Disentanglement and Generalization Under Correlation Shifts

Christina M. Funke*[1], Paul Vicol*[2, 3], Kuan-Chieh Wang[2, 3], Matthias Kümmerer†[1], Richard Zemel†[2, 3], Matthias Bethge†[1]
[1]University of Tübingen, [2]University of Toronto, [3]Vector Institute, *shared first, †shared senior

Motivation & Summary

- Exploiting correlations between factors of variation can increase performance on noisy data.
- But correlations are often not robust: they may change between domains, datasets, or applications
- Minimizing the MI between latent subspaces fails when attributes are correlated.
- We enforce subspace independence conditioned on available attributes, which removes only dependencies that are not due to the correlations structure in the data.

Problem Setup

- We have noisy data $\mathbf{x} = g(\mathbf{s})$ where $\mathbf{s} = (s_1, s_2, \dots, s_K)$ are the underlying factors of variation, which may be correlated
- **Goal:** Find a mapping to a latent space, $f(\mathbf{x}) = \mathbf{z} = (\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_K)$ such that we can recover the GT attributes via linear functions $\hat{s}_k = \mathbf{R}_k \mathbf{z}_k \approx s_k$.
- **Goal:** Learn a model robust to correlation shifts: if we train on data where $\mathrm{corr}(s_i, s_j) > 0$, then we want the resulting model to perform well on uncorrelated data $\mathrm{corr}(s_i, s_j) = 0$, or anticorrelated data, $\mathrm{corr}(s_i, s_j) < 0$.

Objective Functions for Disentanglement

1. **Base:** minimizing a supervised loss L (e.g., MSE or cross-entropy), $\sum_{i=1}^K L(\hat{s}_i, s_i)$
2. **Base+MI:** minimizing the unconditional mutual information between subspaces in addition to the supervised loss, $\sum_{i=1}^K L(\hat{s}_i, s_i) + I(\mathbf{z}_1, \dots, \mathbf{z}_K)$
3. **Base+CMI:** minimizing the conditional mutual information between subspaces conditioned on observed attributes, in addition to the supervised loss, $\sum_{i=1}^K [L(\hat{s}_i, s_i) + I(\mathbf{z}_i; \mathbf{z}_{-i} \mid s_i)]$

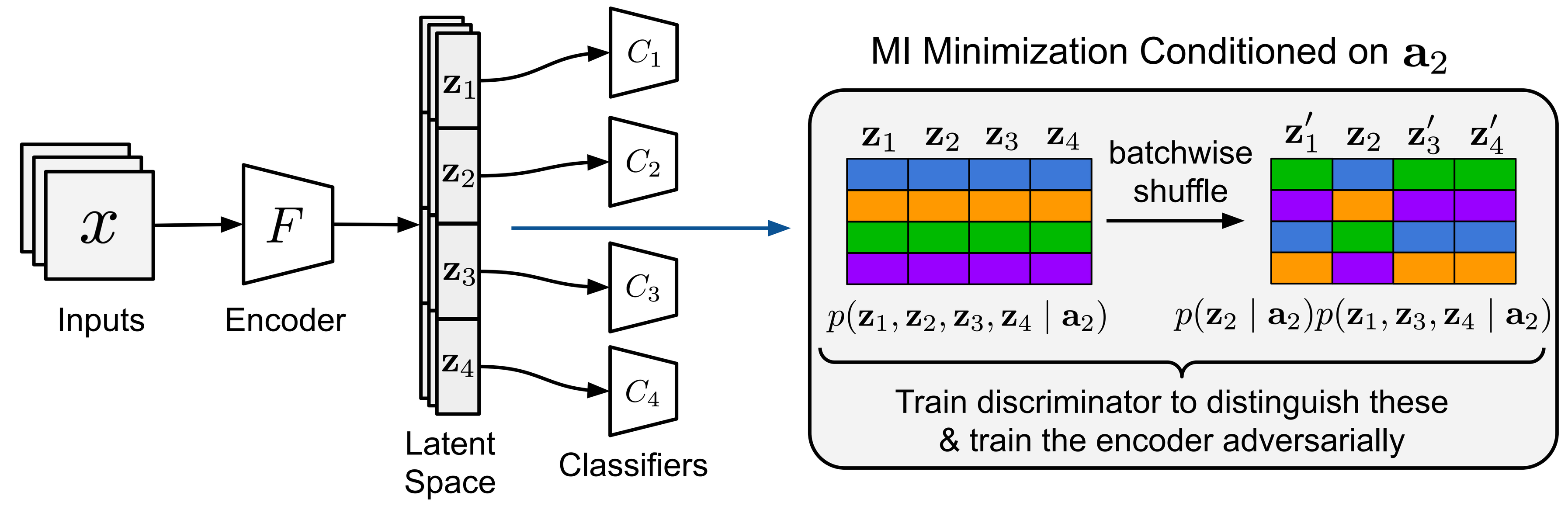
Disentanglement with Correlated Variables

- Consider a linear generative model with correlated Gaussian source variables $\mathbf{s}$, given by:

$$\mathbf{x} = \mathbf{A}\mathbf{s} + \mathbf{n} \quad , \quad \mathbf{s} \sim \mathcal{N}(0, \mathbf{C}_s) \quad , \quad \mathbf{n} \sim \mathcal{N}(0, \mathbf{C}_n)$$

where $\mathbf{C}_s$ and $\mathbf{C}_n$ are covariances for the source and noise variables.

Adversarial Minimization of Conditional Mutual Information



- For most tasks, there is no closed form for MI/CMI. We propose an adversarial approach to minimize CMI, based on batchwise shuffling of latent subspaces.

Full Supervision Does Not Yield Disentanglement

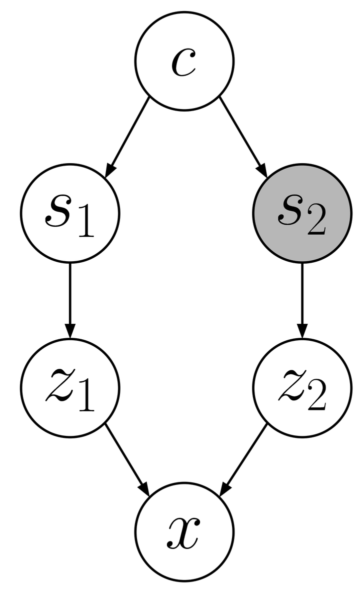
	Base	Base + MI	Base + CMI
VE, Training (Corr = 0.8)	91.9%	69.8%	90.9%
VE, Test (Corr = 0)	87.6%	65.0%	90.9%
M (where $\hat{s} = Mx$)	$\begin{pmatrix} 0.81 & 0.14 \\ 0.14 & 0.81 \end{pmatrix}$	$\begin{pmatrix} 1.07 & -0.46 \\ -0.46 & 1.07 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$

- Performance drops when the correlation between s_1 and s_2 shifts at test time.
- → Tries to make use of the assumed correlation between s_1 and s_2 to counteract information lost due to noise, but this correlation is no longer present.

Unconditional Disentanglement Fails Under Correlation Shift

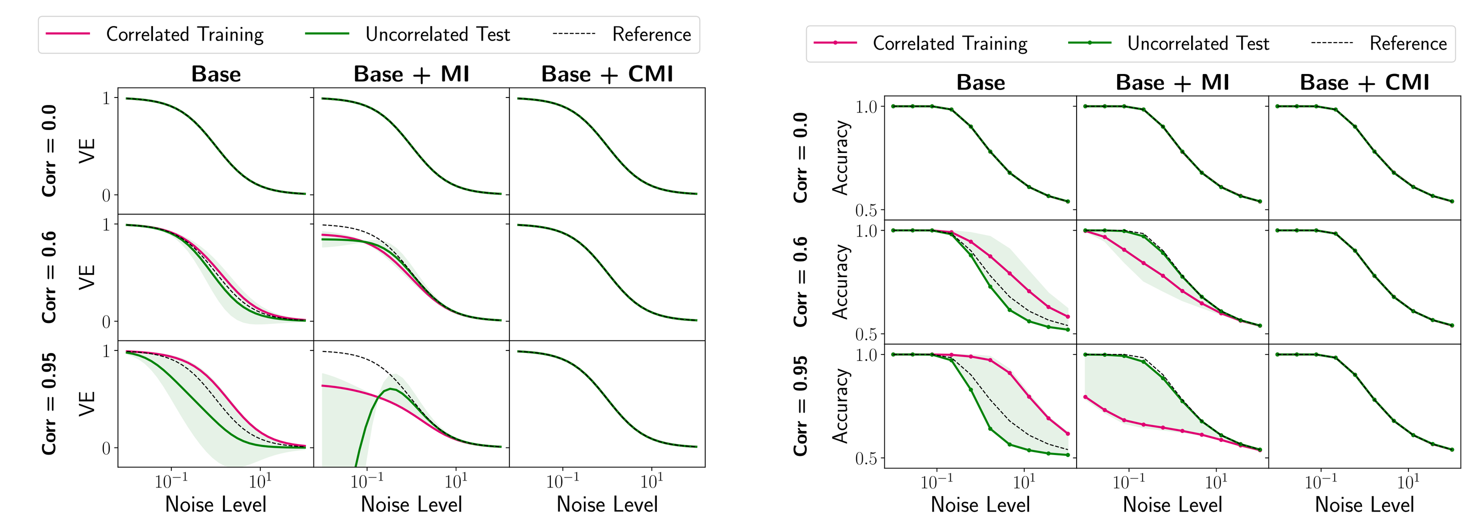
- There is correlation between the sources s_1 and s_2 and therefore $I(s_1; s_2) > 0$.
- By enforcing independence, at least one of the subspaces cannot contain all relevant information about its target value
- The optimal solution under the constraint of minimal MI, $I(\mathbf{z}_1; \mathbf{z}_2) = 0$, fails to model the in-distribution correlated training data.

Conditional Disentanglement is Robust to Correlation Shift



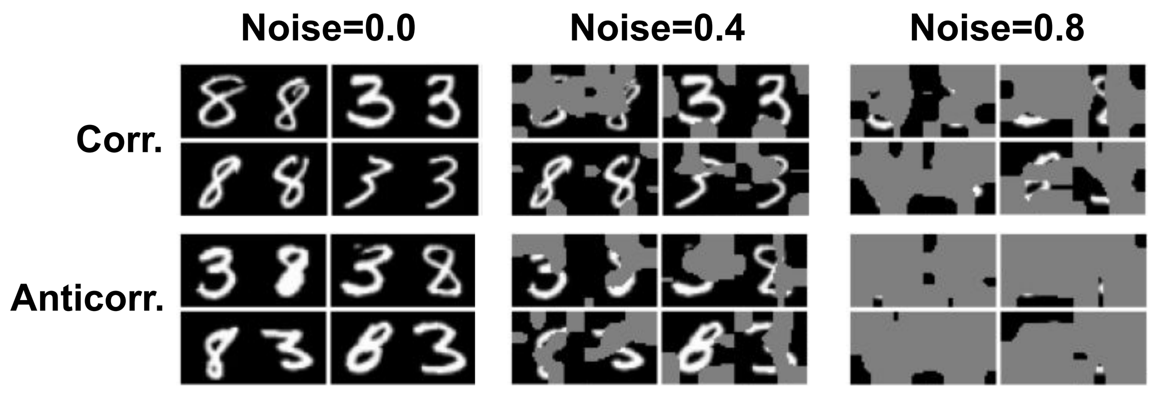
- $\mathbf{z}_1$ and $\mathbf{z}_2$ are independent *conditioned on either of s_1 or s_2*.
- Enforcing independence *conditioned on each of the source variables is sufficient to yield a robust disentangled representation:* $\mathrm{I}(\mathbf{z}_1; \mathbf{z}_2 \mid s_1) = \mathrm{I}(\mathbf{z}_1; \mathbf{z}_2 \mid s_2) = 0$
- We desire that $\mathbf{z}_1$ and $\mathbf{z}_2$ share *as little information as possible* (given the GT correlation), to improve robustness to shifts.
- $\mathbf{z}_1$ necessarily contains information about s_2; we enforce that it does not contain *any more information about $\mathbf{z}_2$ than necessary* via $I(\mathbf{z}_1; \mathbf{z}_2 | s_2)$

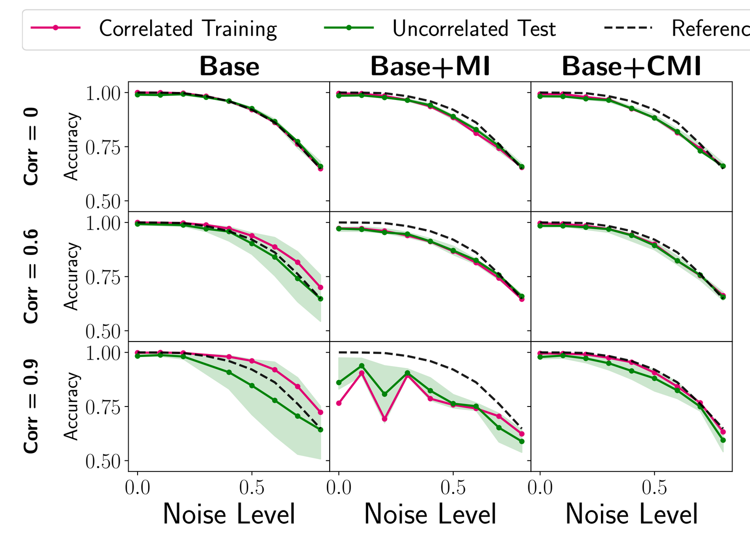
Linear Examples



- Impact of the correlation strength and noise level
- Magenta: performance on correlated training data; Green: performance on test data with a range of correlation shifts

Occluded Multi-Digit MNIST





- **Task:** predict left/right labels
- We used classes 3 and 8 which can become ambiguous under occlusions

Correlated CelebA

- We used attributes `Male` and `Smiling` that we know *a priori* are not causally related.
- Minimizing CMI has a larger effect for stronger correlations, but does not harm performance for low corr.