InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets

Project Proposal for Implementation

Slovyagina Anna
Nasykhova Anastasia
Tarasova Sofia

October 20, 2025

Why InfoGAN?

- **Problem**: Traditional GANs learn entangled representations
- **Limitation**: Hard to control specific features in generated images
- **Solution**: InfoGAN learns disentangled, interpretable representations
- **Key Advantage**: Completely unsupervised approach

Real-world Applications

Data augmentation, feature manipulation, image editing, style transfer

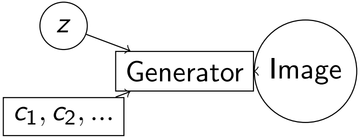
The Big Idea

Standard GAN:

- Input: Random noise z
- Output: $G(z)$
- Problem: Features are entangled

InfoGAN:

- Input: Noise z + Latent codes c
- Output: $G(z, c)$
- Solution: Maximize mutual information $I(c; G(z, c))$



Mutual Information Intuition

What is Mutual Information?

$I(X;Y) = H(X) - H(X|Y)$

- Measures how much information variable Y provides about X
- If X and Y are independent: $I(X;Y) = 0$
- If X and Y are perfectly correlated: $I(X;Y)$ is maximized

InfoGAN Goal

High mutual information between latent codes c and generated images $G(z, c)$
$\Rightarrow$ Changes in c should produce predictable changes in output

The Challenge: Computing Mutual Information

Problem: $I(c; G(z, c))$ requires posterior $P(c|x)$ which is intractable

Solution: Variational Lower Bound

$$I(c; G(z, c)) \geq L_I(G, Q) \tag{1}$$

$$L_I(G, Q) = \mathbb{E}_{c \sim P(c), x \sim G(z,c)}[\log Q(c|x)] + H(c) \tag{2}$$

Where:

- $Q(c|x)$ is auxiliary network approximating $P(c|x)$
- $H(c)$ is entropy of latent codes (constant for fixed distribution)
- Lower bound becomes tight when $Q(c|x) \approx P(c|x)$

InfoGAN Objective Function

Complete Objective

$$\min_{G,Q} \max_{D} V_{InfoGAN}(D, G, Q) = V(D, G) - \lambda L_I(G, Q)$$

Where:

- $V(D, G)$ is standard GAN objective
- $\lambda L_I(G, Q)$ is mutual information regularization
- λ is hyperparameter (typically $\lambda = 1$ for discrete codes)

Implementation Details:

- Q shares layers with discriminator D (minimal overhead)
- Categorical codes: softmax output
- Continuous codes: Gaussian distribution (mean + std)

What InfoGAN Can Discover

MNIST Dataset:

- Digit type (0-9)
- Rotation angle
- Stroke width
- Writing style

Face Datasets:

- Pose/azimuth
- Lighting direction
- Facial expressions
- Glasses presence
- Hair style

3D Objects:

- Rotation/viewpoint
- Object width/size
- Color variations
- Shape deformations

Complex Scenes:

- Background context
- Lighting conditions
- Object arrangements
- Style variations

Project Roadmap

Phase 1: Foundation

- Implement basic GAN architecture
- Add auxiliary network $Q(c|x)$
- Implement variational lower bound $L_I(G, Q)$

Phase 2: Testing

- Train on MNIST dataset
- Verify mutual information maximization
- Test disentanglement quality

Technical Requirements

Hardware:

- GPU with 8GB+ VRAM
- CUDA support
- Sufficient RAM for data loading

Software:

- PyTorch/TensorFlow
- Python 3.8+
- Standard ML libraries

Architecture Components:

- Generator network $G(z, c)$
- Discriminator network $D(x)$
- Auxiliary network $Q(c|x)$
- Loss computation modules

Evaluation Tools:

- Visualization utilities
- Quantitative metrics
- Comparison frameworks

Potential Difficulties & Solutions

Training Stability

Challenge: GANs are notoriously difficult to train

Solution: Use proven techniques (DCGAN, batch normalization, careful hyperparameters)

Hyperparameter Tuning

Challenge: λ parameter needs careful tuning

Solution: Start with paper recommendations, systematic grid search

Evaluation Metrics

Challenge: How to quantify ”disentanglement” quality?

Solution: Visual inspection, downstream task performance, mutual information estimation

How We'll Measure Success

Qualitative Metrics

- Visual quality of generated images
- Interpretability of learned latent codes
- Smooth interpolation between code values
- Disentanglement of different factors

Quantitative Metrics

- Mutual information lower bound convergence
- Classification accuracy using learned codes
- FID/IS scores for image quality
- Comparison with baseline GAN

Success Criteria: Reproduce key results from original paper on MNIST and demonstrate clear disentanglement of at least 2-3 interpretable factors

Why This Project Matters

- **Theoretical Significance:** Bridges information theory and deep generative models
- **Practical Impact:** Enables controllable image generation without supervision
- **Learning Opportunity:** Deep dive into GANs, information theory, and representation learning
- **Future Applications:** Foundation for more advanced disentanglement methods

Key Takeaway

InfoGAN shows that adding a simple information-theoretic regularization to GANs can lead to dramatically more interpretable representations, opening new possibilities for controllable generation and representation learning.

Questions?

Thank you for your attention!

Ready to implement InfoGAN and explore the fascinating world of disentangled representation learning.