Analysis and applications of deep generative models

Pengchuan Zhang
Deep Learning Group, MSR AI

Joint work with Qiang Liu, Denny Zhou, Tao Xu, Qiuyuan Huang, Xiaodong He, Ka Chun Lam, Shumao Zhang, Thomas Hou and other colleagues

Inverse Problem and Machine Learning, Feb 10, 2018

What: density/distribution estimation

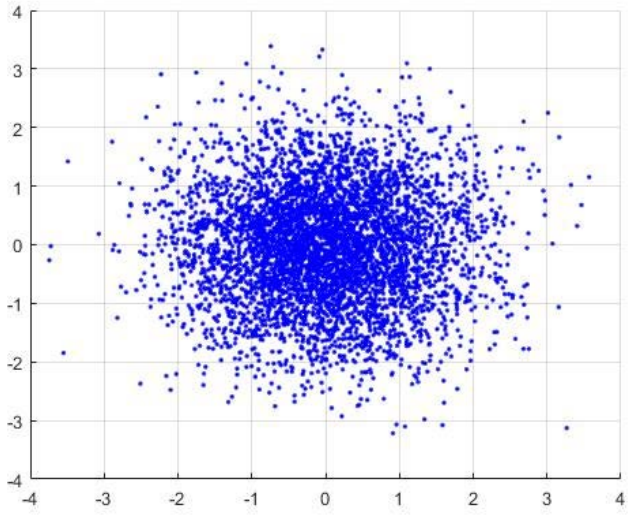
- Target distribution μ
- Learn a generative model $G(z)$ with $z \sim N(0, I)$ to approximate μ
- Two examples
  - Generating images
  - Find a transport map from prior to posterior

Fake

Fake

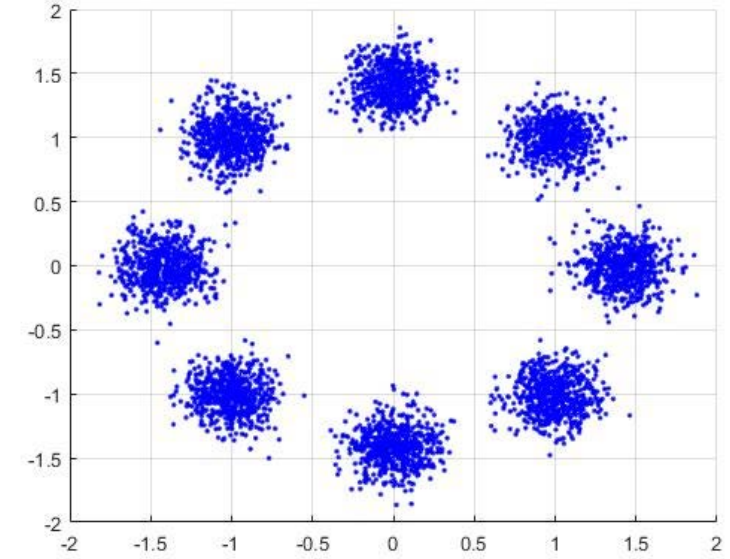
What: density/distribution estimation

- Target distribution μ
- Learn a generative model $G(z)$ with $z \sim N(0, I)$ to approximate μ
- Two examples
  - Generating images
  - Find a transport map from prior to posterior



$$p(x|d) \propto l(d|x)\, p_0(x)$$

$$p_{x|d} \approx G_*\, p_0$$



Why we care

- Learn a latent factor model
  - Dimension reduction, high-level abstractions, unsupervised learning
  - Causality: “What I cannot create, I do not understand.” – Richard Feynman

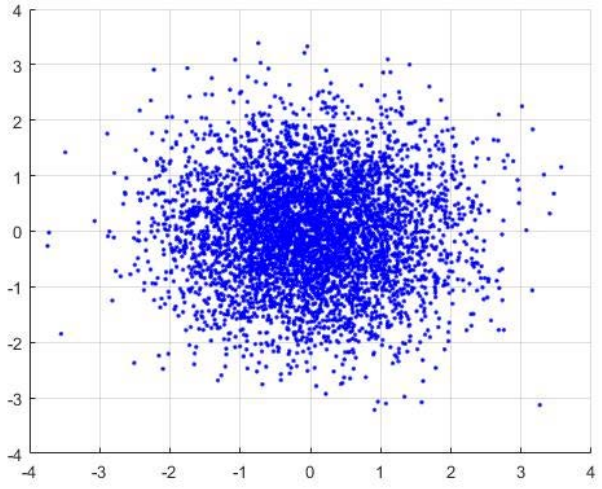
a herd of cows that are grazing on the grass

an old clock next to a light post in front of a steeple

A stop sign flying in the sky

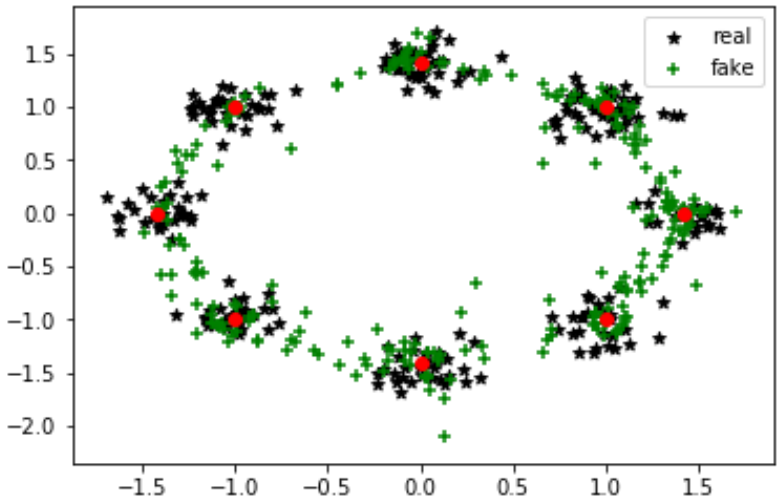
Why we care

- Learn a latent factor model
  - Dimension reduction, high-level abstractions, unsupervised learning
  - Causality: “What I cannot create, I do not understand.” – Richard Feynman
- Learn a transport map (sampling without MCMC)
  - MCMC: Slow mixing rate, curse of dimensionality, correlated samples
  - Uncertainty quantification, especially in high dimensional case



$$p(x|d) \propto l(d|x)\, p_0(x)$$

$$p_{x|d} \approx G_*\, p_0$$



How to learn a generative model $x = G_w(z)$?

$$\min_w D(\, G_{w,*}\, \mu_z\, , \mu)$$

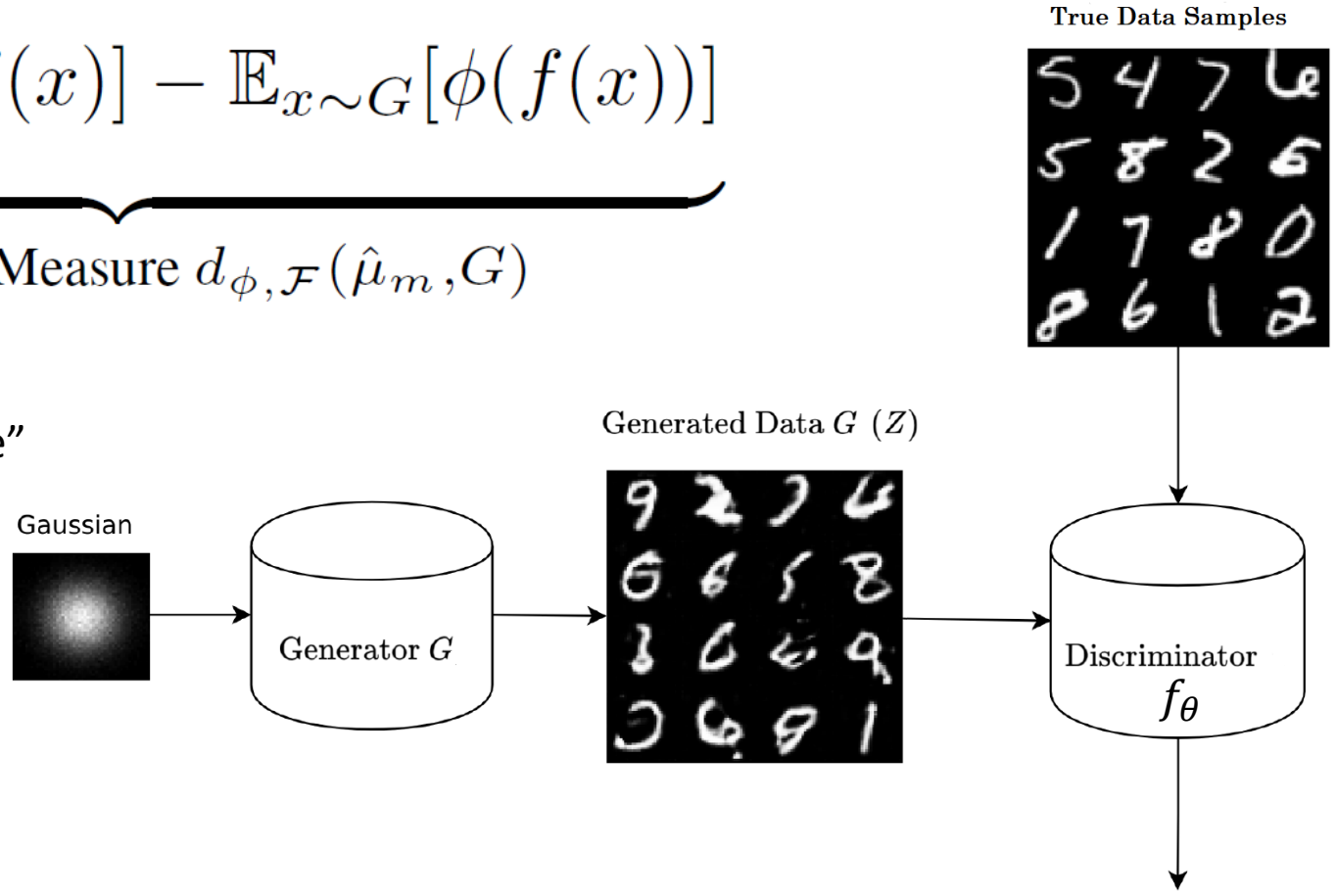
- $G_{w,*}\, \mu_z$
  - Easy to sample from $G_{w,*}\, \mu_z$
  - Density function is difficult to evaluate, even does not exist (maximize likelihood)
- Case 1: only finite samples from μ are available
  - Easy to sample, no access to density ρ_μ
  - VAEs (maximize lower bound of log likelihood), GANs (generalized moment matching)
  - Analysis and applications of GANs (part 1)
- Case 2: density function ρ_μ is available, i.e., $p(x|d) \propto l(d|x)\, p_0(x)$
  - Access to ρ_μ, difficult to sample
  - Training algorithms and preliminary results (part 2)
- Conclusions

Part 1
Analysis and applications of GANs

What are Generative Adversarial Networks (GAN)?

$$\min_{G \in \mathcal{G}} \underbrace{\max_{f \in \mathcal{F}} \mathbb{E}_{x \sim \hat{\mu}_m}[f(x)] - \mathbb{E}_{x \sim G}[\phi(f(x))]}_{\text{Discrepency Measure } d_{\phi,\mathcal{F}}(\hat{\mu}_m, G)}$$

- $\phi = id$: "distance"
- ϕ convex function: "divergence"



Goodfellow et al. (2014)

Gap between theory and practice

$$\min_{G \in \mathcal{G}} d_{\phi, \mathcal{F}}(\hat{\mu}_m, G)$$

GAN variants	Discriminator set	Discrepancy measure
Wasserstein GAN	$\{f : \|\|f\|\|_{Lip} \leq 1\}$	Wasserstein distance
Energy-based GAN	$\{f : 0 \leq f \leq C\}$	Total variation distance
Vanilla GAN	$\{f : f < 0\}$	Jensen-Shannon divergence
f-gan	Other non-parametric function classes	Phi-divergence
Neural GANs	Neural networks f_θ	Neural distance/divergence

Cake of non-parametric discriminator set

- Discrimination: When the neural distance/divergence converges to its minimum, does the learned distribution converge to the target distribution?
- Generalization: Given only finite number of samples, is GAN only memorizing the samples? Or it can learn the underlying target distribution?

Discrimination of GANs with neural distance

Theorem 2.1. *For any target distribution μ, the neural distance is discriminative, i.e.,*

$$d_{\mathcal{F}}(\mu, G) = 0 \text{ implies } \mu = G$$

if and only if

$$\text{span}(\mathcal{F}) \text{ is dense in bounded continuous function space,}$$

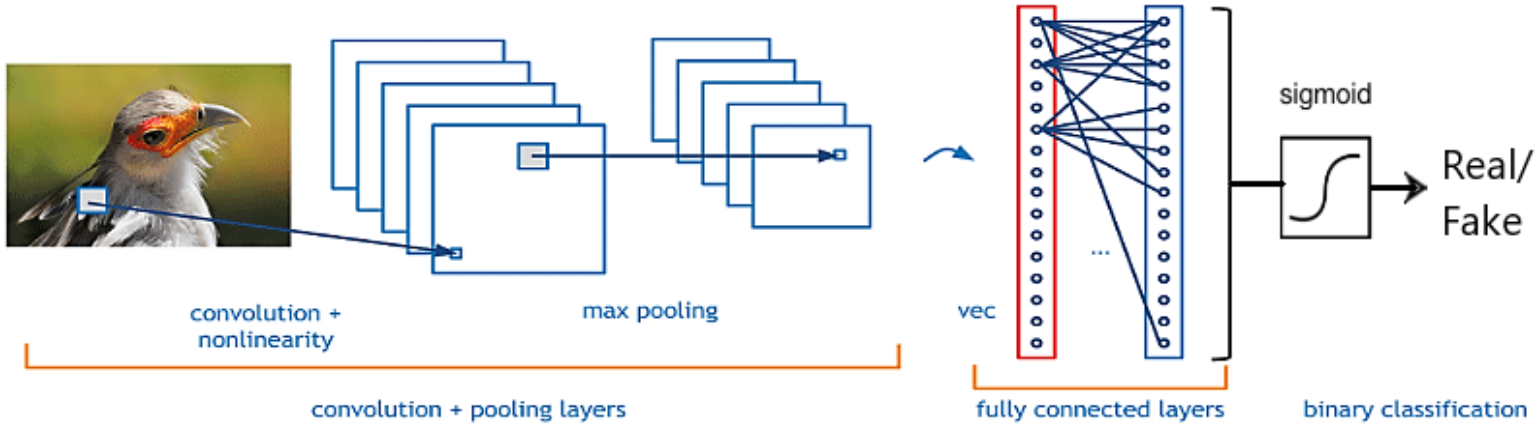
where $\text{span}(\mathcal{F})$ is the subspace consisting of all linear combinations of functions in $\mathcal{F}$.

Under the same condition, $d_{\mathcal{F}}(\mu, G_m) \to 0$ implies that G_m weakly converges to μ.

- Much weaker compared with the previous beliefs: the discriminator set should be dense in the original non-parametri
  - Satisfied even by neural networks with a single neuron
  - Nearly all neural GANs with (leaky) ReLu activations in
- GANs with neural distance are essentially doing m discriminators
  - Moment matching on F implies moment matching on

Discrimination of GANs with neural divergence

- GANs with neural divergence, e.g., neural f-GANs, are discriminative if span of **discriminators without the last nonlinear activation** is dense in the bounded continuous function space.



  - GANs with neural divergence are doing moment matching on the feature vectors.
  - Neural f-GAN and neural WGAN are doing moment matching on the same features, and sharing the same conditions for discrimination.

Generalization of GANs

$$\min_{G \in \mathcal{G}} d_{\mathcal{F}}(\hat{\mu}_m, G)$$

Theorem 3.1. *Assume that all discriminators are bounded by Δ, i.e., $\|f\|_\infty \le \Delta$ for any $f \in \mathcal{F}$. Let $\hat{\mu}_m$ be an empirical measure of an i.i.d. sample of size m drawn from the target distribution μ. Assume $G_m \in \mathcal{G}$ satisfies*

$$d_{\mathcal{F}}(\hat{\mu}_m, G_m) \le \inf_{G \in \mathcal{G}} d_{\mathcal{F}}(\hat{\mu}_m, G) + \epsilon.$$

Then with probability at least $1 - \delta$, we have

$$d_{\mathcal{F}}(\mu, G_m) \le \underbrace{\inf_{G \in \mathcal{G}} d_{\mathcal{F}}(\mu, G)}_{\text{Modeling error}} + \underbrace{R_m^{(\mu)}(\mathcal{F}) + 2\Delta\sqrt{\frac{2\log(1/\delta)}{m}}}_{\text{Generalization error}} + \underbrace{\epsilon}_{\text{Optimization error}},$$

where $R_m^{(\mu)}(\mathcal{F}) := \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left|\frac{2}{m}\sum_i \tau_i f(X_i)\right|\right]$ *is the Rademacher complexity of* $\mathcal{F}$.

- The generalization error is bounded independent of the hypothesis set $\mathcal{G}$.
  - Big difference from the supervised learning
- We also give generalization error under other metrics, like KL divergence and Wasserstein distance.
- Similar theorem for neural divergence.

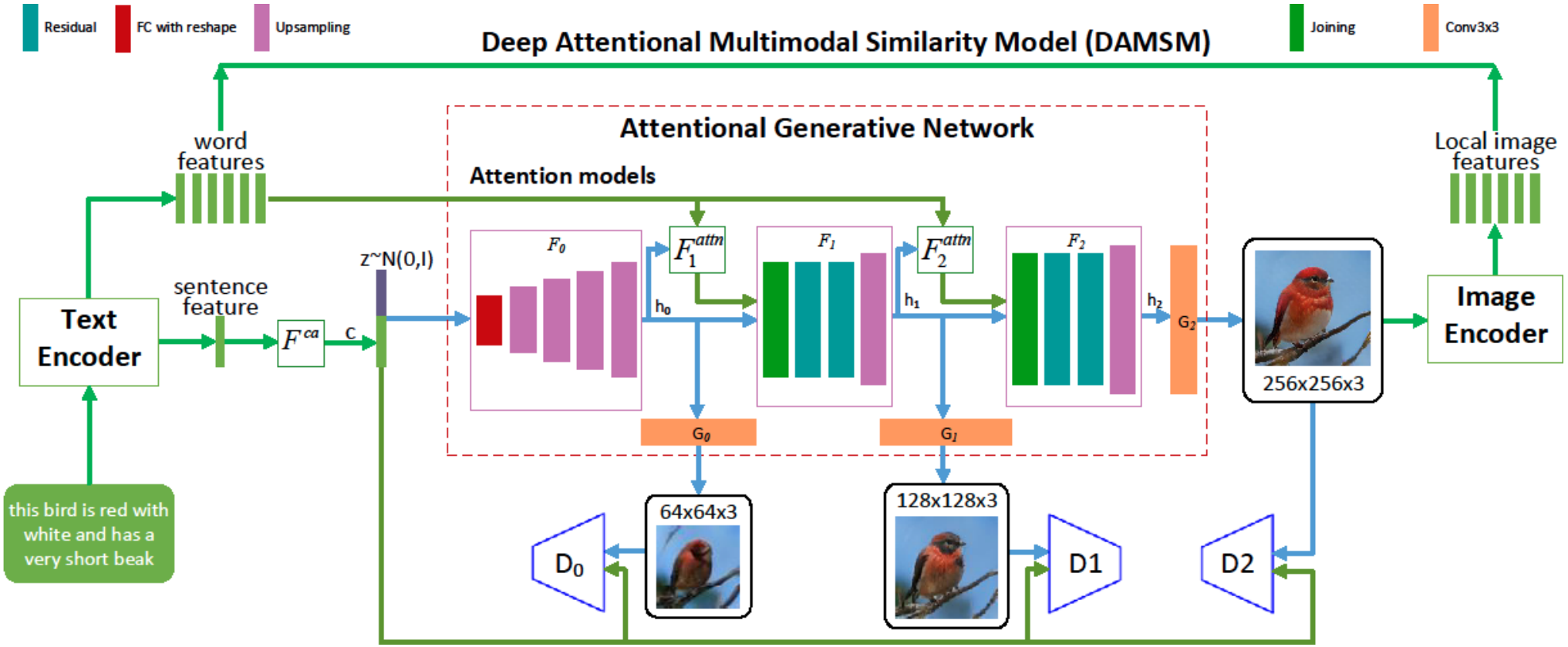
Generalization of GANs, examples

GAN variants	Discriminator set	Generalization error	Sample size	Tight or not
Wasserstein GAN	$\{f: \|\|f\|\|_{Lip} \leq 1\}$	$O(m^{-1/d})$	$O(\epsilon^{-d})$	Yes
MMD-GAM	$\{f \in H: \|\|f\|\|_H \leq 1\}$	$O(m^{-1/2})$	$O(\epsilon^{-2})$	Yes
Energy-based GAN	$\{f: 0 \leq f \leq C\}$	$O(1)$	∞	Yes
Vanilla GAN	$\{f: f < 0\}$	$O(1)$	∞	Yes
Neural GANs	Neural networks f_θ	$O(m^{-1/2})$	$O(\epsilon^{-2})$	Yes

- Our bound is tight w.r.t. the order of sample size m
- Most GANs with their original discriminator sets do not generalize.
- Several GANs in practice, like WGAN with weight clipping, already choose their discriminator sets at the sweet point, where both the discrimination and generalization hold.

AttnGAN – Application to text-to-image synthesis

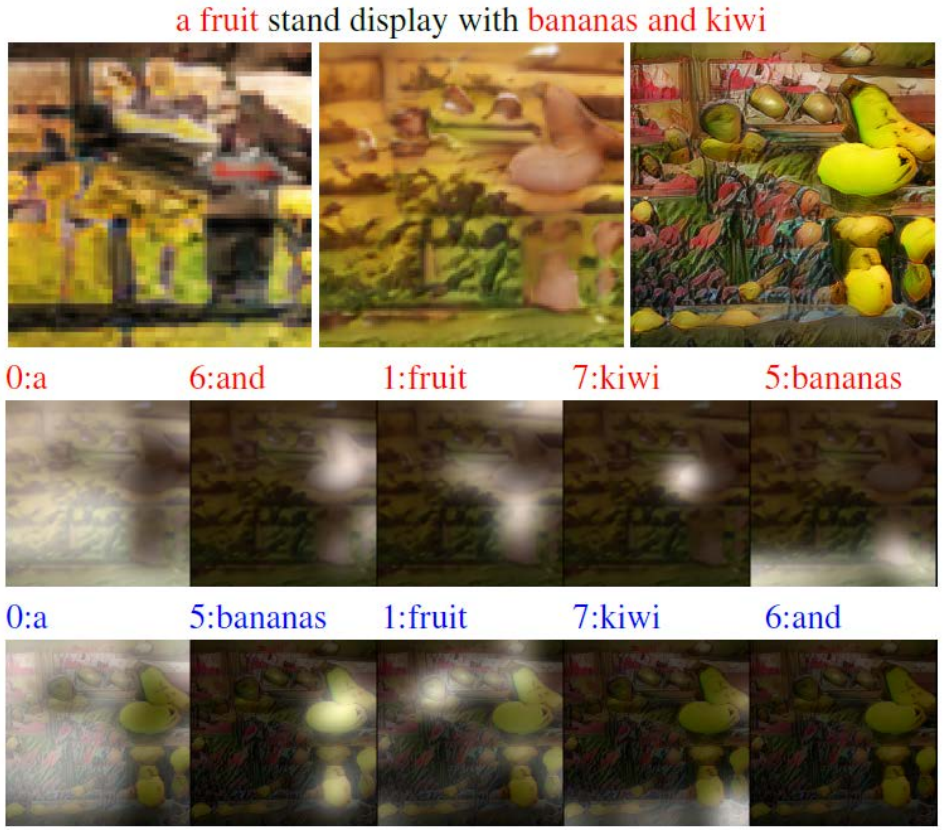
An illustration of the proposed AttnGAN.



In the figure, the generator tries to fool the discriminator, and tries to match the given text content.

Results

- AttnGAN significantly outperforms the previous state of the art, by boosting the best inception score by 14.12% on the CUB dataset, and 170.25% on the COCO dataset.
- For the first time, AttnGAN is able to select the condition at the word level for generating different parts of the image.



Diversity of the generated images

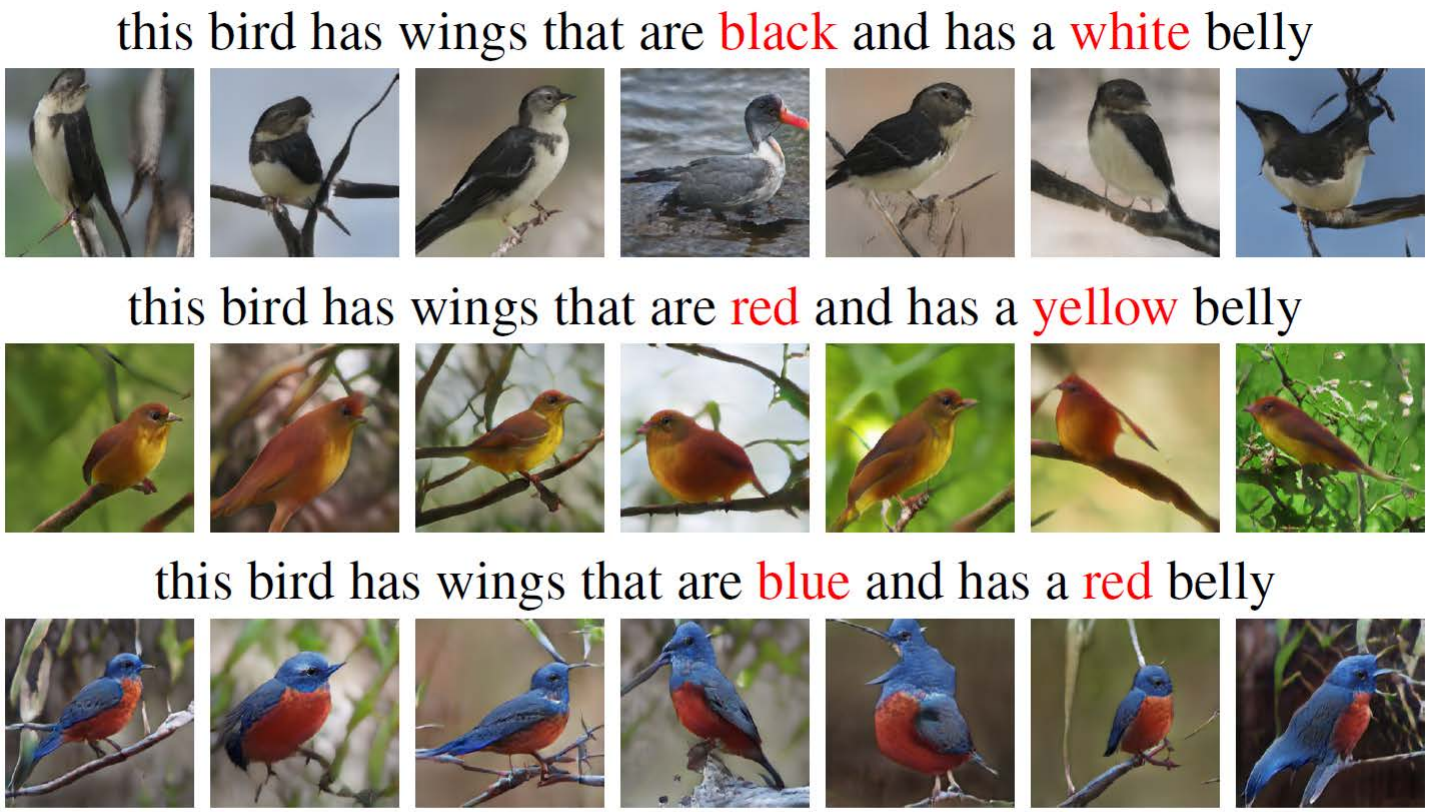
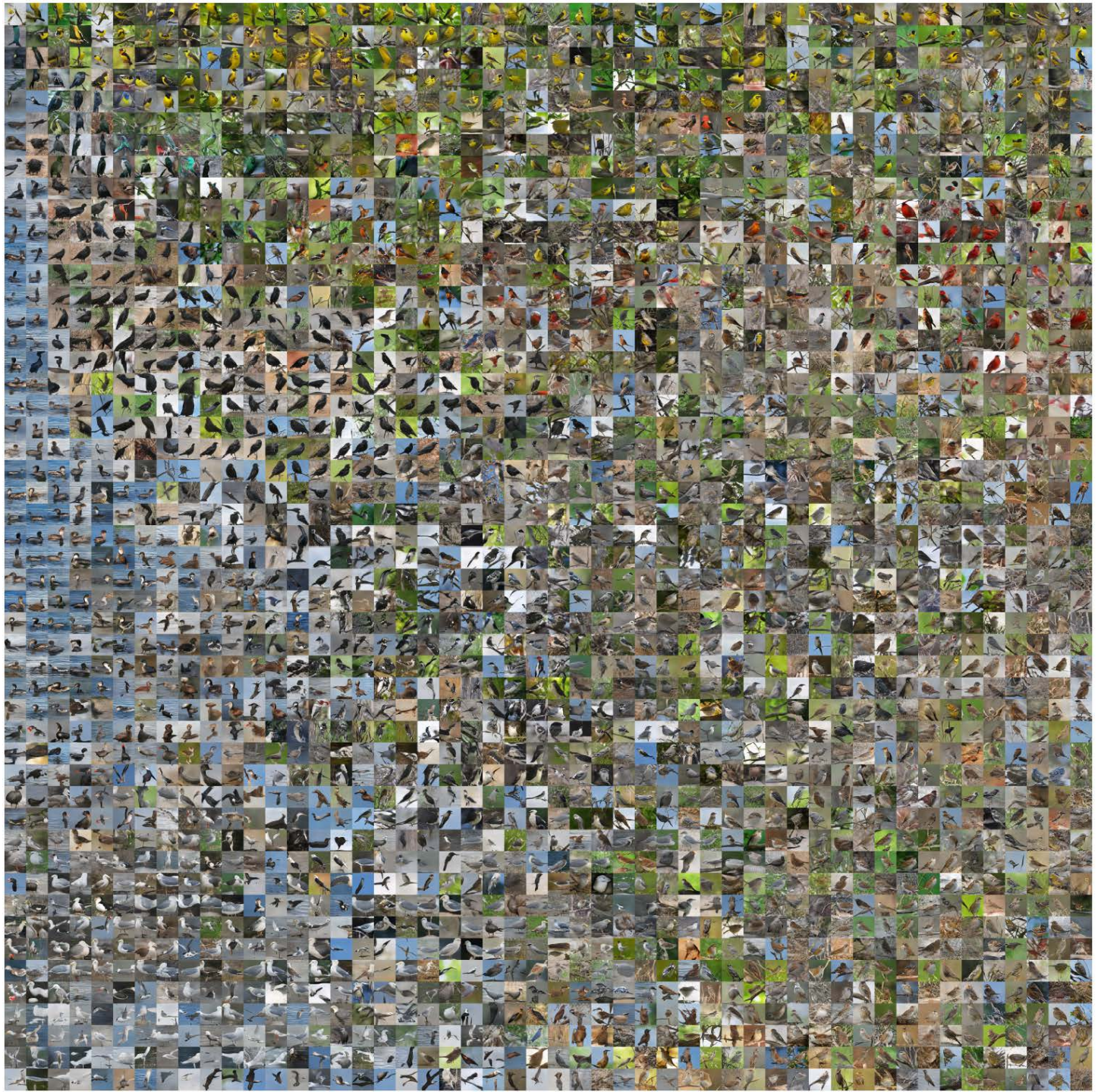


Figure 5. Example results of our AttnGAN model trained on CUB while changing some most attended words in the text descriptions.

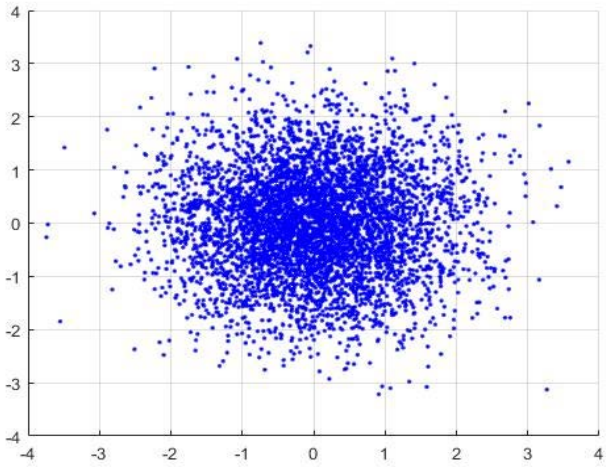
Utilizing t-SNE to embed a large number of images generated by the AttnGAN



Part 2
Deep generative models in inverse problems
Algorithm and preliminary results

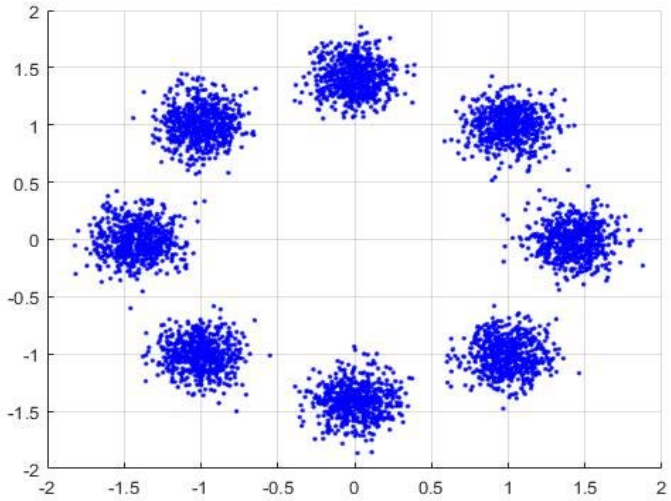
Bayesian inference with transport map

- MCMC
  - Slow mixing rate, curse of dimensionality, correlated samples
- Transport map
  - Chorin & Tu, 2009, implicit sampling: U-shape assumption
  - Moselhy & Marzouk, 2012, measure-preserving maps: polynomial basis, curse of dimensionality
  - Many follow-ups and other related work
- Deep generative model as transport map
  - High capacity, dimension-independent models (making use of the data structure)
  - Density is not computable, maximal likelihood does not work



$$p(x|d) \propto l(d|x)\, p_0(x)$$

$$p_{x|d} \approx G_*\, p_0$$



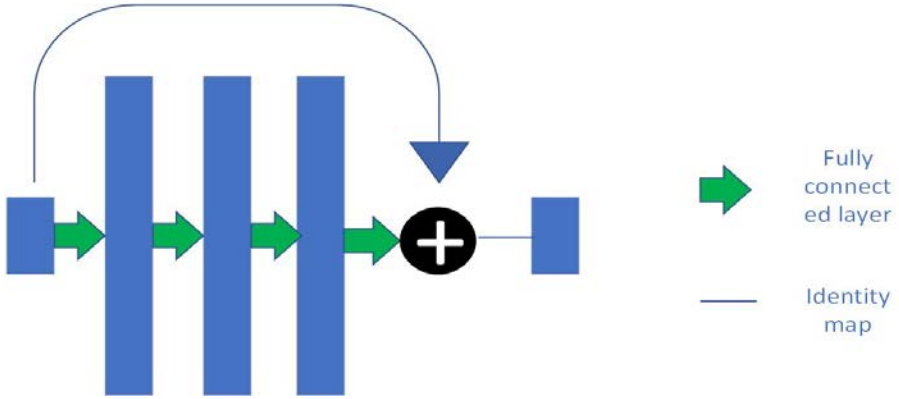
Training algorithm

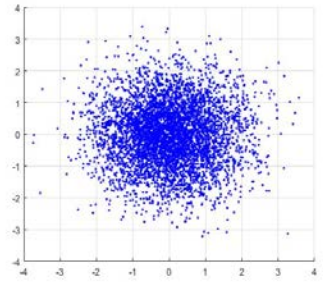
$$\min_{w} KL(\, G_{w,*}\, \rho_z \,,\, p_{x|d}\,)$$

1. Draw random $\{z_i\}_{i=1}^{m}$, calculate $x_i = G_w(z_i)$
2. Compare $\{x_i\}_{i=1}^{m}$ with target distribution $p_{x|d}$, do update:
   $x_i \leftarrow x_i + \epsilon\, v^*(x_i)$
3. Using chain rule to update the generative model:
   $w \leftarrow w + \frac{\epsilon}{m} \sum_{i=1}^{m} v^*(x_i)\, \partial_w G(z_i)$

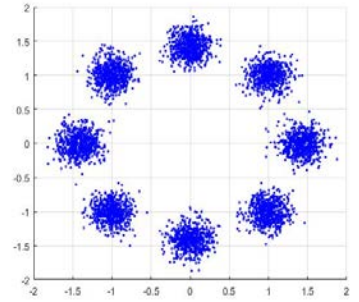
Toy example: 8-gaussian mixture

- Structure of the generator

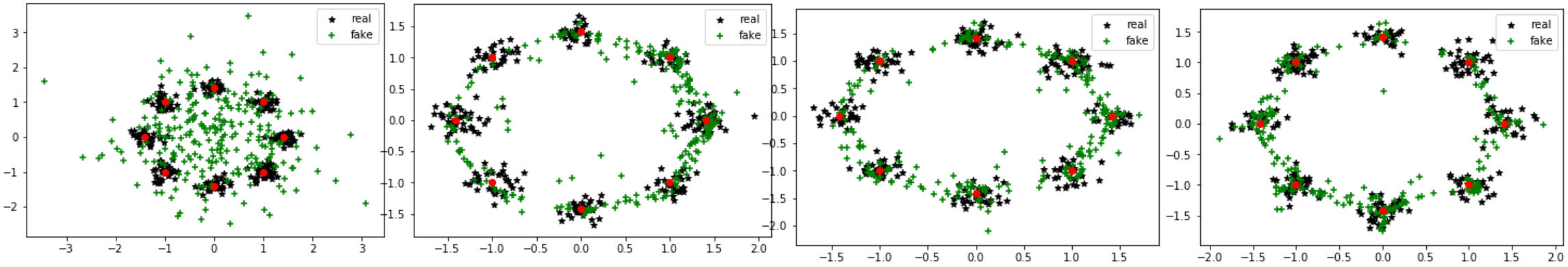




$$p_{x|d} \approx G_* \, p_0$$



- Trained generators (iteration 0, 5e3, 1e4, 1.5e4)



Elliptic equation

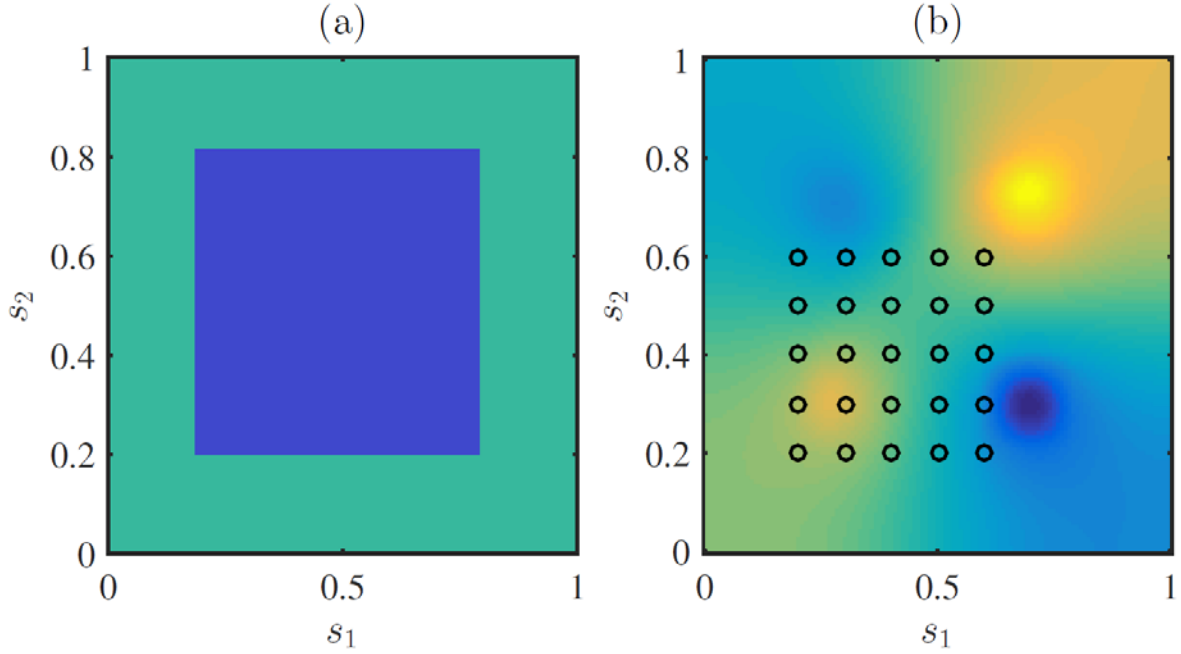
- Forward model

$$\begin{cases} -\nabla_s \cdot (\exp(u(s))\nabla_s p(s)) & = f(s), \quad s \in \Omega \\ \langle \exp(u(s))\nabla p(s),\ \boldsymbol{n}(s) \rangle & = 0, \quad s \in \partial\Omega \end{cases}$$

  - u: parameters to estimate, p: measurements, f: given force
  - Discretized on a 40*40 uniform grid
  - Prior of u(s): $N(0, \Sigma)$
  - Noisy measurements of u

(a). Ground truth log permeability field u(s)

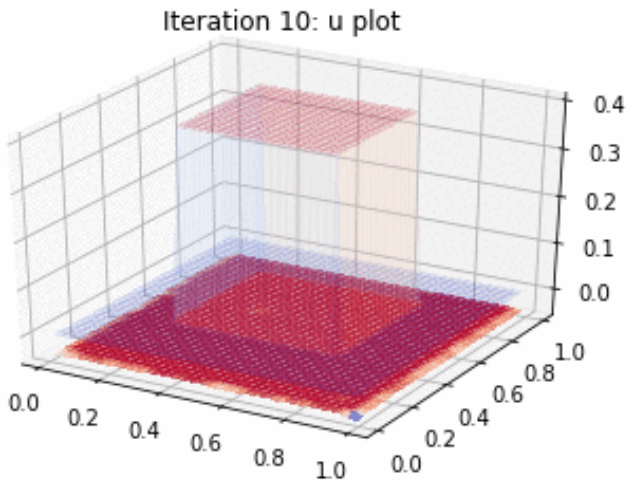
(b). Pressure field p(s) and measurement locations



Elliptic equation

- Structure of the generator
  - U-net (Ronneberger et al, 2015)

- Evolution of a random sample



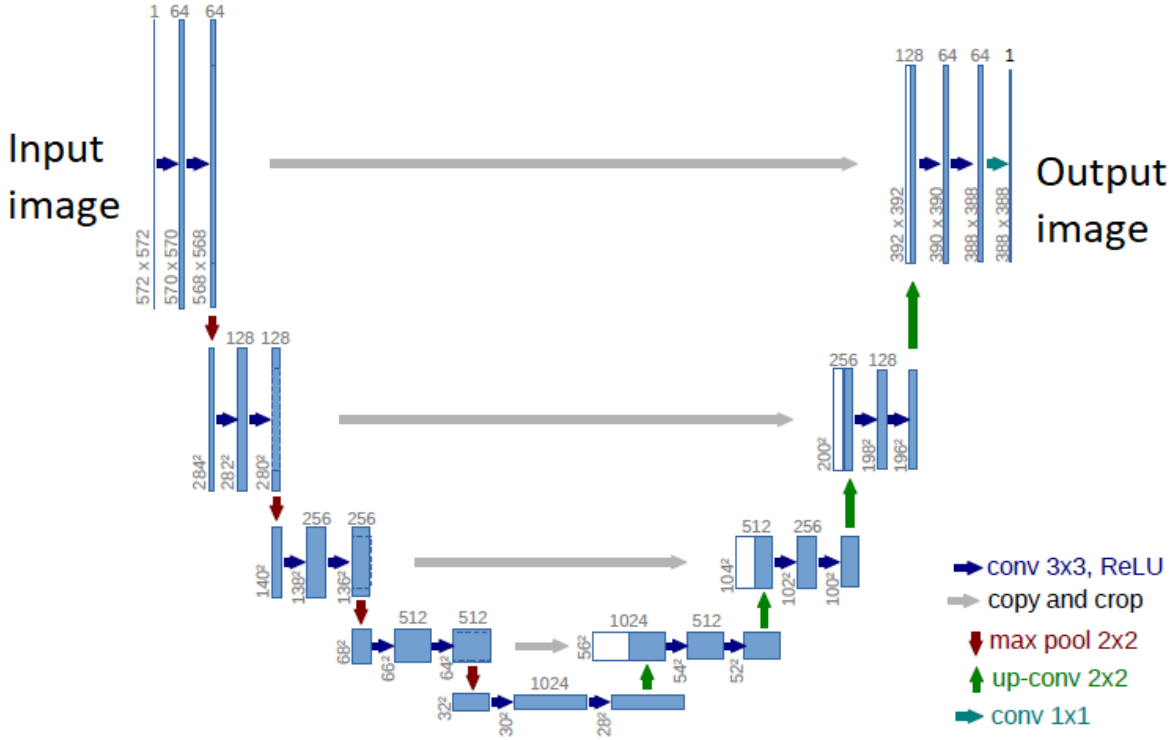


Fig. 1. U-net architecture (example for 32x32 pixels in the lowest resolution). Each blue box corresponds to a multi-channel feature map. The number of channels is denoted on top of the box. The x-y-size is provided at the lower left edge of the box. White boxes represent copied feature maps. The arrows denote the different operations.

Challenges

- Evaluation criteria
  - Accuracy in mean and variance (asymptotically unbiased)
  - Sample correlation
  - Mixing rate
  - Sample diversity
- Generator structures
  - Taking the regularity of the forward model into account

Conclusions

- We provided the necessary and sufficient conditions for GANs to be discriminative/consistent.
- We provided a general and tight bound for the generalization error in GANs.
- We proposed the AttnGAN for text-to-image generation task, and outperformed previous state-of-the-art results.
- We proposed to learn a deep generative model as the transport map from prior and posterior in Bayesian inverse problems.
- We showed our preliminary results, and pointed out challenges we met.

QnA

References

- Pengchuan Zhang, Qiang Liu, Dengyong Zhou, Tao Xu, and Xiaodong He. On the discrimination-generalization tradeoff in GANs. ICLR, 2018.
- Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. arXiv preprint arXiv:1711.10485, 2017.