※特に断りのない限り，図表は論文からの引用

Zero-Shot Text-to-Image Generation (DALL・E ver. 1)

Aditya Ramesh [1] **Mikhail Pavlov** [1] **Gabriel Goh** [1] **Scott Gray** [1]
Chelsea Voss [1] **Alec Radford** [1] **Mark Chen** [1] **Ilya Sutskever** [1]

OpenAI

紹介者：豊田工業大学 知能数理研究室　　辻村 有輝
Toyota Technological Institute

プロンプトからの高品質な画像の生成

(a) a tapir made of accordion. a tapir with the texture of an accordion.

(b) an illustration of a baby hedgehog in a christmas sweater walking a dog

プロンプト＋画像の上半分を入力にして画像の下半分を生成（公式サイトから引用）

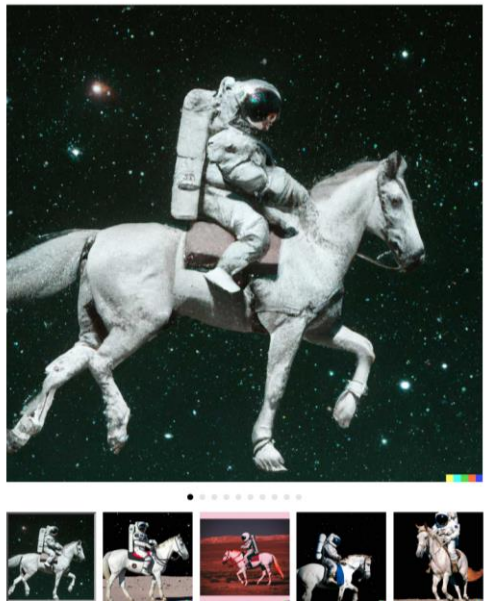
概要

- AutoregressiveなTransformerベースのモデルで**高品質なText-to-Imageモデル**を提案
  - **アプローチのシンプルさがウリ**とのこと
- Transformerの生成対象はDiscrete VAE (dVAE) によって圧縮した画像表現
- 学習はdVAEの学習➡Transformerの学習の2 stage

- 計算効率の最適化も頑張っている

Text-to-Imageモデル

- 自然言語から成るプロンプトを入力に，それに即した画像を生成するモデル
- 最近かなり流行っていますね



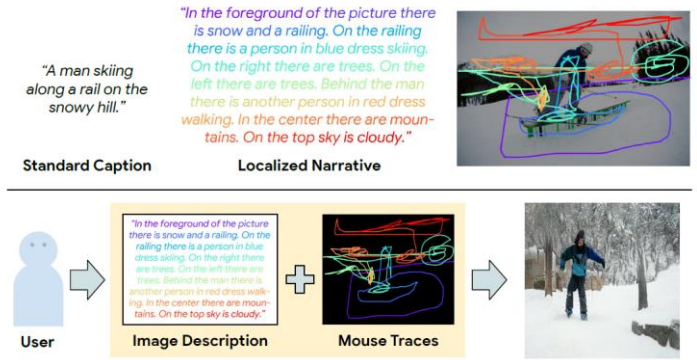
DALL-E 2（公式サイトから引用）
DiffusionベースになったDALL-E
利用には登録とその承認待ちが必要

Midjourney（生成例）
Discordで生成を試せる

Stable Diffusion（GitHubリポジトリから引用）
DiffusionモデルによるText-to-Image
学習済みモデルがオープンに配布されている

貢献

- **シンプルなアプローチで高品質な画像生成**を行った (と言いながら論文からはかなりの苦労が感じられるが・・・)
  - Text-to-Imageモデルは年々複雑になっていく，にもかかわらず生成物にはまだまだartifactができてしまう
  - 先行研究の例 †1 （図は論文から引用）



説明文と対応する画像位置情報を生成先マスクとして利用

This is a collage of two photos. Here we can see a woman playing in the ground and she is holding a racket with her hands. And there is a mesh.

In this image a lady wearing green cap, blue jacket is skiing. She is holding two sticks. The ground is full of snow. In the background there are trees ...

生成例

- 大規模データ（**250M個の画像・テキストペア**）で学習を行った
  - これまでよく使われてきたMS-COCO (ラベル付き画像200K) などと比較して．
  - ごく最近のモデルで使われるデータはさらに桁が増えて5Bといった数字になっている…

†1: Koh et al., 2021, Text-to-Image Generation Grounded by Fine-Grained User Attention.

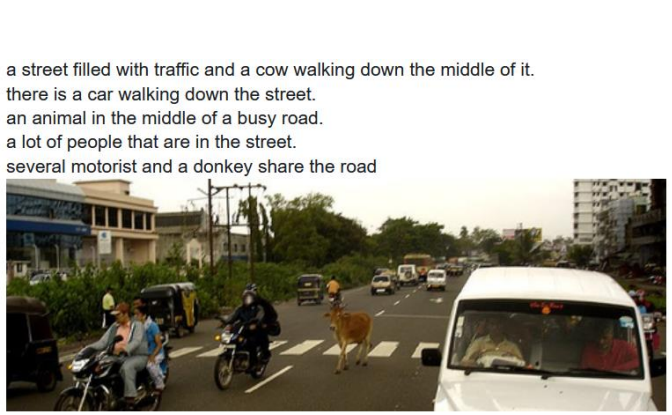
アイディア

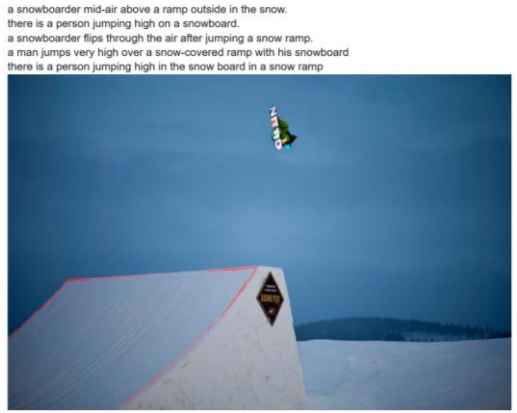
- 最近は言語・画像・音声でauto-regressiveなTransformerモデルが強い！
  ➡Image-to-TextタスクにもTransformerベースのモデルを使う

- Transformerモデルの学習には適切なスケールのデータセットが要る！
  ➡よく使われるデータセットよりもずっと大きなサイズのデータセットを作る

手法｜データセット作成

- これまでの良く使われていたデータセット
  - MS-COCO：数百K個の画像とそのキャプションのペア

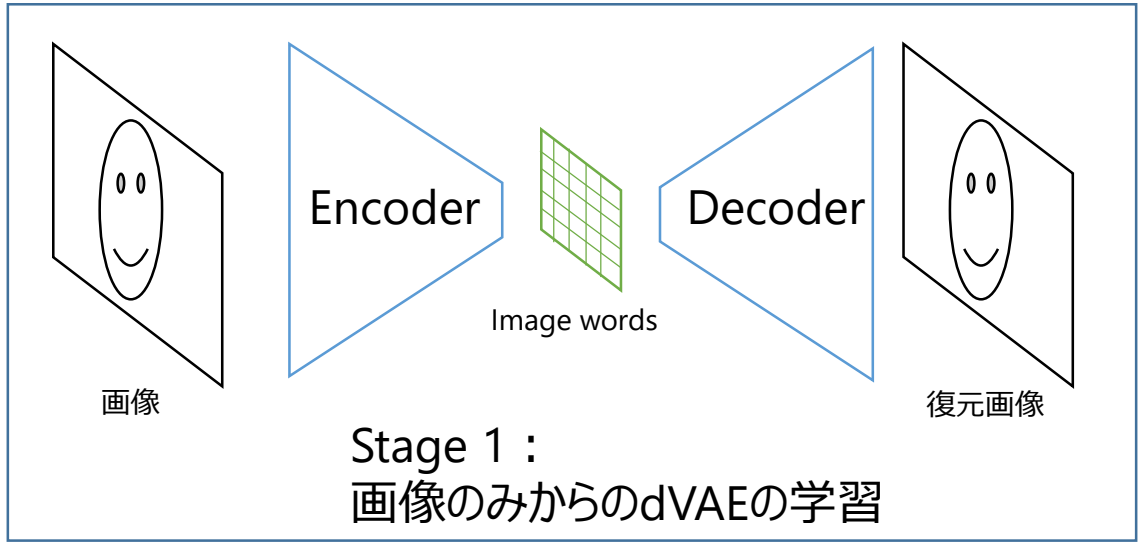


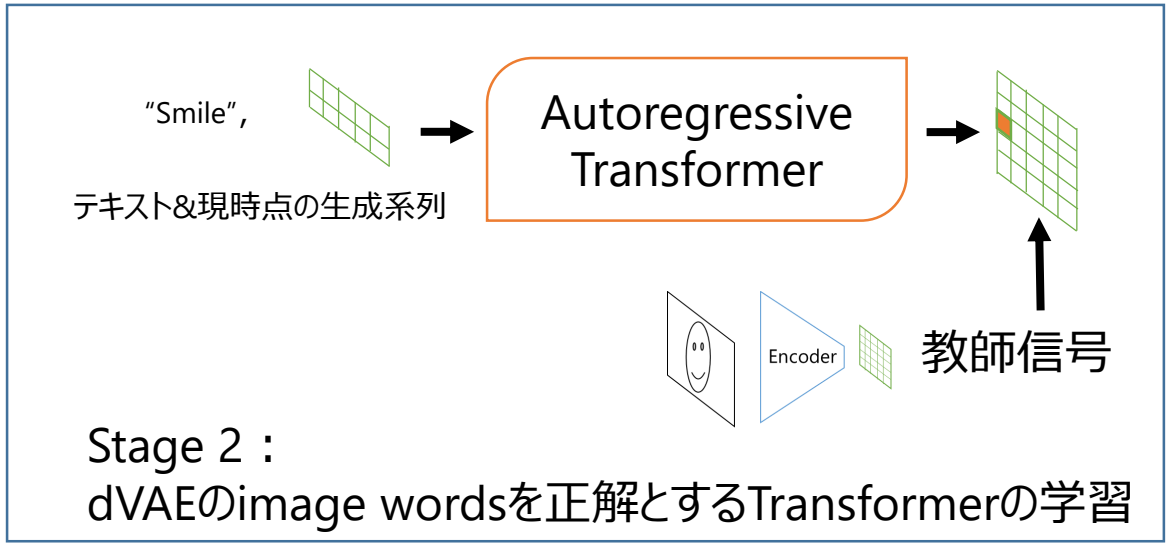


- ネット上から収集した250Mの画像・テキストペア
  - ここで作成したデータはMS-COCOの画像データを一部含むがキャプションは別のもの(評価時にMS-COCOのキャプションデータを使うため，それがリークしていないという意味)

手法｜モデルの全体構造

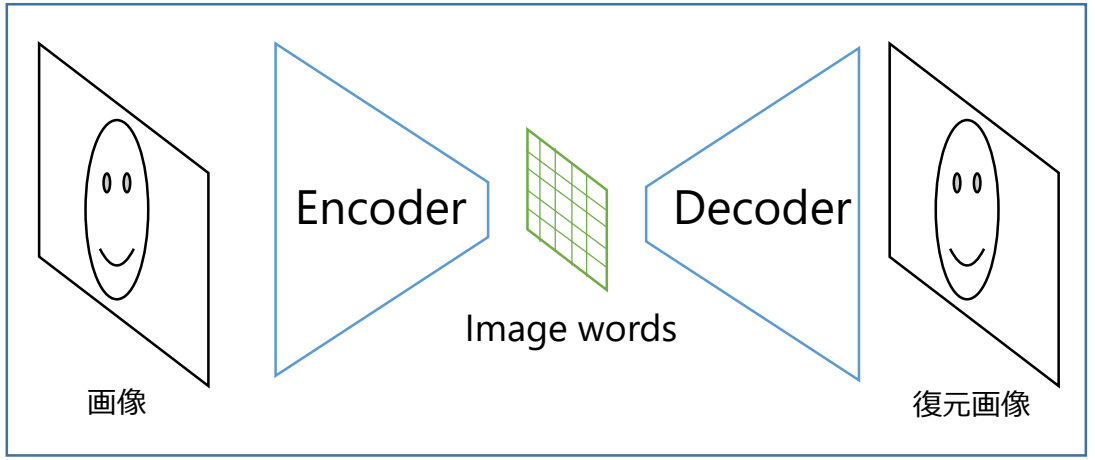
- モデルは大きく分けて二要素で，dVAE ➡ Transformerの順に別々に学習
  - Discrete VAE (dVAE)
    - 大きな画像 (256x256) から小さな圧縮表現 (32x32) を得るのが目的
    - 圧縮表現の各グリッドをimage wordと呼んでいる
  - Auto-regressive Transformer
    - テキストからのdVAEの表現ベクトルの生成



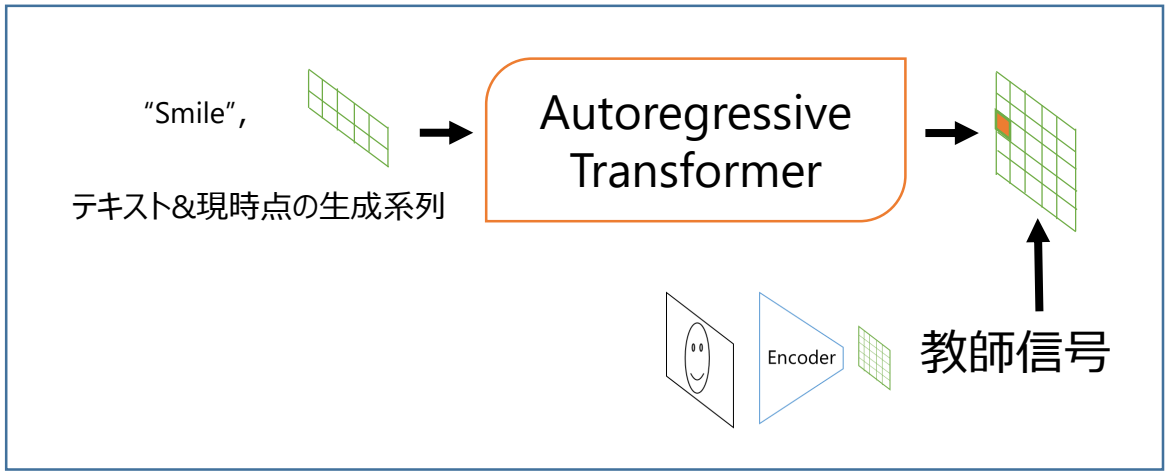


手法｜dVAEモデル

- 画像↔圧縮表現 (image words) のエンコード・デコードを学習
  - 目的：大きなサイズの画像を圧縮
    - オリジナル画像を生成するのはメモリ・学習難度的にかなり大変なため
  - 入力画像サイズ：256x256x3 (RGB)
  - 圧縮表現：32x32x1(離散値, 8192vocab)
    - Gumbel trickにより離散化
    - Image wordsと呼ぶ
    - 元から192分の1に圧縮
    - Transformerでの利用時は32x32➡flattenして1024の1Dな系列とみなす
  - モデル構造はResNetベース



手法｜Transformerモデル

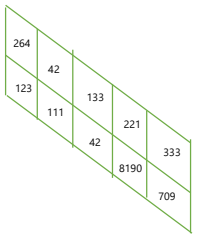


- Auto-regressive Transformerによるテキストからの dVAEの表現ベクトルの予測
  - Decoder-only Sparse Transformer †2 (12B parameters)
  - 入力はテキスト (BPE-encoded) とそれまでの生成image words
    - Text words : vocab_size=16,384, max_length=256 tokens
    - Image words: vocab_size= 8,192, length=1024 tokens
      - Transformer学習時はgumbel noiseを加えずVAEからargmaxでサンプリング

†2: Child et al, 2019, Generating long sequences with sparse transformers.

Transformerモデルの生成ステップ例

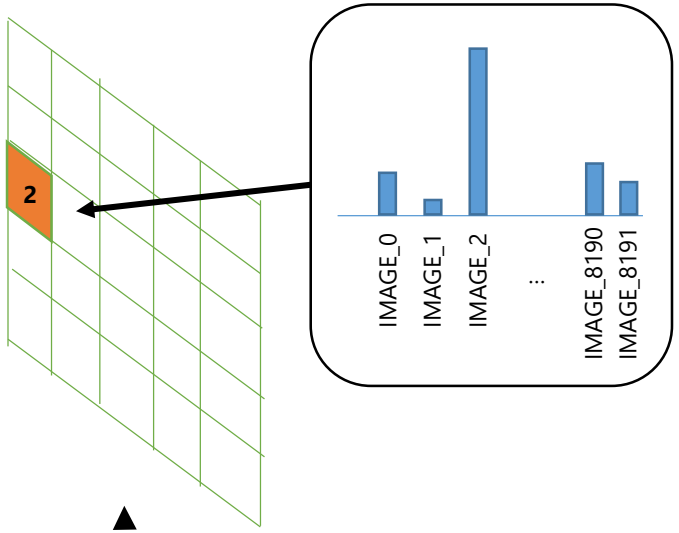
[BOS]
A
very
nice
picture
.
[EOS]
IMAGE_264
IMAGE_42
…
IMAGE_8190
IMAGE_709

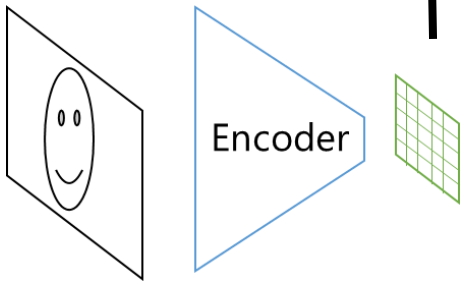


現時点の
生成系列

Autoregressive
Transformer

次の位置のwordを予測する



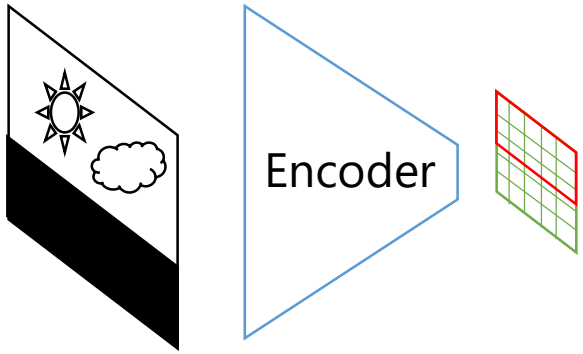


プロンプトに対応する画像をエンコードした
表現を教師信号にする

- テキスト部分の生成も学習対象

上部分だけ描かれた画像からエンコードしてきた表現を使えばImage-to-Imageができる

上半分だけ描かれた画像を
エンコードし上半分16x32
（最初の512系列分）の
グリッドのimage wordsを
Transformerの入力にする



[BOS]
same
view
on
the
top
[EOS]
IMAGE_xxx
IMAGE_yyy
…
IMAGE_zzz

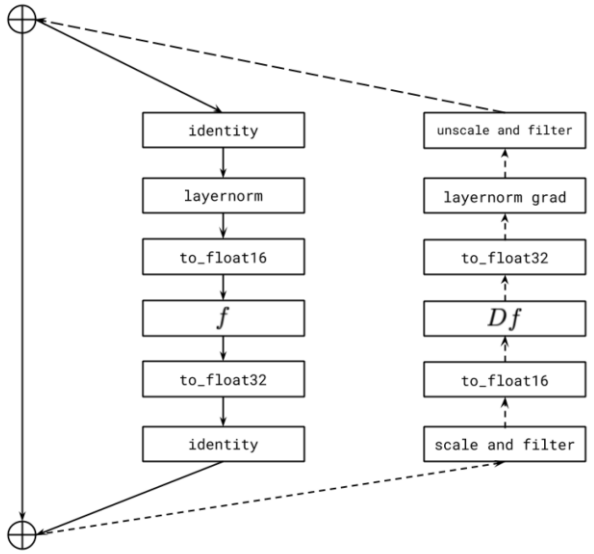
Autoregressive
Transformer

以降は通常通り
生成

画像でコントロールされたimage wordsからの生成が行われる

手法｜MixedPrecision

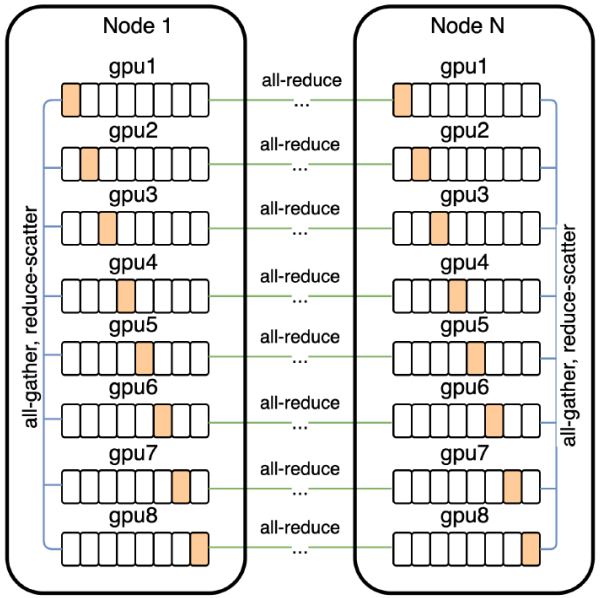
- MixedPrecisionの利用
  - メモリ効率・計算速度のために半精度浮動小数点を利用したいが，全ての計算を半精度にすると勾配情報がアンダーフローする
  - 特に深い層のresidual blockで勾配ノルムがアンダーフローしてゼロになりやすい
    ➡residual blockの勾配計算時にアンダーフローしないようにスケールする
    - また，勾配と順伝播時の活性化後の値は単精度で管理



手法｜Distributed Optimization

- MixedPrecisionを利用してもまだ実験環境（V100, 16GB）に載りきらない
  ➡分散学習を行う
  - メモリが24GBが必要だった

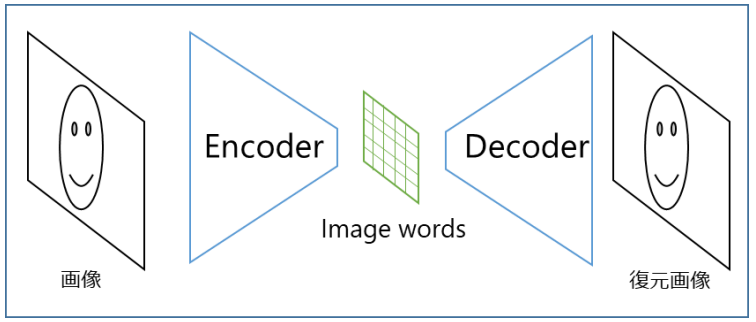
- 8つのGPUに分散して配置
  - 各層の計算中に次の層のパラメータをprefetchし高速化
    - 逆伝播時も同様にprefetch処理を行う
  - 勾配情報の集約がボトルネックになるので圧縮し通信



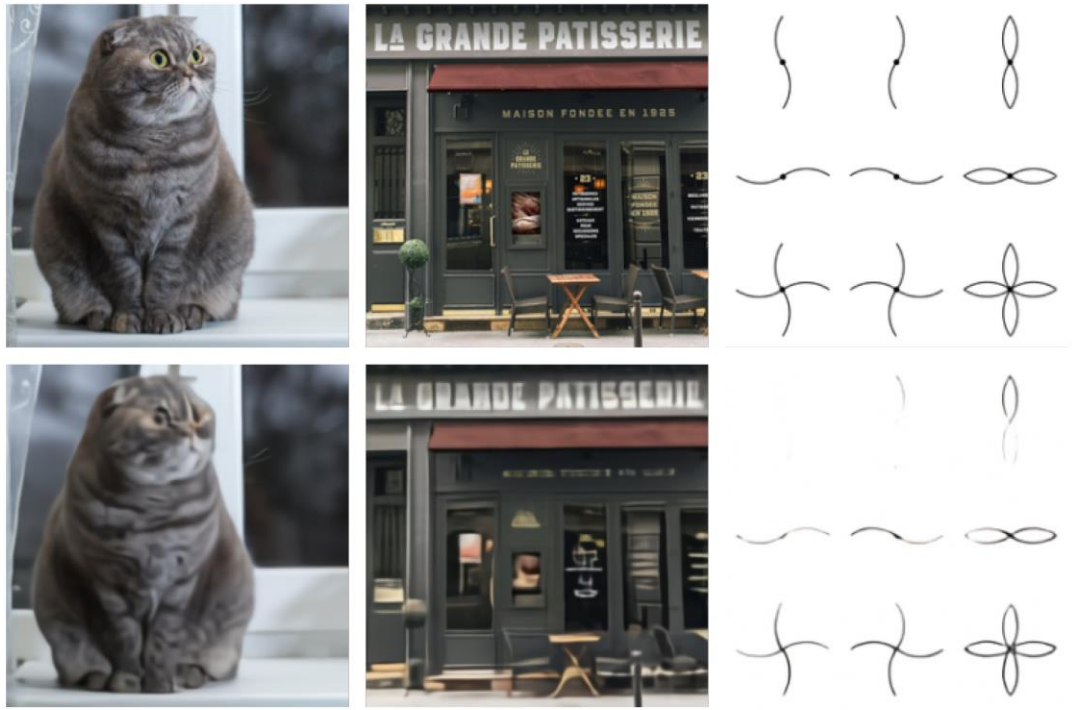
学習の設定

- dVAE，Transformerともに同じデータセットで学習
- dVAEの学習
  - V100 x 64
  - batch_size=512で3M updates
- Transformerの学習
  - V100 x 1024
  - batch_size=1024で430K updates
  - Validation setは606Kペアを用意
    - 最終的に430K updatesを通してoverfittingしなかった

実験結果｜dVAE



- かなり復元できる
  - 詳細な情報は落ちるものの主要な特徴はキープ (...?)
  - よく見ると別物になっている個所もあるが...



実験設定｜Text-to-Image

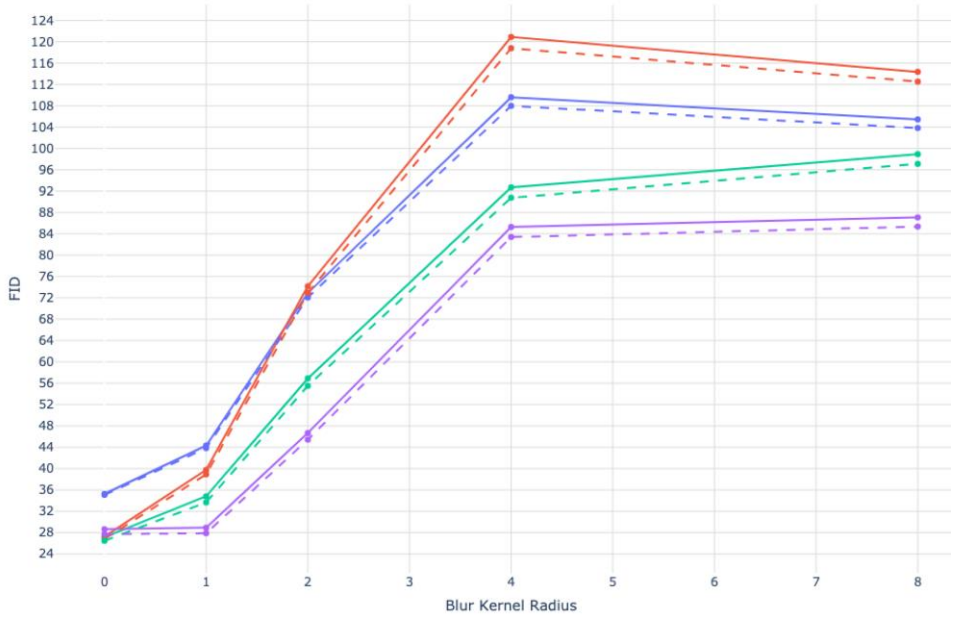
- データセット：MS-COCO, CUB-200
  - CUB-200: 200 subcategoryに分かれた鳥の画像・説明文ペアのデータセット
  - **個々のデータセットでfine-tuningしない (zero-shot)**
  - オリジナル画像が一部データセットに含まれるのでそれを除外して訓練した場合も計測
- 評価指標：IS，FID
  - 共にImageNetで学習したInception Networkを用いて計算されるスコア
    - IS：高いほど良い．生成画像の個々のInception Networkによる予測がシャープかつ，全体として多様なほど高くなるスコア
    - FID：小さいほど良い．生成画像とオリジナル画像それぞれの表現ベクトルの分布間距離をFréchet Distanceで計測することで得られるスコア
- 生成は複数回行い，CLIPモデル†3でテキスト・生成画像をそれぞれエンコードして最もテキストと高い類似度となった生成画像を評価対象とした
- ガウシアンフィルターをかけた場合の性能変化も報告←dVAEの画像劣化をシミュレート

†3: Radford et al., 2021, Learning Transferable Visual Models From Natural Language Supervision.

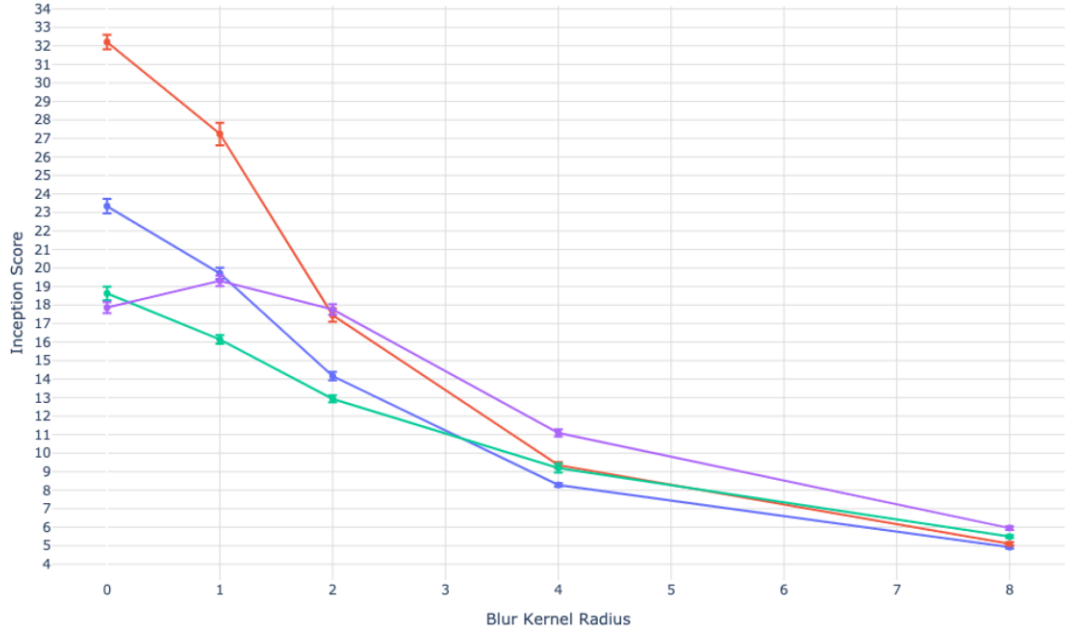
実験結果｜Text-to-Image

- MS-COCO

FID（低いほど良い）



IS（高いほど良い）



AttnGAN DM-GAN DF-GAN DALL-E

実験結果｜Text-to-Image

- MS-COCO（サンプリング数に対する比較）

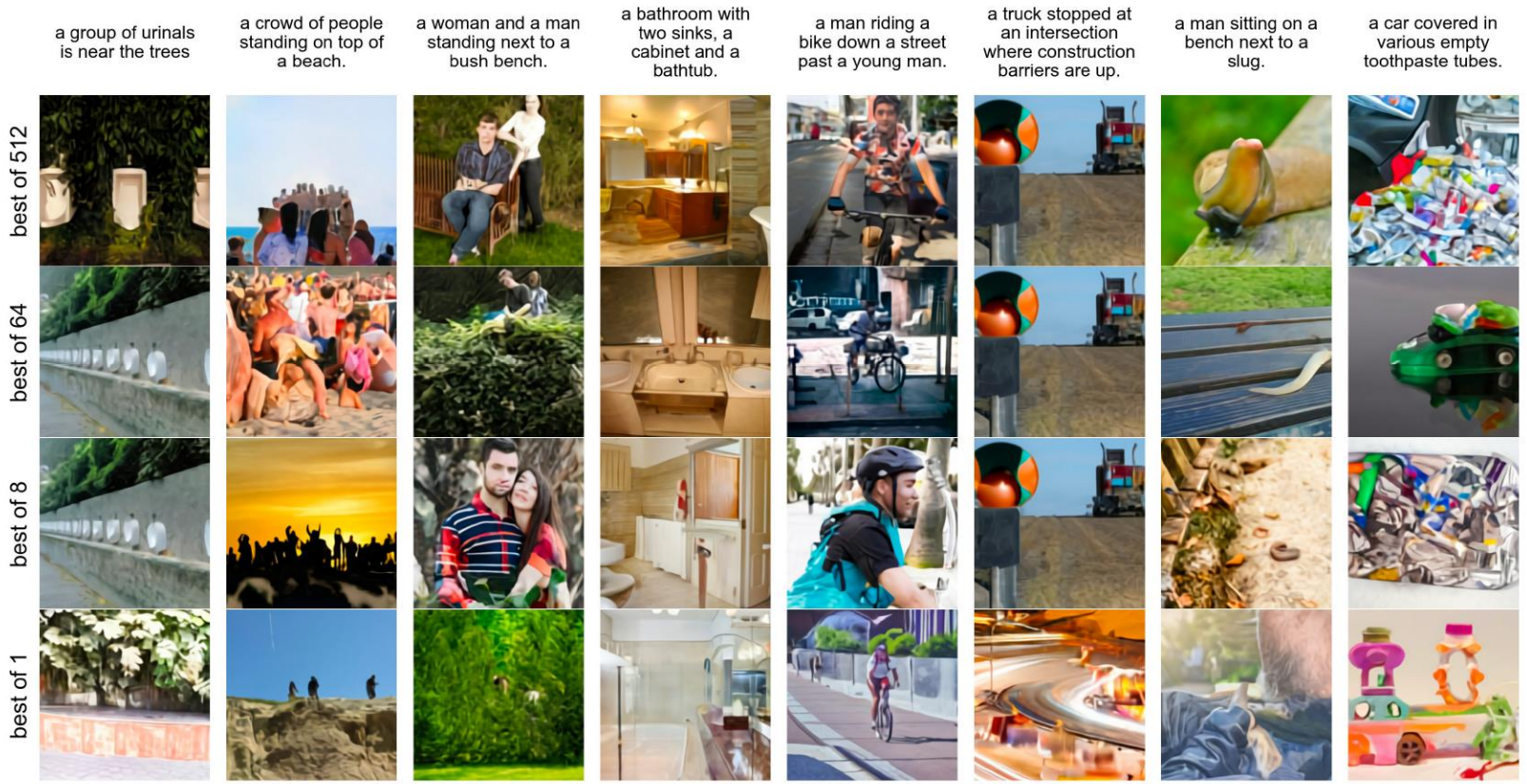
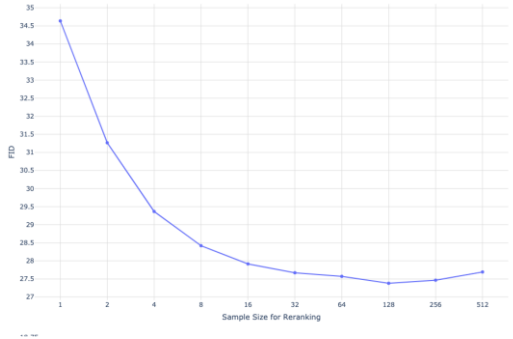
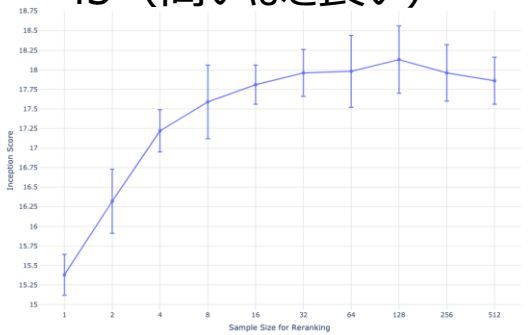


Figure 6. Effect of increasing the number of images for the contrastive reranking procedure on MS-COCO captions.

FID（低いほど良い）



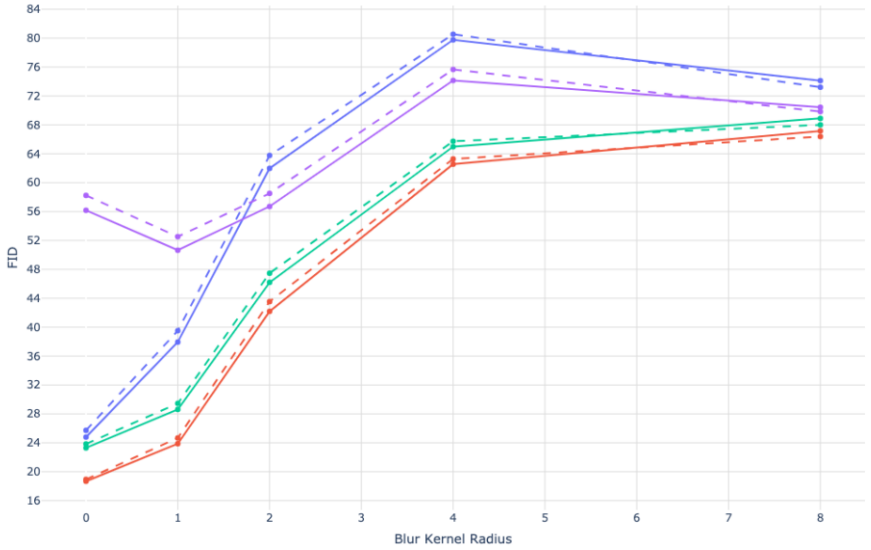
IS（高いほど良い）



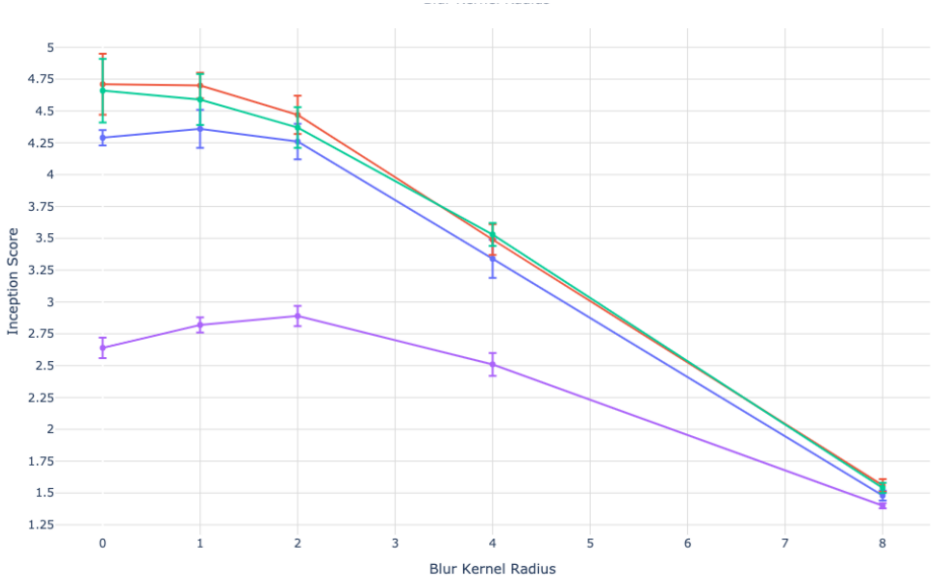
実験結果｜Text-to-Image

- CUB
  - あまりスコアがよくない．鳥のデータに特化している分fine-tuningなしではうまくいかなかったのではと考察されている

FID（低いほど良い）



IS（高いほど良い）



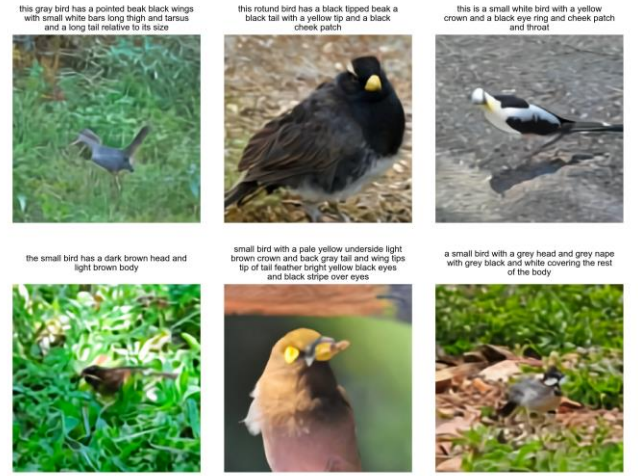
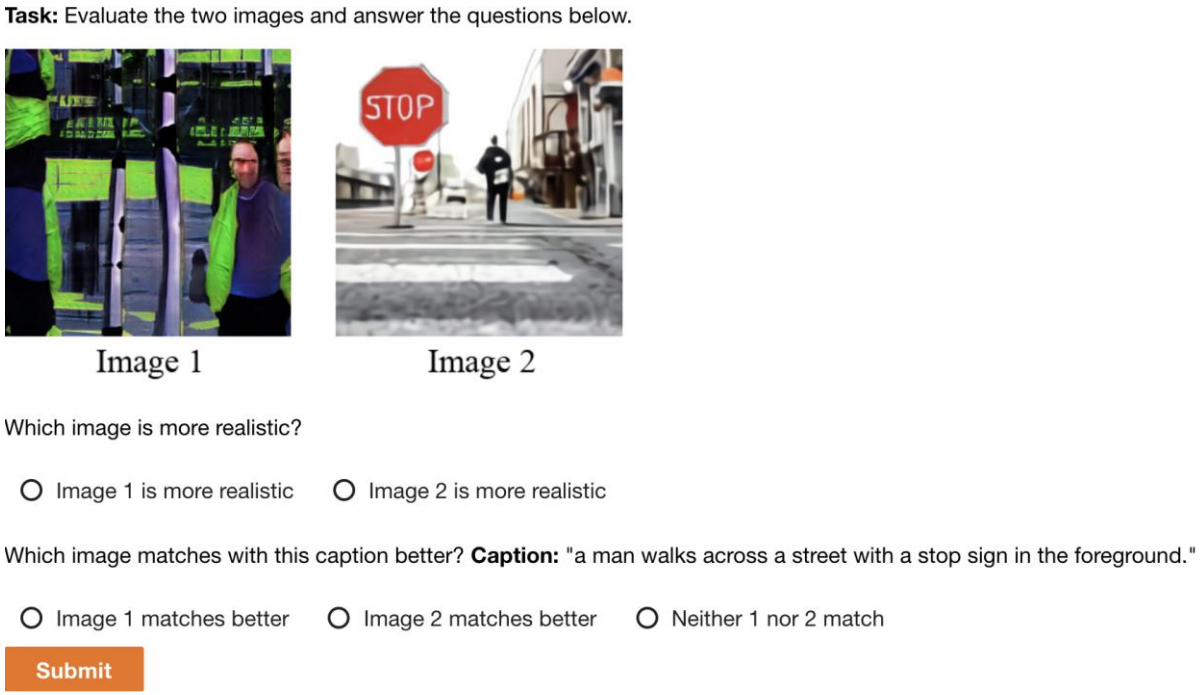


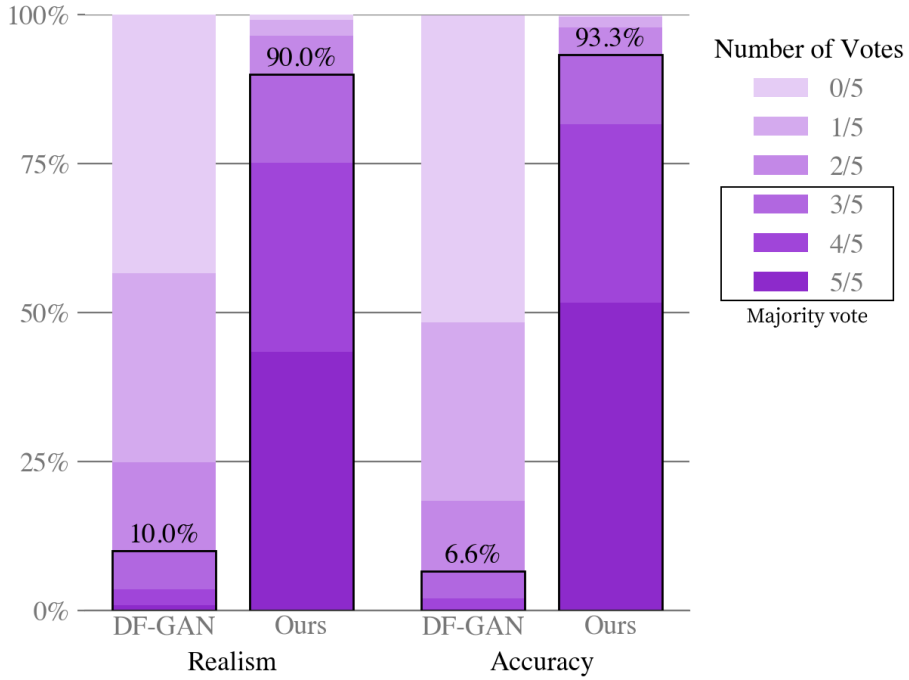
Figure 8. Zero-shot samples from our model on the CUB dataset.

AttnGAN　DM-GAN　DF-GAN　DALL-E

実験結果｜Human evaluation

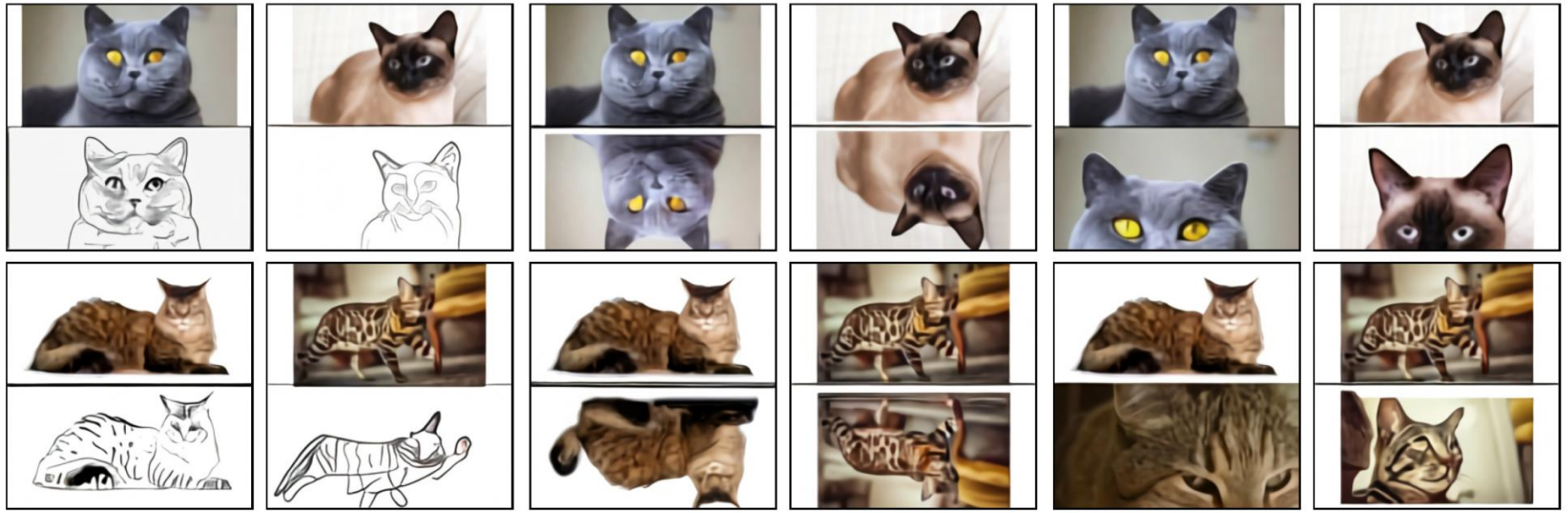
- DF-GANモデルとMS-COCOデータセットのキャプションからの生成で比較
  - DF-GANは論文内で報告されたInception score/FIDスコアが高いモデル
- Evaluatorに生成例のどちらがrealisticか/キャプションに即しているかを投票してもらう





実験結果｜Image-to-Image

- キャプションと上半分だけある画像のエンコード表現を与えた状態から残りを生成開始
- 明確にこの問題を学習していない（zero-shot）にも関わらずうまくいっている



(a) “the exact same cat on the top as a sketch on the bottom”

(b) “the exact same photo on the top reflected upside-down on the bottom”

(c) “2 panel image of the exact same cat. on the top, a photo of the cat. on the bottom, an extreme close-up view of the cat in the photo.”

実験結果｜Image-to-Image

- キャプションと上半分だけある画像のエンコード表現を与えた状態から残りを生成開始
- 明確にこの問題を学習していない（zero-shot）にも関わらずうまくいっている

(d) “the exact same cat on the top colored red on the bottom”

(e) “2 panel image of the exact same cat. on the top, a photo of the cat. on the bottom, the cat with sunglasses.”

(f) “the exact same cat on the top as a postage stamp on the bottom”

まとめ

- シンプルな構造で高品質なText-to-Imageモデルを提案
  - 32x32のグリッドに分けた画像表現をAutoregressiveなTransformerで生成
  - 使い方次第でText+Image入力からの生成もできる
- 正解となる画像表現はDiscrete VAE（dVAE）で学習
  - dVAEは画像サイズを256x256x16bitRGB➡32x32x8192に圧縮する目的
  - 推論時に生成した表現系列をデコードするのにも使う
- 学習はdVAEの学習➡Transformerの学習の2 stage

- かなり計算効率の最適化を頑張っている

所感

- 実際シンプルなアプローチなのか？
  - いろいろと工夫が感じられるが・・・
    - dVAE
      - 離散表現の利用
      - 目的関数のチューニング
      - Encoder終端/Decoder始端の1x1 Conv. が学習の安定化のために重要
    - Transformer
      - テキスト・画像両方の生成を学習，損失は重み付け (テキスト : 画像＝1:7)
  - とはいえモデル構造・パイプラインが相対的にシンプルなのはその通りだと感じる

- “DALL-E”のリポジトリ名にしながら公開しているのがdVAE部分だけなのはどうなのか？