The role of disentanglement in generalisation

Milton L. Montero[1, 2] Casimir J.H. Ludwig[1] Rui Ponte Costa[2]
Gaurav Malhotra[1] Jeffrey S. Bowers[1]

[1]School of Psychological Science
[2]Computational Neuroscience Unit, Department of Computer Science
University of Bristol, United Kingdom

May 3, 2021

Disentanglement in latent variable models

- **Disentangled representations** encode separate factors of variation of the data in separate variables of latent representation.

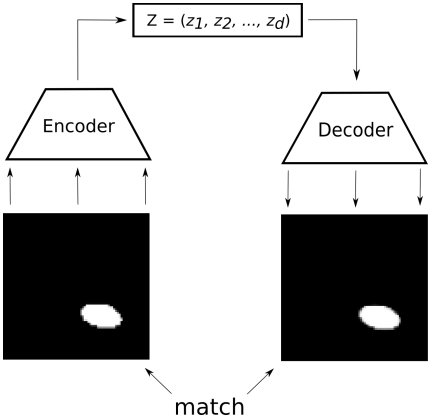
Disentanglement in latent variable models

- **Disentangled representations** encode separate factors of variation of the data in separate variables of latent representation.
- These "should capture the compositional structure of the world" (Duan et al., 2020).

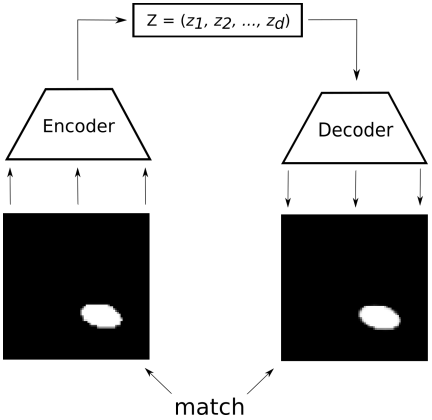
Disentanglement in latent variable models

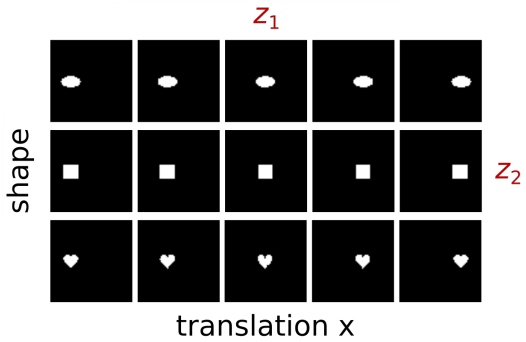
- **Disentangled representations** encode separate factors of variation of the data in separate variables of latent representation.
- These "should capture the compositional structure of the world" (Duan et al., 2020).

Learning disentangled representations



Learning disentangled representations

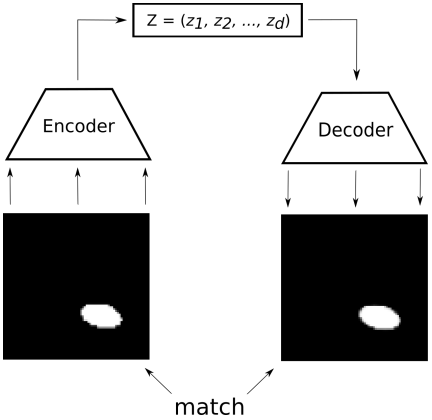


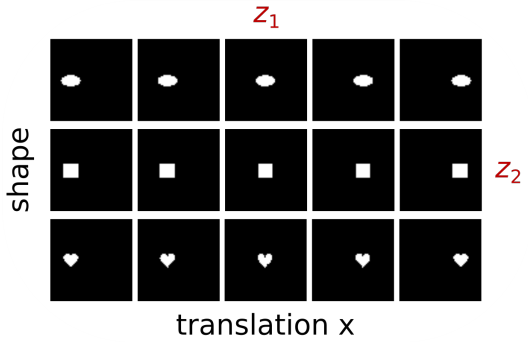


Disentanglement and generalisation

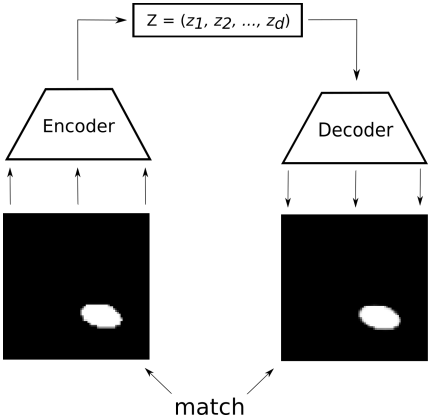
Can disentangled representations support out-of-training-distribution generalisation?

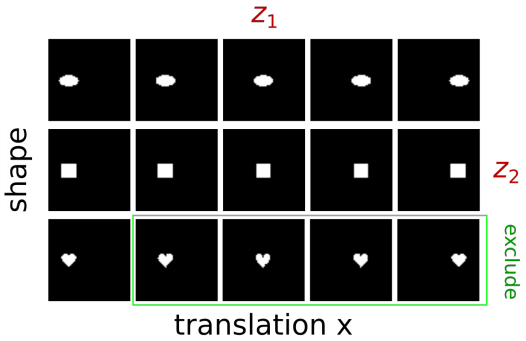
Testing generalisation



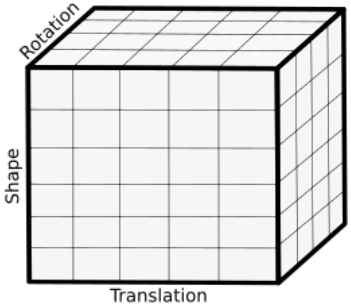


Testing generalisation

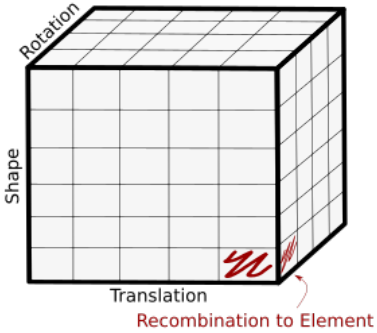




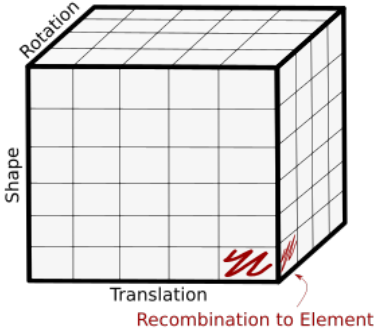
Testing generalisation

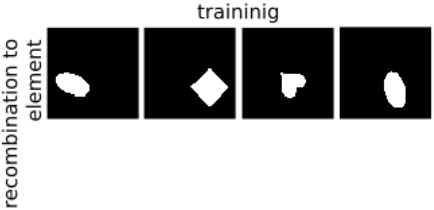


Testing generalisation

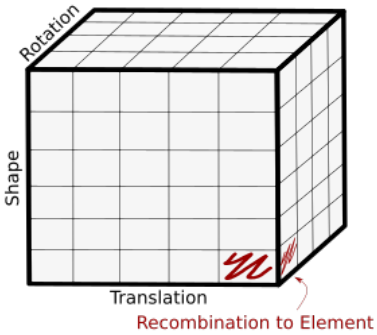


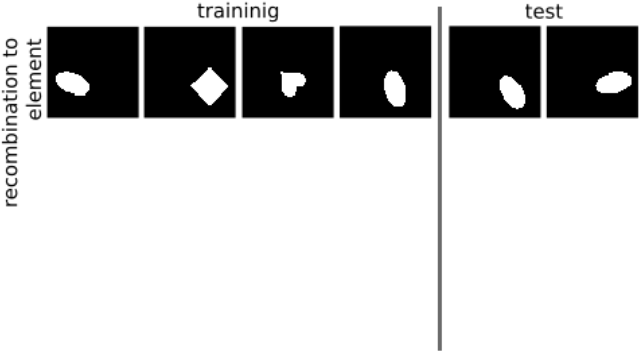
Testing generalisation



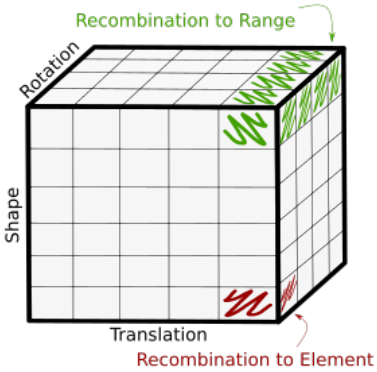


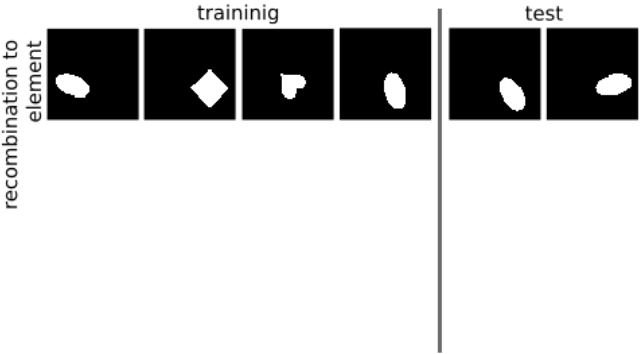
Testing generalisation



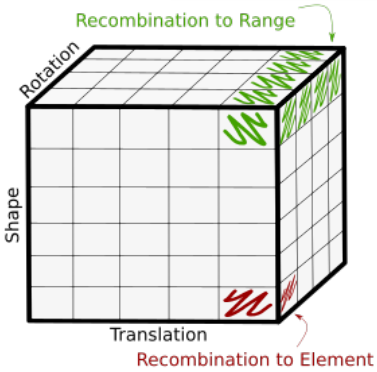


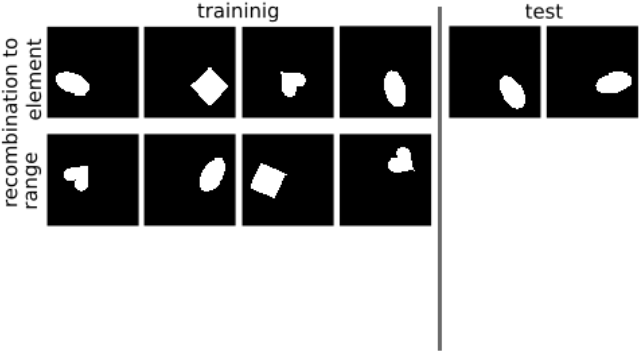
Testing generalisation



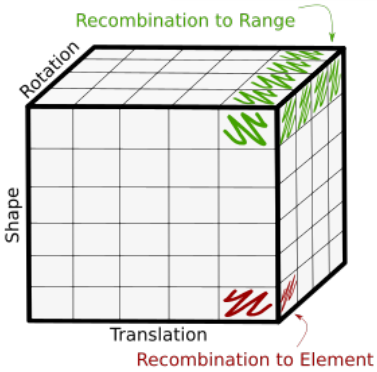


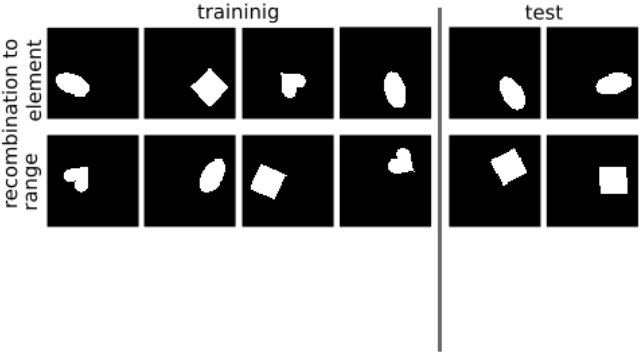
Testing generalisation



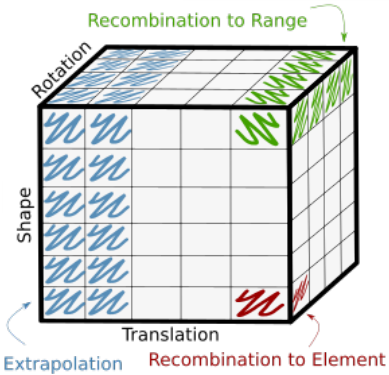


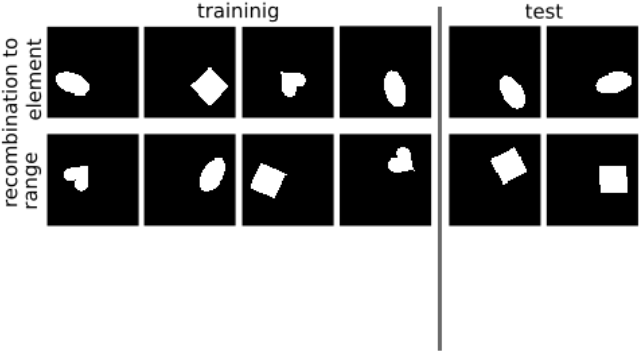
Testing generalisation



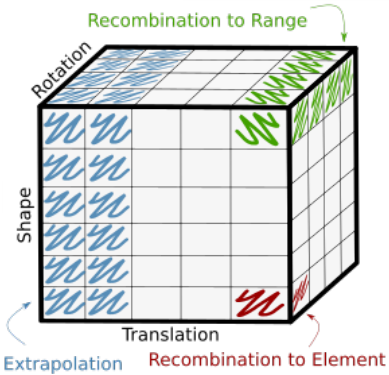


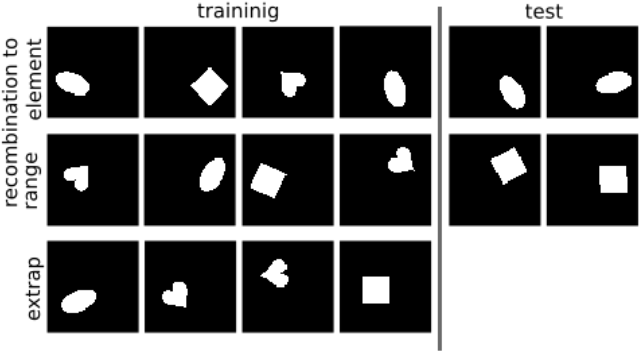
Testing generalisation



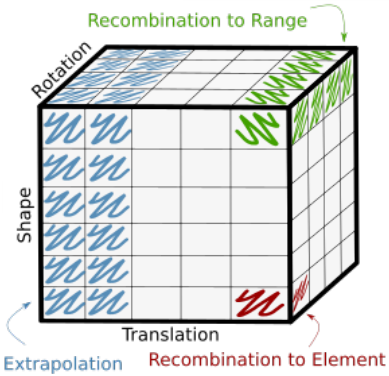


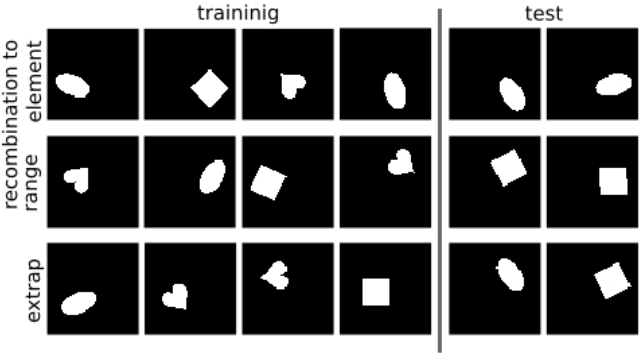
Testing generalisation



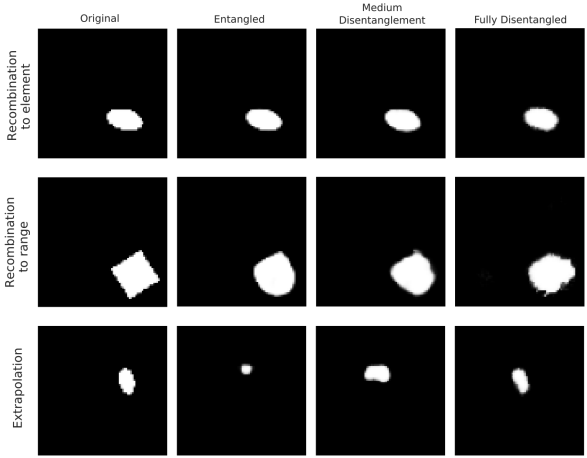


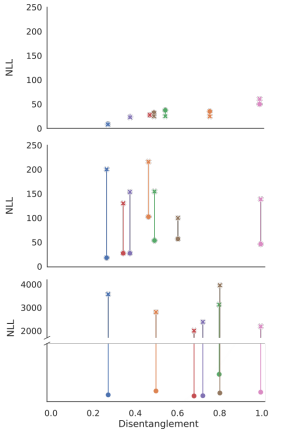
Testing generalisation



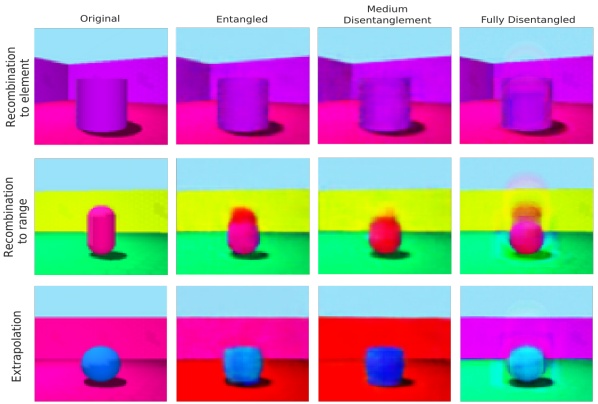


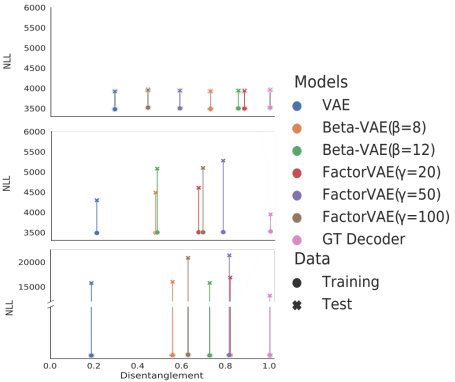
Results: dSprites





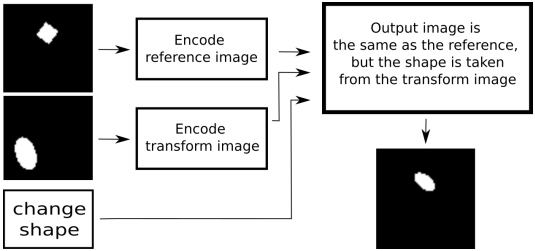
Results: 3DShapes



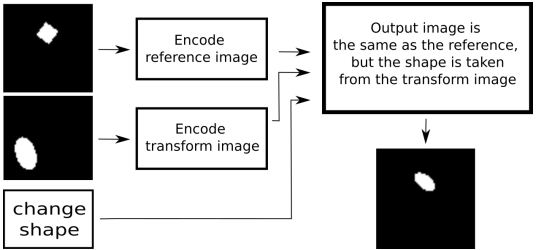


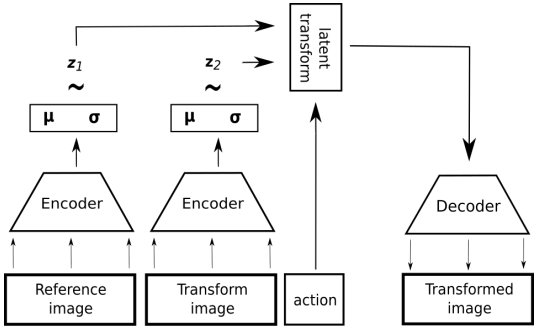
Can we push the models to learn compositional representations by chaning the task?

Composition task and model

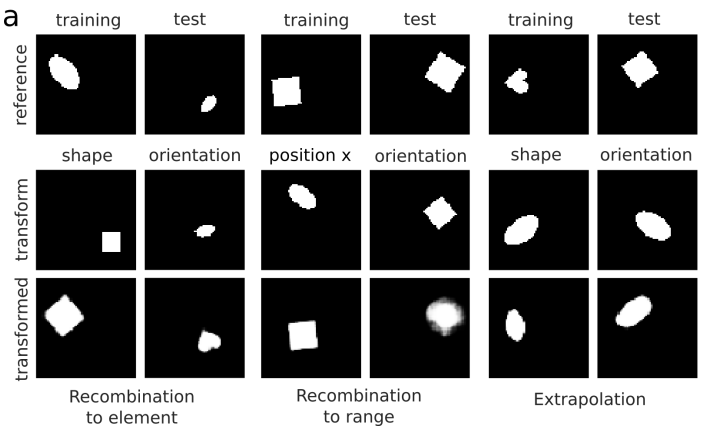


Composition task and model





Compositon task: results



Conclusions

1. There are several ways that we can define generalisation outside the training space.

Conclusions

1. There are several ways that we can define generalisation outside the training space.
2. Disentanglement, does not seem to allow the models to generalise to these settings any more than non-entangled ones.

Conclusions

1. There are several ways that we can define generalisation outside the training space.
2. Disentanglement, does not seem to allow the models to generalise to these settings any more than non-entangled ones.
3. We can obtain high disentanglement by training a model to manipulate the world.

Conclusions

1. There are several ways that we can define generalisation outside the training space.
2. Disentanglement, does not seem to allow the models to generalise to these settings any more than non-entangled ones.
3. We can obtain high disentanglement by training a model to manipulate the world.
4. This works adds to oher contributions that highlight the need for better definitions, metrics, models and training settings .

Acknowledgements

Many thanks to my co-authors:
Jeff Bowers, Gaurav Malhotra, Rui Ponte Costa & Casimir Ludwig