Transferring disentangled representations: bridging the gap between synthetic and real images

Jacopo Dapueto, Nicoletta Noceti, Francesca Odone **Contacts:** jacopo.dapueto@edu.unige.it

Introduction and motivations

Disentangled representation learning (DRL) aims to **identify** and **disentangle** underlying Factors of Variation (FoVs).

A good disentangled representation should be:

1. **Modular:** a FoV affects only a partition of the representation.
2. **Compact:** the partitions should be as small as possible.
3. **Explicit:** we should be able to retrieve all the informative FoVs from it.

How to proceed?

- ➔ *Fully unsupervised DRL has been shown unsatisfactory.*
- ➔ Real data can be described with many factors and so FoV annotation is an uncertain or unfeasible process.

Contributions

- **A novel classifier-free and interpretable metric:** OMES
- A methodology for DR transfer to Target datasets **without FoV annotation**

OMES: *Overlap Multiple Encoding Scores*

Intervention-based metric measuring the quality of factor encoding in a representation while providing information about its structure.
What type of Intervention? An association matrix (Factors, Dimensions) ***S*** is computed from the *correlation* of couples of images ($\boldsymbol{R^1}$, $\boldsymbol{R^2}$) differing in *K=1* factors.

$$OMES(S) = \frac{1}{n}\sum_{j=1}^{n} \alpha \, \underbrace{\mathrm{OS}(S,j)}_{\text{modularity}} + (1-\alpha)\, \underbrace{\mathrm{MES}(S,j)}_{\text{compactness}}$$

Algorithm 1 Compute association matrix S between dimensions and FoVs

Require: $D_\Phi = [\boldsymbol{R}^1, \boldsymbol{R}^2, \boldsymbol{k}], n$ ▷ n number of FoVs
Ensure: $\boldsymbol{R}^1, \boldsymbol{R}^2 \in \mathbb{R}^{N\times m}, \boldsymbol{k} \in \mathbb{R}^N$ ▷ $m = |\mathcal{A}|$, N number of pairs in D

$S \leftarrow \mathrm{ZEROS}(m, n)$
for $j = 1$ to n **do**
  $\boldsymbol{R}^1_j = \boldsymbol{R}^1[\boldsymbol{k} == j, :]$ and $\boldsymbol{R}^2_j = \boldsymbol{R}^2[\boldsymbol{k} == j, :]$
  for $h = 1$ to m **do**
    $PC = PearsonCorr(\boldsymbol{R}^1_j[:, h], \boldsymbol{R}^2_j[:, h])$
    $S[h, j] = 1 - abs(PC)$
  end for
end for
return S ▷ $m \times n$ matrix

OMES properties and its interpretation

- ➔ The most used metrics in the literature (DCI [1] & MIG [2]) are either based on classifiers, or based on Mutual Information Estimation.
- ➔ Differently, OMES is classifier-free and based on Correlation, so it does not depend on the choice of hyperparameters.
- ➔ Our intervention provides *more guarantees on disentanglement* properties.
- ➔ *OMES provides info on the structure of the representation (**S**) and allows us to compute the overall score and a score for each FoV separately.*

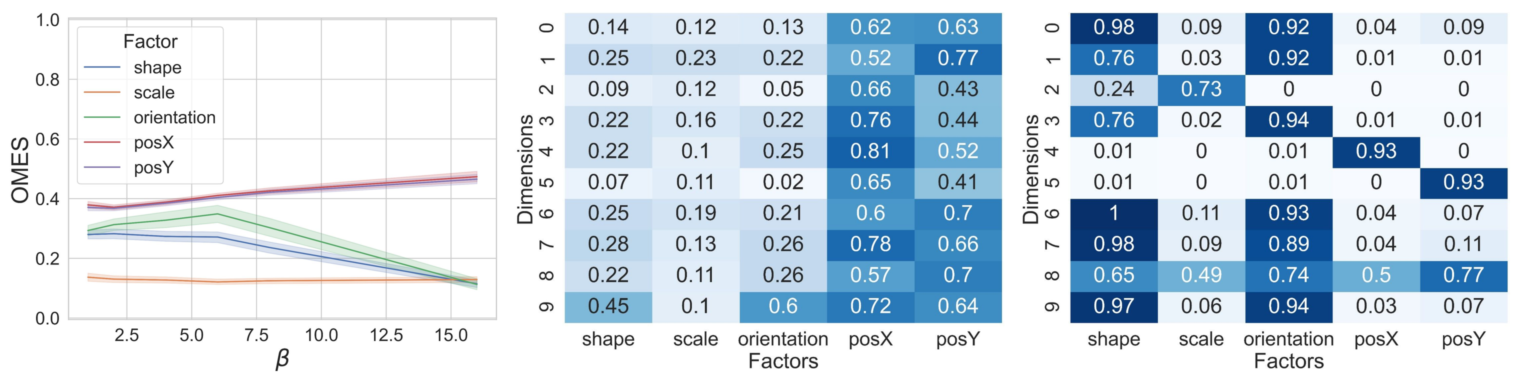

Fig. 1: Dataset *Noisy-dSprites*. **Left: OMES** Scores for each FoV, for different values of β keeping the different FoV separated (α is fixed to 0.5). **Center:** Association matrix S of an unsupervised model (β= 6). **Right:** Association matrix S of a weakly-supervised model.

Transferring disentangled representations

Methodology:

(1) *Learn* disentangled representation on the Source Dataset with **weak supervision** (Ada-GVAE [3]);

(2) *Perform* **unsupervised** *transfer* on Target Dataset w/ or w/o fine-tuning;

(3) Analyze the transferred representation w.r.t. the desired properties.

Dataset	Real	3D	Occlusions	#FoV	Independence	Complete annotation	Resolution	#Images
dSprites	✗	✗	✗	5	✓	✓	64 × 64	737K
Noisy-dSprites	✗	✗	✗	5	✓	✓	64 × 64	737K
Color-dSprites	✗	✗	✗	6	✓	✓	64 × 64	4,4M
Noisy-Color-dSprites	✗	✗	✗	6	✓	✓	64 × 64	4,4M
Shapes3D	✗	✓	✓	6	✓	✓	64 × 64	480K
Isaac3D	✗	✓	✓	9	✓	✓	128 × 128	737K
Coil100-Augmented	✓	✓	✓	4	✓	✓	128 × 128	1,1M
RGB-D Objects	✓	✓	✓	3*	✗	✗	256 × 256	35K

Tab.1: Summary of the datasets and their properties. * in the *#FoV* refers to the possible presence of hidden factors.

Protocol: We trained 20 different models (10 random seeds × 2 values of β) for each Source dataset, for 400.000K iterations. Transfer is performed with the same β of the model for 50K iterations.

- Explicitness is evaluated with FoVs classification with GBT.
- We exploit the interpretability of OMES to evaluate compactness and modularity the single FoV.

Experiments

We considered different couples (Source,Target) to cover different challenges (Resolution, Occlusions, etc.) and scenarios (syn2syn, syn2real, real2real) **incrementally adding complexity**.

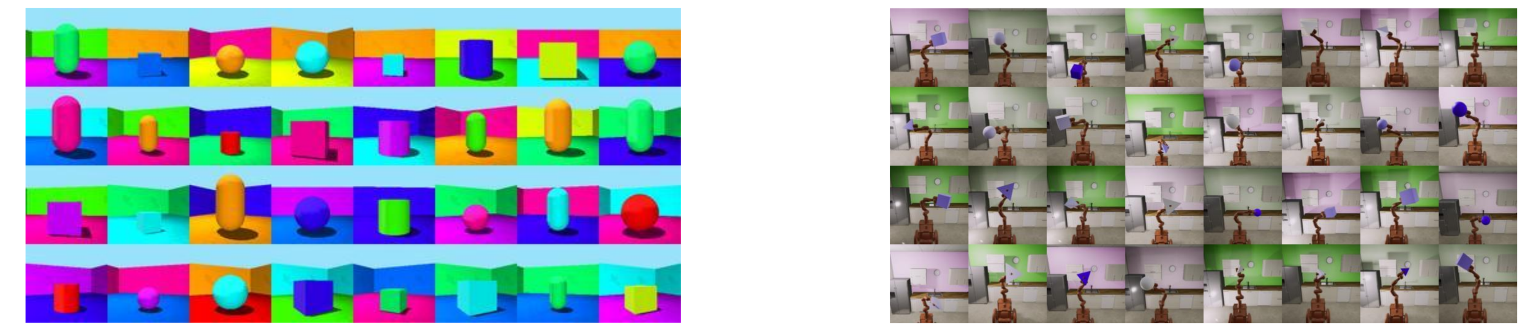

Fig. 2: Example of transfer from synthetic to real data. **Left:** Shapes3D (Source Dataset) **Right:** Isaac3D (Target Dataset)

Pruned	Mean accuracy on FoVs(%)									
	Object shape (3)	Object scale (4)	Camera height (4)	X-movement (8)	Y-movement (5)	Light intensity (4)	Light y-direction (6)	Object color (4)	Wall color (4)	All
✗	34.9 (+5.1)	54.0 (+34.6)	39.2 (+17.4)	33.9 (+29.6)	23.0 (+4.3)	83.6 (+14.0)	85.4 (+12.7)	29.8 (+13.9)	78.1 (+18.9)	51.3 (+16.7)
✓	33.8 (+3.6)	40.1 (+24.1)	33.5 (+11.1)	24.7 (+15.3)	21.6 (+2.7)	69.4 (+17.3)	67.5 (+14.4)	27.0 (+7.3)	61.5 (+16.6)	42.1 (+12.5)

Pruned	Modularity(%)		Compactness(%)	
	Our (OS)	DCI	Our (MES)	MIG
✗ ✓	25.1 (+9.7)	6.3 (+16.1)	21.2 (+10.2)	2.2 (+5.9)

Tab.2: Transfer from Shapes3D (Source) to Isaac3D (Target). Average classification accuracy over the 20 models of the GBT classifier, before and after fine-tuning. The latter is reported in parenthesis in terms of gain or loss w.r.t. the performance before the fine-tuning. All is the average performance of all FoVs. The column Pruned highlights the two different representation modalities: if the classifier is trained on the whole representation, or using only one dimension.

Conclusions

- ➔ If source and target have common FoVs with similar appearance we can obtain good performances on a real Target dataset, even if the source synthetic dataset is much simpler.
- ➔ ***One could design synthetic data to disentangle specific factors of interest preserving the disentanglement properties, even with fine-tuning***

Future directions

- We will explore quantitative methods to assess the distance between Source and Target datasets;
- We will target more specific applications, such as biomedical image classification or action recognition from videos.

References

[1] Eastwood, C., Williams, C.K.(2018). " framework for the quantitative evaluation of disentangled representations." In: International conference on learning representations

[2] Chen, R.T., et al.(2018). "solating sources of disentanglement in variational autoencoders.". In: Advances in neural information processing systems.

[3] Locatello, F., et al (2020). "Weakly-supervised disentanglement without compromises". In: International Conference on Machine Learning. pp. 6348–6359.