Google Research

Unsupervised Disentanglement without Autoencoding: Pitfalls and Future Directions

Andrea Burns, Aaron Sarna, Dilip Krishnan, Aaron Maschinot

BOSTON UNIVERSITY

Motivation: Disentangle without VAEs

We would like to learn disentangled visual representations for improved interpretability, data efficiency, and generalization. Prior work has focused on generative methods for disentanglement, which do not scale well to large datasets.

Thus, we explore regularization methods with contrastive learning, which could result in disentangled representations that are powerful enough for large scale datasets and downstream applications.

Representation — 1 entangled representation

color | texture | ... | shape — K disentangled subembeddings

Loss Formulation

Standard Contrastive [1]: Maximize Mutual Information (MI) between views:

$$\mathcal{L}_{\text{InfoMax}} = -\sum_{i \in I} \log \frac{\exp(z_i \cdot z_j / \tau)}{\sum_{a \in A} \exp(z_i \cdot z_a / \tau)}$$

Extend to maximizing information between views of subembeddings, where the k^{th} subembedding ($k \in K$) maps to a slice of the representation. In all of our experiments, K = 2.

$$\mathcal{L}_{\text{SubInfoMax}} = \sum_{k \in K} \left(-\sum_{i \in I} \log \frac{\exp(z_{i,k} \cdot z_{j,k} / \tau)}{\sum_{a \in A} \exp(z_{i,k} \cdot z_{a,k} / \tau)} \right)$$

We add a regularization term *R* to this to obtain our final objective

$$\mathcal{L}_{\text{Disentanglement}} = \mathcal{L}_{\text{SubInfoMax}} + \lambda R$$

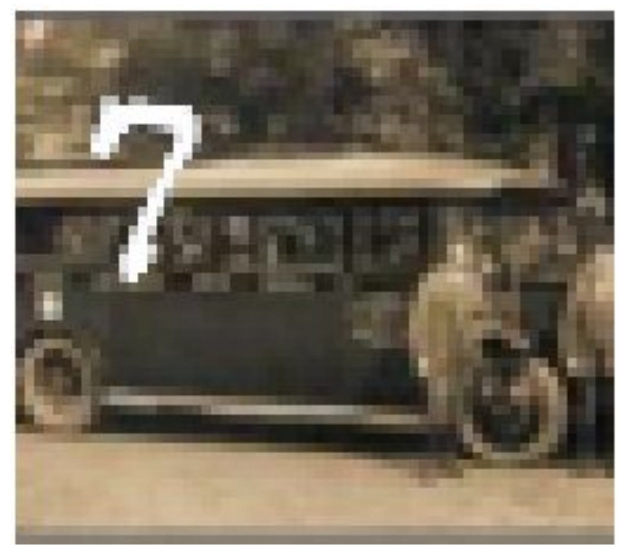
Experimental Setup

1. Train encoder + projection network via contrastive loss with regularizer
2. Use resulting subembeddings to train different linear classifiers

MNIST/STL-10 Dataset

Overlay MNIST digit on STL-10 images & Vary one factor in a view pair while two are fixed:

- Digit Class (DC)
- Digit Location (DL)
- Background Class (BC)



View 0
(digit 7, background car, location 0)

View 1
(digit 7, background car, location 1)

Want to disentangle the two fixed factors of variation in a view pair (DC-BC)

Then perform classification on those two tasks

Regularization Methods

Approach: Minimize MI between subembeddings of the same view

$$R_{\text{InfoMin}} = \sum_{k \neq k'} \sum_{i \in I} \log \frac{\exp(z_{i,k} \cdot z_{i,k'} / \tau)}{\sum_{a \in A} \exp(z_{i,k} \cdot z_{a,k} / \tau)}$$

Approach: Enforce orthogonality to approximate linear independence
Ablations: R_{Ortho} with + without permutation matrix *P*

$$R_{P(\text{Ortho})} = \sum_{k \neq k'} \sum_{i} \sum_{j} \frac{|P(z_{i,k}) \cdot z_{j,k'}|}{\|z_{i,k}\| \|z_{j,k'}\|}$$

Approach: Minimize element-wise dependencies between subembeddings using the Hessian of the loss

$$R_{\text{Hess}} = \sqrt{\sum_{i \in [k]} \sum_{j \in [k']} \left| \frac{\partial \mathcal{L}_{\text{InfoMax}}}{\partial z_i} \cdot \frac{\partial \mathcal{L}_{\text{InfoMax}}}{\partial z_j} \right|^2}$$

Ideal Disentanglement

Requirements

[R1] A subembedding captures only one factor of variation; Random downstream performance on other tasks

[R2] The full representation performs no better than any subembedding;

[R3] Regularization doesn't hurt performance; No loss of information

Classification Input	DC	BC
r	97.3	64.5
r_0	97.3	10.1
r_1	11.7	64.5
$\lvert C(r_0) - C(r_1) \rvert$	85.6	54.4

[R3]

[R2]

[R1]

Results

There are trade-offs in optimization, degree of disentanglement, and absolute task performance:

R_{InfoMin}
- Optimization difficulties prevent disentanglement

$R_{P(\text{Ortho})}$
- Absolute downstream task performance is highest, *i.e.*, [R3] is most satisfied
- [R1] is not very satisfied, which allows for [R2] to be partially satisfied, but is not the desired disentanglement
- Permutation matrix improves [R3] but not [R1], [R2]

R_{Hess}
- Element-wise constraint improves [R1] significantly, but worsens [R2], [R3]

References

1. Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. "A simple framework for contrastive learning of visual representations." *arXiv preprint arXiv:2002.05709* (2020).