Disentangled Face Representations in Deep Generative Models and the Human Brain

Paul Soulos and Leyla Isik

psoulos1@jh.edu lisik@jh.edu

JOHNS HOPKINS UNIVERSITY

Introduction

- How are various features coded across the face network?
- Neural networks are good models of fMRI brain data but are difficult to interpret
- Can disentangled generative models help us understand the representations used during face processing?

Conclusion

- Disentangled generative models performs as well as standard generative models and discriminative models
- The disentangled dimensions are interpretable and provide us with a method to inspect voxel responses
- We find that low-level dimensions appear more posterior while high-level dimensions appear more anterior
- Future work will investigate the role of entangled dimensions in identity coding

References

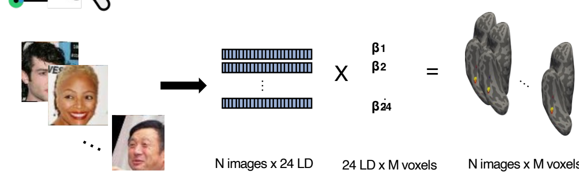
1 Kim & Mnih *ICML* (2018)
2 VanRullen & Reddy *Commun Biol* 2 (2019)

paulsoulos.com/posters/ccn22.pdf

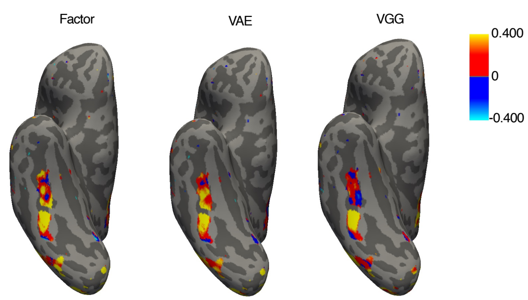
Disentangled Generative Models

- We trained FactorVAE[1] with 24 dimensions on CelebA
- 16 dimensions are interpretable by human raters (disentangled)
- 8 are not interpretable (entangled)

Encoding Model and Performance

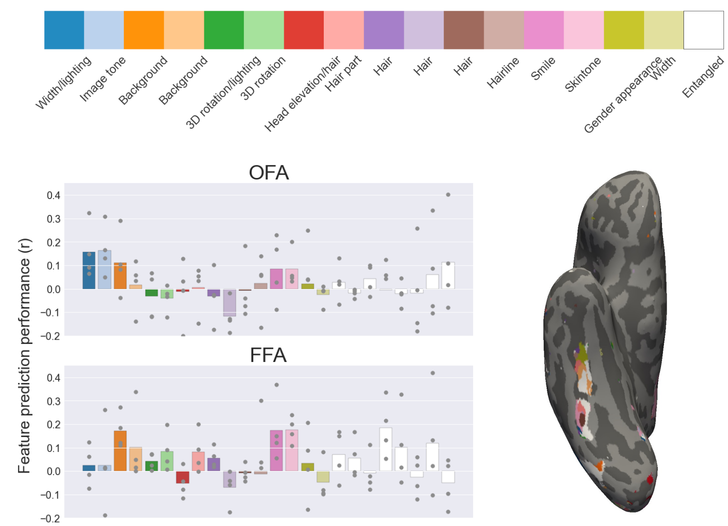


- Four participants saw 8000 face images[2]
- We fit a linear encoding model between model representations and fMRI responses
- FactorVAE performs as well as VAE and VGG in OFA and FFA. No models predict activity in STS well.



Voxel Selectivity

- We perform encoding model prediction for each dimension
- Higher level identity relevant dimensions are represented in more anterior regions



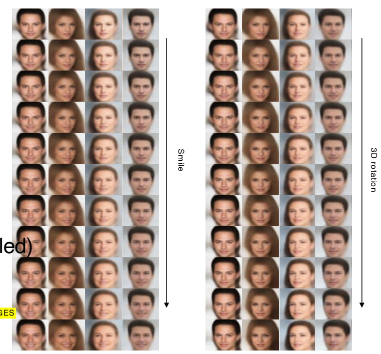
Disentangled Face Representations in Deep Generative Models and the Human Brain

Paul Soulos and Leyla Isik *psoulos1@jh.edu lisik@jh.edu*

Johns Hopkins University

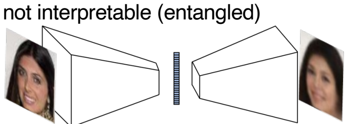
Introduction

- How are various features coded across the face network?
- Neural networks are good models of fMRI brain data but are difficult to interpret
- Can disentangled generative models help us understand the representations used during face processing?



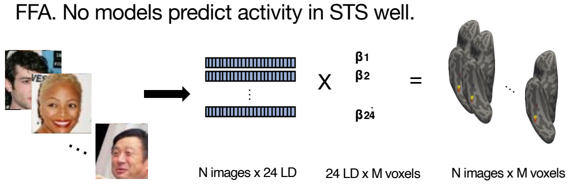
Disentangled Generative Models

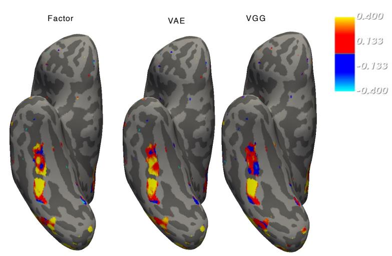
- We trained FactorVAE with 24 dimensions on CelebA
- 16 dimensions are interpretable by human raters (disentangled)
- 8 are not interpretable (entangled)



Encoding Model and Performance

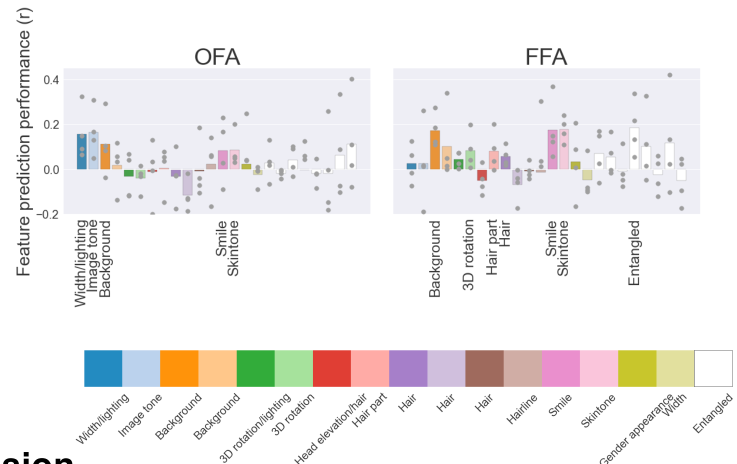
- Four participants saw 8000 face images [1]
- We fit a linear encoding model between model representations and fMRI responses
- FactorVAE performs as well as VAE and VGG in OFA and FFA. No models predict activity in STS well.





Voxel Selectivity

- We perform encoding model prediction for each dimension
- Higher level identity relevant dimensions are represented in more anterior regions



Conclusion

- Disentangled generative models performs as well as standard generative models and discriminative models
- The disentangled dimensions are interpretable and provide us with a method to inspect voxel responses
- We find that low-level dimensions appear more posterior while high-level dimensions appear more anterior
- Future work will investigate the role of entangled dimensions in identity coding

Disentangled Face Representations in Deep Generative Models and the Human Brain

Paul Soulos and Leyla Isik
psoulos1@jh.edu lisik@jh.edu
JOHNS HOPKINS UNIVERSITY

Introduction

- How are various features coded across the face network?
- Neural networks are good models of fMRI brain data but are difficult to interpret
- Can disentangled generative models help us understand the representations used during face processing?

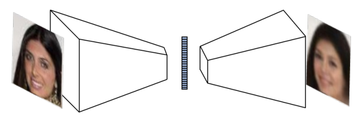
Conclusion

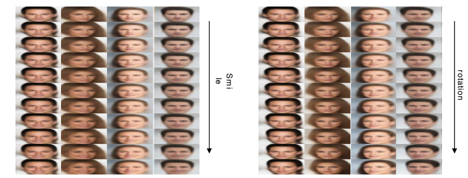
- Disentangled generative models performs as well as standard generative models and discriminative models
- The disentangled dimensions are interpretable and provide us with a method to inspect voxel responses
- We find that low-level dimensions appear more posterior while high-level dimensions appear more anterior
- Future work will investigate the role of entangled dimensions in identity coding

References

Disentangled Generative Models

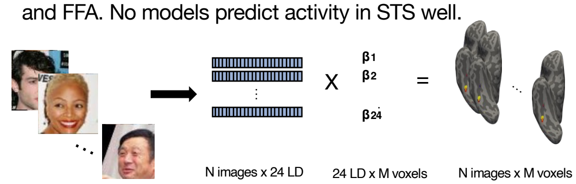
- We trained FactorVAE with 24 dimensions on CelebA
- 16 dimensions are interpretable by human raters (disentangled)
- 8 are not interpretable (entangled)

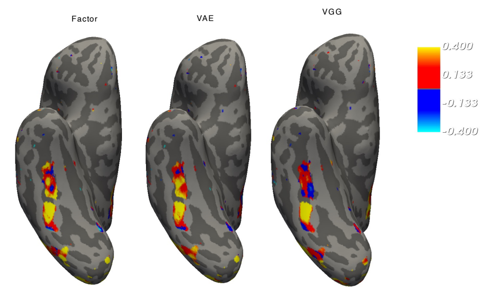




Encoding Model and Performance

- Four participants saw 8000 face images [1]
- We fit a linear encoding model between model representations and fMRI responses
- FactorVAE performs as well as VAE and VGG in OFA and FFA. No models predict activity in STS well.





Voxel Selectivity

- We perform encoding model prediction for each dimension
- Higher level identity relevant dimensions are represented in more anterior regions