ROBUST MULTI-VIEW REPRESENTATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-view data has become ubiquitous, especially with multi-sensor systems like self-driving cars or medical patient-side monitors.

We look at modeling multi-view data through robust representation learning, with the goal of leveraging relationships between views and building resilience to missing information. We propose a new flavor of multi-view AutoEncoders, the Robust Multi-view AutoEncoder, which explicitly encourages robustness to missing views. The principle we use is straightforward: we apply the idea of drop-out to the level of views. During training, we leave out views as input to our model while forcing it to reconstruct all of them. We also consider a flow-based generative modeling extension of our approach in the case where all the views are available.

We conduct experiments for different scenarios: directly using the learned representations for reconstruction, as well as a two-step process where the learned representation is subsequently used as features for the data for a down-stream application. Our synthetic and real-world experiments show promising results for the application of these models to robust representation learning.

1 INTRODUCTION

Multi-view machine learning, or multi-modal machine learning, involves learning on data which has multiple, potentially asynchronous, observation models. For example, videos have both an audio and visual channel which provide complementary information. Images and their captions provide different views of the same data.

The applications of multi-view machine learning are numerous. One simple setting would be to use the multiple views of a dataset to build a more robust classifier than one that can be trained on any single view alone. This is the basis for the co-training learning paradigm. Co-training is a semi-supervised learning framework which uses two or more complementary views to jointly train classifiers over each view. It does so by first building view-specific classifiers and iteratively expands the labeled dataset by adding the unlabeled points based on the confidence of these individual classifiers.

Other common applications are cross-modal translation/retrieval. For example, given an image, we would like to generate an appropriate text caption. On the other hand, given a text description, we could also retrieve the most appropriate image from a given dataset. Language translation, and cross-modal sequence-to-sequence translation in general, is another domain for the application of multi-modal machine learning.

In this paper, we are interested in exploring multi-view representation learning from the perspective of understanding and exploiting the relationships between the multiple views. The idea is that multi-view data doesn't just provide us with multiple sets of ”features” for the data, it also provides structural information in the form of the interactions and redundancies between views. To this end, we propose a multi-view AutoEncoder based approach for **Robust Multi-view Representation Learning**.

The essential idea is straightforward: we apply the idea of drop-out at the level of the views themselves. We would like to encourage robustness to missing views, so we explicitly force a random subset of views to go missing during training while having the model reconstruct all. This encourages the model to leverage local redundancies and relationships between views to build this robustness.

Our method has a two-tiered structure where we have one encoder for each view giving us intermediate codes which are then concatenated and fed into a shared encoder for the final representation. This separation is done so as to allow the model flexibility to learn a good ”translation” of each view which is more conducive to shared representation learning. These individual encoders may be Auto-Encoders themselves or a similar trainable feature extraction method.

Our generative modeling extension looks at using flow-based approaches to learn the latent distribution of the data. Flow-based models work under the philosophy that a *good representation* of the data is one in which the data distribution is simple. They typically consist of a sequence of trainable invertible transforms into a latent space, where they maximize its likelihood over a simple, or trainable, base distribution. We apply the two-tiered structure to this as well, to build a flow-based Multi-view AutoEncoder.

Notation: The data is represented by the rows of X_i for each view i of K views, drawn from the data-distribution $p_{\mathcal{X}}$. We represent any intermediate latent spaces with L_i for view i, and the shared latent space as L_{all}. We assume that our data is centered.

2 Related Work

Learning over multiple modalities is often difficult, due to heterogeneous sources of data, different levels of noise or missing data in some views. This makes it imperative to extract meaningful information from the different views in a robust fashion. Representation learning is thus one of the core directions of multi-modal machine learning research. It is common as an intermediate step before learning over a down-stream task.

Many such learning methods are tailored to certain domains, wherein they exploit the structure available specific to the data. For example, Audio-Video Speech Recognition (AVSR) has been the subject of research for many years now. Traditionally, deep neural networks are used to handle visual, textual and acoustic data Ngiam et al. (2011), Ouyang et al. (2014), Wang et al. (2015) where the model projects the modalities into a joint space Antol et al. (2015), Mroueh et al. (2015), Ouyang et al. (2014), Wu et al. (2014). This representation is then used for the relevant learning task.

Multi-view AutoEncoders are also extended to learn latent representations over multi-modal data. Ngiam et al. (2011) learn modality specific AEs and then fuse together the latent states into a final shared representation. Silberer & Lapata (2014) use auto-encoders for semantic concept grounding, with the addition of a loss-term for object-label prediction. Wang et al. (2015) fine-tunes the representation learned by the generic AEs on a given task.

Neural networks have their advantages; given domain-specific architectures and the potential for pre-training, they often show superior performance on certain tasks. But they need a lot of data and are not always able to gracefully handle missing data from modalities. Another popular multi-modal representation learning approach is based on graphical models. Unsupervised methods such as Deep Boltzman Machines (DBN) Srivastava & Salakhutdinov (2012), Kim et al. (2013), Huang & Kingsbury (2013) are often used for multi-modal representation learning.

Our work, however, tries to remain agnostic to the application domain. We also try to explicitly reason about the *local* relationships between views, where we look at structure that may only exist between subsets of views and not shared between all the views.

3 Approaches

3.1 Robust Multi-view AutoEncoder

The typical strategies for training a Multi-view AutoEncoder (MVAE) have the same concerns outlined in the previous section; they try to directly learn a shared embedding space which best reconstructs all views (Ye et al. (2016), Wang et al. (2015)). This often means learning a single bottle neck representation shared across all views. Such an architecture implicitly captures the intersection of information across all views. We train an MVAE which learns to be robust to missing views by trying to capture the union of information across the views instead.

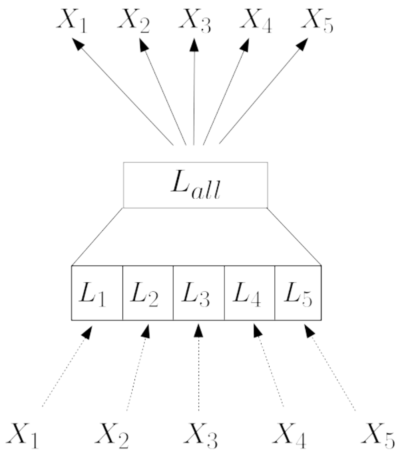
X_1 X_2 X_3 X_4 X_5
L_{all}
L_1 L_2 L_3 L_4 L_5
X_1 X_2 X_3 X_4 X_5

Figure 1: This is the outline of the Robust Multi-view AutoEncoder for 5 views. Each bottom arrow represents the encoder for its respective view, and each upper arrow represents the decoder. L_i is the encoding for view i and L_S is the global latent representation. The bottom arrows are dotted to represent potential drop-out during training time.

In our proposed architecture, called the Robust MVAE (RMAE), as shown in 1, we have two levels of encoding. The first is at the view level, where every view has its own individual encoder network. These encoders produce the latent embeddings L_i for their respective views. Then, we compose this with an additional encoder, called the meta-encoder, which operates on top of these embeddings to produce the final global latent representation L_S. This is then used as input to the decoders for the reconstruction of different views.

Here, our idea for ”robustness” of a representation is the ability of faithful reconstruction of all views given that an arbitrary subset of views are missing at input. For this, we borrow from the idea of dropout; every batch, we drop a different, random subset of views while forcing the reconstruction of all views. In this way, the training encourages the latent representation to exploit redundancy of information across different views.

This is similar to the Variational Auto-Encoder with Arbitrary Conditioning (VAEAC) Ivanov et al. (2019), which is a generative model for estimating arbitrary missing feature values in data (eg. in-painting). While theirs is a single-view approach, similar to RMAE, they also consider sampling ”dropped” features from some prior distribution. However, our approach allows us to learn view-specific encoders, since we can exploit the view-structure in our data. In our proposed work, we will look at generative modeling of multi-view data in a similar fashion to the VAE-AC, as well as using flow-based models.

To emulate dropout more appropriately, we perform a relative scaling of the input to the encoders based on number available views. In dropout with probability p, the output of the used units are scaled by $\frac{1}{p}$ to compensate for the missing unites. Similarly, we scale the available views by $\frac{K}{K_a}$ where K_a is the number of available views during that iteration. However, unlike in unit-level dropout, we also do this during test time when we have missing views.

We can also change where the view-dropout takes place; we can either zero out the input X_i or we can zero out the latent encoding L_1. The former method is similar to encouraging every individual view encoder to output an informative ”mean” embedding which works well in lieu of missing data. The latter localizes the ”robustness” of the encoding to the meta-encoder level.

4 Generative Modeling Extension

Now, we try to look at a more natural way to represent multi-view data, namely as an underlying data generation process with the views as observation models into it. In this paper, we consider flow-based models like RealNVP Dinh et al. (2016) and NICE Dinh et al. (2014) to extend the RMAE.

4.1 Background: Flow-based Generative modeling

We will quickly provide some background on Flow-based generative modeling. The philosophy behind these approaches is that a ”good” representation of the data is one in which the data has a simple distribution. To achieve this, they learn an invertible encoding $q(\dot{})$ of the data into a space where this is the case. The pdf of the (single-view) latent distribution can be represented in terms of the data distribution (or vice-versa), using the change-of-variables theorem:

$$p_{\mathcal{X}}(x) = \left|\det \frac{dq}{dx}\right| p_L(q(x)) \tag{1}$$

The invertible function q is designed to have a triangular Jacobian, allowing efficient computation of the determinant. Even with this restriction, we can design q which have a lot of representative power. This is typically achieved by composing sequence of invertible transforms, each with appropriate Jacobians. These transforms are usually ”coupling” transforms , which allow us to represent useful inter-dependence structure between the transformed covariates.

Equation 1 is gives the likelihood the training procedure optimizes. The latent distribution can either be a simple distribution like a standard Gaussian, or a trainable one like a mixture model and/or an AutoRegressive model.

4.2 Flow-Based RMAE

For our flow-based model, we primarily consider the case where we have all the view data available during training and testing. Our architecture remains largely the same as the one represented in Figure 1. The individual view-encoders can be flow-based models or independently trained auto-encoders; the intermediate shared representation is then fed through a flow-based invertible encoding to give the final encoding.

In the case of missing views, we can apply the idea behind AC-Flow Li et al. (2019). Here, we would condition on an arbitrary subset of missing views to learn a flexible flow-transform for each view. We leave this for the immediate next step for our future work.

5 Experiments

For our experiments, we first look at some synthetic data to demonstrate the applicability of our approaches. We follow this up with some real world experiments where we use the learned representation as features for down-stream classification tasks.

5.1 Synthetic Experiments

We design our synthetic datasets with structured redundancies between views. We look at the case where, for K views, we need $K-1$ views to reconstruct the last. We achieve this by having each view i sharing half of its features with view $i-1$, and the other half with view $i+1$, with view 0 and K wrap around to each other. The views are then independently transformed (so as to not have the exact same observable covariates), and perturbed by independent noise.

5.1.1 RMAE

Here are the baseline methods we compare against for the RMAE:

- *Multi-view Feature Concatention (CAT)*
  A simple alternative method for representation learning is to simply concatenate the multiple available views into a single feature vector. While this approach preserves all the information as contained by the different views, it is not robust to missing views, since it does there is no inter-relationship modeling.

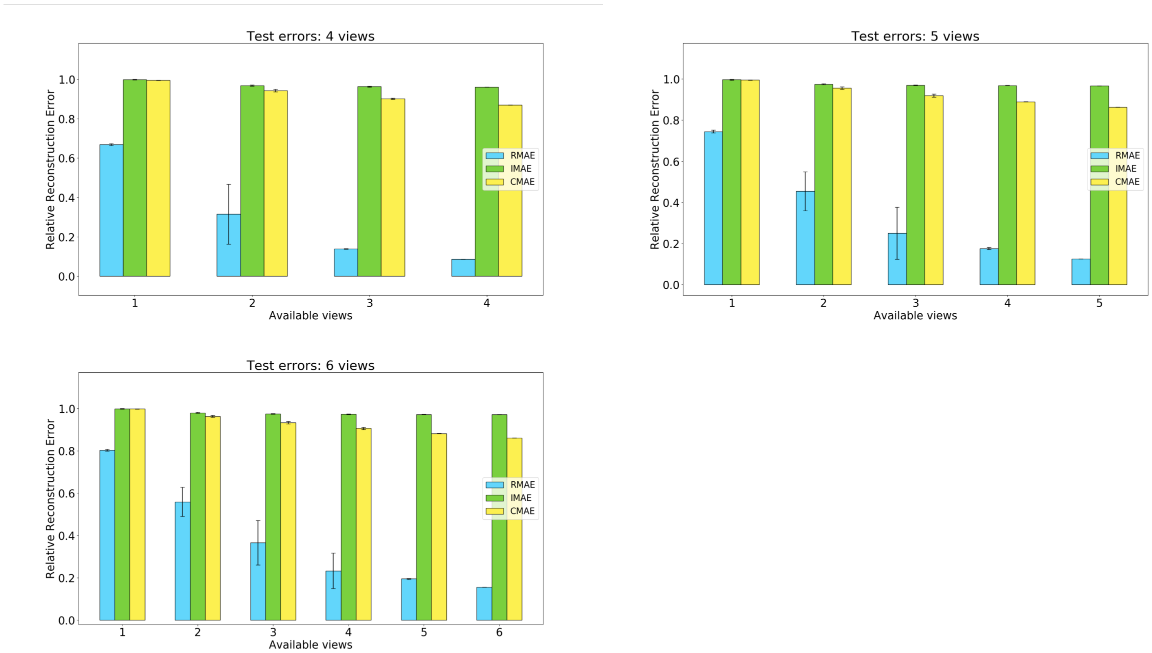

Figure 2: Train and test reconstruction error vs. number of views for different AE competitors. This plot also shows the 1-sigma confidence interval for different choices of available views.

- *Intersection Multi-view Auto-Encoder (IMAE)*

  This baseline is an alternative multi-view AutoEncoder approach which forces all the views to share a single bottle-neck representation. The final code is represented as the average code from all the views. This method tends to extract the intersection of the information contained between the latent spaces, and is thus not as robust to relationships and redundancies local to only a subset of views.

- *Concatenation AutoEncoder (CMAE)*

  Here, we first concatenate the multi-view features just as the first baseline, but train an AutoEncoder above this to learn inter-view relationships. This approach is the middleground between the RMAE and just simple feature concatenation; and can be seen as skipping the first level of view-specific encoders in the RMAE framework.

  While this approach is the closest in spirit to the RMAE, it lacks the initial feature transformation/encoding which often helps unravel inter-view relationship structure.

We look at $K = 4, 5, 6$ for our experiments, with each view having 6 dimensions (3 shared with the previous view and the rest shared with the next view). Figure 2 shows the reconstruction error for all the views, against the number of available views. The trend is intuitive for the RMAE; the reconstruction error improves as we include more views. The confidence intervals in the bar plots represents the variance over choices of input views, for a given number of available views. In the case of CMAE and IMAE, these intervals are small because there is little difference over choices of different views – the representation does not learn robust relationships across different views for reconstruction.

We also have single-view error matrices for $K = 6$ shown in Figure 3. Here, an element (i, j) of the matrix represents the error from using a single view i for the reconstruction of a single view j. Note: We only show this for $K = 6$ but the others look the same. We expect the banded diagonal structure, since by design, each of our views is constructed to share features only with the next and previous views. Again, RMAE demonstrates that it is better able to uncover these local relationships between the views.

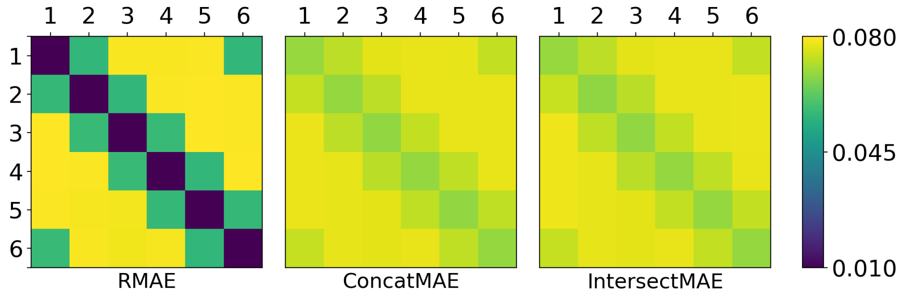

Figure 3: [6-view problem] One-to-one single-view reconstruction for different AE competitors.

5.2 Flow-based RMAE

Here, we look at the how well our generative model can recover the underlying data distribution of our synthetic datasets. We compare the relative performance of using a simple gaussian underlying distribution vs. an AutoRegressive model for our flow-based approaches. In these experiments, we train our models on the training set and look at how well their generated samples compare with the test set. Here, we look at $K = 3$ with each view having 6 dimensions.

Figure 4 shows samples drawn from the simple base distribution and the learned AR base distribution, as compared to true samples from the test-set. We project the data into 2D space using a `umap` embedding to better visualize the differences between the samples. Here, we see that the AR base model is able to better represent the data, as compared to the simple gaussian base distribution.

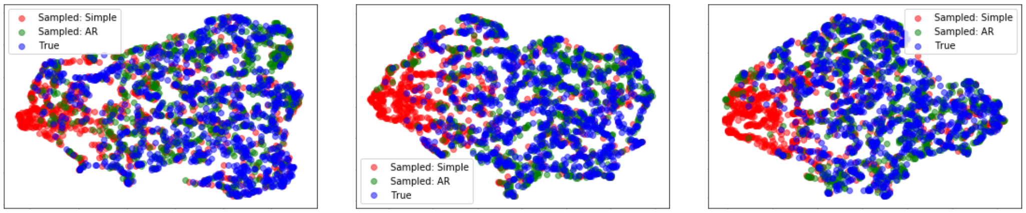

Figure 4: This figure shows samples as generated from the simple gaussian and AR base distributions as they compare with test samples from the true distribution.

5.2.1 Real-world Experiments

We only consider RMAE for our current real-world experiments. Here, we look at the usefulness of the learned representation for down-stream classification tasks on real-world datasets. Each method is trained similarly with view-dropouts as before, and the test-time latent representation is used as features for the downstream tasks.

The datasets we consider are:

- *3 Sources News Dataset*[1] This dataset consists of featurized news articles from three sources: BBC, Guardian, Reuters.
- *NUS-Wide-Lite* Chua et al. (July 8-10, 2009) This is an image dataset where each image is associated with one or more of 81 concepts like ”lake” or ”person”; the views are different image featurizations.
- *N-MNIST* Basu et al. (2017) This dataset consists of three noisy versions of the original MNIST dataset as the different views.

[1] http://mlg.ucd.ie/datasets/3sources.html

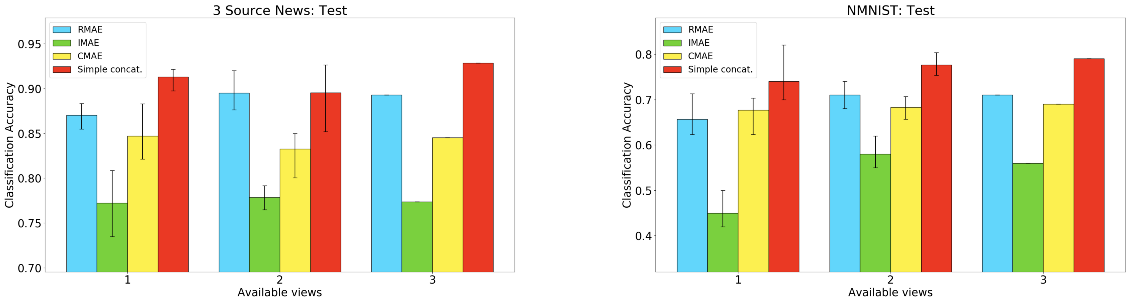

Figure 5: [3-Source News + NMNIST] Plots for accuracy vs. number of available views for the different approaches.

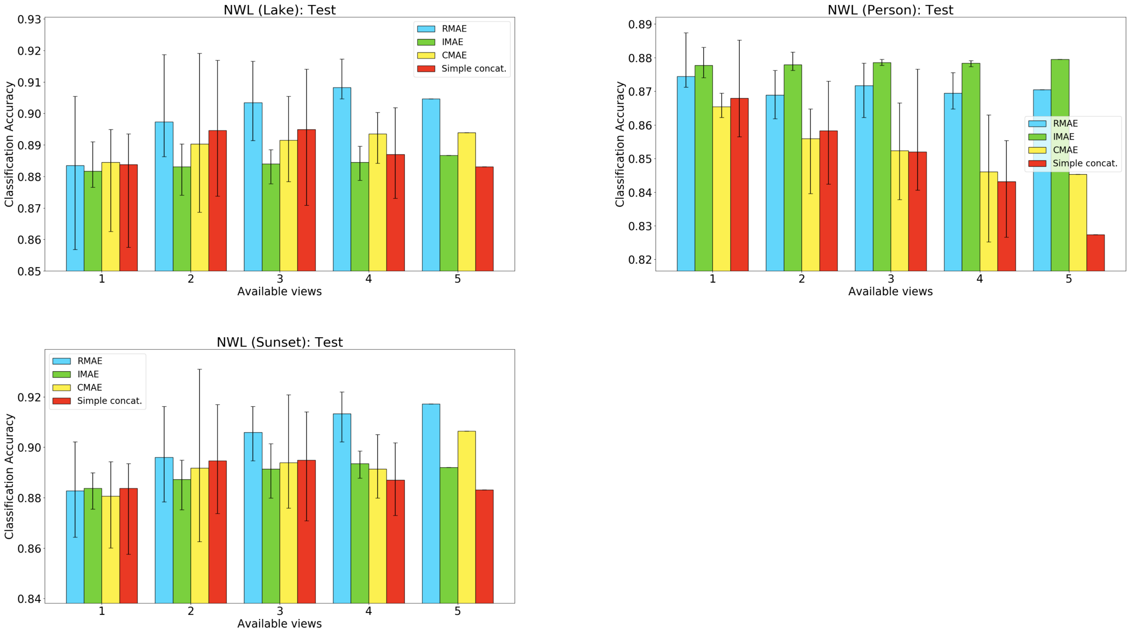

Figure 6: [NUS-WIDE-Lite: Sunset] Plots for accuracy vs. number of available views for the different approaches.

In Figures 5 and 6, we see that different methods win out over different datasets. Leveraging local relationships may be detrimental in cases where the global shared structure is the most important for solving the task. However, we note that RMAE consistently is either the best or the second best for all the datasets, over most of view-subsets. This shows us that there is indeed some generalizable structure that the RMAE is able to uncover which is useful for the tasks at hand.

This can likely be improved for all methods by simultaneously training the representation learning as well as the downstream tasks. Currently, the representations learned are agnostic to the task at hand, and are specifically tailored to reconstruction. Incorporating the application in the training process would likely help improve the performance of the learned representations on the classification tasks.

6 Conclusion

In this paper, we considered the problem of Robust Multi-view Representation Learning where we sought to leverage relationships between views to learn representations of multi-view data. We proposed two methods, one based on view-dropout and its flow-based generative modeling extension. Synthetic and real world experiments show promising results for their application to missing data reconstruction as well as other down-stream learning tasks.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, 2015.

Saikat Basu, Manohar Karki, Sangram Ganguly, Robert DiBiano, Supratik Mukhopadhyay, Shreekant Gayaka, Rajgopal Kannan, and Ramakrishna Nemani. Learning sparse feature representations using probabilistic quadtrees and deep belief nets. *Neural Processing Letters*, 45(3):855–867, 2017.

Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yan-Tao Zheng. Nus-wide: A real-world web image database from national university of singapore. In *Proc. of ACM Conf. on Image and Video Retrieval (CIVR'09)*, Santorini, Greece., July 8-10, 2009.

Laurent Dinh, David Krueger, and Yoshua Bengio. Nice: Non-linear independent components estimation, 2014.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp, 2016.

Jing Huang and Brian Kingsbury. Audio-visual deep learning for noise robust speech recognition. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 7596–7599. IEEE, 2013.

Oleg Ivanov, Michael Figurnov, and Dmitry Vetrov. Variational autoencoder with arbitrary conditioning. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=SyxtJh0qYm.

Yelin Kim, Honglak Lee, and Emily Mower Provost. Deep learning for robust feature generation in audiovisual emotion recognition. In *2013 IEEE international conference on acoustics, speech and signal processing*, pp. 3687–3691. IEEE, 2013.

Yang Li, Shoaib Akbar, and Junier B. Oliva. Flow models for arbitrary conditional likelihoods, 2019.

Youssef Mroueh, Etienne Marcheret, and Vaibhava Goel. Deep multimodal learning for audio-visual speech recognition. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2130–2134. IEEE, 2015.

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. 2011.

Wanli Ouyang, Xiao Chu, and Xiaogang Wang. Multi-source deep learning for human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2329–2336, 2014.

Carina Silberer and Mirella Lapata. Learning grounded meaning representations with autoencoders. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 721–732, 2014.

Nitish Srivastava and Ruslan Salakhutdinov. Learning representations for multimodal data with deep belief nets. In *International conference on machine learning workshop*, volume 79, 2012.

Daixin Wang, Peng Cui, Mingdong Ou, and Wenwu Zhu. Deep multimodal hashing with orthogonal regularization. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

Zuxuan Wu, Yu-Gang Jiang, Jun Wang, Jian Pu, and Xiangyang Xue. Exploring inter-feature and inter-class relationships with deep neural networks for video classification. In *Proceedings of the 22nd ACM international conference on Multimedia*, pp. 167–176, 2014.

TengQi Ye, Tianchun Wang, Kevin McGuinness, Yu Guo, and Cathal Gurrin. Learning multiple views with orthogonal denoising autoencoders. In *International Conference on Multimedia Modeling*, pp. 313–324. Springer, 2016.