Generative Adversarial Text to Image Synthesis

Amit Manchanda
14116013

Anshul Jain
14116016

Under the guidance of
Dr. Vinod Pankajakshan
(Assistant Professor, ECE, IIT Roorkee)

Content

- Objective
- Background
  - GANs
  - Text Embeddings
- Methodology and Results
  - Vanilla GANs
  - WGANs
  - Attention GANs
- Future Scope of Research
- Conclusion

Objective

Translating text in form of single statement human written descriptions directly into image.

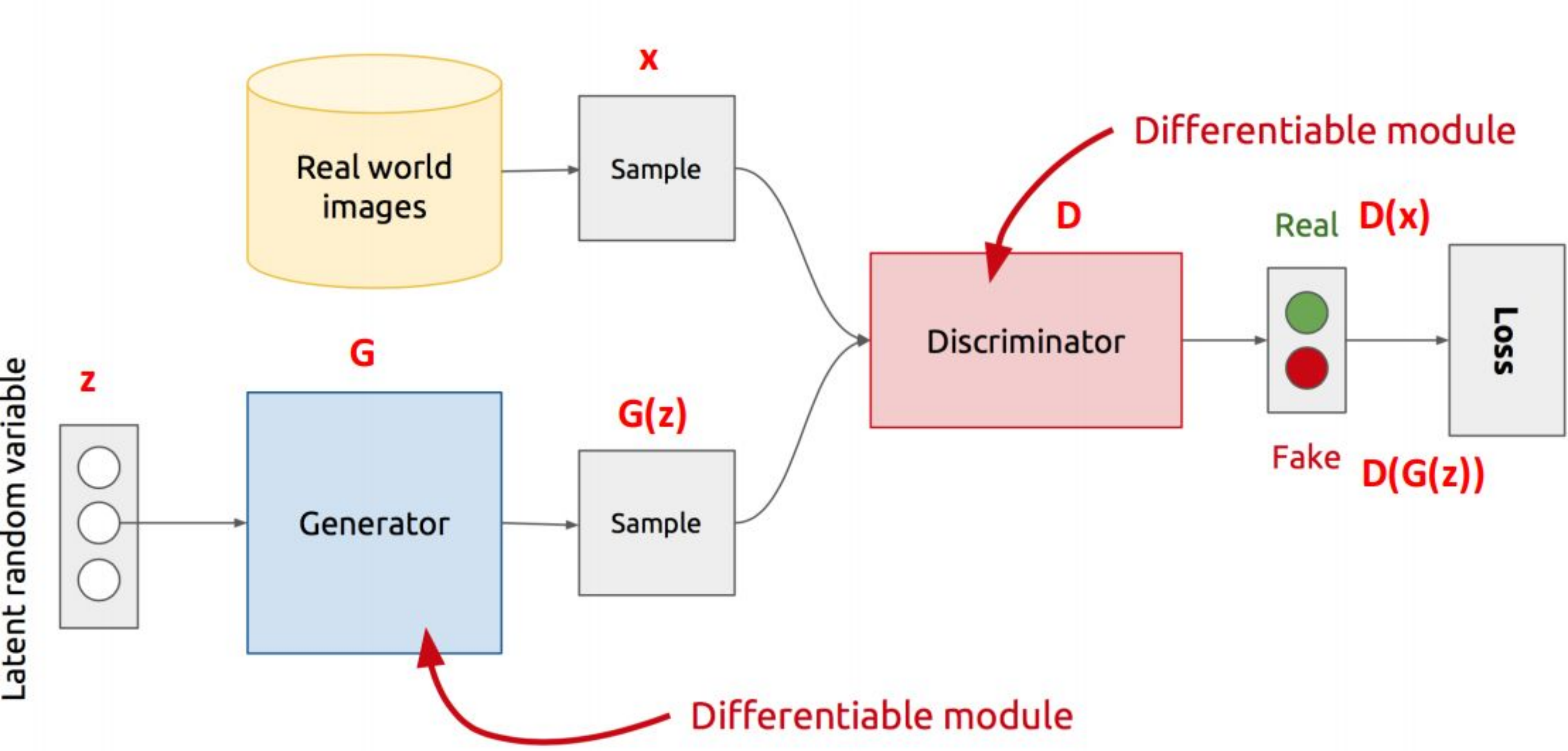
this bird has yellow belly breast throat eyebrow with black and grey wings and tail

the blue backed white bellied baby bird has a very fat little belly

this is a very small bird with a white belly and side the bird s head and wings are black

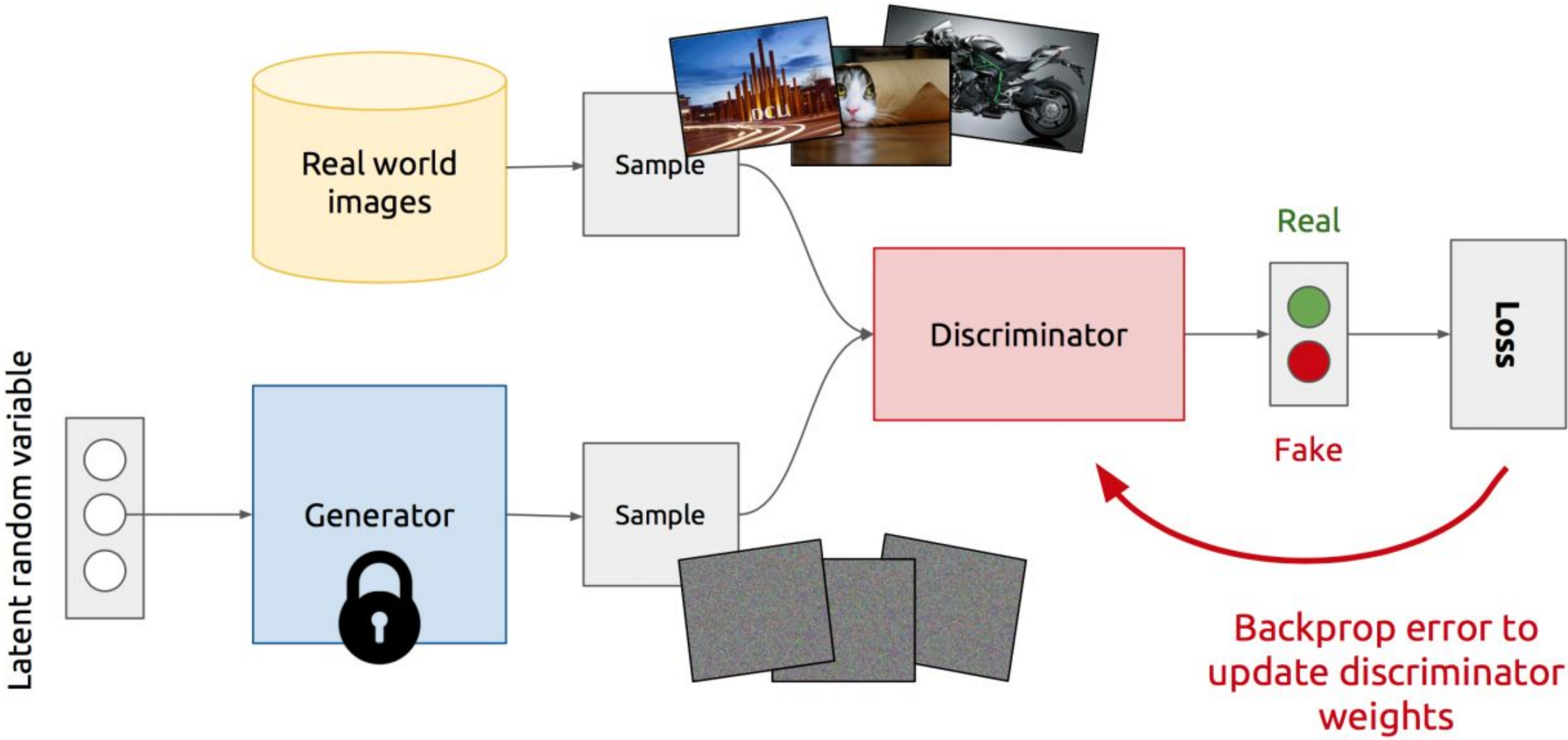
Background

Background: GANs



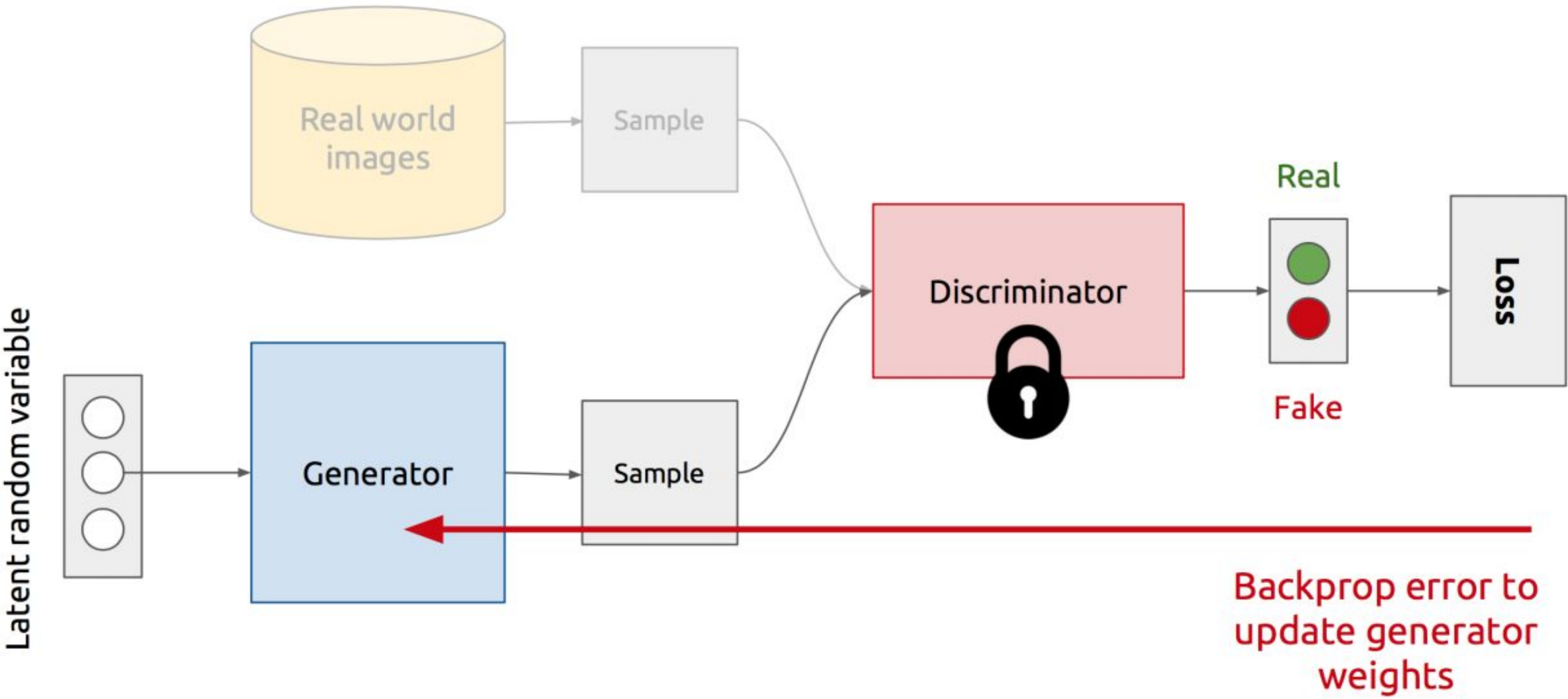
Source : http://slazebni.cs.illinois.edu/spring17/lec11_gan.pdf

Background: GANs (continued)



Source : http://slazebni.cs.illinois.edu/spring17/lec11_gan.pdf

Background: GANs (continued)

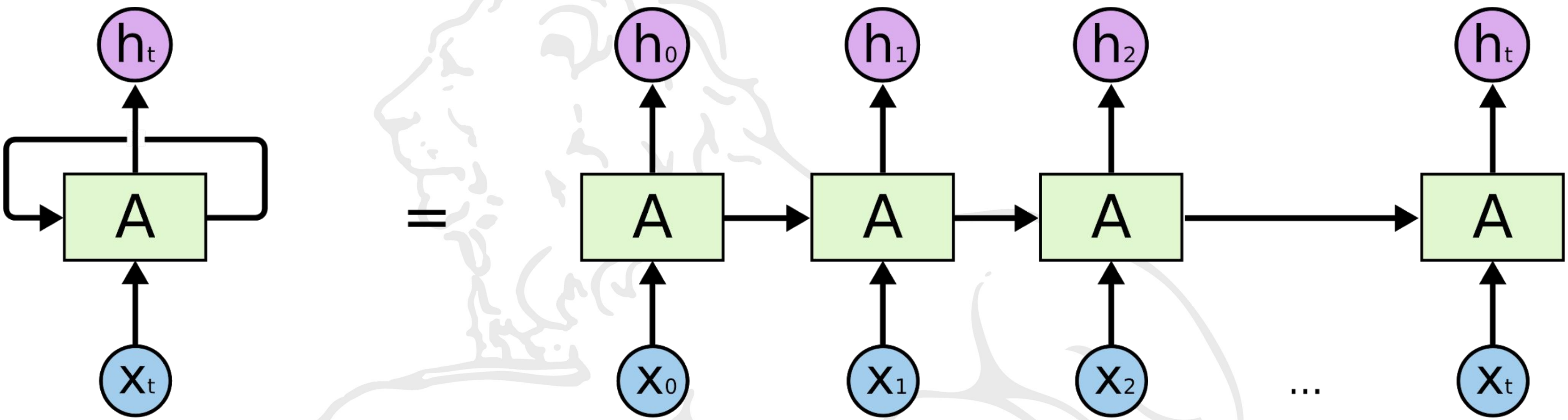


$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{data}(x)} \log D(x) + \mathbb{E}_{z\sim p_z(z)} \log(1 - D(G(z)))$$

Source : http://slazebni.cs.illinois.edu/spring17/lec11_gan.pdf

Background: Text Embedding

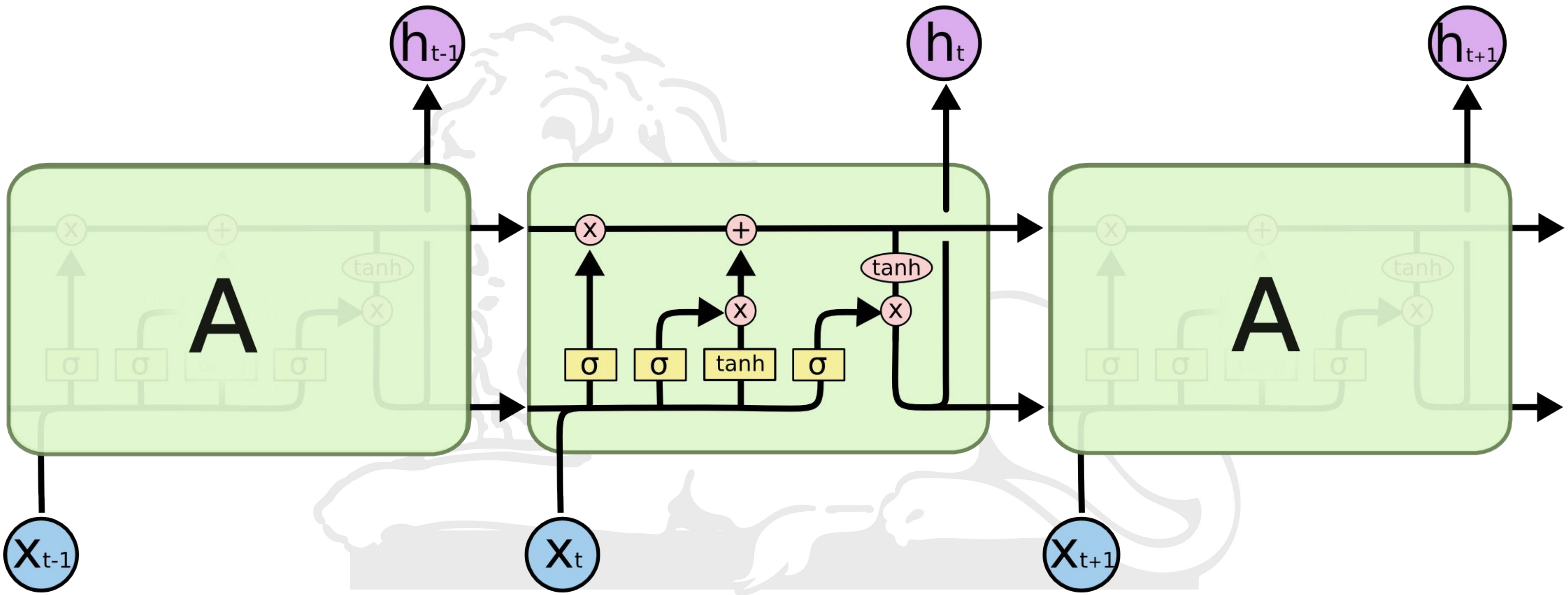
Recurrent Neural Network



Source: http://colah.github.io/posts/2015-08-Understanding-LSTMs/

Background: Text Embedding (continued)

Long Short Term Memory Network



Source: http://colah.github.io/posts/2015-08-Understanding-LSTMs/

Background: Text Embedding (continued)

- Skip-thought Vectors :

  Consists of an encoder-decoder model which generates the surrounding sentences based on the given sentence.

- The following objective function is to be optimized.

$$\sum_t logP(w_{i+1}^t | w_{i+1}^{<t}, h_i) + \sum_t logP(w_{i-1}^t | w_{i-1}^{<t}, h_i)$$

Datasets

Datasets

We used Caltech-UCSD Birds(CUB) dataset and Oxford-102 flowers dataset.

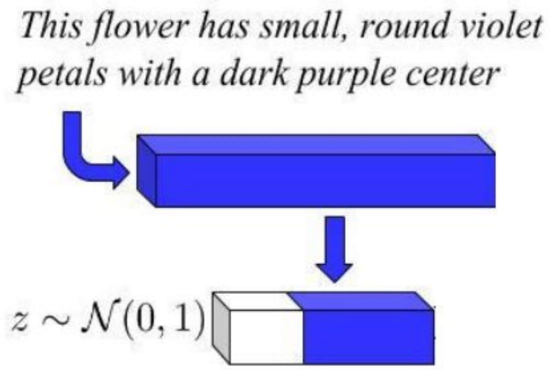
- CUB dataset contains 11,788 birds images of 200 categories.
- Oxford-102 dataset contains 8,189 images from 102-different flower categories.

CUB	Train	Test
#samples	8,855	2,933
caption/images	10	10

Oxford-102	Train	Test
#classes	82	20
#samples	6,142	2,047
caption/images	5	5

Methodology and Results

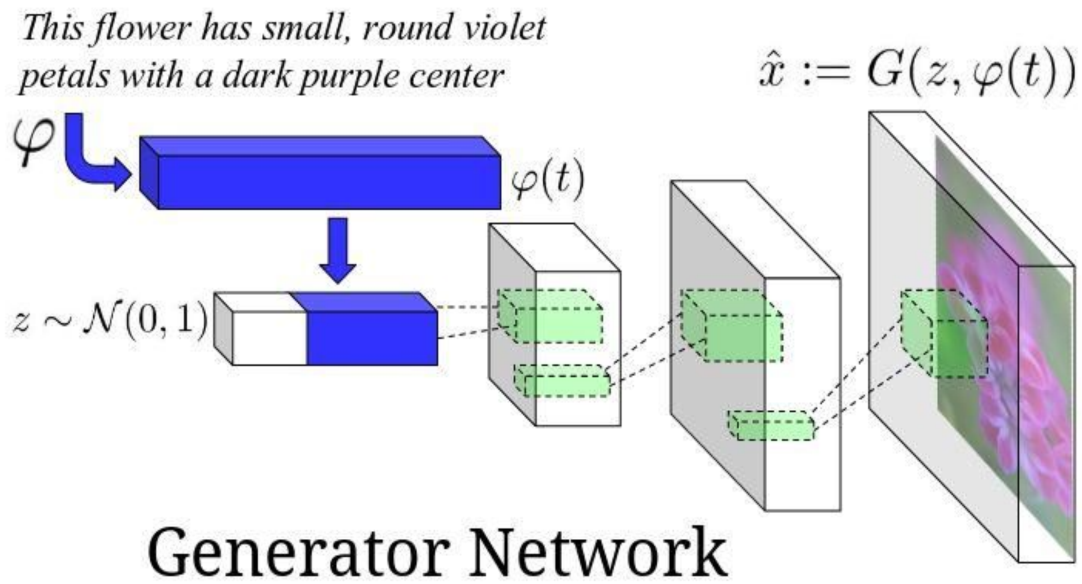
Vanilla GANs



Source: [1]

Vanilla GANs (continued)

$s_f \leftarrow \mathrm{D}(\hat{\mathrm{x}}, \mathrm{h})$ fake image, right text



$$L_g \leftarrow log(s_f)$$

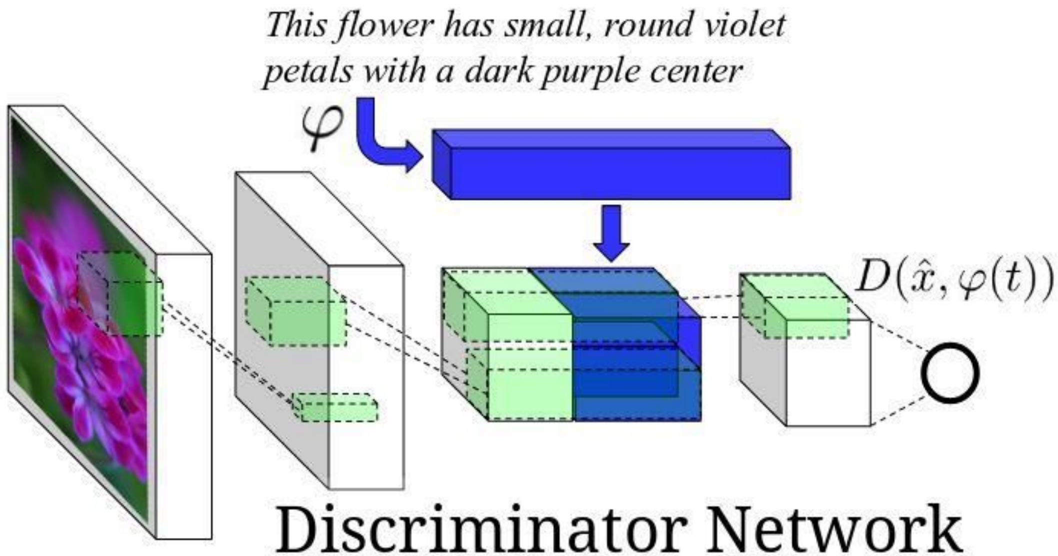
Source: [1]

Vanilla GANs (continued)

$s_r \leftarrow$ D(x, h) real image, right text

$s_w \leftarrow$ D(x, $\hat{\text{h}}$) real image, wrong text

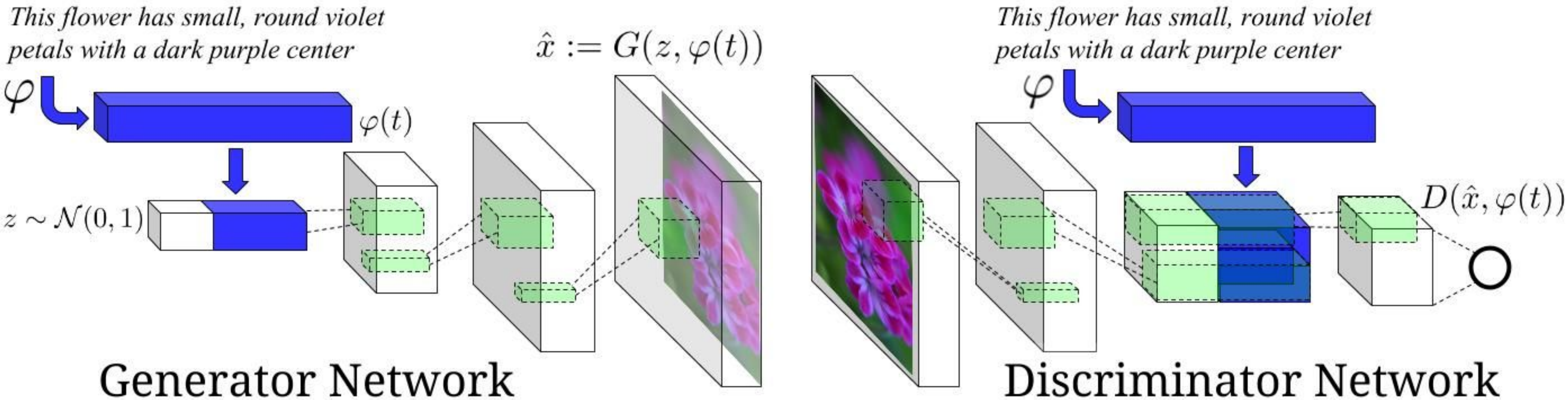
$s_f \leftarrow$ D($\hat{\text{x}}$, h) fake image, right text



$$L_d \leftarrow log(s_r) + (log(1 - s_w) - log(1 - s_f))/2$$

Source: [1]

Vanilla GANs (continued)



$$\min_G \max_D V(D,G) = \mathbb{E}_{x \sim p_{data}(x)} \log D(x|c) + \mathbb{E}_{z \sim p_z(z)} \log(1 - D(G(z|c)))$$

Source: [1]

Vanilla GANs (continued)

the flower has abundance of yellow petals and brown anthers

flower is purple and pink in petal and features a dark dense core

this flower has petals that are red and bunched together

the petals of this flower are white and the pistil is a golden yellow

the petals of the flower are pink in color and have a yellow center

this flower is yellow in color, and has petals that are uneven along the edge

Vanilla GANs (continued)

WGANs

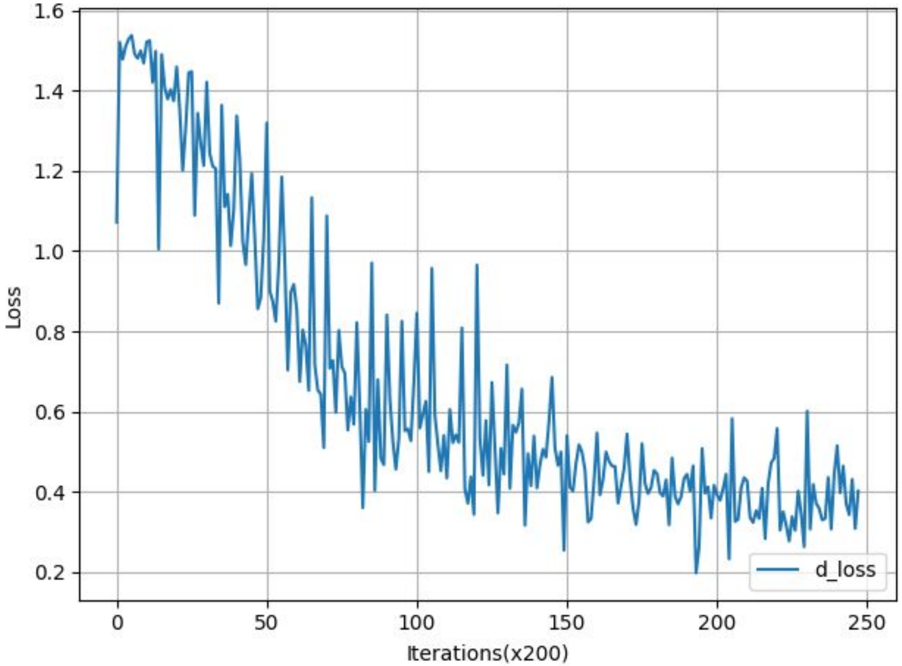
- Minimize the distance between real distribution and model distribution.
- Uses Earth-Mover or Wasserstein distance.
- We want to model a distribution P_θ as a generator network g dependent on parameter θ.

$$\nabla_\theta W(P_r, P_\theta) = \nabla_\theta(\mathbb{E}_{x\sim p_r}[f_w(x)] + \mathbb{E}_{z\sim p_z}[f_w(g_\theta(z))])$$

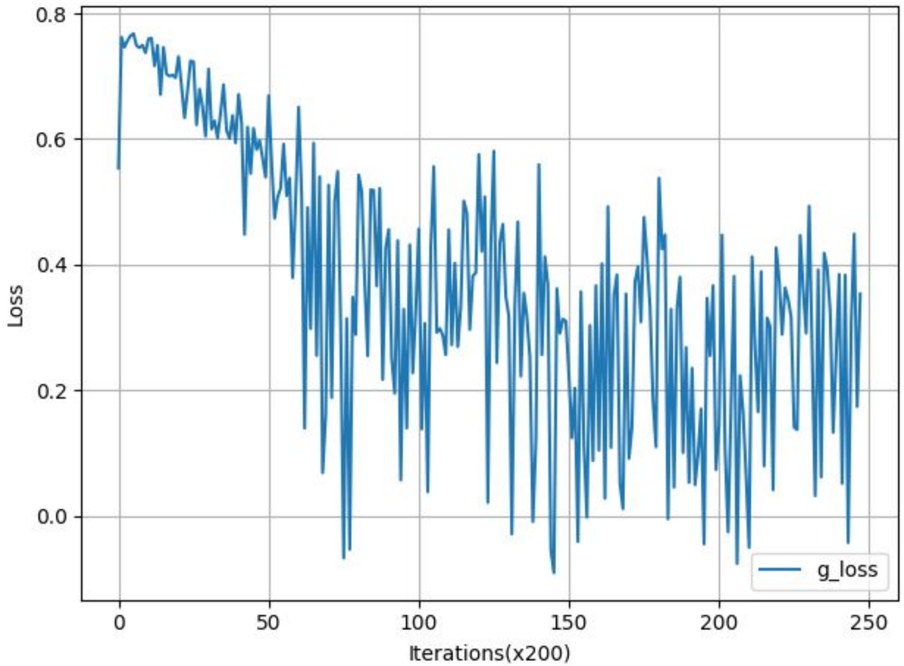
$$= -\mathbb{E}_{z\sim p_z}[\nabla_\theta f_w(g_\theta(z))]$$

where f_w is the critic.

WGANs (continued)



Wasserstein Loss



Generator Loss

WGANs (continued)

grey and lemon colored bird with black cheek patch.

this bird has a black crown as well as a green belly.

this beautiful gold and gray colored bird had a sharp pointed beak and black tail

this bird has a belly that is black with orange cheek patches

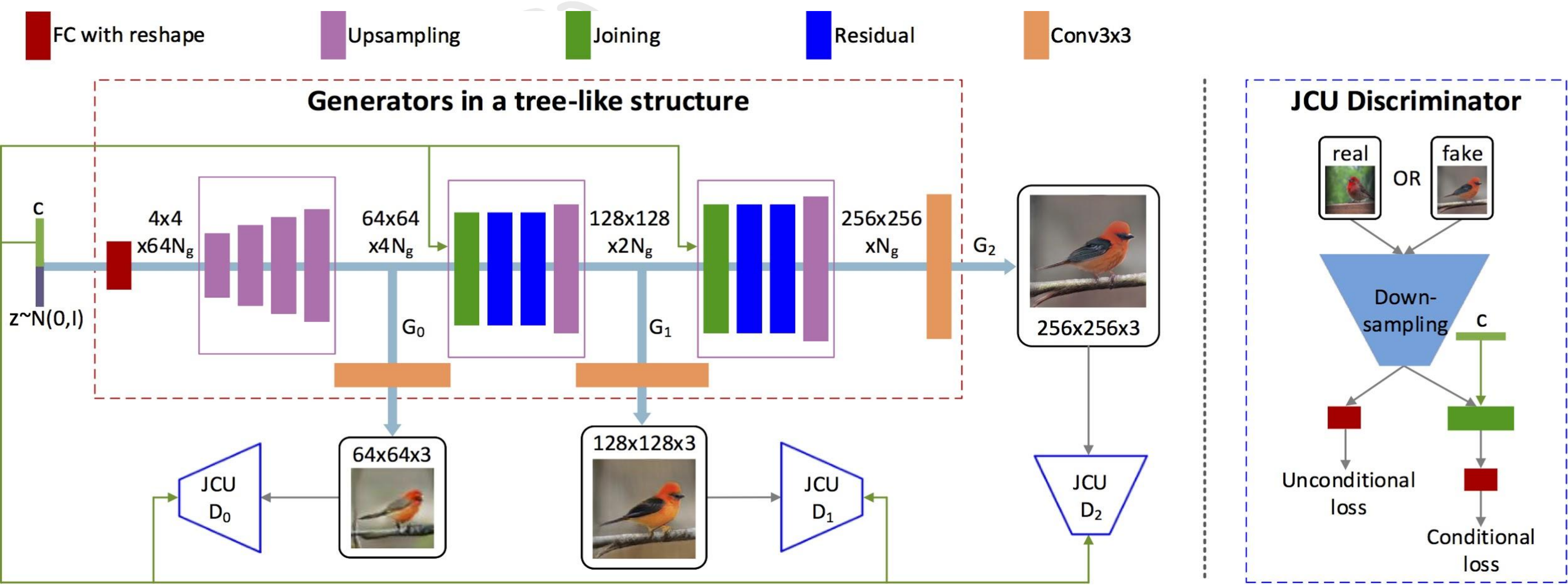
a fluffy bird with shades of browns and grays and a speck on white on it's tail.

a brown and gray bird with a short bill and an orange spot on it's crown.

a beautiful bird with black and white wings and a red head with a sharp pointy bill.

WGANs (continued)

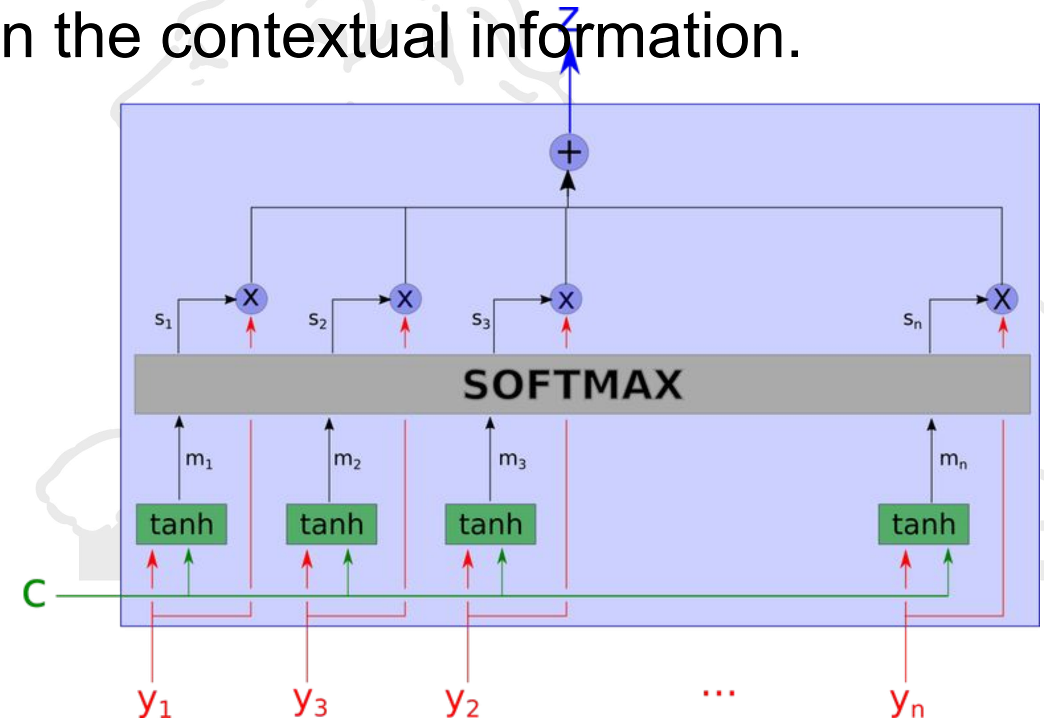
StackGAN : A multi stage generation process.



Source: [3]

Attention GANs : Attention Mechanism

- Motivated by the human tendency to focus on certain words.
- Model takes n inputs along with context and returns a weighted sum of inputs.
- Focus on the contextual information.



Attention Model

Attention GANs (continued)

Deep Attentional Multimodal Similarity Model (DAMSM)

- Text Encoder:
    - uses bi-directional LSTM to extract feature vectors
    - Global sentence vector is generated in the last state.
- Image Encoder:
    - uses part of Inception-v3 trained on ImageNet.
    - Global feature vector is taken from last pooling layer.
- DAMSM loss calculated to find similarity between image and sentence.

Attention GANs (continued)

Attention Generative Network

- Model has m generator discriminator pairs.
- Each generator takes hidden state h_i as input and produces an intermediary image.

$$\hat{x}_i = G_i(h_i)$$

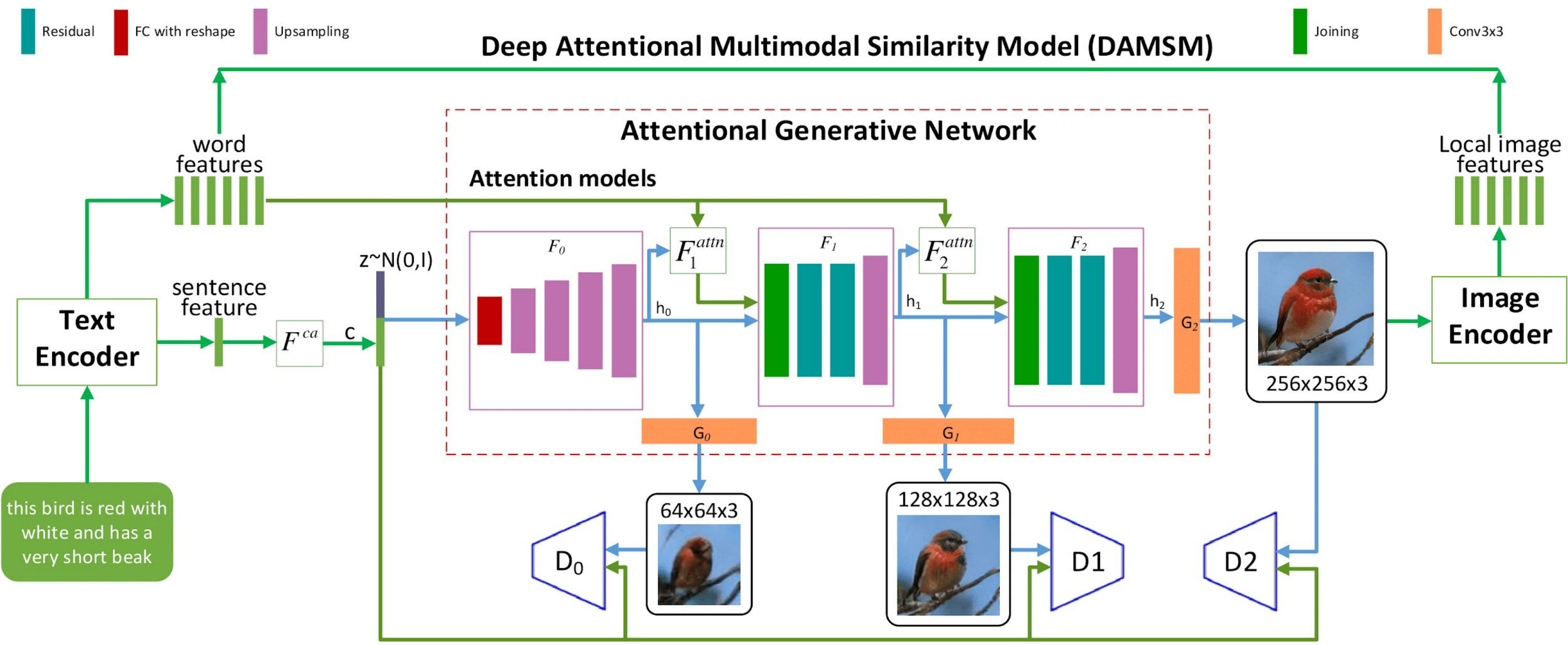
- Hidden states are defined as follows :

$$h_0 = F_0(z, F^{ca}(\bar{e}))$$

$$h_i = F_i(h_{i-1}, F_i^{attn}(e, h_{i-1})) \text{ for } i = 1, 2, \ldots, m-1$$

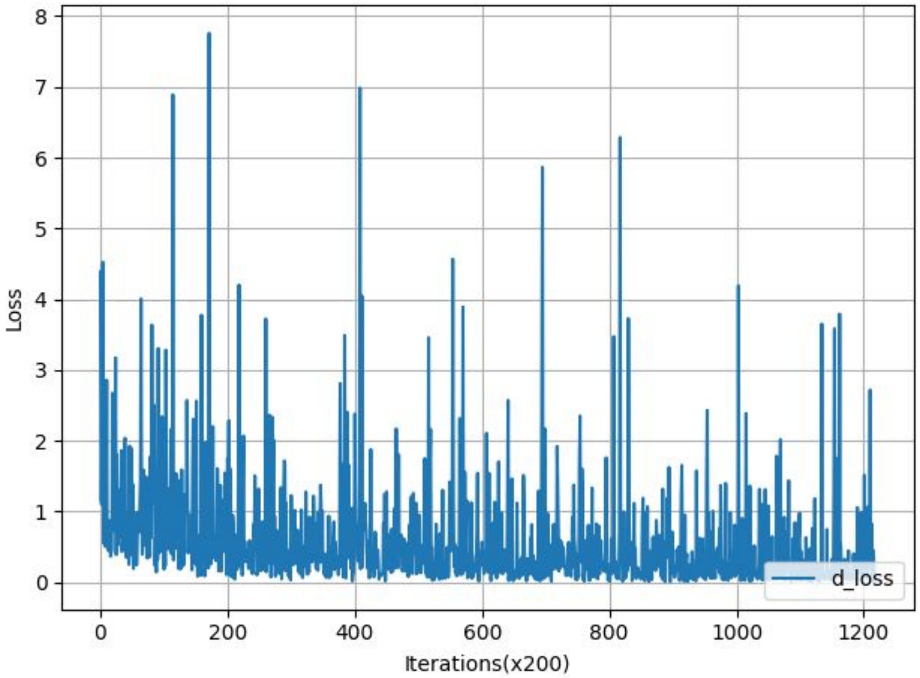
- The word context vectors from attention mechanism is used to generate images for next stage.

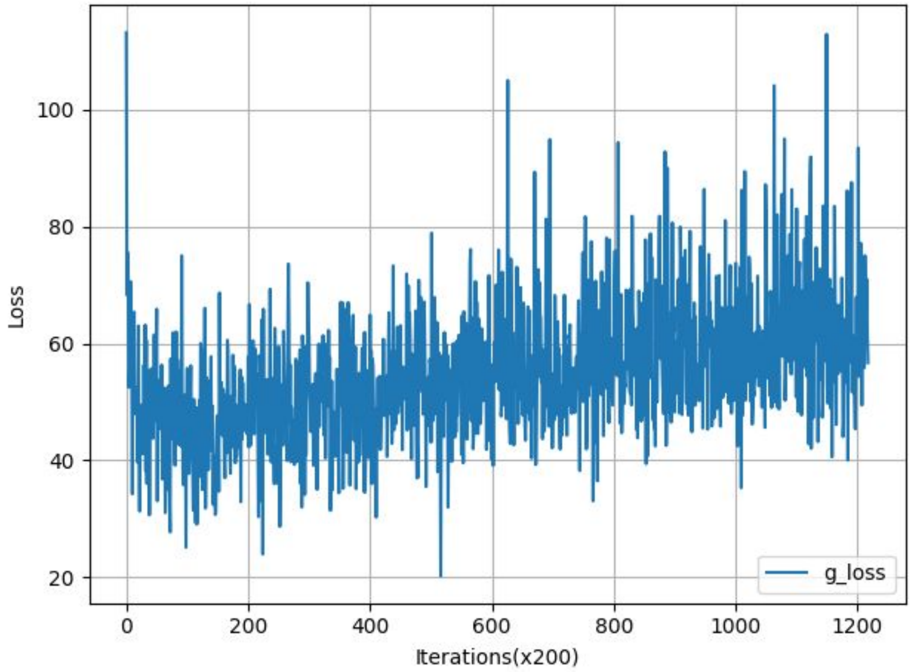
Attention GANs (continued)



Source: [4]

Attention GANs (continued)





Discriminator Loss

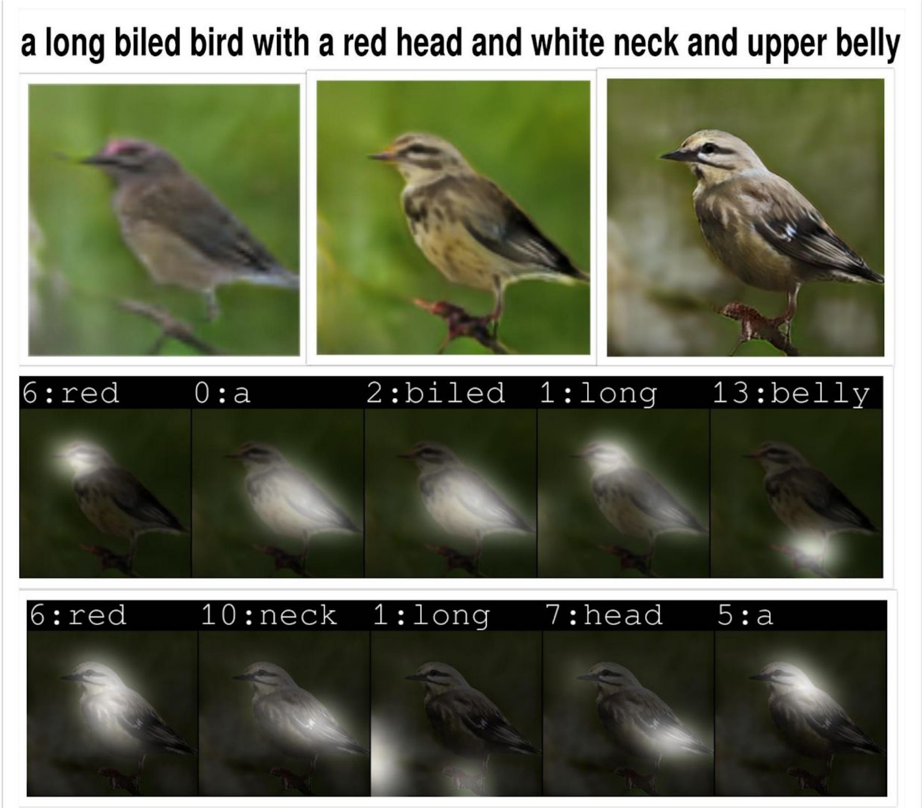
Generator Loss

Attention GANs (continued)

this medium sized perching bird has a grey body with barred light grey chest and a bright red head and neck

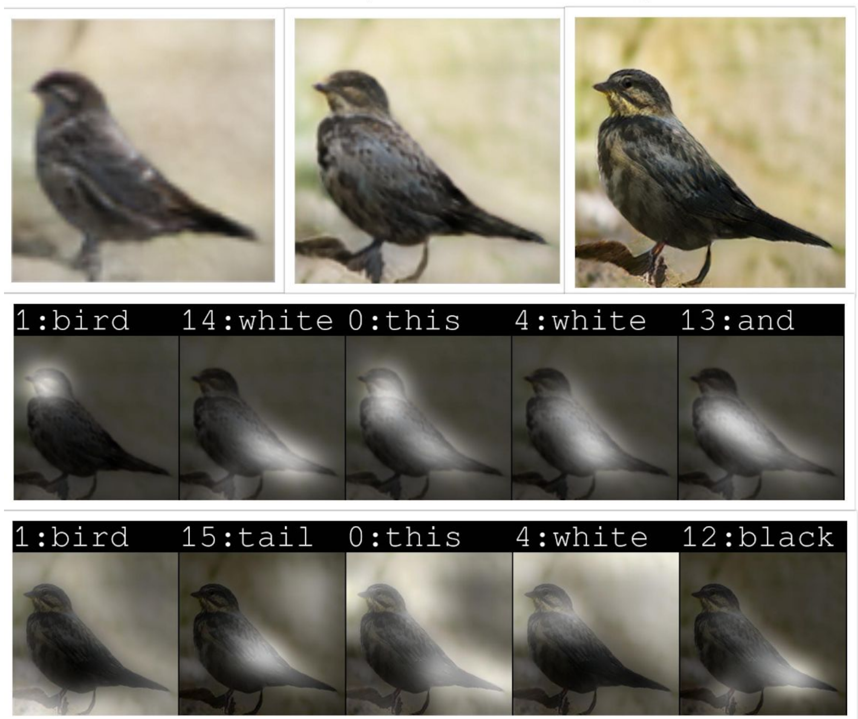
this bird has wings that are blue and has a white belly

Attention GANs (continued)

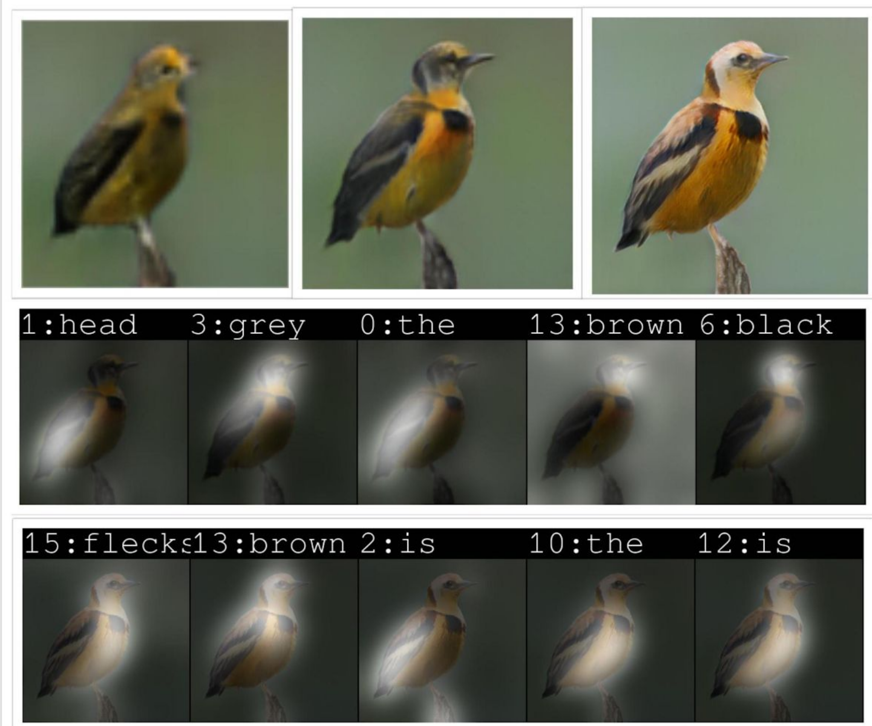


Attention GANs (continued)

this bird has a white breast belly and abdomen and a long black and white



the head is grey with a black crown and throat the body is brown with flecks of red



Attention GANs (continued)

this bird has wings that are 1 and has a 2 belly

1 - red 2 - white 1 - black 2 - black 1 - red 2 - black





1 - black 2 - blue 1 - red 2 - red 1 - blue 2 - red

Attention GANs (continued)

Future Scope of Research

- Divide the image generation process into individual object generation.
- WGAN with attention mechanism.
- Training on MS-COCO dataset to produce generalized images.
- Object oriented learning.

Conclusions

- Successfully implemented a model for synthesizing images using text descriptions.
- Generated images of size 256 × 256 and photorealistic quality.
- Implemented image-word loss, DAMSM, to be used for training the model.
- Explored conditional WGAN.

References

1. S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, “Generative adversarial text to image synthesis,” in Proceedings of the 33rd International Conference on Machine Learning - Volume 48, ICML’16, pp. 1060–1069, JMLR.org, 2016.
2. M. Arjovsky, S. Chintala, and L. Bottou, “Wasserstein generative adversarial networks,” in Proceedings of the 34th International Conference on Machine Learning, vol. 70 of Proceedings of Machine Learning Research, pp. 214–223, PMLR, 06–11 Aug 2017.
3. H. Zhang, T. Xu, H. Li, S. Zhang, X. Huang, X. Wang, and D. N. Metaxas, “Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks,” in ICCV, 2017.
4. T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He, “Attngan: Finegrained text to image generation with attentional generative adversarial networks,” CoRR, vol. abs/1711.10485, 2017.
5. I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, “Generative adversarial nets,” in NIPS, 2014.

Thank You