Relating by Contrasting: A Data-efficient Framework for Multimodal Generative Models

Yuge Shi  
Department of Engineering Science  
University of Oxford  
yshi@robots.ox.ac.uk

Brooks Paige  
University College London &  
Alan Turing Institute  
b.paige@ucl.ac.uk

Philip H.S. Torr  
Department of Engineering Science  
University of Oxford  
philip.torr@eng.ox.ac.uk

N. Siddharth  
Department of Engineering Science  
University of Oxford &  
Alan Turing Institute  
nsid@robots.ox.ac.uk

Abstract

Multimodal learning for generative models often refers to the learning of abstract concepts from the commonality of information in multiple modalities, such as vision and language. While it has proven effective for learning generalisable representations, the training of such models often requires a large amount of “related” multimodal data that shares commonality, which can be expensive to come by. To mitigate this, we develop a novel contrastive framework for generative model learning, allowing us to train the model not just by the commonality between modalities, but by the distinction between “related” and “unrelated” multimodal data. We show in experiments that our method enables data-efficient multimodal learning on challenging datasets for various multimodal variational autoencoder (VAE) models. We also show that under our proposed framework, the generative model can accurately identify related samples from unrelated ones, making it possible to make use of the plentiful unlabeled, unpaired multimodal data.

1 Introduction

To comprehensively describe concepts in the real world, humans collect multiple cognitive signals of the same object such as image, sound, text and video. We refer to each of these media as a modality, and a collection of different media featuring the same underlying concept is characterised as multimodal data. Learning from multiple modalities has been shown to yield more generalisable representations (Guo et al., 2019; Yildirim, 2014; Zhang et al., 2020), as different modalities are often complimentary in content while overlapping for common abstract concept.

Despite the motivation, it is worth noting that the multimodal framework is not exactly data-efficient—constructing a suitable dataset requires a lot of “annotated” unimodal data, as we need to ensure that each multimodal pair is related in a meaningful way. The situation is worse when we consider more complicated multimodal settings such as language–vision, where one-to-one or one-to-many correspondence between instances of the two datasets are required, due to the difficulty in categorising data such that commonality amongst samples is preserved within categories. See Figure 1 for an example from the CUB dataset (Welinder et al., 2010a); although the same species of bird is featured in both image-caption pairs, their content differs considerably.
It would be unreasonable to apply the caption from one to describe the bird depicted in the other, necessitating one-to-one correspondence between images and captions.

However, the scope of multimodal learning has been limited to leveraging the commonality between “related” pairs, while largely ignoring “unrelated” samples potentially available in any multimodal dataset—constructed through random pairing between modalities [Figure 3]. We posit that if a distinction can be established between the “related” and “unrelated” observations within a multimodal dataset, we could greatly reduce the amount of related data required for effective learning.

While different forms of dependencies on $r$ can be taken, here we make the simplifying assumption that the Pointwise Mutual Information (PMI) between $x$ and $y$ should be high when $r = 1$, and low when $r = 0$. Intuitively, this can be achieved by adopting a max-margin metric. We therefore propose to train the generative models with a novel contrastive-style loss (Hadsell et al., 2006; Weinberger and Saul, 2009), and demonstrate the effectiveness of our proposed method from a few different perspectives: Improved multimodal learning: showing improved multimodal learning for various state-of-the-art multimodal generative models on two challenging multimodal datasets. This is evaluated on four different metrics (Shi et al., 2019) summarised in § 4.2. Data efficiency: learning generative models under the contrastive framework requires only 20% of the data needed in baseline methods to achieve similar performance—holding true across different models, datasets and metrics; Label propagation: the contrastive loss encourages a larger discrepancy between related and unrelated data, making it possible to directly identify related samples using the PMI between observations. We show that these data pairs can be used to further improve the learning of the generative model.

2 Related Work

Contrastive loss Our work aims to encourage data-efficient multimodal generative-model learning using a popular representation learning metric—contrastive loss (Hadsell et al., 2006; Weinberger and Saul, 2009). Although learning robust generalisable representations have been an important desideratum in generative-model learning (Bengio et al., 2013; Tenenbaum and Freeman, 2000), most applications of contrastive methods skip the generative component, and directly learn representations for specific tasks, as seen in contrastive predictive coding for time series data (van den Oord et al., 2018), image classification (Hénaff et al., 2019), noise contrastive estimation for vector embeddings of words (Mnih and Kavukcuoglu, 2013), as well as a range of frameworks such as DIM (Hjelm et al., 2018), MoCo (He et al., 2019), SimCLR (Chen et al., 2020) for more general visual-representation learning. These approaches, while effective for the particular task at hand, fall short in learning representations that are generalisable to different tasks.

Contrastive methods have also been employed under a generative-model setting, but typically on generative adversarial networks (GANs) to either preserve or identify factors-of-variations in their inputs. For instance, SiGAN (Hsu et al., 2019) uses a contrastive loss to preserve identity for face-image hallucination from low-resolution photos, while (Yildirim et al., 2018) uses a contrastive loss to disentangle the factors of variations in the latent code of GANs. We here employ a contrastive loss in a distinct setting of multimodal generative model learning, that, as we will show with our experiments and analyses promotes better, more robust representation learning.

Multimodal VAEs We also demonstrate that our approach is applicable across different approaches to learning multimodal generative models. To do so, we first summarise past work on multimodal VAE into two categories based on the modelling choice of approximated posterior $q_b(z|x,y)$:

Explicit joint models where $q_b$ is explicitly modelled as one joint encoder $q_b(z|x,y)$. Example work in this area include JMVAE (Suzuki et al., 2017), triple ELBO (Vedantam et al., 2018) and MFM (Tsai et al., 2019). Since the joint encoder require multimodal pair $(x,y)$ as input, these approaches typically require additional modelling components and/or inference steps
to deal with missing modality at test time; in fact, all three approaches propose to train unimodal
VAEs on top of the joint model that handles data from each modality independently.

**Factorised joint models** where \( q_\theta \) is factorised as some combination \( f \) of two unimodal encoders,
\( \text{i.e. } q_\theta(z|x,y) = f(q_{\phi_x}(z|x), q_{\phi_y}(z|y)) \).
Multimodal VAEs of this type were first seen in Wu and Goodman (2018), proposing the MVAE
model with \( f \) defined as a product of experts (PoE), i.e.
\( q_\theta(z|x,y) = q_{\phi_x}(z|x)q_{\phi_y}(z|y)p(z) \),
allowing for cross-modality generation without extra modelling components. Particularly,
the MVAE was constructed to cater to multimodal settings where data was not guaranteed to be
organised into related sets, and where additional modalities were taken to be, in terms of information
content, subsets of a primary data source—such as images and their class labels.

Alternately, Shi et al. (2019) explored an approach that explicitly leveraged the availability of
related/paired data, motivated by arguments from embodied cognition of the world. They propose
the MMVAE model, which additionally differs from the MVAE model in its choice of posterior
approximation—where \( f \) is modelled as the mixture of experts (MoE) of unimodal posteriors—to
ameliorate shortcomings to do with precision miscalibration of the PoE. Furthermore, Shi et al.
(2019) also posit four criteria that a multimodal VAE should satisfy, which we adopt in this work
to evaluate the performance of our models.

## 3 Methodology

Given data over observations from two modalities \((x, y)\), one can learn a multimodal VAE by optimising
\( p_\theta(x, y, z) = p(z)p_{\phi_x}(z|x)p_{\phi_y}(y|z) \), where \( p_\theta(\cdot|z) \) are deep neural networks (decoders)
parametrised by \( \Theta = \{ \theta_x, \theta_y \} \). To maximise joint marginal likelihood \( \log p_\theta(x, y) \), one approxi-
mates the intractable model posterior \( p_\theta(z|x, y) \) with a variational posterior \( q_\theta(z|x, y) \), allowing
us to optimise an variational evidence lower bound (ELBO), defined as
\[
\log p_\theta(x, y) \geq \mathbb{E}_{z \sim q_\theta(z|x,y)} \left[ \log \frac{p_\theta(z, x, y)}{q_\theta(z \mid x, y)} \right] = \text{ELBO}(x, y). \tag{1}
\]

The remaining question here the approximated posterior \( q_\theta(z|x,y) \) be modelled. As mentioned in
§2, there are two schools of thinking, namely explicit joint model such as JMVAE (Suzuki et al.
2017) and factorised joint model including MVAE (Wu and Goodman 2018) and MMVAE (Shi et al.
2019). In this work we demonstrate the effectiveness of our approach for all these models.

### 3.1 Contrastive loss for "relatedness" learning

Unlike previous work where it is always assumed that \( x \) and \( y \) are related (Figure 2a), we introduce
an extra random variable \( r \) that dictates the relatedness of multimodal observations (Figure 2b), where
\( r = 1 \) when \( x, y \) shares a pre-determined commonality (such as digit), and \( r = 0 \) otherwise. To motivate our approach, we begin with the following hypothesis regarding the relationship between
pointwise mutual information of related and unrelated observations of different modalities:

**Hypothesis 3.1.** Let \((x, y) \sim p_\theta(x, y)\) be a related data pair from two modalities, and let \( y' \) denote
a data point not related with \( x \). Then we assume the pointwise mutual information \( I(x, y) > I(x, y') \).

Here we take \( p_\theta \) to refer to the multimodal objective in Eq. (1), i.e. conditioning on \( r = 1 \). Pointwise
mutual information provides a measure of the statistical dependence between two values \( a, b \), and for a
joint distribution \( p(a,b) \) is defined as \( I(a,b) = \log \frac{p(a,b)}{p(a)p(b)} \); mutual information is the expected value
of the pointwise mutual information under \( p(a,b) \). The assumption in Hypothesis 3.1 should be fairly
uncontroversial: we say simply that under the joint distribution for related data \( p_\theta(x, y) \), dependence
measured between related points \( x, y \) is stronger than between unrelated points \( x, y' \).

In practical terms, we can use this formulation of dependence to motivate a contrastive loss. For
related \( x, y \) and unrelated \( x, y' \), we can expand the inequality \( I(x, y) > I(x, y') \) as
\[
\log p_\theta(x, y) - \log p_\theta(x) - \log p_\theta(y) > \log p_\theta(x, y') - \log p_\theta(x) - \log p_\theta(y').
\]

Re-arranging terms, we introduce an explicit slack variable \( C > 0 \), and write
\[
\log p_\theta(x, y) - \log p_\theta(x, y') = \left[ \log p_\theta(y) - \log p_\theta(y') \right] + C.
\tag{2}
\]

1. marginal prevalence term
2. unrelated instance gap
Intuitively, this can be understood as decomposing the difference between the two joint probabilities \( \log p_{\Theta}(x, y) \) and \( \log p_{\Theta}(x, y') \) as two terms: 1. \( [\log p_{\Theta}(y) - \log p_{\Theta}(y')] \), which captures the difference in the joint probabilities due to the marginal contribution of the different values of \( y \) and \( y' \), a term irrelevant to whether the data are related; and 2. \( C \), which captures the remaining gap in joint probabilities due to evaluating related as opposed to unrelated points.

The above suggests that to achieve data-efficient multimodal learning by utilising the relatedness between multimodal pairs, one may maximise term (2) in (2) through maximising the difference between the joint marginal likelihood of related pairs \( p_{\Theta}(x, y) \) (Figure 3a) and that of unrelated pairs \( p_{\Theta}(x, y') \) (Figure 3b). This poses multimodal learning as a max-margin optimisation problem, for which a natural choice of objective is the contrastive (triplet) loss \( \mathcal{L}_c \) that takes the following form:

\[
\mathcal{L}_c(x, y, y') = d(x, y) - d(x, y') + m. \tag{3}
\]

Intuitively, \( \mathcal{L}_c \) attempts to make distance \( d \) between a positive pair \((x, y)\) smaller than the distance between a negative pair \((x, y')\) by margin \( m \). We can then adopt this loss to our relatedness learning objective by replacing \( d \) by (negative) joint marginal likelihood \( \log p_{\Theta} \), which is approximated by ELBO as shown in (1). Further, we can follow the practice in Song et al. (2016) and take the LogSumExp among \( N \) negative samples \( y_i' \in \{Y\}_{i=1}^N \), giving us the following objective

\[
\mathcal{L}_c(x, Y) = -\log p_{\Theta}(x, y) + \log \sum_{i=1}^N p_{\Theta}(x, y_i'). \tag{4}
\]

Also, since the loss is asymmetric, one can average over \( \mathcal{L}_c(y, X) \) and \( \mathcal{L}_c(x, Y) \) so that negative samples in both modalities are accounted for. We can now write our final contrastive objective:

\[
\mathcal{L}_c(x, y) = \frac{1}{2} \{ \mathcal{L}_c(x, Y) + \mathcal{L}_c(y, X) \} = -\log p_{\Theta}(x, y) + \frac{1}{2} \left( \log \sum_{x_i' = 1}^N p_{\Theta}(x_i', y) + \log \sum_{y_i' = 1}^N p_{\Theta}(x, y_i') \right) \\
\approx -\text{ELBO}(x, y) + \frac{1}{2} \left( \log \text{sumexp}_{x' \in X} \text{ELBO}(x', y) + \log \text{sumexp}_{y' \in Y} \text{ELBO}(x, y') \right) \tag{5}
\]

Note that since only ELBO terms are needed in (5), the contrastive learning framework can be directly applied to any multimodal generative model without needing extra components. In our experiments, we take \( N = 5 \) negative samples for each modality. We also show that it is possible to apply the contrastive loss for cases where number of modalities considered is greater than 2. See Appendix B for more details.

**Dissecting \( \mathcal{L}_c \)** Although (5) directly maximises ELBO like the VAE through (1), \( \mathcal{L}_c \) by itself is not an effective objective for VAE training since \( \mathcal{L}_c \) also contains term (2) that minimises ELBO, which can overpower the effect of (1) during training.

We provide an intuition to this phenomenon using a simple Gaussian example, illustrated in Figure 4. Here, we evaluate the log likelihood of \( \log p(x, y) \) in column 2, 3, 4 (green) on a Gaussian distribution \( \mathcal{N}((m_x, m_y); c) \), with the images in the first column of Figure 4 as means and with constant variance \( c \). We can see that while achieving high log likelihood \( \log p(x, y) \) on this distribution requires close imitation of \( (m_x, m_y) \) (second column), both unrelated digits (third column) and noise (last column) can lead to equally low joint log likelihoods. This indicates that the generative model need not generate valid, unrelated images to minimise (2)—generating noise would have the same effect on log likelihood. As a result, the model can trivially minimise minimise \( \mathcal{L}_c \) by generating noise that minimises (2) instead of accurate reconstruction that maximises (1).

This learning dynamic is verified empirically, as we show in Figure 9 in Appendix C if we optimise \( \mathcal{L}_c \) by itself, while the loss approaches 0 within the first 100 iterations, both (1) and (2) takes on extremely low values, resulting in a model that generates random noise.
Final objective  To mitigate the above effect, we need to ensure that when optimising the objective in (5), minimising \( \ell _2 \) does not overpower maximising \( \ell _1 \). We therefore introduce a hyperparameter \( \gamma \) on \( \ell _1 \) to upweight the maximisation of ELBO. Our final objective is therefore:

\[
\mathcal{L}(x, y) = -\gamma \text{ELBO}(x, y) + \frac{1}{2} \left( \log \text{sumexp} \text{ELBO}(x', y) + \log \text{sumexp} \text{ELBO}(x, y') \right),
\]

where \( \gamma > 1 \). In our experiments we simply take \( \gamma = 2 \) (upweight ELBO in \( \mathcal{L} \) by 1), however as observed in our ablation studies (see Appendix D), larger \( \gamma \) encourages better quality of generation and more stable training in some cases, while models trained with smaller \( \gamma \) are better at predicting “relatedness” between multimodal samples. We also note that optimising (6) maximises \( \text{PTC}(x, y) = \log \left( \frac{p_\Theta(x, y)}{p_\Theta(x)p_\Theta(y)} \right) \); see Appendix A for a proof.

4 Experiments

As stated in § 1, we analyse the suitability of contrastive learning for multimodal generative models from three perspectives—improved multimodal learning (§ 4.3), data efficiency (§ 4.4) and label propagation (§ 4.5). We now introduce the datasets and metrics used for our experiments.

4.1 Datasets

MNIST-SVHN  The dataset is designed to separate conceptual complexity, i.e. digit, from percep- tual complexity, i.e. color, style, size. Each data pair contains 2 samples of the same digit, one from each dataset (see examples in Figure 3a). We construct the dataset such that each instance from one dataset is paired with 30 instances of the same digit from the other dataset. Although both datasets are simple and well-studied, the many-to-many pairing between samples creates matching of different writing styles vs. backgrounds and colors, making it a challenging multimodal dataset.

CUB Image-Captions  We also consider a more challenging language-vision multimodal dataset, Caltech-UCSD Birds (CUB) (Reed et al., 2016; Welinder et al., 2010b). The dataset contains 11,788 photos of birds, paired with 10 captions describing the bird’s physical characteristics, collected through Amazon Mechanical Turk (AMT). See CUB image-caption pair in Figure 1.

4.2 Metrics

Shi et al. (2019) proposed four criteria for multimodal generative models (Figure 5, left), that we summarise and unify as metrics to evaluate these criteria for different generative models (Figure 5, right). We now introduce each criterion and its corresponding metric in detail.

(a) Latent factorisation
(b) Coherent joint generation
(c) Coherent cross generation
(d) Synergy

Figure 5: Left of each pair: Four criteria for multi-modal generative models; image adapted from Shi et al. (2019). Right of each pair: Four metrics to evaluate the model’s performance on criterion in corresponding row.

(a) Latent accuracy (Figure 5a)  Criterion: latent space should factor into “shared” and “private” subspaces for samples from two modalities. To examine this criterion, we fit a linear classifier on the samples from \( z \sim q_\phi(z|x, y) \) to see if it is possible to correctly classify the information shared between the two modalities. For data in MNIST-SVHN, this can be the digit label as shown in Figure 5a (right). To evaluate on this metric, we check if \( \hat{l}_z \) is the same as the digit label of the original inputs \( x \) and \( y \). The intuition here is, if we can extract the commonality between \( x \) and \( y \) from latent representation using a linear transform, the latent space would have factorised as desired.
Finding: Contrastive learning improves multimodal learning across all models and datasets.

4.3 Improved multimodal learning

Finding: Contrastive learning improves multimodal learning across all models and datasets.

MNIST-SVHN See Table 1 for results on the full MNIST-SVHN dataset. Note that for MMVAEs, since the joint posterior \( q_\phi \) is factorised as the mixture of unimodal posteriors \( q_\phi \), and \( q_\theta \), the model never directly takes sample from the explicit form of the joint posterior. Instead, it takes equal number of samples from each unimodal posterior, reflective of the equal weighting of the mixture. As a result, it is not meaningful to compute synergy coherence for MMVAE as it is exactly the same as the coherence of any single-way generation.

From Table 1, we see that our approach improves multimodal learning performance significantly for all three generative models evaluated on the metrics. The results showcase the robustness of our approach from the perspectives of modelling choice and metric of interests. The only caveat to this is the SVHN latent accuracy and SVHN=>MNIST cross coherence for MVAE. We believe that this is due to precision-miscalibration of the experts due to difference in complexity of the input modalities, which, as pointed out in both Shi et al. (2019); Wu and Goodman (2018), is a problem with the PoE factorisation and cannot be remedied by a contrastive loss.

Table 1: Evaluation of baseline MMVAE, MVAE, JMVAE and their contrastive variation (c-<MODEL>) on MNIST(8)-SVHN(S) dataset, using 100% (top) and 20% (bottom) of data respectively.

| Data          | Methods          | Latent accuracy (%) | Joint Coherence (%) | Cross coherence (%) | Synergy coherence (%) |
|---------------|------------------|---------------------|---------------------|---------------------|-----------------------|
|               |                  | M               | S                  | S→M                | M→S                   | joint→M               | joint→S               |
| 100% of data used | MMVAE            | 92.48 (±0.17)    | 79.03 (±1.37)      | 42.32 (±1.09)      | 70.77 (±0.61)          | 85.50 (±3.05)          | —                    | —                    |
|               | c-MMVAE          | 93.97 (±0.06)    | 81.87 (±0.82)      | 43.94 (±0.91)      | 79.66 (±0.99)          | 92.67 (±1.29)          | —                    | —                    |
|               | MVAE             | 91.65 (±0.17)    | 13.09 (±0.62)      | 37.59 (±0.66)      | 10.32 (±0.21)          | 22.73 (±1.42)          | 93.87 (±0.91)         | 83.30 (±3.17)         |
|               | M→S              | 93.25 (±0.65)    | 36.70 (±1.14)      | 40.62 (±2.12)      | 8.61 (±0.41)           | 36.1 (±0.11)           | 93.55 (±1.71)         | 87.24 (±2.19)         |
| 20% of data used | JMVAE            | 91.65 (±0.10)    | 73.02 (±0.88)      | 42.74 (±1.08)      | 69.51 (±1.20)          | 86.75 (±2.03)          | —                    | —                    |
|               | c-JMVAE          | 84.45 (±0.7)     | 57.98 (±1.27)      | 42.18 (±1.50)      | 49.63 (±2.70)          | 54.98 (±2.62)          | 85.77 (±0.66)         | 68.15 (±1.38)         |
|               | MVAE             | 82.68 (±0.67)    | 76.89 (±1.34)      | 38.63 (±0.98)      | 6.84 (±2.24)           | 60.64 (±2.2)           | 88.29 (±0.44)         | 75.00 (±0.69)         |
|               | M→S              | 91.64 (±0.10)    | 73.02 (±0.88)      | 42.74 (±1.08)      | 69.51 (±1.20)          | 86.75 (±2.03)          | —                    | —                    |
|               | JMVAE            | 88.54 (±0.17)    | 68.90 (±1.50)      | 37.41 (±0.66)      | 59.32 (±2.02)          | 76.33 (±2.23)          | —                    | —                    |
|               | c-JMVAE          | 91.64 (±0.10)    | 73.02 (±0.88)      | 42.74 (±1.08)      | 69.51 (±1.20)          | 86.75 (±2.03)          | —                    | —                    |
|               | MVAE             | 90.29 (±0.37)    | 12.76 (±0.96)      | 30.03 (±1.11)      | 9.98 (±0.61)           | 18.49 (±0.67)          | 92.70 (±1.12)         | 35.85 (±1.60)         |
|               | M→S              | 93.75 (±0.42)    | 48.59 (±0.97)      | 37.37 (±1.56)      | 10.58 (±0.14)          | 27.30 (±0.87)          | 90.92 (±1.52)         | 82.64 (±1.42)         |
|               | JMVAE            | 77.53 (±0.11)    | 52.55 (±2.16)      | 26.37 (±0.44)      | 42.58 (±3.10)          | 41.44 (±2.26)          | 85.07 (±0.34)         | 51.95 (±2.20)         |
|               | c-JMVAE          | 78.32 (±0.62)    | 67.91 (±2.19)      | 39.03 (±1.96)      | 55.81 (±2.41)          | 57.89 (±2.90)          | 79.35 (±1.34)         | 73.28 (±0.68)         |

CUB Following Shi et al. (2019), for the images in CUB, we observe and generate in feature space instead of pixel space by preprocessing the images using a pre-trained ResNet-101 (He et al. 2016). A nearest-neighbour lookup among all the features in the test set is used to project the feature generations of the model back to image space. This helps circumvent CUB image complexities to some extent—as the primary goal here is to learn good models and representations of multimodal data, rather than a focus on pixel-level image quality of generations.
Finding: Generative models learned contrastively are good predictors of “relatedness”, enabling label propagation and matching baseline performance on full datasets, using only 10% of data.

### 4.4 Data Efficiency

Finding: Contrastive learning on 20% of data matches baseline models on full data.

We plot the quantitative performance of MMVAE with and without contrastive learning against the percentage of the original dataset used, as seen in Figure 6. We observe that performance of contrastive MMVAE (c-MMVAE, red) is consistently better than the baselines (MMVAE, blue), and that baseline performance using all related data is matched by the contrastive MMVAE using just 10—20% of data. The partial datasets used here are constructed by first taking \( n \) of data. The partial datasets used here are constructed by first taking \( n \) of data. The percentage of the original dataset used, as seen in Figure 6. We observe that performance of contrastive MMVAE (c-MMVAE, red) is consistently better than the baselines (MMVAE, blue), and that baseline performance using all related data is matched by the contrastive MMVAE using just 10—20% of data. The partial datasets used here are constructed by first taking \( n \) of data.

Table 2: Evaluation of baseline MMVAE, MVAE, JMVAE and their contrastive variation (c-<MODEL>) on CUB image-caption pairs, using 100% (top) and 20% (bottom) of data respectively.

| Data         | Methods          | Joint Coherence (CCA) | Cross Coherence (CCA) | Synergy Coherence (CCA) |
|--------------|------------------|-----------------------|-----------------------|-------------------------|
| 100% of data | MMVAE 0.212 (\( \pm 3.46 \times 10^{-2} \)) | 0.154 (\( \pm 7.06 \times 10^{-3} \)) | 0.244 (\( \pm 5.83 \times 10^{-3} \)) | -                      |
|              | c-MMVAE 0.314 (\( \pm 3.12 \times 10^{-2} \)) | 0.188 (\( \pm 4.02 \times 10^{-3} \)) | 0.334 (\( \pm 2.26 \times 10^{-3} \)) | -                      |
|              | MVAE 0.174 (\( \pm 6.93 \times 10^{-3} \)) | 0.100 (\( \pm 6.93 \times 10^{-3} \)) | 0.115 (\( \pm 6.93 \times 10^{-3} \)) | 0.110 (\( \pm 6.93 \times 10^{-3} \)) |
|              | c-MVAE 0.344 (\( \pm 1.17 \times 10^{-3} \)) | 0.156 (\( \pm 4.15 \times 10^{-3} \)) | 0.216 (\( \pm 5.73 \times 10^{-3} \)) | 0.231 (\( \pm 5.36 \times 10^{-3} \)) |
|              | JMVAE 0.220 (\( \pm 1.72 \times 10^{-3} \)) | 0.157 (\( \pm 5.98 \times 10^{-3} \)) | 0.191 (\( \pm 5.12 \times 10^{-3} \)) | 0.212 (\( \pm 3.21 \times 10^{-2} \)) |
|              | c-JMVAE 0.255 (\( \pm 1.34 \times 10^{-3} \)) | 0.149 (\( \pm 2.53 \times 10^{-3} \)) | 0.226 (\( \pm 7.46 \times 10^{-3} \)) | 0.202 (\( \pm 3.12 \times 10^{-3} \)) |
| 20% of data  | MMVAE 0.117 (\( \pm 3.51 \times 10^{-2} \)) | 0.094 (\( \pm 2.12 \times 10^{-3} \)) | 0.153 (\( \pm 7.47 \times 10^{-3} \)) | -                      |
|              | c-MMVAE 0.205 (\( \pm 1.65 \times 10^{-2} \)) | 0.136 (\( \pm 1.15 \times 10^{-3} \)) | 0.251 (\( \pm 3.80 \times 10^{-3} \)) | -                      |
|              | MVAE 0.125 (\( \pm 2.30 \times 10^{-2} \)) | 0.030 (\( \pm 3.56 \times 10^{-3} \)) | 0.079 (\( \pm 7.37 \times 10^{-3} \)) | 0.101 (\( \pm 3.90 \times 10^{-3} \)) |
|              | c-MVAE 0.220 (\( \pm 1.72 \times 10^{-3} \)) | 0.072 (\( \pm 8.05 \times 10^{-3} \)) | 0.184 (\( \pm 5.96 \times 10^{-3} \)) | 0.191 (\( \pm 5.12 \times 10^{-3} \)) |
|              | JMVAE 0.127 (\( \pm 3.76 \times 10^{-2} \)) | 0.118 (\( \pm 3.82 \times 10^{-3} \)) | 0.154 (\( \pm 8.34 \times 10^{-3} \)) | 0.181 (\( \pm 3.26 \times 10^{-2} \)) |
|              | c-JMVAE 0.269 (\( \pm 1.20 \times 10^{-2} \)) | 0.134 (\( \pm 2.45 \times 10^{-3} \)) | 0.210 (\( \pm 2.35 \times 10^{-3} \)) | 0.192 (\( \pm 4.14 \times 10^{-3} \)) |
Here, we show that our contrastive framework encourages a larger discrepancy between the PMI of related vs. unrelated data, as set out in hypothesis 3.1, allowing one to first train the model on a small subset of related data, and subsequently construct a classifier using PMI that identifies related samples in the remaining data. We now introduce our pipeline for label propagation in details.

**Pipeline** As showing in Figure 7, we first construct a full dataset by randomly matching instances in MNIST and SVHN, and denote the related pairs by \( F_r \) (full, related). We further assume access to only \( n \% \) of \( F_r \), denoted as \( S_r \) (small, related), and denote the rest as \( F_m \), containing a mix of related and unrelated pairs. Next, we train a generative model \( g \) on \( S_r \). To find a relatedness threshold, we construct a small, mixed dataset \( S_m \) by randomly matching samples across modalities in \( S_r \). Given relatedness ground-truth for \( S_m \), we can compute the PMI \( I(x, y) = \log p_{\theta}(x, y) - \log p_{\theta}(x)p_{\theta}(y) \) for all pairs \((x, y)\) in \( S_m \) and estimate an optimal threshold. This threshold can now be applied to the full, mixed dataset \( F_m \) to identify related pairs giving us a new related dataset \( F_{\hat{r}} \), which can be used to further improve the performance of the generative model \( g \).

**Results** In Figure 8 (a-e), we plot the performance of baseline MMVAE (blue) and contrastive MMVAE (red), trained with (darker lines) and without (lighter lines) label propagation. Here, the x-axis is the proportion in size of \( S_r \) to \( F_r \), i.e. the percentage of related data used to pretrain the generative model before label propagation. We compare these results to MMVAE trained on all related data \( F_r \) (yellow, dotted) as a "best case scenario" of these training regimes.

Clearly, label propagation using a contrastive model is helpful, and in general the improvement is greater when less data is available. Figure 8 also shows that when \( S_r \) is 10% of \( F_r \), c-MMVAE (red, dark) is competitive with the performance of baseline MMVAE trained on \( F_r \) (yellow, dotted).

For baseline MMVAE, however, label propagation hurts performance no matter the size of \( S_r \), as shown by the blue curves in Figure 8 (a-e). This can be explained by Figure 8 where we compute the precision, recall, and \( F_1 \) score of relatedness prediction on \( F_m \), for models trained on 10% of all related data. We also compare to a simple label-propagation baseline, where the relatedness of \( F_m \) is predicted using a siamese network (Hadsell et al., 2006) trained on the same 10% dataset. We notice that the precision, recall and \( F_1 \) score of MMVAE is significantly lower than contrastive MMVAE, indicating that the PMI evaluated on MMVAE is a poorer indicator of relatedness than that from the contrastive MMVAE. Notably, the Siamese baseline in Figure 8 is a competitive predictor of relatedness, with higher precision than both MMVAE and contrastive MMVAE, but lower recall than contrastive MMVAE. However, note that with the contrastive MMVAE, relatedness can be predicted without additional training and only requires a simple threshold computation directly computed using the generative model. The fact that the contrastive MMVAE’s relatedness-prediction performance matches that of a discriminative model, while the baseline MMVAE’s does not, strongly supports the view that the contrastive loss encourages generative models to utilise and better learn the relatedness between multimodal pairs.

5 Conclusion

We introduced a contrastive-style objective for multimodal VAE, aiming at reducing the amount of multimodal data needed by exploiting the distinction between "related" and "unrelated" multimodal pairs. We showed that this objective improves multimodal training, drastically reduce the amount of multimodal data needed, and establishes a strong sense of "relatedness" for the generative model. These findings hold true across a multitude of datasets, models and metrics. The positive results of our method indicates that it is beneficial to utilise the relatedness information when training on multimodal data, which has been largely ignored in previous work. While we propose to utilise it implicitly through contrastive loss, future work may consider relatedness as a random variable in the graphical model and see if explicit dependency on relatedness can be useful.
Figure 8: Model performance with (dark) and without (light) label propagation using MMVAE vs. c-MMVAE.

Acknowledgements

YS, NS, and PHST were supported by the ERC grant ERC-2012-AdG 321162-HELIOS, EPSRC grant Seebibyte EP/M013774/1 and EPSRC/MURI grant EP/N019474/1, with further support from the Royal Academy of Engineering and FiveAI. YS was additionally supported by Remarkdip through their PhD Scholarship Programme. BP is supported by the Alan Turing Institute under the EPSRC grant EP/N510129/1. Special thanks to Elise van der Pol for helpful discussions on contrastive learning.
Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.

T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.

W. Guo, J. Wang, and S. Wang. Deep multimodal representation learning: A survey. *IEEE Access*, 7:63373–63394, 2019.

R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’06)*, volume 2, pages 1735–1742, 2006.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

K. He, H. Fan, Y. Wu, S. Xie, and R. B. Girshick. Momentum contrast for unsupervised visual representation learning. *arXiv preprint arXiv:1911.05722*, 2019.

R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, P. Bachman, A. Trischler, and Y. Bengio. Learning deep representations by mutual information estimation and maximization. In *ICLR 2019 : 7th International Conference on Learning Representations*, 2019.

C.-C. Hsu, C.-W. Lin, W.-T. Su, and G. Cheung. Sigan: Siamese generative adversarial network for identity-preserving face hallucination. *IEEE Transactions on Image Processing*, 28(12):6225–6236, 2019.

O. J. Hénaff, A. Srinivas, J. D. Fauw, A. Razavi, C. Doersch, S. M. A. Eslami, and A. van den Oord. Data-efficient image recognition with contrastive predictive coding. *arXiv preprint arXiv:1905.09272*, 2019.

D. Massiceti, P. K. Dokania, N. Siddharth, and P. H. Torr. Visual dialogue without vision or dialogue. In *NeurIPS Workshop on Critiquing and Correcting Trends in Machine Learning*, 2018.

A. Mnih and K. Kavukcuoglu. Learning word embeddings efficiently with noise-contrastive estimation. In *Advances in Neural Information Processing Systems 26*, pages 2265–2273, 2013.

S. E. Reed, Z. Akata, B. Schiele, and H. Lee. Learning deep representations of fine-grained visual descriptions. *CoRR*, abs/1605.05395, 2016.

Y. Shi, S. Narayanaswamy, B. Paige, and P. Torr. Variational mixture-of-experts autoencoders for multi-modal deep generative models. In *NeurIPS 2019 : Thirty-third Conference on Neural Information Processing Systems*, pages 15718–15729, 2019.

H. O. Song, Y. Xiang, S. Jegelka, and S. Savarese. Deep metric learning via lifted structured feature embedding. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4004–4012, 2016.

M. Suzuki, K. Nakayama, and Y. Matsuo. Joint multimodal learning with deep generative models. In *International Conference on Learning Representations Workshop*, 2017.

J. B. Tenenbaum and W. T. Freeman. Separating style and content with bilinear models. *Neural Computation*, 12(6):1247–1283, 2000.

Y. H. Tsai, P. P. Liang, A. A. Bagherzade, L.-P. Morency, and R. Salakhutdinov. Learning factorized multimodal representations. In *International Conference on Learning Representations*, 2019.

A. van den Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

R. Vedantam, I. Fischer, J. Huang, and K. Murphy. Generative models of visually grounded imagination. In *International Conference on Learning Representations*, 2018.

K. Q. Weinberger and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10(9):207–244, 2009.

P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010a.

P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010b.
M. Wu and N. Goodman. Multimodal generative models for scalable weakly-supervised learning. In *Advances in Neural Information Processing Systems*, pages 5580–5590, 2018.

G. Yildirim, N. Jetchev, and U. Bergmann. Unlabeled disentangling of gans with guided siamese networks. 2018.

I. Yildirim. From perception to conception: Learning multisensory representations. 2014.

C. Zhang, Z. Yang, X. He, and L. Deng. Multimodal intelligence: Representation learning, information fusion, and applications. *IEEE Journal of Selected Topics in Signal Processing*, pages 1–1, 2020.
Appendix:

A Connection of final objective to Pointwise Mutual Information

Here, we show that minimising the objective in (6) maximises the PMI between \( x \) and \( y \):

\[
\mathcal{L}(x, y) = -\gamma \text{ELBO}(x, y) + \frac{1}{2} \left( \log \sum_{x' \in X} \text{ELBO}(x', y) + \log \sum_{y' \in Y} \text{ELBO}(x, y') \right)
\]

\[
\approx -\gamma \log p_{\Theta}(x, y) + \frac{1}{2} \left( \log \sum_{x' \in X} p_{\Theta}(x', y) + \log \sum_{y' \in Y} p_{\Theta}(x, y') \right)
\]

\[
= - (\gamma - \frac{1}{2}) \log p_{\Theta}(x, y) - \frac{1}{2} \log \frac{p_{\Theta}(x, y)}{\sum_{y' \in Y} p_{\Theta}(x, y') \sum_{x' \in X} p_{\Theta}(y, x')}
\]

\[
\approx - (\gamma - \frac{1}{2}) \log p_{\Theta}(x, y) - \frac{1}{2} \log \frac{p_{\Theta}(x, y)}{p_{\Theta}(x)p_{\Theta}(y)} \tag{7}
\]

We see in (7) that minimising \( \mathcal{L} \) can be decomposed to maximising both the joint marginal likelihood \( p_{\Theta}(x, y) \) and PMI(\( x, y \)). Note that since \( \gamma > 1 \), we can be sure that the joint marginal likelihood weighting \( \gamma - \frac{1}{2} \) is non-negative.
B Generalisation to $M > 2$ Modalities

In this section we show how the contrastive loss generalise to cases where number of modalities considered $M$ is greater than 2.

Given observations from $M$ modalities $D = \{X_1, X_2, \ldots, X_m, \ldots, X_M\}$, where $X_m$ denotes unimodal dataset of modality $m$ of size $N_m$, i.e. $X_m = \{x^{(i)}_m\}_{i=1}^{N_m}$. Similar to (4), we can write the assymetrical contrastive loss for any observation $x^{(i)}_m$ from modality $m$, where negative samples are taken for all $(M - 1)$ other modalities:

$$
\mathcal{L}_C(x^{(i)}_m, D_{\bar{m}}) = -\log p_{\Theta}(x^{(i)}_m) + \log \sum_{d=1}^{M} \sum_{j=1 \atop d \neq m}^{N} p_{\Theta}(x^{(i)}_1:\{d-1\},(d+1):M, x^{(j)}_d) .
$$

(8)

We can therefore rewrite (5) as:

$$
\mathcal{L}_C(x^{(i)}_1:M) = \frac{1}{M} \sum_{m=1}^{M} \mathcal{L}_C(x^{(i)}_m, D_{\bar{m}})
$$

(9)

$$
= -\log p_{\Theta}(x^{(i)}_1:M) + \frac{1}{M} \sum_{m=1}^{M} \left( \log \sum_{d=1 \atop d \neq m}^{M} \sum_{j=1}^{N} p_{\Theta}(x^{(i)}_1:(d-1),(d+1):M, x^{(j)}_d) \right),
$$

(10)

where $N$ is the number of negative samples, all $\log p_{\Theta}(x^{(i)}_1:M)$ are approximated by the following joint ELBO for $M$ modalities:

$$
\log p_{\Theta}(x^{(i)}_1:M) \geq \mathbb{E}_{z \sim q_{\Phi}(z|x_1:M)} \left[ \log \frac{p_{\Theta}(z, x^{(i)}_1:M)}{q_{\Phi}(z|x_1:M)} \right] = \text{ELBO}(x^{(i)}_1:M).
$$

(11)

While the above gives us the true generalisation of (5), we note that the number of times where ELBO needs to be evaluated in (10) is $O(M^2N)$, making it difficult to implement this objective in practice, especially on more complicated datasets. We therefore propose a simplified version of the objective, where we estimate the second term of (10) with $N$ sets of random samples from all modalities. Specifically, we can precompute the following $M \times N$ random index matrix $J$:

$$
J = \begin{bmatrix}
    j_{11} & j_{12} & \cdots & j_{1N} \\
    j_{21} & j_{22} & \cdots & j_{2N} \\
    \vdots & \vdots & \ddots & \vdots \\
    j_{M1} & j_{M2} & \cdots & j_{MN}
\end{bmatrix},
$$

(12)

where each entry of $J$ is a random integer taken from range $[1, N_m]$. We can then replace the second term of (10) random samples selected by the indices in $J$, giving us

$$
\mathcal{L}_C(x^{(i)}_1:M) \approx -\log p_{\Theta}(x^{(i)}_1:M) + \log \sum_{n=1}^{N} p_{\Theta}(x_1^{(J_{1n})}, x_2^{(J_{2n})}, \ldots, x_M^{(J_{Mn})}),
$$

(13)

The number of times ELBO needs to be computed is now $O(N)$, and is no longer relevant to the number of modalities $M$.

We can now also generalise the final objective in (6) to $M$ modalities:

$$
\mathcal{L}(x^{(i)}_1:M) = -\gamma \text{ELBO}(x^{(i)}_1:M) + \log \text{sumexp} \ \text{ELBO}(x_1^{(J_{1n})}, x_2^{(J_{2n})}, \ldots, x_M^{(J_{Mn})}).
$$

(14)
C The Ineffectiveness of training with $\mathcal{L}_C$ only

We demonstrate why training with the contrastive loss proposed in (5) is ineffective, and why additional ELBO term is needed for the final objective. As we show in Figure 9 when training with $\mathcal{L}_C$ only, while the contrastive loss (green) quickly drops to zero, both term (1) and (2) in (5) also reduces drastically. This means the joint marginal likelihood of any generation $\log p_\Theta(x, y)$ is small regardless the relatedness of $(x, y)$.

In comparison, we also plot the training curve for model trained on the final objective in (6), which upweights term (1) in (5) by $\gamma$. We see in Figure 10 that by setting $\gamma = 2$, the joint marginal likelihood (yellow and blue curve) improves during training, while $\mathcal{L}_C$ (green curve) gets minimised.

Figure 9: First 300 iterations of training using contrastive loss $\mathcal{L}_C$ only.

Figure 10: First 300 iterations of training with final loss $\mathcal{L}$, where $\gamma = 2$. 
D Ablation study of $\gamma$

In §3 we specified that $\gamma$ needs to be greater than 1 to offset the negative effect of minimising ELBO through term $(\gamma)$ in (6). Here, we study the effect of $\gamma$ in details.

Figure 11 compares latent accuracy, cross coherence and joint coherence of MMVAE on MNIST-SVHN dataset trained on different values of $\gamma$. Note that here we only consider cases where $\gamma \geq 1$, since the minimum value of $\gamma$ is 1. In this case, the loss reduces to the original contrastive objective in (5).

A few interesting observations from the plot are as follows: First, when $\gamma = 1$, the model is trained using the contrastive loss only, and as we showed is an ineffective objective for generative model learning. This is verified again in Figure 11 when $\gamma = 1$, both coherences and latent accuracies take on extremely low values; interestingly, there is a significant boost of performance across all metrics by simply increasing the value of $\gamma$ from 1 to 1.1; after that, as the value of $\gamma$ increases, performance on most metrics decreases monotonically (joint coherence being the only exception), and eventually converges to baseline MMVAE (dotted lines in Figure 11). This is unsurprising, since the final objective in (6) reduces to the original joint ELBO as $\gamma$ approaches infinity.

![Figure 11: Performance on different metrics for different values of $\gamma$. Dotted lines represents the performance of baseline MMVAE.](image)

Figure 11 seems to suggest that 1.1 is the optimal value for hyperparameter $\gamma$, however close inspection of the qualitative generative results shows that this might not be the case. See Figure 12 for a comparison of the model’s generation between MMVAE models trained on (from left to right) $\gamma = 1.1$, $\gamma = 2$ and $\gamma = +\infty$ (i.e. original MMVAE). Although $\gamma = 1.1$ yields model with high coherence scores, we can clearly see from the left-most column of Figure 12 that the generation of the model seems deprecated, especially for the SVHN modality, where the backgrounds of model’s generation appear to be unnaturally spotty and deformed. This problem is mitigated by increasing $\gamma$ — as shown in Figure 11 the image generation quality of $\gamma = 2$ (middle column) is not visibly different from that of $\gamma = +\infty$ (right column).

To verify this observation, we also compute the marginal log likelihood $\log p_{\Theta}(x, y)$ to quantify the quality of generations. We compute this for all $\gamma$s considered in Figure 11 and take the average over the entire test set. From the results in Figure 13 we can see a significant increase of the log likelihood between $\gamma = 1.1$ to $\gamma = 1$. This gain in image generation quality then slows down as $\gamma$ further increases, and as all other metrics converges to the original MMVAE model.
\[ \gamma = 1.1 \quad \gamma = 2 \quad \gamma = +\infty \]

(a) Joint generation, MNIST

(b) Joint generation, SVHN

(c) Reconstruction, MNIST

(d) Reconstruction, SVHN

(e) Cross generation, S→M

(f) Cross generation, M→S

Figure 12: Generations of MMVAE model trained using the final contrastive objective, with (from left to right) \( \gamma = 1.1, 2 \) and \( +\infty \). Note in (c), (d), (e), (f), the top rows are the inputs and the bottom rows are their corresponding reconstruction/cross generation.

Figure 13: Performance on different metrics for different values of \( \gamma \). Dotted lines represent the performance of baseline MMVAE.
We use architectures listed in Table 3 for the unimodal encoder and decoder for MMVAE, MVAE and JMVAE. For JMVAE we use an extra joint encoder, the architecture of which is described in Table 4.

| Encoder | Decoder |
|---------|---------|
| Input ∈ $\mathbb{R}^{1\times28\times28}$ | Input ∈ $\mathbb{R}^{L}$ |
| FC. 400 ReLU | FC. 400 ReLU |
| FC. L, FC. L | FC. L x 28 x 28 Sigmoid |

(a) MNIST dataset

| Encoder |
|---------|
| Input ∈ $\mathbb{R}^{3\times32\times32}$ |
| 4x4 conv. 32 stride 2 pad 1 & ReLU |
| 4x4 conv. 64 stride 2 pad 1 & ReLU |
| 4x4 conv. 128 stride 2 pad 1 & ReLU |
| 4x4 conv. L stride 1 pad 0, 4x4 conv. L stride 1 pad 0 |

Decoder

| Input ∈ $\mathbb{R}^{L}$ |
| 4x4 upconv. 128 stride 1 pad 0 & ReLU |
| 4x4 upconv. 64 stride 2 pad 1 & ReLU |
| 4x4 upconv. 32 stride 2 pad 1 & ReLU |
| 4x4 upconv. 3 stride 2 pad 1 & Sigmoid |

(b) SVHN dataset.

| Encoder | Decoder |
|---------|---------|
| Input ∈ $\mathbb{R}^{2048}$ | Input ∈ $\mathbb{R}^{L}$ |
| FC. 1024 ELU | FC. 256 ELU |
| FC. 512 ELU | FC. 512 ELU |
| FC. 256 ELU | FC. 1024 ELU |
| FC. L, FC. L | FC. 2048 |

(c) CUB image dataset.

| Encoder |
|---------|
| Input ∈ $\mathbb{R}^{1590}$ |
| Word Emb. 128 |
| 4x4 conv. 32 stride 2 pad 1 & BatchNorm2d & ReLU |
| 4x4 conv. 64 stride 2 pad 1 & BatchNorm2d & ReLU |
| 4x4 conv. 128 stride 2 pad 1 & BatchNorm2d & ReLU |
| 1x4 conv. 256 stride 1x2 pad 0x1 & BatchNorm2d & ReLU |
| 1x4 conv. 512 stride 1x2 pad 0x1 & BatchNorm2d & ReLU |
| 1x4 conv. 128 stride 1 pad 0, 4x4 conv. L stride 1 pad 0 |

Decoder

| Input ∈ $\mathbb{R}^{L}$ |
| 4x4 upconv. 512 stride 1 pad 0 & ReLU |
| 1x4 upconv. 256 stride 1x2 pad 0x1 & BatchNorm2d & ReLU |
| 1x4 upconv. 128 stride 1x2 pad 0x1 & BatchNorm2d & ReLU |
| 4x4 upconv. 64 stride 2 pad 1 & BatchNorm2d & ReLU |
| 4x4 upconv. 32 stride 2 pad 1 & BatchNorm2d & ReLU |
| 4x4 upconv. 1 stride 2 pad 1 & ReLU |
| Word Emb. 1590 |

(d) CUB-Language dataset.

Table 3: Unimodal encoder and decoder architectures.
Encoder

Input $\in \mathbb{R}^{3\times32\times64}$
- 4x4 conv. 32 stride 2 pad 1 & ReLU
- 4x4 conv. 64 stride 2 pad 1 & ReLU
- 4x4 conv. 128 stride 2 pad 1 & ReLU
- 1x4 conv. 128 stride 1x2 pad 0x1 & ReLU
- 4x4 conv. L stride 1 pad 0, 4x4 conv. L stride 1 pad 0

(a) MNIST-SVHN dataset.

Encoder

Input $\in \mathbb{R}^{1\times32\times192}$
- 4x4 conv. 32 stride 2 pad 1 & BatchNorm2d & ReLU
- 4x4 conv. 64 stride 2 pad 1 & BatchNorm2d & ReLU
- 4x4 conv. 128 stride 2 pad 1 & BatchNorm2d & ReLU
- 1x4 conv. 256 stride 1x2 pad 0x1 & BatchNorm2d & ReLU
- 1x4 conv. 512 stride 1x2 pad 0x1 & BatchNorm2d & ReLU
- 4x4 conv. L stride 1 pad 0, 4x6 conv. L stride 1 pad 0

(b) CUB Image-Caption dataset.

Table 4: Joint encoder architectures.