Bias and Generalization in Deep Generative Models

Shengjia Zhao*, Hongyu Ren*, Arianna Yuan, Jiaming Song, Noah Goodman and Stefano Ermon

*equal contribution
Success in Generative Modeling of Images

Training Images

Generated Images

Brock A, et al. "Large scale gan training for high fidelity natural image synthesis."
Goal: Understanding Generalization

How do generative models generalize?
Generalization Example: Object Count

All training images have 2 objects

How many are in the generated images?

A

2

2

2

B

2

3

1
Empirical Study of Generalization: Method

- Design datasets
- Train generative models (VAE, GAN, PixelCNN)
- Observe generalization behavior
- Find common patterns
Generalization Example: Object Count

All training images have 2 objects

How many are in the generated images?

A

B

2 2 2

2 2 2

2 2 2

2 3 2
Generalization in Feature Space: Numerosity

Training Distribution

Generated Distribution (Observed)

Generates a log-normal shaped distribution
Multiple Numerosities

Training Distribution

Frequency

# Objects

2 7
Multiple Numerosities: Only 2

Training Distribution

Generated Distribution
Multiple Numerosities: Only 7

Training Distribution

Generated Distribution
Multiple Numerosities: Additive Hypothesis

Training Distribution

Generated Distribution
(Observed)
Additive Hypothesis with 2 and 4 Objects

Training Distribution

Generated Distribution (Hypothesized)
Actual Result: Prototype Enhancement

Training Distribution

Generated Distribution (Observed)

3 objects most likely, even though no training image contains 3 objects!
Prototype Enhancement

Similar pattern for other features: color, size, location

Training Distribution

Generated Distribution (Observed)
Multiple Features

Training Images

Generated Images

A

B

Shape

Color

Shape

Color
Memorization vs. Generalization

Training Distribution

Generated Distribution

Shape
Color

Shape
Color

- Observed in Training Set
- Not Observed in Training Set
Memorization vs. Generalization

Training Distribution

- Shape
  - Color

Generated Distribution

- Shape
  - Color

- Shape
  - Color

- Shape
  - Color

Legend:
- Observed in Training Set
- Not Observed in Training Set
Different Setups, Similar Results

- Different features (shape, color, size, numerosity, etc.)
- Different models: (VAE, GAN, PixelCNN, etc.)
- Different architectures (fully connected, convolutional, etc.)
- Different hyper-parameters (network size, learning rate, etc.)
Conclusion

• New methodology: design datasets to probe generative models
• Observed common patterns across different setups

Welcome to our poster session for further discussions!

Tuesday 5-7pm @ Room 210 & 230 AB #6

Code available at github.com/ermongroup/BiasAndGeneralization