ORACLE

機械学習入門：次世代AI「GAN」による仮想現実の生成

StackGANの概要とデモンストレーション

2021年 5月31日

日本オラクル
テクノロジー事業戦略統括 戦略ビジネス本部
デジタル・トランスフォーメーション推進室
データアナリスト　横山 慎一郎

“Flying birds”

input

output

テキストから画像を生成するGAN

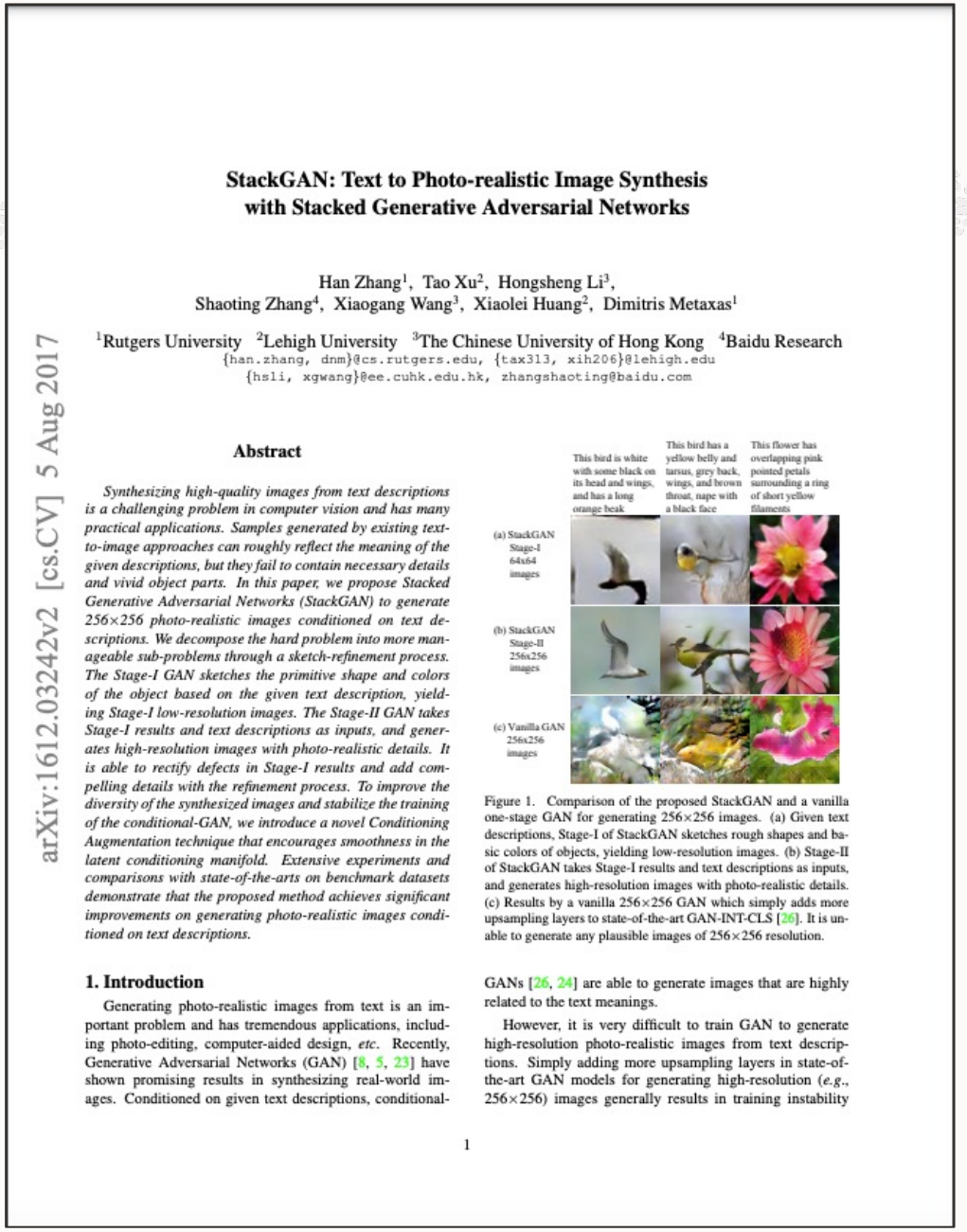
	生成できる解像度	References
GAN-INT-CLS	64	Generative Adversarial Text to Image Synthesis, ICML 2016
GAWWN	128	Learning What and Where to Draw, NIPS 2016
StackGAN	256	Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks, ICCV 2017
TAC-GAN	128	Text Conditioned Auxiliary Classifier Generative Adversarial Network, arXiv 2017
AttnGAN	256	Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks, CVPR 2018
FusedGAN	64	Semi-supervised FusedGAN for Conditional Image Generation, arXiv 2018
HDGAN	512	Photographic Text-to-Image Synthesis with a Hierarchically-nested Adversarial Network, arXiv 2018

参考: http://akmtn.hatenablog.com/entry/2018/03/25/182759

StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks

- 著者
  - Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, Dimitris Metaxas
  - Rutgers University, Lehigh University, The Chinese University of Hong Kong, Baidu Research

- 内容
  - 先行研究では、テキストから画像を生成することは、大まかな特徴は表現できても、細部まで表現できなかった。
  - そこでGANを利用して高精度の画像生成モデルを提唱。
  - StackGANでは2段階のアプローチにより構成される。
    - Stage Ⅰ: テキストから低解像度の画像を生成。
    - Stage Ⅱ: Ⅰの生成画像とテキストから高解像度の画像を生成。
  - 結果、先行研究のテキストから画像を生成する技術より、高精度の画像を生成できた。



StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks

Han Zhang[1], Tao Xu[2], Hongsheng Li[3],
Shaoting Zhang[4], Xiaogang Wang[3], Xiaolei Huang[2], Dimitris Metaxas[1]

[1]Rutgers University [2]Lehigh University [3]The Chinese University of Hong Kong [4]Baidu Research
{han.zhang, dnm}@cs.rutgers.edu, {tax313, xih206}@lehigh.edu
{hsli, xgwang}@ee.cuhk.edu.hk, zhangshaoting@baidu.com

Abstract

Synthesizing high-quality images from text descriptions is a challenging problem in computer vision and has many practical applications. Samples generated by existing text-to-image approaches can roughly reflect the meaning of the given descriptions, but they fail to contain necessary details and vivid object parts. In this paper, we propose Stacked Generative Adversarial Networks (StackGAN) to generate 256×256 photo-realistic images conditioned on text descriptions. We decompose the hard problem into more manageable sub-problems through a sketch-refinement process. The Stage-I GAN sketches the primitive shape and colors of the object based on the given text description, yielding Stage-I low-resolution images. The Stage-II GAN takes Stage-I results and text descriptions as inputs, and generates high-resolution images with photo-realistic details. It is able to rectify defects in Stage-I results and add compelling details with the refinement process. To improve the diversity of the synthesized images and stabilize the training of the conditional-GAN, we introduce a novel Conditioning Augmentation technique that encourages smoothness in the latent conditioning manifold. Extensive experiments and comparisons with state-of-the-arts on benchmark datasets demonstrate that the proposed method achieves significant improvements on generating photo-realistic images conditioned on text descriptions.

Figure 1. Comparison of the proposed StackGAN and a vanilla one-stage GAN for generating 256×256 images. (a) Given text descriptions, Stage-I of StackGAN sketches rough shapes and basic colors of objects, yielding low-resolution images. (b) Stage-II of StackGAN takes Stage-I results and text descriptions as inputs, and generates high-resolution images with photo-realistic details. (c) Results by a vanilla 256×256 GAN which simply adds more upsampling layers to state-of-the-art GAN-INT-CLS [26]. It is unable to generate any plausible images of 256×256 resolution.

1. Introduction

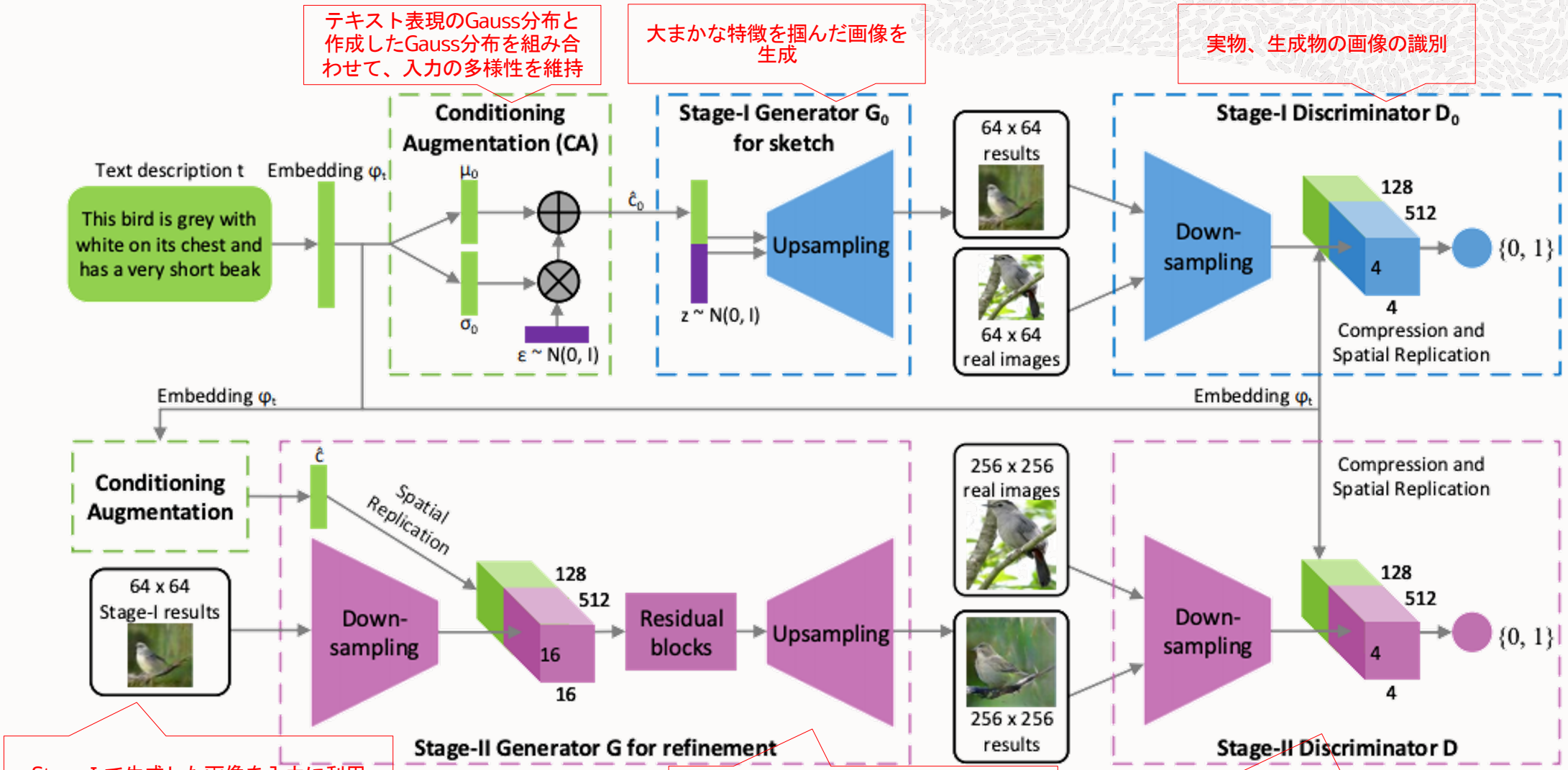
Generating photo-realistic images from text is an important problem and has tremendous applications, including photo-editing, computer-aided design, *etc*. Recently, Generative Adversarial Networks (GAN) [8, 5, 23] have shown promising results in synthesizing real-world images. Conditioned on given text descriptions, conditional-GANs [26, 24] are able to generate images that are highly related to the text meanings.

However, it is very difficult to train GAN to generate high-resolution photo-realistic images from text descriptions. Simply adding more upsampling layers in state-of-the-art GAN models for generating high-resolution (*e.g.*, 256×256) images generally results in training instability

1

https://arxiv.org/pdf/1612.03242.pdf

アーキテクチャ



参考: StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks

実装方法

- パラメータ
  - Up-sampling block
    - 3 x 3 , ストライド1の畳み込みを適用
    - 最終層以外は畳み込み層の後にバッチ正則化とReLU関数を適用
    - 128 x 128 StackGANモデルでは、2つの残差ブロックを利用
    - 256 x 256 StackGANモデルでは、4つの残差ブロックを利用
  - Down-sampling block
    - 4 x 4, ストライド2の畳み込みを適用
    - 最初の層以外はバッチ正則化とLeakyReLUを適用
  - Vector dimension etc.
    - Ng = 128 (Conditioningベクトルの次元)
    - Nz = 100 (ノイズベクトルの次元)
    - Mg = 16 (Ⅱ. Generator: 圧縮するサイズ)
    - Md = 4 (Ⅰ. Discriminator: 圧縮するサイズ)
    - Nd = 128 (Ⅰ. テキストEmbeddingを圧縮する次元)
    - W0 = H0 = 64 (Stage Ⅰ で生成される画像サイズ. W0 x H0)
    - W = H = 256 (Stage Ⅱ で生成される画像サイズ. W x H)

- 学習
  - Stage Ⅰ のG, Dの学習
    - 600 epoch
    - Stage Ⅱ は固定
  - Stage Ⅱ のG, Dの学習
    - 600 epoch
    - Stage Ⅰ は固定
  - Optimizer（最適化アルゴリズム）
    - バッチサイズ64のADAMを利用
    - 学習率0.0002で100 epoch毎に半減させる

参考: StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks

GAN-INT-CLS, GAWWNとの比較結果

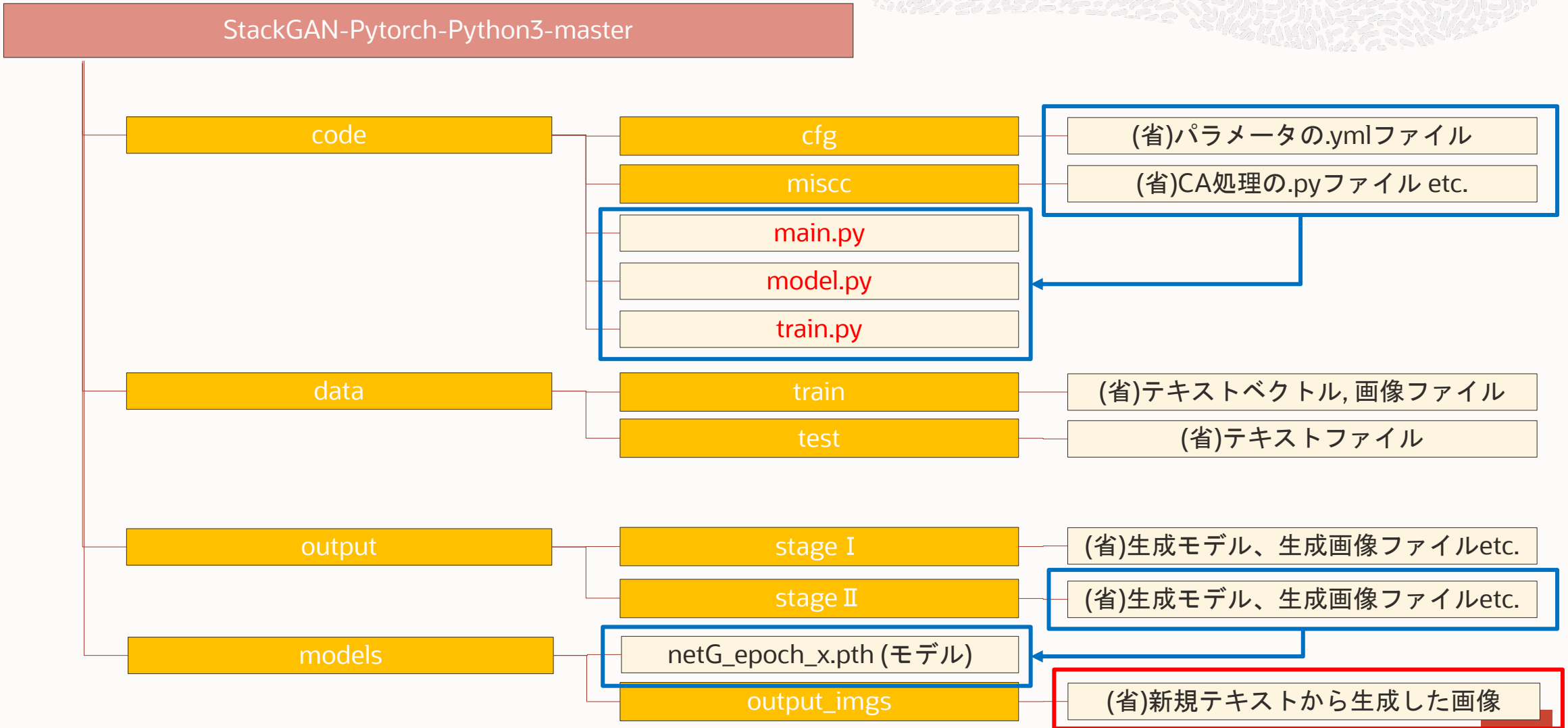
Metric	Dataset	GAN-INT-CLS	GAWWN	Our StackGAN
Inception score	CUB	2.88 ± .04	3.62 ± .07	**3.70 ± .04**
	Oxford	2.66 ± .03	/	**3.20 ± .01**
	COCO	7.88 ± .07	/	**8.45 ± .03**
Human rank	CUB	2.81 ± .03	1.99 ± .04	**1.37 ± .02**
	Oxford	1.87 ± .03	/	**1.13 ± .03**
	COCO	1.89 ± .04	/	**1.11 ± .03**

参考: StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks

デモ

- https://github.com/anthonyftwang/StackGAN-Pytorch-Python3　を参考に、構造やコードの概要について説明していく。
- 学習済みモデルを利用して、テキスト表現から画像生成してみる。
- StackGANの応用を考えていく（ざっくり）

構造




ORACLE