian as the distribution parameterized by mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\sigma}$ and can achieve (ii) by minimizing the KL divergence among the distributions $q_{\boldsymbol{\phi}}(\boldsymbol{z})$ and $p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})$, where the KL divergence is given by,

$$\begin{aligned} \mathsf{KL}\left(q_{\boldsymbol{\phi}}(\boldsymbol{z}) \| p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{y})\right) &= \mathbb{E}_{\boldsymbol{z}\sim q_{\boldsymbol{\phi}}(\boldsymbol{z})}\left[\log \frac{q_{\boldsymbol{\phi}}(\boldsymbol{z})}{p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{y})}\right] \\ &\overset{(i)}{=} \mathbb{E}_{\boldsymbol{z}\sim q_{\boldsymbol{\phi}}(\boldsymbol{z})}\left[\log \frac{q_{\boldsymbol{\phi}}(\boldsymbol{z})}{p(\boldsymbol{z})p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{y}|\boldsymbol{z})/p(\boldsymbol{x}, \boldsymbol{y})}\right] \\ &= \mathbb{E}\Big[\log q_{\boldsymbol{\phi}}(\boldsymbol{z}) - \log p(\boldsymbol{z}) - \log p(\boldsymbol{x}|\boldsymbol{z}) \\ &\qquad - \log p(\boldsymbol{y}|\boldsymbol{z}) + \log p(\boldsymbol{x}, \boldsymbol{y})\Big] \\ &\overset{(ii)}{=} \mathbb{E}\Big[\log q_{\boldsymbol{\phi}}(\boldsymbol{z}) - \log p(\boldsymbol{z}) - \log p(\boldsymbol{x}|\boldsymbol{z}) \\ &\qquad - \log p(\boldsymbol{y}|\boldsymbol{z})\Big] + \log p(\boldsymbol{x}, \boldsymbol{y}), \end{aligned} \tag{4}$$

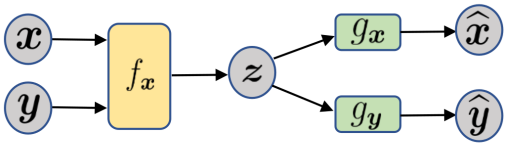


*Figure 6.* Architectural design of multimodal variational autoencoder (M-VAE). $\boldsymbol{z}$ is sampled from Gaussian distribution as discussed in Sec. 3.2. The optimization w.r.t. KL divergence term in (4) is done by the encoder $f_x$ and the maximization of posteriors $p(\boldsymbol{x}|\boldsymbol{z})$ and $p(\boldsymbol{y}|\boldsymbol{z})$ is done by reducing the reconstruction loss among original inputs $(\boldsymbol{x}, \boldsymbol{y})$ and $(\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{y}})$.

where $(i)$ follows from (3) and $(ii)$ uses the fact that term $p(\boldsymbol{x}, \boldsymbol{y})$ is independent of $\boldsymbol{\phi}$. Consequently, the (4) can we rewritten as,

$$\log p(\boldsymbol{x}, \boldsymbol{y}) = \mathbb{E}\Big[\log p(\boldsymbol{z}) + \log p(\boldsymbol{x}|\boldsymbol{z}) + \log p(\boldsymbol{y}|\boldsymbol{z}) - \log q_{\boldsymbol{\phi}}(\boldsymbol{z})\Big] + \mathsf{KL}\left(q_{\boldsymbol{\phi}}(\boldsymbol{z}) \,\|p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{y})\right) \quad (5)$$

where the left-hand side term in the log-likelihood function we wish to minimize, the first term under expectation on right-hand side, generally known as the evidence lower-bound (ELBO), is a lower-bound on the complete joint log-likelihood $\log p(\boldsymbol{x}, \boldsymbol{y})$, and $D_{\mathsf{KL}}(\cdot)$ can be interpreted as regularization and its value gives the quality of lower bound.

Please note that (4) can be generalized to more than 2 modalities. In the following subsection, we discuss how can we realize the additional terms added by generalization of VAE to M-VAE during the implementation.

### 3.3. Implementing Multimodal VAE (M-VAE)

If we compare the (4) with the unimodal VAE, it can be observed that addition of new mode $\boldsymbol{y}$ leads to the term $p(\boldsymbol{y}|\boldsymbol{z})$. Intuitively it means that we are additionally trying to maximize the expectation of recovering $\boldsymbol{y}$ given the latent variable $\boldsymbol{z}$. This is realized in the architecture of multimodal VAE (M-VAE; see Fig. 6) by making the following changes:

- add the mode $\boldsymbol{y}$ as an input to the encoder $f_x$ to learn the common cause latent variable $\boldsymbol{z}$. Encoder $E$ minimizes the KL divergence.
- add another decoder $g_y$, that tries to reconstruct the input $\boldsymbol{y}$ similar to the decoder $g_x$ reconstructing $\boldsymbol{x}$.

We evaluate the M-VAE model in Section. 4.

### 3.4. Alignment using Wasserstein Distance

We next turn to a practical constraint that is encountered while collecting multimodal data. Often the collected data is far from perfect and has unobserved modalities, e.g., there might be an image without its caption, or a caption with corresponding image missing or corrupted in storage. We thus propose an adaptation of our M-VAE architecture by factorizing the encoder network $q_\phi(z|x, y)$ into individual components $q_{\phi_x}(z|x)$ and $q_{\phi_y}(z|y)$, i.e. we wish to arrive at the latent representation $z$ even if only one of the two is modalities is fed into the corresponding encoder (see Fig. 7). However we would want the representations $q_{\phi_x}(z|x)$ and $q_{\phi_y}(z|y)$ to be aligned or as close to each other as possible if corresponding $x$ and $y$ come from the same source, hence we have the "consensus among distribution" module in the architecture shown in Fig. 6. This can be achieved by using an alignment loss during the training. We achieve this by using Wasserstein distance (Villani, 2009) among two distributions since the Wasserstein distance between two Gaussian distributions can be calculated in closed form using the mean and variance formula:

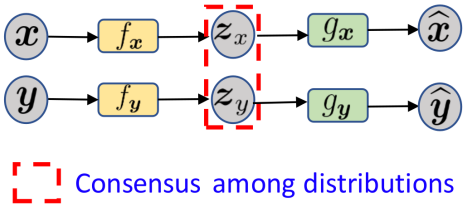


*Figure 7.* Architectural design of multimodal Wasserstein variational autoencoder (MW-VAE). $\boldsymbol{z}_x$ and $\boldsymbol{z}_y$ are pushed together using Wasserstein distance as shown in (6). The optimization w.r.t. KL divergence term and reconstruction loss are same as explained in M-VAE. At the end of the training either of $\boldsymbol{z}_x$ or $\boldsymbol{z}_y$ can be used for reconstructing both $\widehat{\boldsymbol{x}}$ and $\widehat{\boldsymbol{x}}$ (corroborated experimentally in Section 4).

$$d^2 = \|\mu_1 - \mu_2\| + \mathrm{Tr}(\Sigma_1 + \Sigma_2 - 2(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2}) \quad (6)$$

where $\mu_1$ and $\mu_2$ are means of the two Gaussians and $\Sigma_1$ and $\Sigma_2$ are their corresponding variances. The fact that variances in VAEs are diagonal further help the computation of second term in (6).

## 4. Experiments

### 4.1. Setup

**Datasets.** In order to test our hypothesis and the designed architecture, we use MNIST (Deng, 2012) dataset. Conventionally MNIST is considered to be a unimodal dataset, however, we consider it as a multimodal dataset for the purpose of hypothesis testing by considering labels as an additional mode, i.e., the image is mapped as mode $\boldsymbol{x}$ (or $\boldsymbol{x}_1$) and the label is mapped as mode $\boldsymbol{y}$ (or $\boldsymbol{x}_2$). We also demonstrate an experiment with MS-COCO(Lin et al., 2014), where image is considered as modality $\boldsymbol{x}$ and the captions are embedded using BERT (Devlin et al., 2019) to be used as mode $\boldsymbol{y}$.

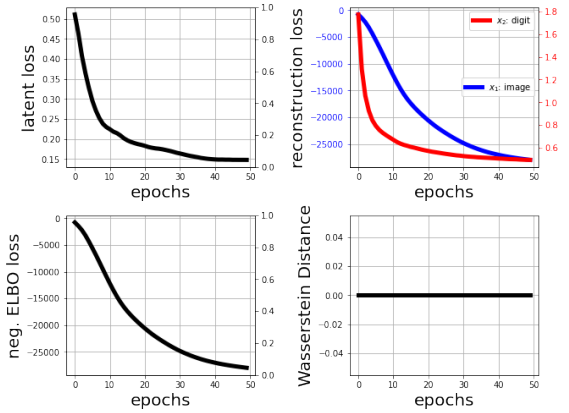


*Figure 8.* Training curves for mVAE-POE. There is a single latent curve since there is single latent code recovered from POE. The resulting code is used for reconstructing both $x_1$ and $x_2$ and thus the 2 reconstruction curves. Finally ELBO loss is the sum of the latent loss and reconstruction losses.


*Figure 9.* Reconstructed samples using mVAE-POE when (a) both the modes are fed to the encoder (corresponding labels generated for line 1 of image: $[7, 0, 3, 1, 3, 3, 7, 2]$), (b) only image is fed to the encoder (corresponding labels generated for line 1 of image: $[7, 7, 7, 7, 7, 7, 7, 7]$), and (c) only labels are fed to the encoder (corresponding labels generated for line 1 of image: $[2, 0, 4, 0, 7, 7, 5, 9]$). It can be seen that under unobserved modes the reconstructed samples are not meaningful implying that the representations of image and corresponding labels aren't aligned.

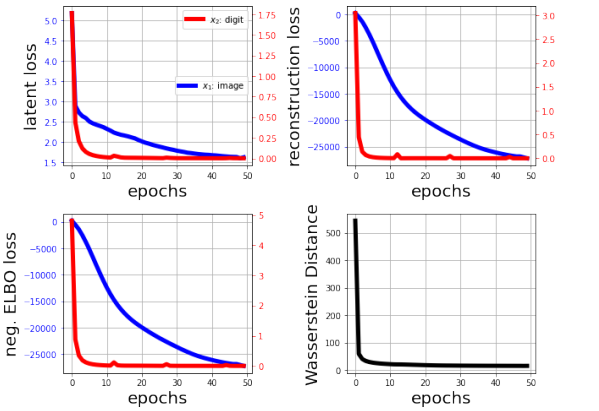


*Figure 10.* Training curves for M-VAE. We observe that the losses w.r.t. digits drop faster than image. This implies that learning the latent representation of digit labels is easier than corresponding images. We also plot the Wasserstein distance among the encoder distributions but it is not used for the optimization process

**Evaluation.** We consider 2 different cases of inference: (i) under fully observed modes, and (ii) under partially observed modes. In the former we feed both the image and label to the encoders and reconstruct them using the latent code, while in the latter, we observe the reconstruction of both the modes using the latent code obtained from encoding only one of the two modes. The models are evaluated in terms of the learnt joint representation. The efficiency of the learnt distribution can be seen directly through the inference under partially observed modalities. We show this in each case by showing a random subset of reconstructed images and labels from the latent codes.

### 4.2. Baselines

**Product of Experts.** We first start with one of the prior works that implemented a multimodal VAE using product of experts (mVAE-POE) (Wu & Goodman, 2018). This architecture considers multiple encoders like our architecture in 7. The latent representations in mVAE-POE are directly multiplied (hence product of experts) since the two distributions are independent under observed datasets. The training curves of mVAE-POE and corresponding samples generated can be seen in Fig. 8 and Fig. 9. It can be seen that the model fails to generalize to a case where all the modes are not observed, hence not learning a meaningful joint representation under partial observation of modes.

**Multimodal VAE without Alignment.** We next consider the case where encoder is factorized as in our architecture shown in Fig. 12, but without the Wasserstein distance optimization. We present the corresponding training curves and samples in Fig. 10 and Fig. 11. We first observe that the losses w.r.t digits drop faster than the ones for images. This implies it is easier for the encoder to find a efficient


*Figure 11.* Reconstructed samples using M-VAE when (a) both the modes are fed to the encoder (corresponding labels generated for line 1 of image: $[0, 5, 4, 2, 4, 1, 8, 6]$), (b) only image is fed to the encoder (corresponding labels generated for line 1 of image are not aligned: $[9, 0, 6, 0, 3, 6, 6, 3]$), and (c) only labels are fed to the encoder (corresponding labels generated for line 1 of image: $[9, 5, 0, 9, 1, 9, 5, 2]$). It can be seen that when image modality is unobserved in (c) reconstructed samples are not meaningful.

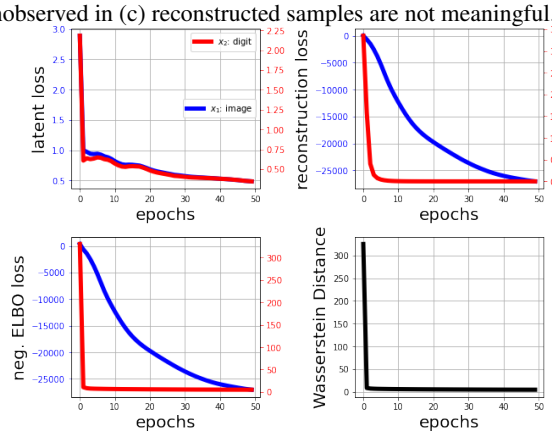


*Figure 12.* Training curves for MW-VAE. We observe that the losses w.r.t. digits drop faster than image in all cases but the latent loss. This is because the latent loss of digit labels now also needs to consider its alignment w.r.t. its corresponding image representation. Also note that the Wasserstein distance (used for optimization) among the encoder distributions converges to a near-zero value (in contrast to Fig. 10 where it converges to a non-zero constant.)

latent representation for the labels. Even though we do not optimize w.r.t the Wasserstein distance, we observe it steadily drops and saturates at a non-zero value as we go through the training. Since the Wasserstein distance is not close to 0 we can conclude that the distributions are not closely aligned. This is expected since we do not explicitly model the alignment loss in M-VAE. The effect of misalignment is also observed in the reconstructed samples shown in Fig. 9(c), i.e., the distribution of only labels is not aligned for reconstruction of images. A similar effect is observed in Fig. 9(b) wherein only the image mode is observed. Here, while the image reconstruction is valid, it is observed that the corresponding labels generated are incorrect.

### 4.3. Multimodal Wasserstein VAE

We next present results from our main architecture, MW-VAE depicted in Fig. 7. The corresponding training curves can be see in Fig. 12. It can again be observed that losses for digits drop faster in all cases except the latent loss. This is


*Figure 13.* Reconstructed samples using MW-VAE when (a) both the modes are fed to the encoder (corresponding labels generated for line 1 of image: $[2, 4, 9, 2, 5, 7, 4, 0]$), (b) only image is fed to the encoder (corresponding labels generated for line 1 of image: $[5, 9, 8, 9, 4, 3, 7, 3]$), and (c) only labels are fed to the encoder (corresponding labels generated for line 1 of image: $[5, 4, 7, 3, 1, 0, 3, 0]$). It can be seen that reconstructed samples are meaningful and aligned with generated labels irrespective of partial observation of input modes implying that the representations of image and corresponding labels are aligned. Observe that while (a)&(b) have handwriting styles (c) does not. This might be because (c) is generated by using labels only (and no images) and hence does not provide information about the handwriting.


*Figure 14.* Reconstructed samples using MW-VAE for MS-COCO dataset when (a) both the modes are fed to the encoder (corresponding labels generated for line 1 of image: $[2, 4, 9, 2, 5, 7, 4, 0]$), (b) only image is fed to the encoder (corresponding labels generated for line 1 of image: $[5, 9, 8, 9, 4, 3, 7, 3]$), and (c) only labels are fed to the encoder (corresponding labels generated for line 1 of image: $[5, 4, 7, 3, 1, 0, 3, 0]$).

because under the MW-VAE formulation the digit encoder not only needs to encode its mode (i.e., digit labels) but also needs to align its representation with the corresponding image representation. It can also be seen that the Wasserstein distance falls to a lower value (since we are optimizing for it) than the M-VAE case in Fig. 10.

The effect of alignment using the optimization w.r.t Wasserstein distance can also be observed in the samples generated (shown in Fig. 13). It can be observed that the model is able to reconstruct meaningful images under all 3 scenarios: (i) both modes are observed, (ii) only image mode is observed, and (iii) only label mode is observed. Also note that while in Fig. 13(a)&(b) images have handwritten style, the images in Fig. 13(c) seem to be lacking such styles. We hypothesize this is a result of the fact that digit label alone as a modality cannot give enough information to the decoder $g_x$ about the writing style of a given sample, i.e., it only contains information w.r.t digit label but not handwriting style.

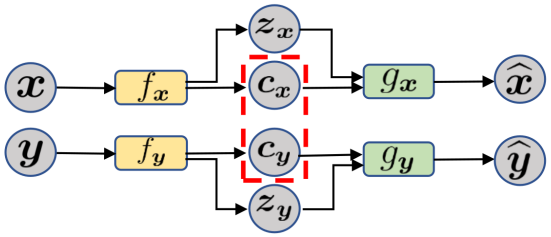


*Figure 15.* Architecture of factorized multimodal Wasserstein variational autoencoder (FMW-VAE). Only $\boldsymbol{c}_x$ and $\boldsymbol{c}_y$ are pushed together using Wasserstein distance (shown in red dashed box).


*Figure 16.* Reconstructed samples using FMW-VAE when (a) both the modes are fed to the encoder (corresponding labels generated for line 1 of image: $[8, 0, 1, 7, 0, 5, 9, 3]$), (b) only image is fed to the encoder (corresponding labels generated for line 1 of image: $[4, 6, 2, 1, 0, 7, 9, 7]$), and (c) only labels are fed to the encoder (corresponding labels generated for line 1 of image: $[4, 8, 6, 1, 0, 3, 7, 0]$). It can be seen that reconstructed samples are meaningful and aligned irrespective of partial observation of input modes implying that the representations of image and corresponding labels are aligned. Observe that (c) also has handwriting style here.

We also repeat this experiment using MS-COCO dataset. Here, the image is considered as mode $\boldsymbol{x}$ and the corresponding captions as $\boldsymbol{y}$. In order to feed the caption into the MW-VAE, we first get a dense representation for the caption by passing it through BERT (Devlin et al., 2019). The recovered BERT-embedding for the caption is then used as the mode $\boldsymbol{y}$. The results corresponding to COCO shown in Fig. 14 are not as representative because of several reasons: (i) the image sizes were downsampled to $32 \times 32$ to fit the GPU requirements, (ii) latent code size is only 256 which is very low and the model sizes for encoder decoder are not very large. However, refining this results is left as a future extension of this work.

### 4.4. Factorizing the Priors in MW-VAE (FMW-VAE)

Seeing the effect of alignment in MW-VAE, we investigate if it's possible to learn a factorized representation in the latent space such that we can add handwriting information to the label information even when the corresponding images are missing. To do this we start by factorizing our priors distribution as shown in Fig. 5(c). We assume that each modality has some information in common with other modalities but also might have independent information. This is realized in the modeling (see Fig. 15) by ensuring that each of our encoders give us two factorized distributions and we enforce Wasserstein distance only on the common latent variable. The samples generated by Factorized MW-VAE are shown in Fig. 16. We generate the handwriting latent code by averaging the latent representations for each digit from the training dataset. This handwriting latent code is then added to the label latent code to generate the samples in Fig. 16(c). Comparing the samples with Fig. 13(c), we can see the difference when handwriting styles are added to the reconstruction. However the current methodology shows some drawbacks. The reconstructions are not perfect when the handwriting code for one digit is applied to another. This points to some possible correlation between the learnt factorized representations. One possible way of solving this could be enforcing a direct disentanglement using auxilliary loss such as minimizing the maximum mean discrepancy (MMD) (Gretton et al., 2006). We defer this to a future work.

## 5. Conclusion

In the paper we started by studying different methodologies that are currently used for multimodal training. We observed that a majority of the solutions lie in the space of transformer-like architectures. We also noted that transformers pose some fundamental limitations in terms of multimodal learning. We further observe that alignment of multiple distributions is possible post-encoding as well using different loss functions such as NCE. However, since all the reviewed methods depend on large models such as BERT, Transformers, or R-CNN, etc, we shift our focus to develop smaller model with more theoretical flavour. This benefits us since we have limitations w.r.t GPU memory. It is also of utility since lightweight representation alignment is faster and environment-friendly. We thus propose an architecture that uses lightweight encoder decoder models and can be independently evaluated both qualitatively using reconstructed samples and quantitatively using log-likelihood scores. Our model, MW-VAE factorizes the encoder for different modalities and aligns the corresponding distributions using Wasserstein distance optimization. We show that our model is better than prior work on multimodal VAE (i.e., mVAE-POE) or a generic joint distribution learner (i.e., M-VAE). We also show the importance of Wasserstein distance optimization by comparing MW-VAE with M-VAE. The factorization of encoders in MW-VAE further helps us do inference under partially observed modes in datasets. This is of particular importance because often data in real world has missing modalities. We also made observations about the nature of information stored in individual modes of MW-VAE, i.e., digit modality does not give information about handwriting style to the decoder. To address this we introduce an inductive bias wherein the latent priors are factorized to facilitate a looser alignment, i.e., the alignment constraint is applied to only a subset of latent factors. Following this we

show the Factorized Multimoald Wasserstein-VAE and its benefits. We notice that while the latent code is factorized it might still encode correlation as explained in Section 4.4. We pose the problem of learning disentangled latent factors such that the correlation between factorized priors is minimized. We leave this as a future extension of this work. In the following section we discuss some of the other possible future directions to this work.

## 6. Future Work

### 6.1. Extension to More than 2 Modes

From the formulation of MW-VAE, we see that Wasserstein distance is an important part of the optimization. However, extension to more than two modes will add combinatorially more terms w.r.t pairwise Wasserstein distance. This is not a scalable solution as the number of modes increase. There is a possibility of circumventing this issue by exploring Wasserstein barycenter problem.

### 6.2. Disentangled Latent Representation

From Fig. 13 we observed that some information such as handwriting style cannot be encoded in label representation. We try to solve this by factorizing the latent and enforcing alignment only on the common factors. However we notice there might be some correlation among the factors and thus propose that use of auxiliary losses such as MMD could be explored to further enforce disentanglement among the factors.

### 6.3. Sequence Modeling and Alignment

In our proof of concept we consider a dataset wherein both image and corresponding label are realized as a single input (even the caption is encoded as a dense representation output from BERT). However a more flexible model could process the caption (sentence) sequence of inputs as a modality. There are several methods proposed for modeling a sequence such as Recurrent Neural Network VAE (RNN-VAE) (Fabius et al., 2015) wherein the hidden state of RNN at the end of the sequence is used as the input to the VAE. It might be possible to combine our MW-VAE with RNN-VAE such that the sequence is encoded with RNN and then the corresponding latent distribution is aligned with the image representation using MW-VAE. If there are multiple sequences (e.g., image with multiple captions as in MS-COCO (Lin et al., 2014)) however we would need to explore these in combination with Barycenter problem mentioned in previous subsection.

### 6.4. Continuous modeling of Sequence

Another interesting modeling idea is to consider a continuous alignment of sequence (e.g., image caption) with non-sequence (e.g., image). It might be possible to model a continuous alignment of caption with the image. For example, a subset of sequence aligns well with a fraction of the image. The question arises if we can extend our multimodal formulation to model the granular dependency among different parts of the caption with different parts of the image. This is a non-trivial extension.

## References


Chen, Y.-C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., and Liu, J. **UNITER : Universal Image-text Representation Learning**. In *European Conference on Computer Vision (ECCV)*, pp. 104–120. Springer, 2020. URL https://www.ecva.net/papers/eccv_2020/papers_ECCV/papers/123750103.pdf.

Deng, L. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics (NAACL)*, 2019. URL https://aclanthology.org/N19-1423.pdf.

Fabius, O., van Amersfoort, J. R., and Kingma, D. P. Variational recurrent auto-encoders. In *ICLR (Workshop)*, 2015.

Girshick, R., Donahue, J., Darrell, T., and Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Computer Vision and Pattern Recognition (CVPR)*, pp. 580–587, 2014. URL https://www.cv-foundation.org/openaccess/content_cvpr_2014/papers/Girshick_Rich_Feature_Hierarchies_2014_CVPR_paper.pdf.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., and Smola, A. A kernel method for the two-sample-problem. *Advances in neural information processing systems*, 19: 513–520, 2006.

Gutmann, M. and Hyvärinen, A. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. pp. 297–304. JMLR Workshop and Conference Proceedings, 2010. URL https://proceedings.mlr.press/v9/gutmann10a/gutmann10a.pdf.

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., and Duerig, T. **Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision**. In *International Conference on Machine Learning (ICML)*, 2021. URL http://proceedings.mlr.press/v139/jia21b/jia21b.pdf.

Kim, W., Son, B., and Kim, I. **ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision**. In *International Conference on Machine Learning (ICML)*, 2021. URL https://arxiv.org/pdf/2102.03334.pdf.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. URL https://arxiv.org/pdf/1312.6114.pdf.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Morgado, P., Vasconcelos, N., and Misra, I. **Audio-Visual Instance Discrimination with Cross-Modal Agreement**. In *Computer Vision and Pattern Recognition (CVPR)*, pp. 12475–12486, June 2021. URL https://openaccess.thecvf.com/content/CVPR2021/papers/Morgado_Audio-Visual_Instance_Discrimination_with_Cross-Modal_Agreement_CVPR_2021_paper.pdf.

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. You only look once: Unified, real-time object detection. In *Computer Vision and Pattern Recognition (CVPR)*, pp. 779–788, 2016. URL https://www.cv-foundation.org/openaccess/content_cvpr_2016/papers/Redmon_You_Only_Look_CVPR_2016_paper.pdf.

Reynolds, D. A. Gaussian mixture models. *Encyclopedia of biometrics*, 741:659–663, 2009.

Rezende, D. and Mohamed, S. Variational inference with normalizing flows. In Bach, F. and Blei, D. (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1530–1538, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/rezende15.html.

Tan, M. and Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning (ICML)*, pp. 6105–6114. PMLR, 2019. URL https://proceedings.mlr.press/v97/tan19a/tan19a.pdf.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. URL https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Villani, C. The wasserstein distances. In *Optimal Transport*, pp. 93–111. Springer, 2009.

Wu, M. and Goodman, N. Multimodal generative models for scalable weakly-supervised learning. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 5580–5590, 2018.