SJSU SAN JOSÉ STATE UNIVERSITY

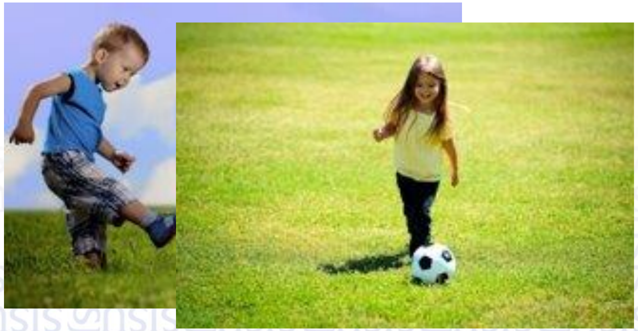
MS Software Engineering
Data Science Cohort 2021

Artificial Intelligence enabled Image Generation from text

Problem

1. To generate image given text description
2. Caption text can consist of multiple objects (Usually <= 5)

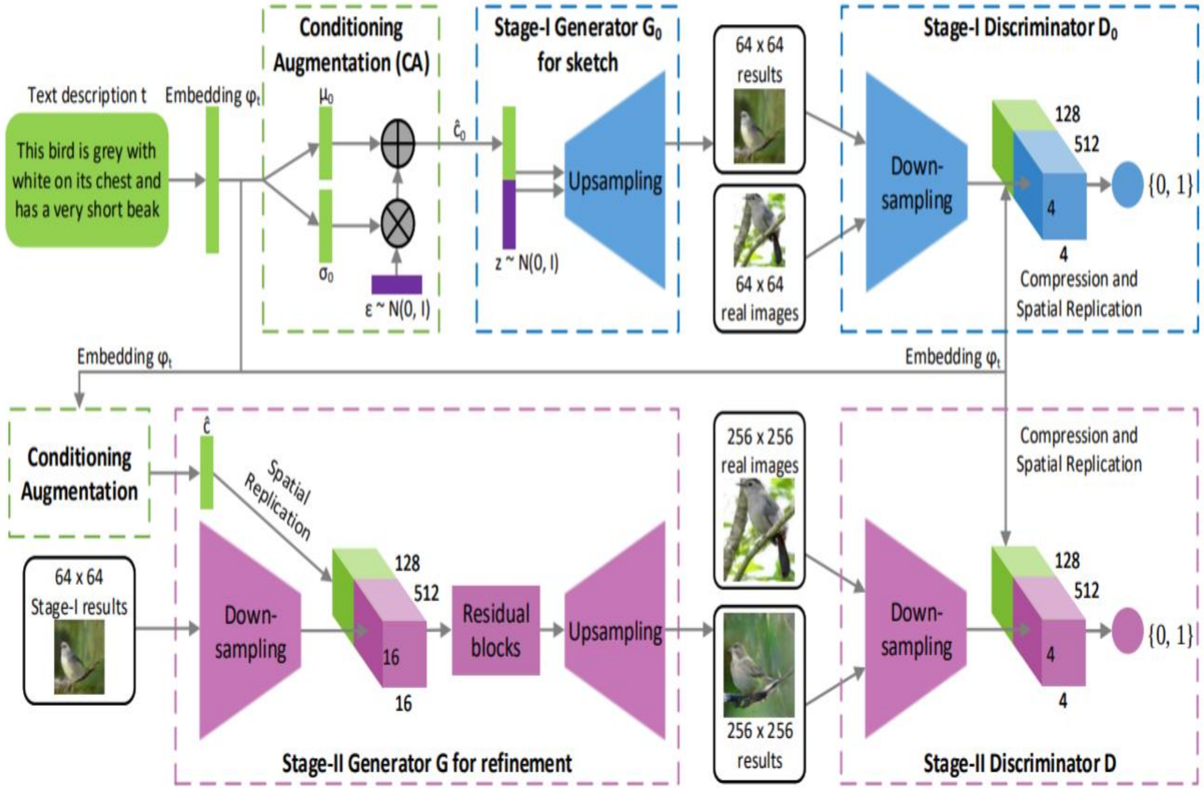
"A child is playing with a ball on a lawn"



Previous Work StackGAN

Two main ideas:

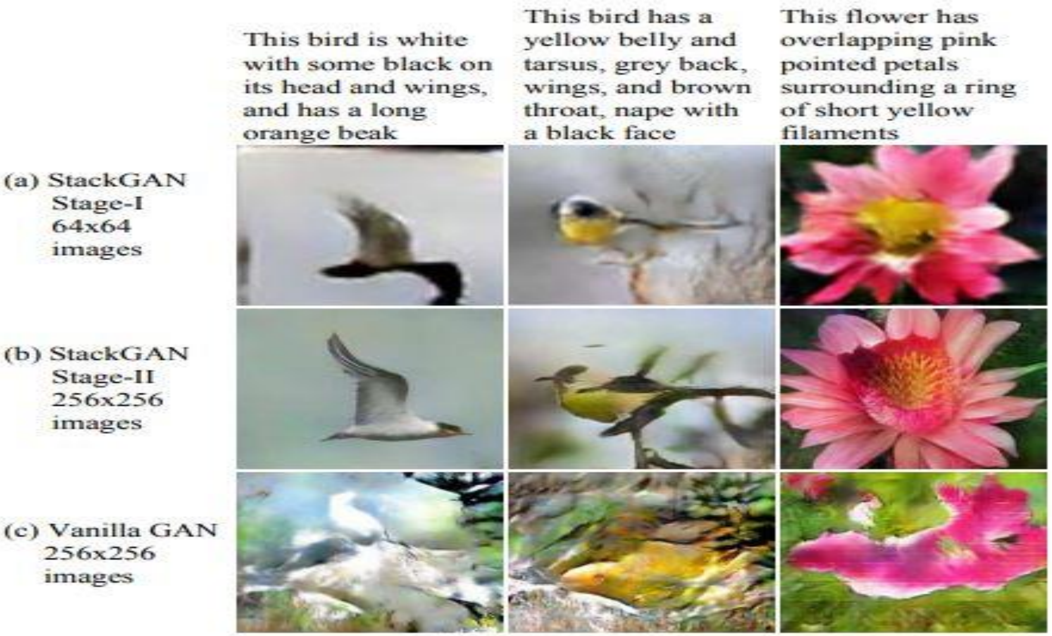
1. Mutli-Scale Architecture
2. Conditioning Augmentation



Reference -https://arxiv.org/abs/1612.03242

Previous Work
StackGAN's Multi-Scale Architecture

1. Stage I produces a 64 x 64 images
2. Stage II takes Stage I's output as conditional input to produce 256 x 256 images



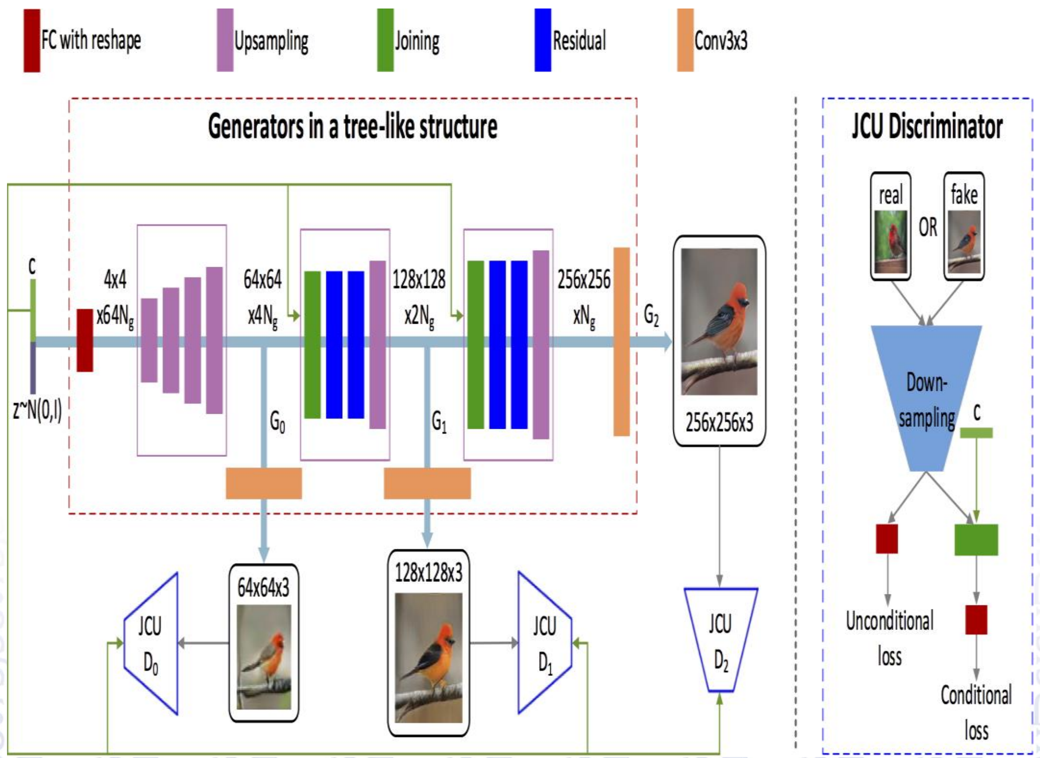
Reference -https://arxiv.org/abs/1612.03242

Previous Work
StackGAN's Stage I and Stage II results

Reference -https://arxiv.org/abs/1612.03242

Previous Work StackGAN++

- Adds multiple Generators-Discriminators pairs.
- Trains all Generators jointly.
- Adds a "color-consistency" regularization for the generator.

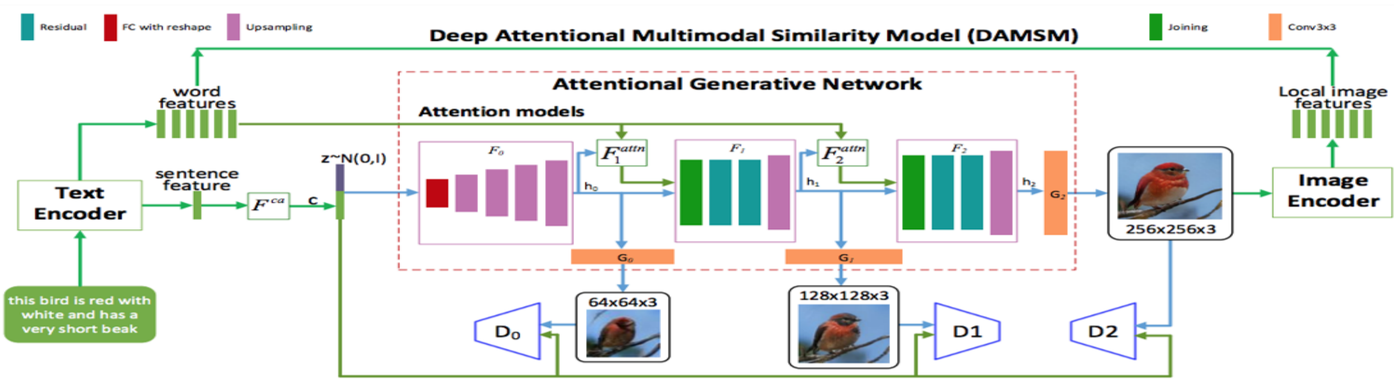


Reference -https://arxiv.org/abs/1710.10916

Previous Work
AttnGAN

The contribution of the AttnGAN can be divided into two parts:

1. Attentional Generator Network
2. Deep AttentionalMultimodal Similarity Model (DAMSM)



Reference - https://arxiv.org/abs/1711.10485

Previous Work
AttnGAN

Reference - https://arxiv.org/abs/1711.10485

Summary of Previous Work

1. GANs can generate realistic looking images adhered to characteristics of textual description.

2. The networks are conditioned on an embedding of textual description.

3. This approach has led to good results on simple, well-structured data sets containing a specific class of objects (e.g., faces, birds, or flowers) at the image center.

What if textual description and images become more complex?

- For instance,
  - More than one object
  - Large variety in background and scenery

Object Pathway-GAN

- Focuses specifically on individual objects.
- Generates these objects at meaningful locations in the image simultaneously generating a background that fits with the overall image description

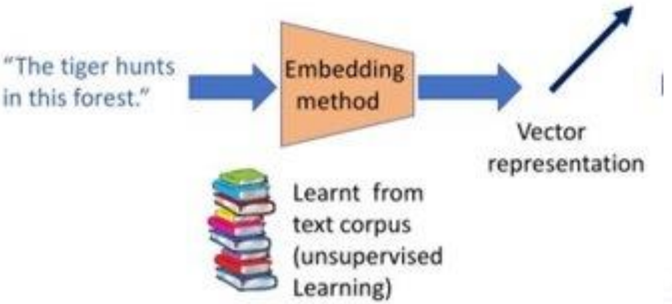
Text Embedding -

Uses a fixed-dimensional vector to represent a small piece of text

- A bi-directional Long Short-Term Memory (LSTM) that extracts semantic vectors from the text description.

- The two hidden states (one in each direction) of the LSTM model are concatenated to represent the semantic meaning of a word.

- The last hidden states of the bi-directional LSTM are concatenated to be the global sentence vector



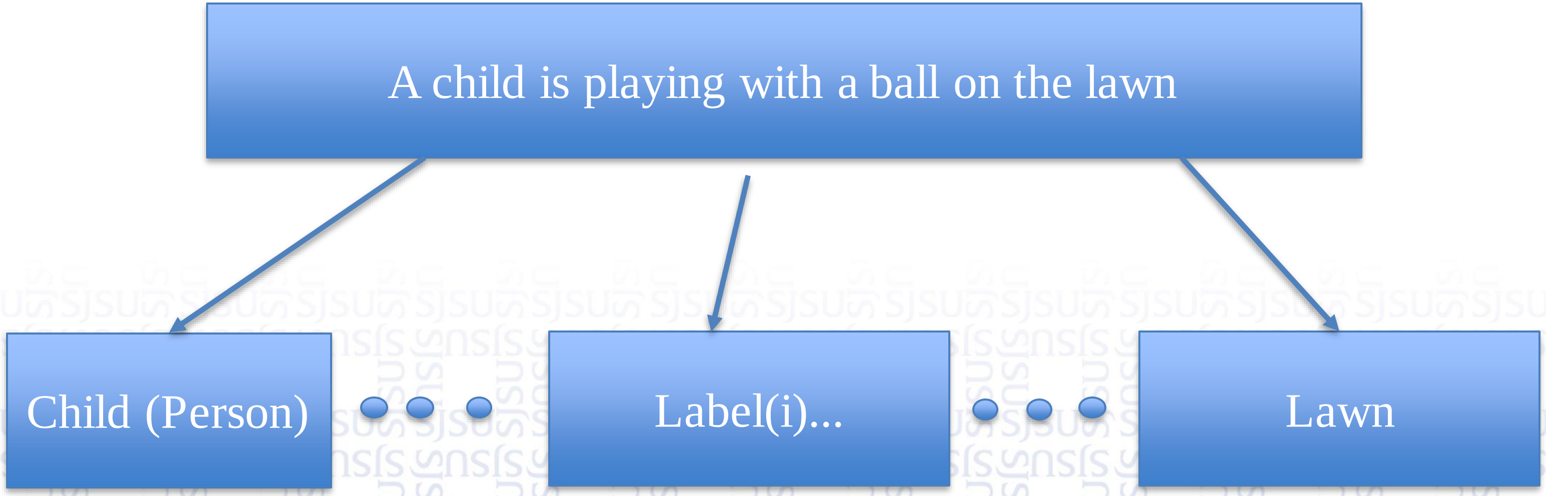
Source - Internet

State-of-the-art
OP-GAN

- This is a conditional-GAN (cGAN) framework, in which both the generator and the discriminator gets additional information, such as labels and captions, as input.
- Hence, this is a supervised image generation model which requires more than image caption to generate image
- Consist of two streams – Object pathway and Global pathway

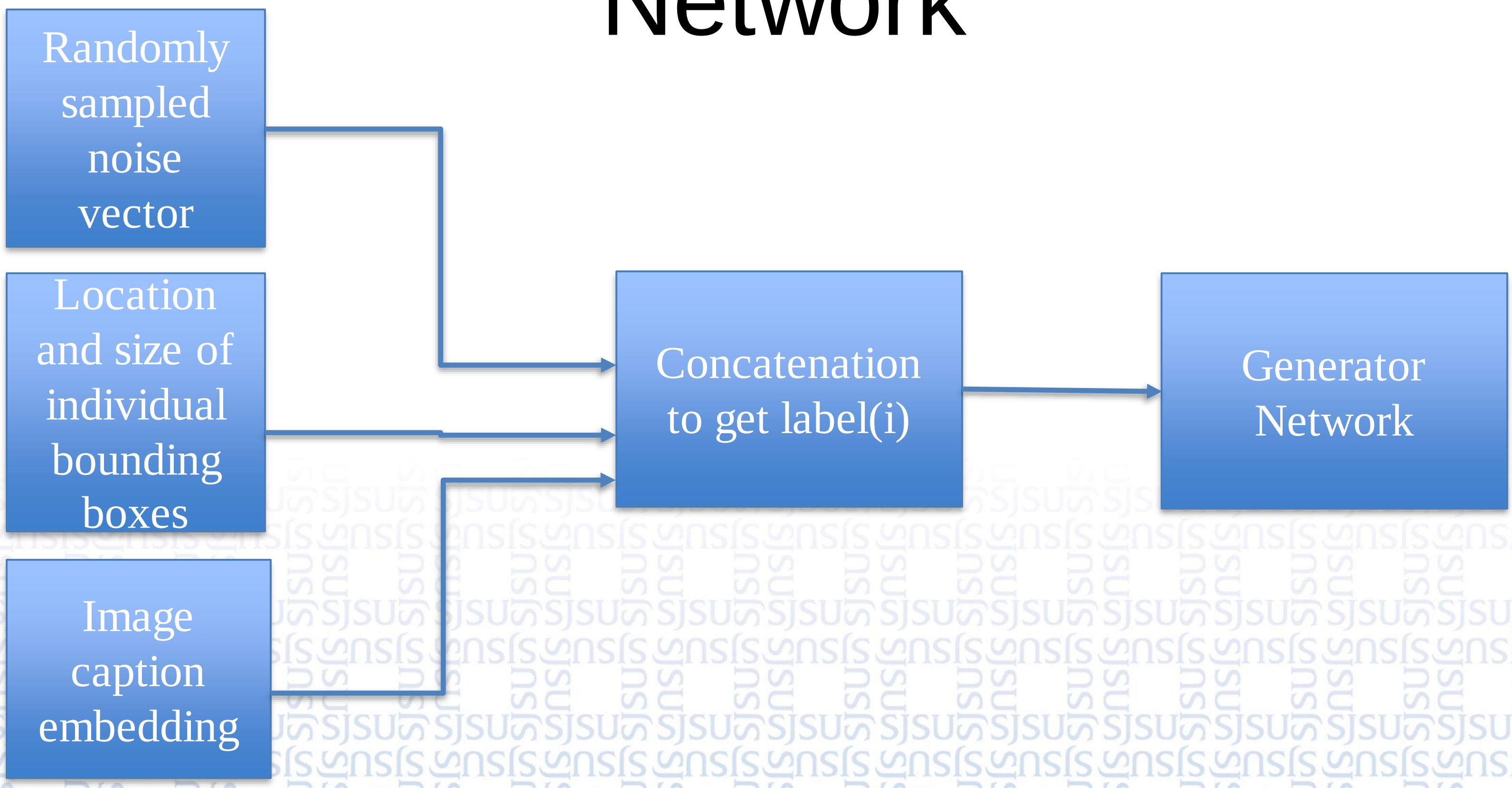
OP-GAN (Continued..)

- Label/Object extraction from caption. The labels are one hot encoded vectors.



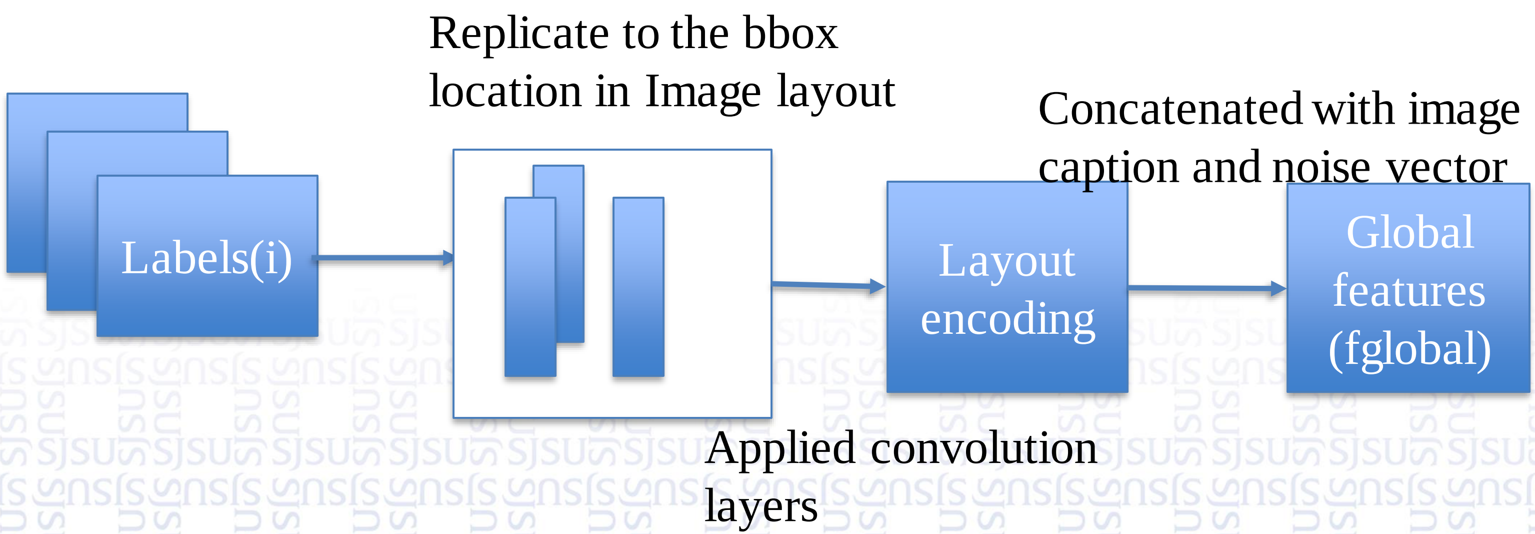
SJSU SAN JOSÉ STATE UNIVERSITY

Preprocessing in Generator Network



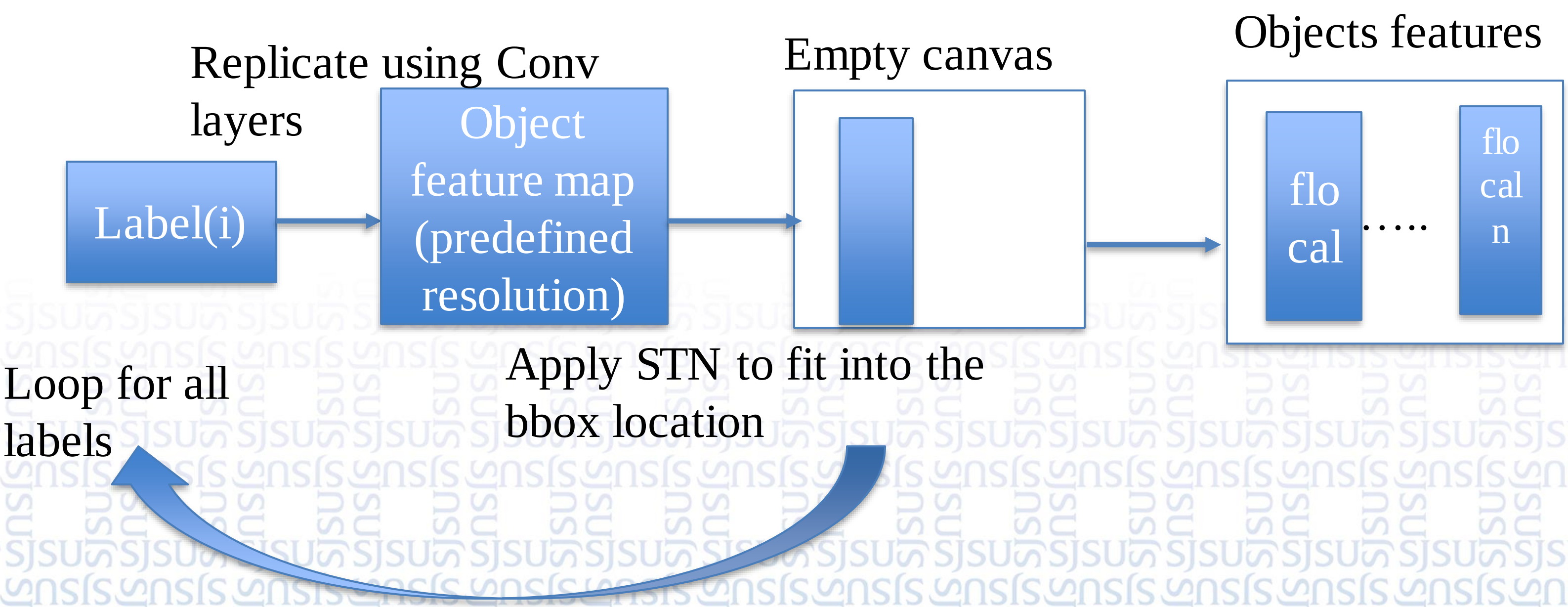
Global Pathway

- Creates general layout for the global scene

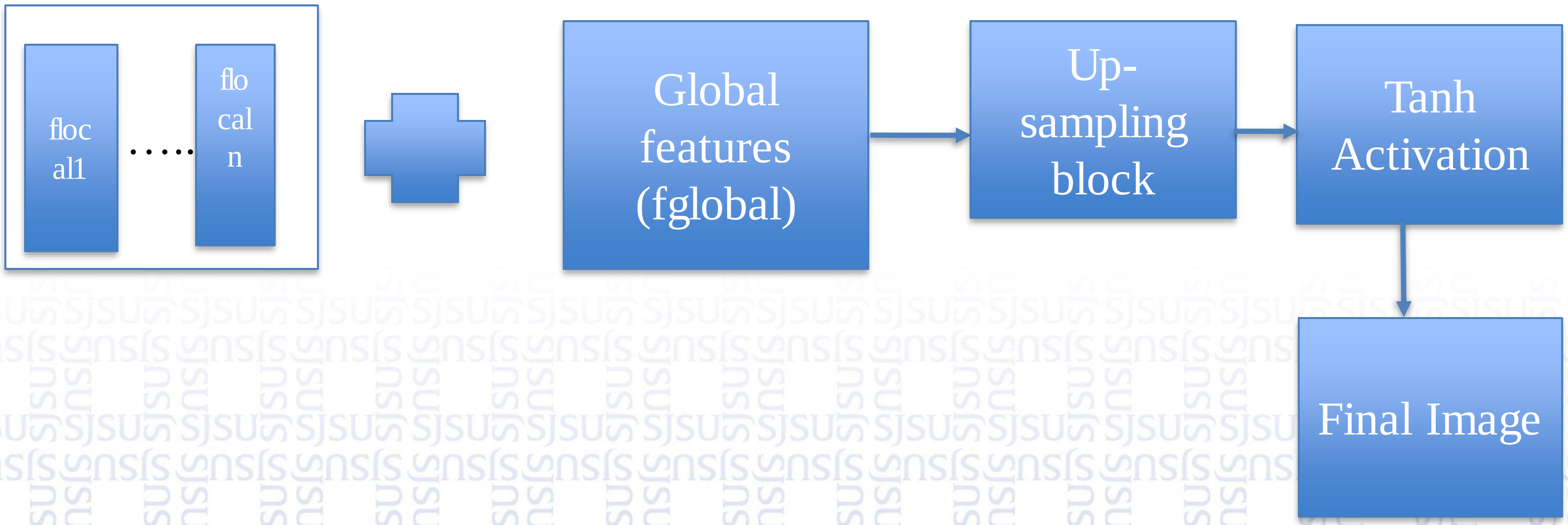


Object Pathway

- Generates features of objects



Final Image

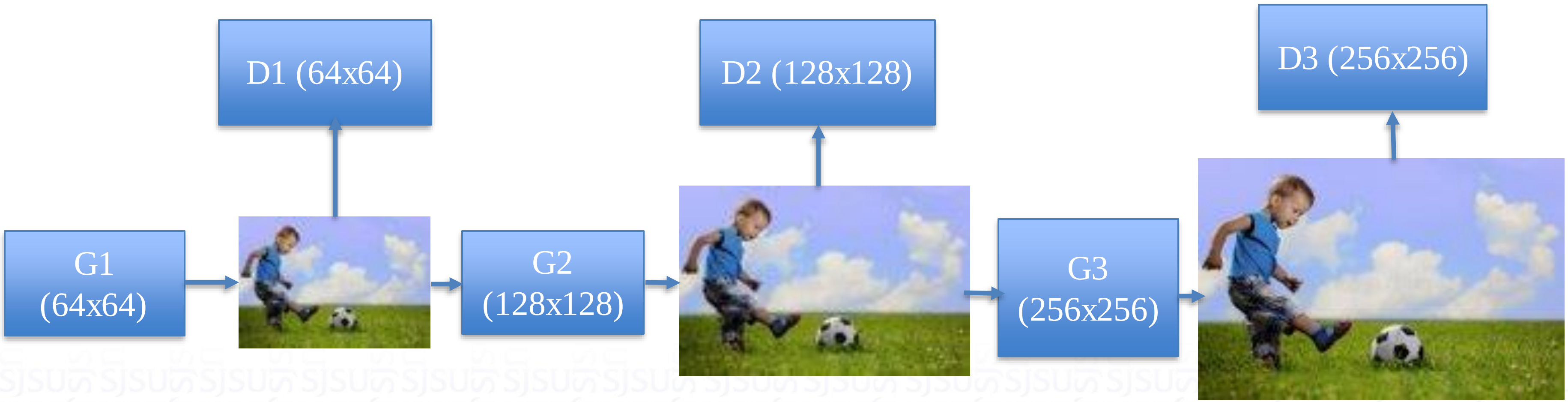


floc
al1
.....
flo
cal
n
Global
features
(fglobal)
Up-
sampling
block
Tanh
Activation
Final Image

Discriminator

- **Input:**
  - Original/Generated image.
  - location and size of bounding box.
  - labels of bounding box.
- **Global Pathway:**
  - Takes in the image as input and generates global features.
- **Object Pathway:**
  - Takes in the image as input and extracts the objects.
  - Concatenates extracted features of objects with bounding box labels.
  - Replicates the object in the bounding box.
  - **Implemented Spectral Normalization.**
- **Output:** Concatenates outputs of Global and Object Pathways and Classifies image as either real or artificial.

High Level architecture of OP-GAN



Challenges in training GAN

- OP-GAN model is big in terms of network size and thereby consuming huge amount of memory for weights.
- Train images are high resolution which further restricts the batch size to much lower value. (Usually GPU memory is 16GB)
- Lesser batch size causes more time to model convergence.
- Discriminator may learn too fast/too slow and cause training instability

Solution - Model Improvement

- Introduced **Spectral normalization** layer in discriminator to replace the 2D batch norm layer.
- This constraints the Lipschitz constant of the convolutional filters.
- Helps to overcome the limitations of lower batch size and improved stability in discriminator.

Challenges to Train GAN model

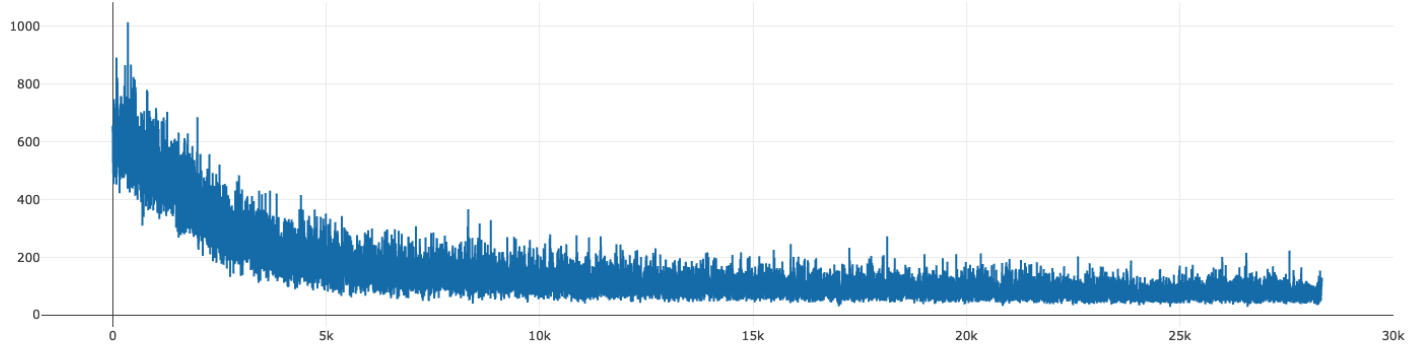
- SJSU HPC. Some limitations are
  - Time limit for normal users in GPU nodes.
  - Condo nodes (has no time limit) are not available to normal users.
  - Too low GPU memory to load a sufficiently bigger batch size
- Microsoft Azure – Provisioned a single node GPU with **NVIDIA Tesla K80.**
- Training time is too high. Take approximately 26 hours to train single epoch.

Solution – Paperspace Gradient

- Provision a single node GPU machine with **Nvidia Tesla v100.**
- Added 100GB SSD to improve the I/O performance.

MLops using MLflow

- Extensively used MLflow's Tracking API to log the metrics and model hyper-parameters



SJSU SAN JOSÉ STATE UNIVERSITY

Evaluation Metrics

Current Evaluation Metrics

- Inception Score and Fréchet Inception Distance.

- IS: This metric focuses on two things, how distinct an object in each image (I.e., whether generated images contain a clearly recognizable object) and variety of objects from overall image generated by GAN (I.e., whether the model can generate images of many different object categories).

Generated image contains single object

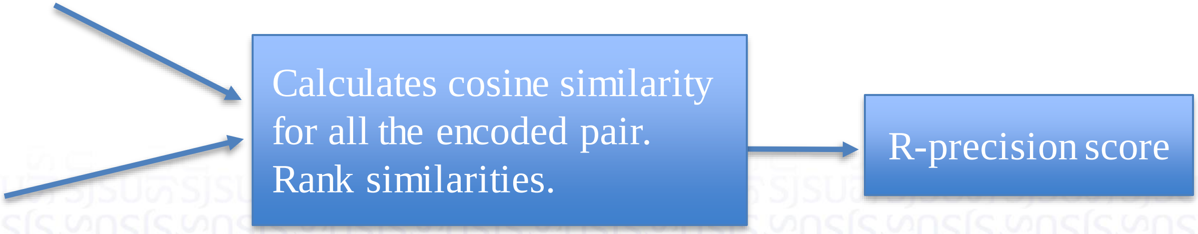
- FID: Calculates the distance between real and generated images.

- Drawbacks:
  - They both does not take image caption into account while evaluating the generated images.
  - They both rely on CNN pretrained on ImageNet dataset, and is completely different from MS-COCO dataset.

Evaluation Metrics

- VS Similarity and R-Precision
  - Visual Semantic Similarity: Measures the distance between generated image and caption.
  - R-precision: Performs ranking of the similarity between real caption and randomly sampled captions for a given generated image.

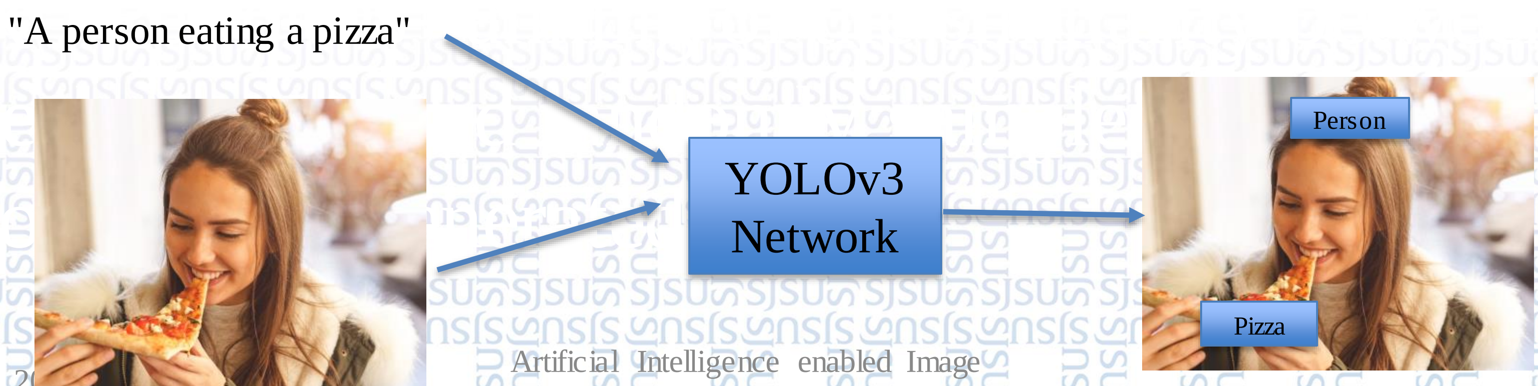
"A person eating pizza"



- Drawback:
  - These two metrics do not evaluate the quality of individual objects.
  - Models might overfit to the evaluation metric during training.
  - Images are evaluated based on background characteristics.

Semantic Object Accuracy

- To address the challenges and issue mentioned above.
  - We use one of the novel evaluation metric based on pretrained object detector network.
- SOA addresses three main problems,
  - Takes image caption into account.
  - Pre-trained object detector on same domain.
  - Focuses on foreground object.



Semantic Object Accuracy

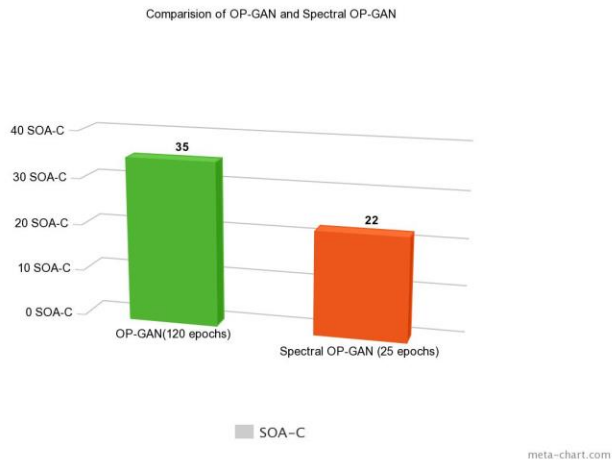
- **How this SOA works?**

$$SOA = \frac{1}{|C|} \sum_{c \in C} \frac{1}{|Ic|} \sum_{i_c \in I_c} \text{YOLOv3(ic)}$$

- For each of the 80 objects (person, dog, chair, bulb etc) that is given as label in MS-COCO dataset, all the captions are retrieved, and three images are generated for each caption.
- Ex. Captions (person, woman, man) -> Label (person) and caption (monitor, tv, screen) -> Label(monitor)
- Generated images are given as input to pretrained object detector to identify the objects in generated images.
- For each label calculate how often the specific object was detected in the generated images for this label. (We get 80 individual results as there are 80 labels in MS-COCO dataset)
- SOA: The SOA score gives us the true positives of a generated image,
  - Accuracy Score : Calculate the average accuracy for overall classes in a dataset.

Semantic Object Accuracy

- Result: We were able to achieve Semantic Object Accuracy score of 22 in 25 epochs compared to previous model which achieves 35 in 120 epochs.



Rank	Model	FID	SOA-C	IS
1	OP-GAN	24.70	35.58	27.88
2	DM - GAN		33.44	30.49
3	AttnGAN		25.88	25.89
4	StackGAN + OP	55.30		12.12

Table 0-1 Model comparison sorted by SOA-C

Demo

For more information:

Sivaranjani Kumar (sivaranjani.kumar@sjsu.edu)
Akshaya Nagarajan(akshaya.nagarajan@sjsu.edu)
Pooja Patil(pooja.patil@sjsu.edu)
Vignesh Kumar Thangarajan(vigneshkumar.thangarajan@sjsu.edu)

Thank You