U.S. ARMY COMBAT CAPABILITIES DEVELOPMENT COMMAND – ARMY RESEARCH LABORATORY

Multi-modal Generative Adversarial Networks Make Realistic and Diverse but Untrustworthy Predictions When Applied to Ill-posed Problems

John S. Hyatt, Michael S. Lee
Computational & Information Sciences Directorate

Approved for public release: distribution unlimited

02/08/2021

Motivation: step-by-step

- Multi-modal Generative Adversarial Networks Make Realistic and Diverse but Untrustworthy Predictions When Applied to Ill-posed Problems
- Generative Models
  - Converts *latent space representation* to *data space*
  - Latent space Z: "string of numbers"
  - Data space X: e.g., image
  - GANs, VAEs, autoregressive models, INNs

Generative Adversarial Networks (GANs)

Two models

- Generator, G
- Discriminator or Critic, C

Trained adversarially

- $G(\mathbf{z})$ generates output in X-space.
- $C(\mathbf{x})$ learns "x is real."
- $C(G(\mathbf{z}))$ learns "$G(\mathbf{z})$ is fake."

Variational Autoencoders (VAEs)

Two models
- Encoder, E
- Decoder, D

Trained sequentially
- $E(\mathbf{x})$ gives a distribution in Z-space.
- Not like an autoencoder, which gives a single point!
- $D(E(\mathbf{x}))$ attempts to reconstruct $\mathbf{x}$.
- After training, $D(\mathbf{z})$ is the generative model.

Motivation: step-by-step

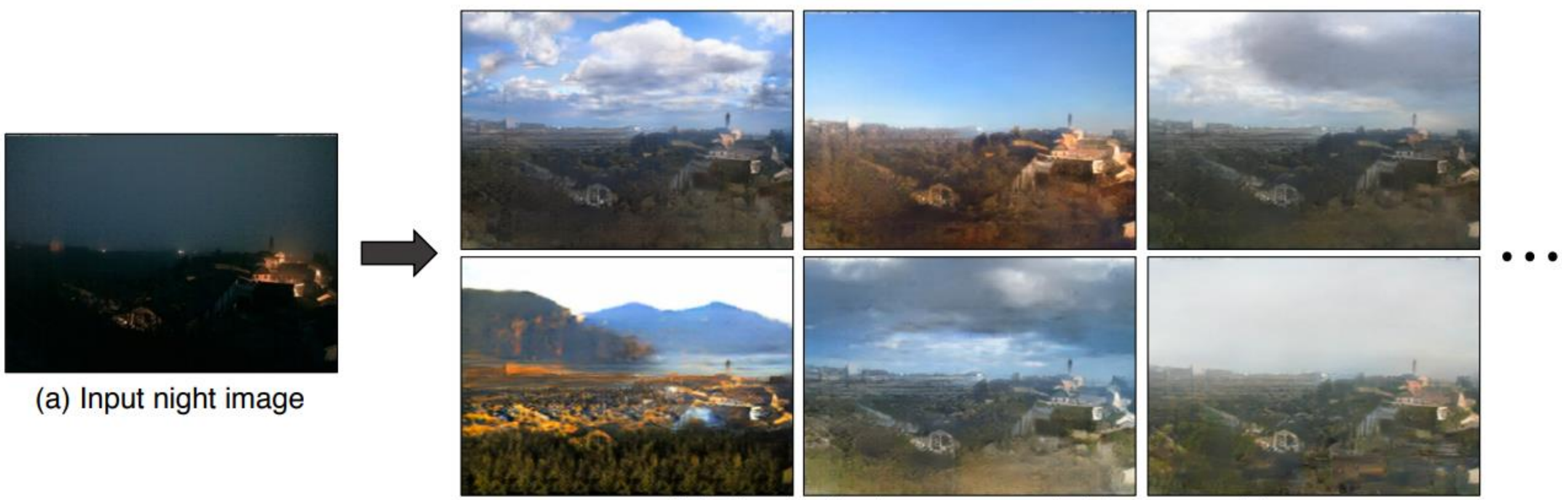
- Multi-modal Generative Adversarial Networks Make Realistic and Diverse but Untrustworthy Predictions When Applied to Ill-posed Problems
- Multi-modality is hard
  - Conditional problems (additional input *Y*)
  - Normal generative models can't combine *Y* & *Z* inputs.
  - Tend to return the same prediction for a given **y**.

Motivation: step-by-step

Multi-modal Generative Adversarial Networks Make Realistic and Diverse but Untrustworthy Predictions When Applied to Ill-posed Problems

Progress has been made in this area

- BicycleGAN: combines (conditional) GAN and VAE
- Produces realistic and diverse outputs:



(a) Input night image

(b) Diverse day images sampled by our model

From: Zhu et al, "Toward Multimodal Image-to-Image Translation," NIPS 2017.

Motivation: step-by-step

Multi-modal Generative Adversarial Networks Make Realistic and Diverse but Untrustworthy Predictions When Applied to Ill-posed Problems

Common in ML:

- Does the problem have a solution? (Out-of-distribution classification)
- Is the solution a continuous function of initial conditions? (Adversarial examples)
- Is the solution unique?

Motivation: step-by-step

- Multi-modal Generative Adversarial Networks Make Realistic and Diverse but Untrustworthy Predictions When Applied to Ill-posed Problems
- "Many" generative problems are aesthetic
  - "Does the result look realistic?"
- Some applications have more stringent criteria:
  - Risk management
  - Uncertainty quantification
  - etc.

What exactly do we want?

A way to bijectively map between X and Z representations for some partial information Y

- Inverse problems / one-to-many maps
- "Extreme" super-resolution (feature generation)
- Denoising audio
- Etc.

Sample $\mathbf{z} \sim p_Z(\mathbf{z}) \rightarrow$ generate $G(\mathbf{y},\mathbf{z}) \rightarrow$ sample $\mathbf{x} \sim p_{X|Y=\mathbf{y}}(\mathbf{x})$

And know that $p_{X|Y=\mathbf{y}}(\mathbf{x})$ really is the right distribution!

X = data space, Z = latent space, Y = conditioning info

BicycleGAN

Zhu et al, "Toward Multimodal Image-to-Image Translation," NIPS 2017.

Three models:

- Generator, $G : Y,Z \to X$
- Encoder, $E : X \to Z$
- Discriminator, C

Optimized on:

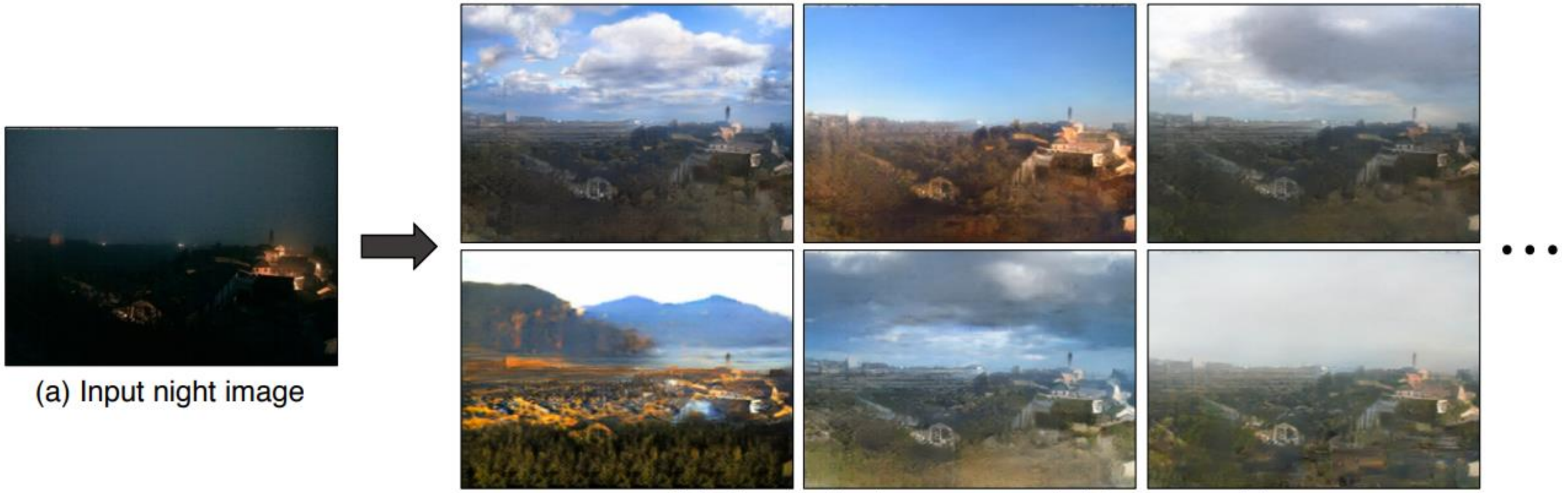
- "Is $G(\mathbf{y},\mathbf{z})$ realistic?"
- "Is $E(G(\mathbf{y},\mathbf{z}))$ close to $\mathbf{z}$?"
- "Is $G(E(\mathbf{x}))$ close to $\mathbf{x}$?"
- "Make E's output resemble a Gaussian, $\mathcal{N}(\mathbf{0},\mathbf{1})$."

BicycleGAN

Asking *G* and *E* to invert each other...sort of

- *E* has no (conditional) *Y* input
- *E* outputs a point cloud rather than a point
- Training isn't symmetric for *G* and *E*

No direct examination of how accurately *E* and *G* learn distributions



(a) Input night image

(b) Diverse day images sampled by our model

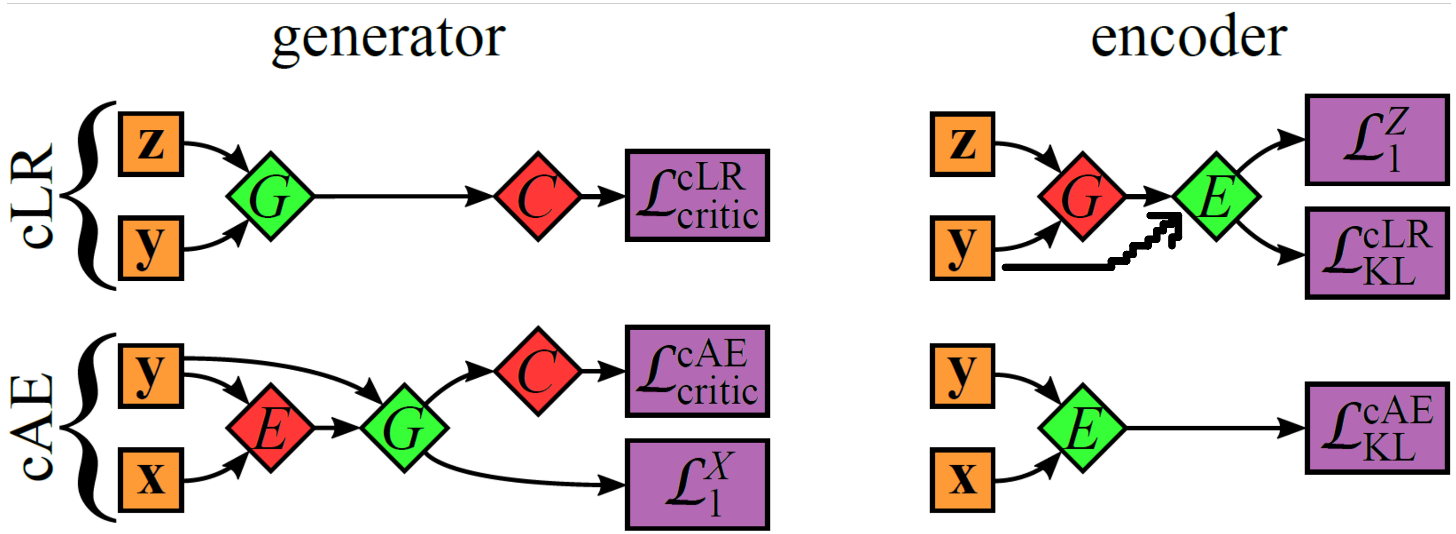
From: Zhu et al, "Toward Multimodal Image-to-Image Translation," NIPS 2017.

Our Tests

BicycleGAN doesn't produce desired distributions (on a much simpler problem).

What if we change BicycleGAN:

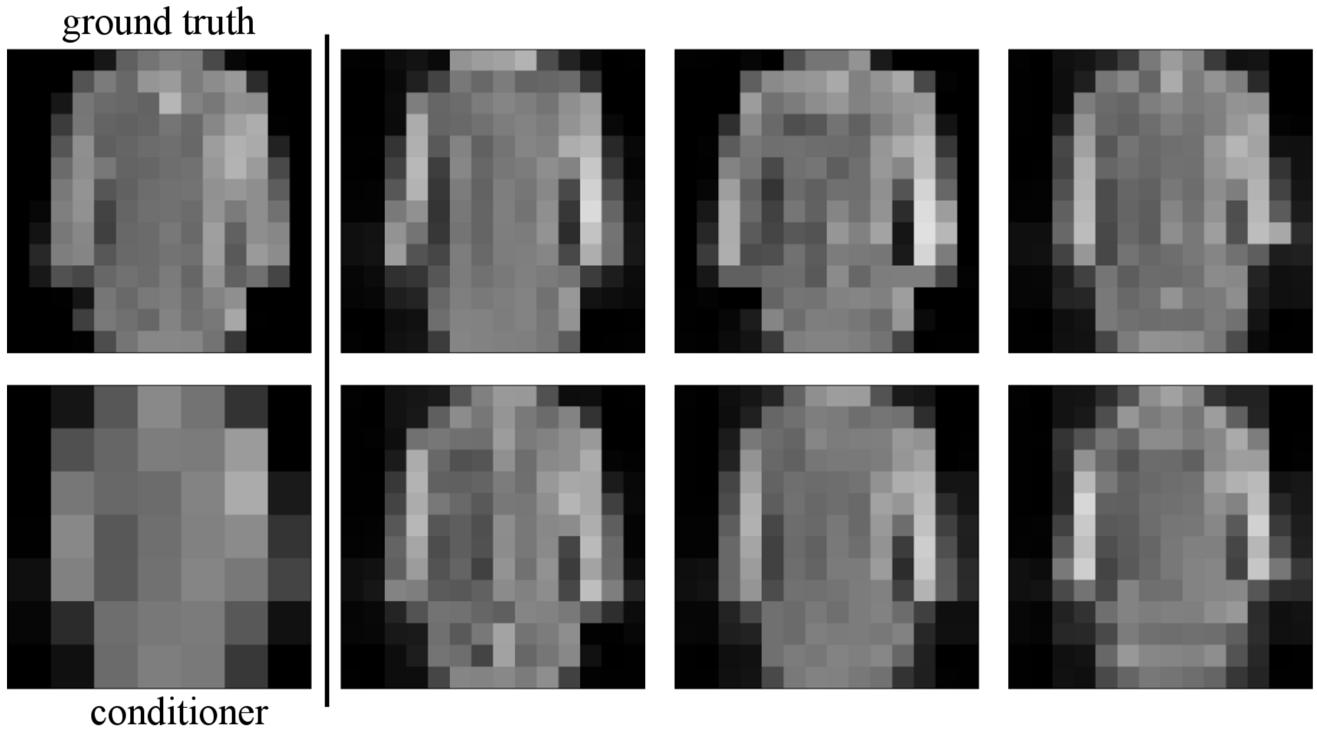
- Change encoder to $E : Y,X \rightarrow Z$
- Make E deterministic rather than variational
- Split cycle consistency losses between E and G



Our Results

E and *G* never truly learn to invert one another.

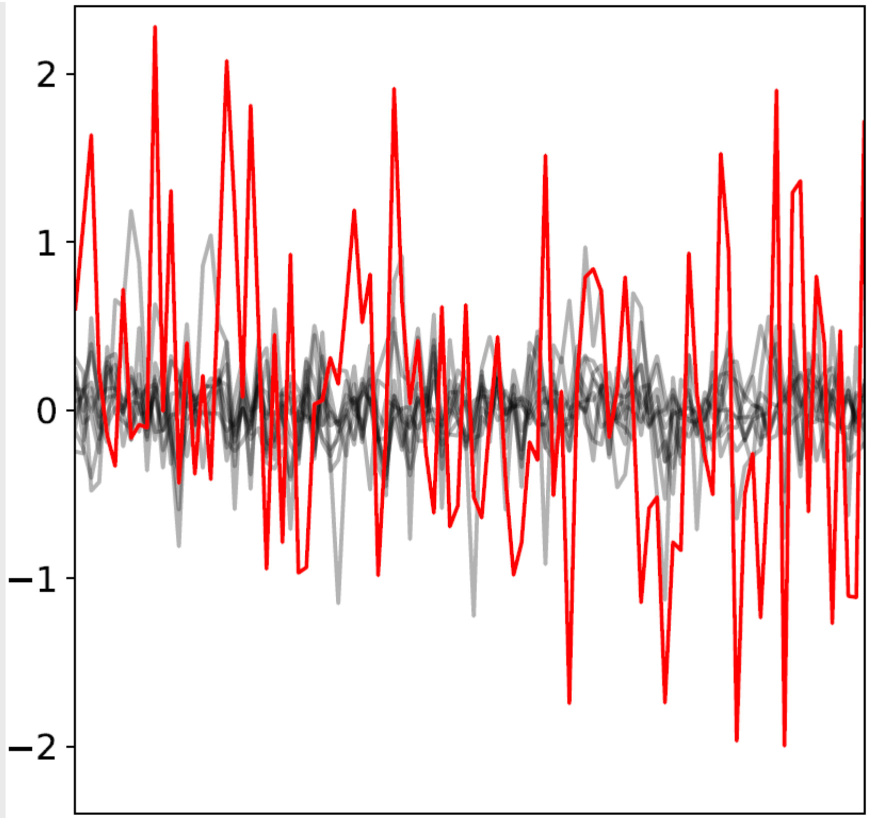
Results are diverse and realistic...



Our Results

...but not representative.

– Looking at *Z*-space representations:



Gaussian (correct distribution)

Learned distribution

Big Takeaways

Big takeaways:

- Realism and diversity ***DO NOT*** mean a model has learned the target distribution!
- Representativeness ***SHOULD NOT*** be taken for granted!
- Encoder/decoder model pairs are hard to train. It may make more sense to look at *explicitly invertible* models

Bonus Slide: Conditional INNs

Invertible Neural Networks are naturally bijective maps.

- Conditional INNs are understudied, but very promising.