Separating Multimodal Modeling from Multidimensional Modeling for Multimodal Learning

Divyam Madaan [1] Taro Makino [1] Sumit Chopra [1 2] Kyunghyun Cho [1 3 4]

Abstract

Supervised multimodal learning is defined as learning to map a set of separate modalities to a target. Despite its intuitive definition, it is unclear whether one should model this problem using a multidimensional model, where the features from all the modalities are concatenated and treated as multidimensional features from a single modality or a multimodal model, where we use the information about the modality boundaries. In this work we formalize the framework for supervised multimodal learning and identify the conditions that favor multimodal modeling over multidimensional modeling. It is advantageous when the dependency across or within modalities shift during test time. Through a series of synthetic experiments, where we fully control the data generation process, we verify the necessity of multimodal modeling for solving a supervised multimodal learning problem. Our proposed framework is agnostic to any assumptions pertaining to model architectures and can have a widespread impact by informing modeling choices when dealing with data from different modalities.

1. Introduction

The problem of supervised multimodal learning involves mapping input data to a target, where the data is generated from multiple domains and the information about the boundaries between different domains (a.k.a., modality grouping) is accessible. This problem has garnered interest in numerous applications, such as autonomous vehicles (Xiao et al., 2020), clinical efficiency (Soenksen et al., 2022), robotics (Driess et al., 2023) and so on. Despite the availability of diverse data sources in numerous approaches (Baltrušaitis et al., 2018; Wang et al., 2020; Akkus et al., 2023), we often observe that they fail to obtain performance that is even on-par with the unimodal predictors (Goyal et al., 2017; Wu et al., 2022; Dancette et al., 2021; Si et al., 2022b).

In this work, motivated by the goal of bringing more structure to solving the multimodal learning problem, we address the question of *whether and when we benefit from the modality grouping information*. We adopt a probabilistic formulation and describe two modeling paradigms for multimodal learning, *multidimensional* and *multimodal* modeling. Multidimensional modeling clubs the input data features from each modality disregarding the modality grouping information. On the other hand, multimodal modeling leverages the modality grouping by separately learning dependencies among features within each modality (a.k.a. intra-modality dependencies) and among features across different modalities (a.k.a. inter-modality dependencies), and subsequently combining them. We demonstrate that when the distributions of the training and the test sets are identical, the generalization performance of the two modeling approaches is indistinguishable. Multimodal modeling proves to be advantageous over multidimensional modeling when encountering specific types of distribution shifts, such as shifts in the inter- and intra-modality dependencies between training and test distributions because we can explicitly control the contribution of these dependencies.

2. What is multimodal learning?

Multimodal learning refers to a problem setup where the input is expressed as a set of observations from different modalities. Unlike conventional unimodal learning, multimodal learning can exploit the information from multiple modalities for modelling purposes. In this work, we are interested in supervised multimodal learning, where the goal is to map this set of multiple modalities as input to the target.

We begin with the dataset $\mathcal{D} = \{(x_i, y_i)_{i=1}^n\}$. Without loss of generality, y_i is the binary output and $x_i \in \mathbb{R}^n$ represents the input attributes, where n is even with two modalities, $x = [x_1, \ldots, x_{n/2}]$ and $x' = [x_{n/2+1}, \ldots, x_n]$. We use $[x'_1, \ldots, x'_{n/2}]$ to indicate the features of the second modality, emphasizing that these features belong to x'.

[1] New York University [2] New York University Grossman School of Medicine [3] Genentech [4] CIFAR. Correspondence to: Divyam Madaan <divyam.madaan@nyu.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

In multimodal learning, a natural question that arises is *whether and when do we benefit from using information about modality boundaries for multimodal learning.* This question is interesting because we can build a multimodal learner that does not exploit the modality grouping information by concatenating all the modalities. To answer this question, we define multidimensional and multimodal modeling in the following section: multidimensional modeling disregards modality grouping, and multimodal modeling exploits this information.

3. Multidimensional modeling vs. multimodal modeling

Several studies have proposed ways to incorporate modality boundary information, primarily by building novel architectures (Katsaggelos et al., 2015; Lin et al., 2017; Cadene et al., 2019; Radford et al., 2021; Wu et al., 2022). We however shift our focus towards the probabilistic formulation rather than the parameterization to study multimodal learning. In doing so, we describe two modeling paradigms for multimodal learning. First, multidimensional modeling disregards the modality grouping information by concatenating the modalities. Second, multimodal modeling incorporates the modality grouping information directly. In this section, we elaborate the details of these two modeling paradigms.

3.1. Multidimensional modelling

Multidimensional modeling considers all the dimensions together as a single modality without exploiting the modality boundaries. In addition to the input x and output y, we introduce a match variable m in order to capture the statistical dependencies across the input features given the label. This match variable is a binary random variable that is conditioned on all the input features and the target. We only observe the instances where the match condition is satisfied (i.e., when $m = 1$):

$$p(x, y, m = 1) = p(y) \left(\prod_{i=1}^{n} p(x_i|y) \right) p(m = 1|x_1, \ldots, x_n, y). \tag{1}$$

We characterize this data generating process as a *multidimensional data generating process*. We use this mechanism to break the conditional independence among the input dimensions given the label, which is often referred to as the 'explaining away' phenomenon. That is,

$$p(x|y) \neq \prod_{i=1}^{n} p(x_i|y). \tag{2}$$

Under this data generating process, the predictive probability over the label y encompasses both the marginal probability of the label and the relationship between the features and the label. This can be expressed formally as follows:

$$\log p(y|x, x') = \log p(y) + \log p(x, x'|y) + \text{const}. \tag{3}$$

The second term on the right hand side cannot be expressed as a sumamtion of $\log p(x|y)$ and $\log p(x'|y)$ because x and x' are not conditionally independent of each other. Thus, this predictive probability clearly shows that it is necessary to model the interaction among all the observational dimensions in order to properly predict the label.

3.2. Multimodal modeling

Multimodal modeling differs from multidimensional modeling by levaraging information about the modality boundaries. From here on, we refer the match variable m from multidimensional modeling as global match variable. We incorporate the modality boundary information by introducing modality specific match variables u and u' that are associated with the modalities x and x' respectively. These match variables are always set to one. As a result, these match variables give rise to two types of dependencies due to the phenomenon of explaining away, as discussed earlier.

The first type of dependency is *inter-modality dependency*, which is induced by the global match variable m between different modalities. We denote it by $p(m = 1|x, x', y)$. The second type of dependency is *intra-modality dependency*, which is induced by the modality-specific match variables u and u' separately within the features of each modality. The degrees of match is captured by $p(u = 1|x, y)$ and $p(u' = 1|x', y)$.

Multimodal modeling captures inter and intra-modality dependencies separately and combines them later. The joint probability including the global and modality-specific match variables, can be expressed as follows:

$$\begin{aligned} p(x, x', y, m = 1, u = 1, u' = 1) = p(y) \\ \times \underbrace{\left(\prod_{i=1}^{\lfloor n/2 \rfloor} p(x_i|y) \right) p(u = 1|x, y)}_{\text{Modality 1}} \\ \times \underbrace{\left(\prod_{i=1}^{\lfloor n/2 \rfloor} p(x'_i|y) \right) p(u' = 1|x', y)}_{\text{Modality 2}} \\ \times p(m = 1|x, x', y). \end{aligned} \tag{4}$$

We refer to this data generating process as the *multimodal data generating process*. The predictive distribution under this data generating process captures the influence of both the individual modalities on the label, as well as the combined impact of multiple modalities on the label.

We can write this as

$$\log p(y|x,x') = \log p(y) + \log p(x|y) + \log p(x'|y) + \log p(x,x'|y) + \text{const.} \quad (5)$$

3.3. Relationship between multidimensional and multimodal modeling

We can define multidimensional modeling and multimodal modeling as two extremes of a spectrum, with the joint probability represented as follows:

$$\begin{aligned} p(x,x',y,m=1,&u=1,u'=1) \propto p(y) \\ &\times \underbrace{\left(\prod_{i=1}^{\lfloor n/2 \rfloor} p(x_i|y)\right) p^{\alpha}(u=1|x,y)}_{\text{Modality 1}} \\ &\times \underbrace{\left(\prod_{i=1}^{\lfloor n/2 \rfloor} p(x'_i|y)\right) p^{\alpha}(u'=1|x',y)}_{\text{Modality 2}} \\ &\times p^{\beta}(m=1|x,x',y). \end{aligned} \quad (6)$$

The presence and influence of the dependencies, both intra-modality and inter-modality, can be regulated by parameters α and β. When α is set to zero, we revert to the joint distribution of multidimensional modeling in Equation (1), where there is no consideration of modality boundaries. Similarly, when β is equal to zero, the joint probability simplifies to an ensemble of unimodal predictors, where there is no dependency across the modalities. On the other hand, when both α and β are greater than zero, we recover the joint distribution of multimodal modeling in Equation (4), which exploits the information about modality boundaries.

4. When do these modeling paradigms differ?

In this section, we distinguish multidimensional and multimodal modeling approaches by highlighting two specific types of shifts related to the presence of modality grouping information. More specifically, we focus on how the inter- and intra-modality dependencies change between training and test times. They may not change, change together or change separately from each other. These shifts can stem from factors such as variations in data collection methods, dissimilar characteristics of training and testing datasets, or external influences that affect the distribution of the global and modality-specific match variables.

To provide readers with a more concrete example, consider visual question answering (VQA), a task that involves answering an open-ended question using information from an associated image. Dancette et al. (2021); Si et al. (2022a) showed that VQA contains exploitable shortcuts associated with either image-only information, text-only information or a combination of both. These unimodal and multimodal shortcuts are decision rules that perform well when there is no distribution shift; they however fail to generalize effectively when faced with certain type of distribution shifts.

4.1. With inter-modality dependency shifts

The change in the conditional distribution of the global match variable $p(m=1|x,x',y)$ (a.k.a multimodal shortcuts) is what we define as the inter-modality dependency shift. Especially, we are interested in the case where the inter-modality dependency that exists during training disappears or weakens during test time. This is equivalent to having a smaller β at the test time in Equation (6).

When β decreases, the dependency between modalities weakens. This is equivalent to changing the term $\log p(m=1|x,x',y)$ in Equation (5). We model this by introducing a parameter ν as a coefficient of the predictive term $\log p(m=1|x,x',y)$ in the predictive probability. This parameter allows us to effectively cope with the shift in the inter-modality dependency. The updated formulation of the predictive distribution for multimodal modeling is

$$\log p(y|x,x') = \log p(y) + \log p(x|y) + \log p(x'|y) + \nu \cdot \log p(x,x'|y) + \text{const.} \quad (7)$$

During the inference phase, when such a distribution shift occurs, we would ideally adjust the value of ν to a smaller value. This is reflected in the vanishing of the predictive term $\log p(x,x'|y)$ towards zero.

In contrast to multimodal modeling, in multidimensional modeling we do not capture the unimodal relationships between the input and output. Thus, we cannot weaken the inter-modality dependency separately from the intra-modality dependencies. This differentiates multimodal modeling from multidimensional modeling.

4.2. With intra-modality dependency shifts

The shift in the dependencies within a modality $p(u=1|x,y)$ and $p(u'=1|x',y)$ (a.k.a unimodal shortcuts), are defined as the intra-modality dependency shifts. Similar to the previous case, we are interested in the case where the intra-modality dependency that exists during training weakens during testing. This is equivalent to having a smaller α at the test time in Equation (4). During inference, to control the strength of the intra-modality shifts, we adjust the values of α towards zero.

As α decreases, the dependencies within the modalities weaken, which can be expressed by reducing the impact of the terms $\log p(x|y)$ and $\log p(x'|y)$ in Equation (6). Thus, to cope with the intra-modality dependency shift, we introduce parameter μ as coefficient of the terms $\log p(x|y)$ and $\log p(x'|y)$, allowing us to adjust the contributions of these terms in the predictive distribution. We do not have prior knowledge about which modality exhibits a stronger dependency, hence we use the same coefficient for both modalities. The updated formulation of the predictive distribution can be expressed as

$$\log p(y|x,x') = \log p(y) + \mu \cdot \log p(x|y) + \mu \cdot \log p(x'|y) + \log p(x,x'|y) + \text{const.} \tag{8}$$

When the intra-modality dependency shifts during the testing time, we adjust μ to a smaller value. to reduce the contribution of the predictive terms $\log p(x|y)$ and $\log p(x'|y)$ for each modality.

Although we model the intra-modality dependency shifts, but it is much more realistic for the inter-modality dependency to shift. This is because the features with a modality are more tightly coupled because they tend to be acquired from a single sensory device. Empirically based on the prior study (Dancette et al., 2021) we do see that in VQA majority of the shortcuts (approximately 90%) are due to the shift in the inter-modality dependencies.

4.3. What if there was no distributional shift?

In scenarios where the distributions remain unchanged between training and test times, there is no distinction between multidimensional and multimodal modeling approaches, in principle. This is because multidimensional modeling subsumes the modality-specific terms in Equation (5) within $\log p(x,x'|y)$. While this observation may appear trivial, we empirically verify it using synthetic data in Section 6, where both approaches achieve identical performance when we solely focus on evaluating their performance on identically distributed test set.

5. Discussion and related work

In the previous section, we examined the circumstances under which multimodal and multidimensional paradigms differ from each other, focusing on intra- and inter-modality distribution shifts in multimodal learning. In this section, we differentiate our work from earlier attempts and delve into some properties of our analysis.

Distribution shifts. Distributional shifts in machine learning are being studied increasingly more so in recent years (Arjovsky et al., 2019; Lipton et al., 2018; Ruan et al., 2022; Castro et al., 2020). These studies however are neither well-specified nor are generically applicable to a broad set of models. Here, we instead focus on carefully defining and studying inter- and intra-modality dependency shifts that are relevant to multimodal learning, where we have modality grouping information. These kind of shifts have been observed in some of the existing VQA benchmarks (Goyal et al., 2017; Dancette et al., 2021; Si et al., 2022a). In this work, rather than concentrating on a particular dataset, we provide a general framework to study and express the shifts in inter- and intra-modality dependencies mathematically.

Selection bias. To capture the dependencies among different dimensions, we use match variables, similar to the use of selection variables in modeling selection bias (Hernán et al., 2004; Cortes et al., 2008; Bareinboim and Pearl, 2016). Selection bias occurs when certain examples are excluded from the dataset (Horwitz and Feinstein, 1978; Heckman, 1979; Robins, 2001). In real-world scenarios, it is commonly observed that the distribution of the selection variables deviates between the training and testing phases (Hernán et al., 2004; Cortes et al., 2008; Bareinboim and Pearl, 2016).

Parametrization-free analysis. Our work stands out from existing research on multimodal learning (Katsaggelos et al., 2015; Lin et al., 2017; Cadene et al., 2019; Radford et al., 2021; Wu et al., 2022) in a significant way – we do not try to come up with a better parametrization of a conditional distribution over the label given multiple modalities, but instead look at the probabilistic formulation, independent of specific parametrization of conditional distributions. By doing so, we illustrate that the prevailing strategy of creating novel architectures may not yield a desired outcome in multimodal learning. We instead look at the probabilistic generative model of multimodal learning and identify a specific case of distributional shift under which multimodal modeling can be more beneficial than multidimensional modeling.

In the absence of inter-modality dependency. If we already know in advance that intra-modality dependencies are much stronger and important than inter-modality dependencies, then the best approach would be to ignore the inter-modality dependency term in Equation (7). When the value of ν approaches zero in an extreme case, multimodal modeling simplifies to an ensemble of unimodal predictors due to the lack of meaningful correlation between the modalities (Wu et al., 2022; Makino et al., 2022).

Consider an example of tiger detection, where $y = 1$ (indicating the presence of a tiger). The first modality is shape information, and the second modality is texture information. When $y = 0$, the first modality would have a random shape, and the second modality a random texture. Conversely, when the label is “tiger”, i.e., $y = 1$, first modality (shape) has a tiger-like shape, regardless of the second modality (texture) and the same applies to the texture modality. This implies that $\beta = 0$ i.e., there is no special dependency between these two modalities.

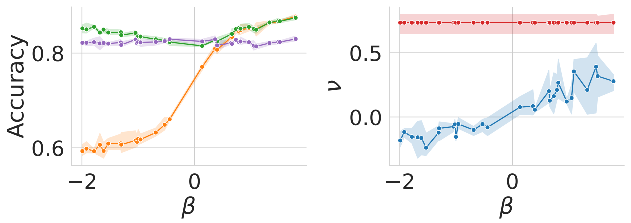

Figure 1: **Results for Inter-modality dependency shift. (Left)** Accuracy comparison between multidimensional, ensemble of unimodal predictors and multimodal modeling with varying β during inference. **(Right)** Change in optimal ν with changing β during inference. Red curve represents the ν obtained during training.

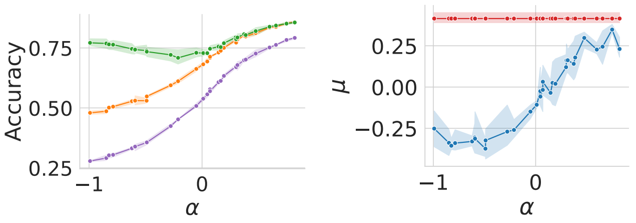

Figure 2: **Results for Intra-modality dependency shift. (Left)** Accuracy comparison between multidimensional, ensemble of unimodal predictors and multimodal modeling with varying α during inference. **(Right)** Change in optimal μ with changing α during inference. Red curve represents the μ obtained during training.

6. Experiments

To differentiate multidimensional and multimodal modeling, we examine the shifts in the inter- and intra-modality dependency during testing. We create a synthetic data generating process that allows us to control the strength of these dependencies and show a clear advantage of multimodal modeling when there is a shift in these dependencies.

6.1. Experimental Setup

The label y is generated from a Bernoulli distribution and six features $x = \{x_0, \dots x_5\}$ are drawn from a normal distribution with a mean of $(0.1i + 0.1) \cdot (y - 0.5)$ for x_i. m follows a Bernoulli distribution, with its mean determined by $\sigma'(\beta x_1 x_2 x_3, 0.1) \cdot y + \sigma'(-\beta x_1 x_2 x_3, -0.3) \cdot (1-y)$. $\sigma'(x, \lambda) = \frac{1}{1+\exp(-(x+\lambda))}$ denotes the shifted sigmoid function. Similarly, u and u' follow Bernoulli distributions with means $\sigma'(\alpha x_0 x_1, 0.3) \cdot y + \sigma'(-\alpha x_0 x_1, -0.5) \cdot (1-y)$. and $\sigma'(\alpha x_4 x_5, 0.2) \cdot y + \sigma'(-\alpha x_4 x_5, -0.4) \cdot (1-y)$ respectively. β and α are used to control the intra- and inter-modality dependencies in the multimodal data generating process. We generate a total of 20,000 samples, with an equal distribution of 10,000 samples for each class. These samples were then divided into three subsets: 60% for training, 20% for validation, and 20% for testing.

6.2. Quantitative results

Performance with inter-modality dependency shift. To simulate the shift in inter-modality dependency, we fix α to one and β to two during the training phase. During testing, we introduce variation in the value of β within the range of $-\beta$ to β, thereby replicating the desired shift. Figure 1 shows the impact of inter-modality dependency shift on different modeling paradigms. As expected, multidimensional modeling experiences a significant decrease in accuracy when this dependency weakens. Conversely, multimodal modeling with optimal ν and an ensemble of unimodal predictors consistently outperforms multidimensional modeling. To verify our hypothesis, the optimal ν was found using random search on the validation set of the shifted distribution. As β increases, optimal ν proportionally increases as well since ν captures the strength of the inter-modality dependency in the predictive distribution.

Performance with intra-modality dependency shift. We replicate the inter-modality dependency shift by keeping α and β constant to one during training. Then, during testing, we vary α within the range of $-\alpha$ to α. The results are demonstrated in Figure 2. The accuracy of the unimodal ensemble experiences a substantial decline as the intra-modality dependency weakens. Similarly, multidimensional modeling's performance is also affected, although it performs relatively better. On the other hand, multimodal modeling consistently achieves higher accuracy across all values of α. As expected, the optimal μ obtained through random search exhibits a consistent upward trend even when intra-modality dependency shifts during testing time.

7. Conclusion

In this work, we focus on supervised multimodal learning, where in addition to the relationship between inputs and outputs, we have the modality grouping information. Taking a probabilistic perspective, we explore two modeling paradigms for multimodal learning: *multidimensional modeling*, which treats the features as a single modality and *multimodal modeling*, which takes into account the modality grouping information. We distinguish these two modeling paradigms by highlighting two specific types of shifts relevant to multimodal learning. Specifically, we show the advantages of multimodal modeling in the presence of shift in the inter- and intra-modality dependencies. Conversely, when these dependencies remain unchanged between training and testing, these modeling paradigms in principle collapse onto each other.

Acknowledgement

This research was supported by Samsung Advanced Institute of Technology (Next Generation Deep Learning: from pattern recognition to AI), Hyundai NGV (Uncertainty in Neural Sequence Modeling in the Context of Dialogue Modeling), NSF Award 1922658, the Center for Advanced Imaging Innovation and Research (CAI2R), a National Center for Biomedical Imaging and Bioengineering operated by NYU Langone Health, and funded by the National Institute of Biomedical Imaging and Bioengineering through award number P41EB017183. This content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies.

References

C. Akkus, L. Chu, V. Djakovic, S. Jauch-Walser, P. Koch, G. Loss, C. Marquardt, M. Moldovan, N. Sauter, M. Schneider, et al. Multimodal deep learning. *arXiv preprint arXiv:2301.04856*, 2023.

M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

T. Baltrušaitis, C. Ahuja, and L.-P. Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 2018.

E. Bareinboim and J. Pearl. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 2016.

R. Cadene, C. Dancette, M. Cord, D. Parikh, et al. Rubi: Reducing unimodal biases for visual question answering. *Advances in neural information processing systems*, 32, 2019.

D. C. Castro, I. Walker, and B. Glocker. Causality matters in medical imaging. *Nature Communications*, 2020.

C. Cortes, M. Mohri, M. Riley, and A. Rostamizadeh. Sample selection bias correction theory. In *Algorithmic Learning Theory: 19th International Conference*, 2008.

C. Dancette, R. Cadene, D. Teney, and M. Cord. Beyond question-based biases: Assessing multimodal shortcut learning in visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

D. Driess, F. Xia, M. S. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023.

Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017.

J. J. Heckman. Sample selection bias as a specification error. *Econometrica: Journal of the econometric society*, pages 153–161, 1979.

M. A. Hernán, S. Hernández-Díaz, and J. M. Robins. A structural approach to selection bias. *Epidemiology*, 2004.

R. I. Horwitz and A. R. Feinstein. Alternative analytic methods for case-control studies of estrogens and endometrial cancer. *New England Journal of Medicine*, 1978.

A. K. Katsaggelos, S. Bahaadini, and R. Molina. Audiovisual fusion: Challenges and new approaches. 2015.

D. Lin, G. Chen, D. Cohen-Or, P.-A. Heng, and H. Huang. Cascaded feature network for semantic segmentation of rgb-d images. In *Proceedings of the IEEE international conference on computer vision*, 2017.

Z. Lipton, Y.-X. Wang, and A. Smola. Detecting and correcting for label shift with black box predictors. In *International conference on machine learning*, 2018.

T. Makino, K. J. Geras, and K. Cho. Mitigating input-causing confounding in multimodal learning via the back-door adjustment. In *NeurIPS 2022 Workshop on Causality for Real-world Impact*, 2022.

A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 2021.

J. M. Robins. Data, design, and background knowledge in etiologic inference. *Epidemiology*, 2001.

Y. Ruan, Y. Dubois, and C. J. Maddison. Optimal representations for covariate shift. In *International Conference on Learning Representations*, 2022.

Q. Si, F. Meng, M. Zheng, Z. Lin, Y. Liu, P. Fu, Y. Cao, W. Wang, and J. Zhou. Language prior is not the only shortcut: A benchmark for shortcut learning in vqa. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, 2022a.

Q. Si, F. Meng, M. Zheng, Z. Lin, Y. Liu, P. Fu, Y. Cao, W. Wang, and J. Zhou. Language prior is not the only shortcut: A benchmark for shortcut learning in VQA. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, 2022b.

L. R. Soenksen, Y. Ma, C. Zeng, L. Boussioux, K. Villalobos Carballo, L. Na, H. M. Wiberg, M. L. Li, I. Fuentes, and D. Bertsimas. Integrated multimodal artificial intelligence framework for healthcare applications. *Nature Digital Medicine*, 2022.

Y. Wang, W. Huang, F. Sun, T. Xu, Y. Rong, and J. Huang. Deep multimodal fusion by channel exchanging. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

N. Wu, S. Jastrzebski, K. Cho, and K. J. Geras. Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks. In *Proceedings of the 39th International Conference on Machine Learning*, 2022.

Y. Xiao, F. Codevilla, A. Gurram, O. Urfalioglu, and A. M. López. Multimodal end-to-end autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, 2020.