the flower has petals that are bright pinkish purple with white stigma
this white and yellow flower have thin white petals and a round yellow stamen
this small bird has a pink breast and crown, and black primaries and secondaries.
this magnificent fellow is almost all black with a red crest, and white cheek patch.

---

**Generative Adversarial Text to Image Synthesis**[^1]

---

**Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, Honglak Lee**

1 University of Michigan, Ann Arbor, MI, USA (UMICH.EDU)
2 Max Planck Institute for Informatics, Saarbrücken, Germany (MPI-INF.MPG.DE)

---

Presented By: Ezgi Ekiz

Outline

• The goal is to synthesize images that are mistakable for real from textual description. The method is built upon:

• Text encoding that captures important visual details

• Generative Adversarial Networks (GAN) and GAN-CLS

• Manifold interpolation

• Style transfer
Similar Work

Similar Work

Similar Work

A stop sign is flying in blue skies.

A herd of elephants flying in the blue skies.

A toilet seat sits open in the grass field.

A person skiing on sand clad vast desert.

Text Feature Representation

- The representation should capture important visual details

- Word/character based convolutional recurrent network is used

\[
\frac{1}{N} \sum_{n=1}^{N} \Delta(y_n, f_v(v_n)) + \Delta(y_n, f_t(t_n))
\]

\[
f_v(v) = \arg \max_{y \in Y} E_{t \sim T(y)} [\phi(v)^T \varphi(t)]
\]

\[
f_t(t) = \arg \max_{y \in Y} E_{v \sim V(y)} [\phi(v)^T \varphi(t)]
\]

Text Feature Representation

Accumulate matching score

The beak is yellow and pointed

Sequential encoding

Convolutional encoding

The beak is yellow and pointed and the wings are blue.
Multimodality

- A mapping between text and pixels should be learned: GAN is used

- In GAN, the generator network tries to fool adversarially trained discriminator network
  - both are conditioned on text
  - Discriminator acts as a smart adaptive loss function
GAN

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

This flower has small, round violet petals with a dark purple center

\[ \varphi \rightarrow \varphi(t) \]

\[ z \sim \mathcal{N}(0, 1) \]

\[ \hat{x} := G(z, \varphi(t)) \]

This flower has small, round violet petals with a dark purple center

\[ \varphi \rightarrow \varphi(t) \]

\[ D(\hat{x}, \varphi(t)) \]

**Generator Network**
- Fully connected layer (dim-reduc.)
- Leaky ReLU
- Concatenation
- Deconvolution

**Discriminator Network**
- Several layers of stride-2 conv. (with spatial batch normalization)
- Leaky ReLU
- Fully connected layer (dim-reduc.) + rectification (text)
- Depth Concatenation
- conv, rectification, conv.

GAN - CLS

• Naive GAN: \(<\text{real img}, \text{matching text}>\): unrealistic images contribute learning, \(<\text{synthetic img}, \text{arbitrary text}>\): wrong class contributes learning

• GAN CLS: GAN + \(<\text{real image}, \text{mismatched text}>\): should be scored as fake, an additional signal provided by discriminator
GAN - CLS

Algorithm 1 GAN-CLS training algorithm with step size $\alpha$, using minibatch SGD for simplicity.

1: Input: minibatch images $x$, matching text $t$, mismatching $\hat{t}$, number of training batch steps $S$
2: for $n = 1$ to $S$ do
3: \hspace{1em} $\hat{h} \leftarrow \varphi(t)$ \{Encode matching text description\}
4: \hspace{1em} $\hat{h} \leftarrow \varphi(\hat{t})$ \{Encode mis-matching text description\}
5: \hspace{1em} $z \sim \mathcal{N}(0, 1)^Z$ \{Draw sample of random noise\}
6: \hspace{1em} $\hat{x} \leftarrow G(z, \hat{h})$ \{Forward through generator\}
7: \hspace{1em} $s_r \leftarrow D(x, h)$ \{real image, right text\}
8: \hspace{1em} $s_w \leftarrow D(x, \hat{h})$ \{real image, wrong text\}
9: \hspace{1em} $s_f \leftarrow D(\hat{x}, h)$ \{fake image, right text\}
10: \hspace{1em} $\mathcal{L}_D \leftarrow \log(s_r) + (\log(1 - s_w) + \log(1 - s_f))/2$
11: \hspace{1em} $D \leftarrow D - \alpha \partial \mathcal{L}_D / \partial D$ \{Update discriminator\}
12: \hspace{1em} $\mathcal{L}_G \leftarrow \log(s_f)$
13: \hspace{1em} $G \leftarrow G - \alpha \partial \mathcal{L}_G / \partial G$ \{Update generator\}
14: end for
GAN - INT

• Based on the observation that interpolations between embeddings tend to be near the data manifold, extra amount of text embeddings can be generated (although they don’t have a matching text/images, they are useful for D)

• A term added to generator objective:

$$\mathbb{E}_{t_1, t_2 \sim p_{data}} \left[ \log(1 - D(G(z, \beta t_1 + (1 - \beta)t_2))) \right]$$
Style Transfer

\[ s \leftarrow S(x), \; \hat{x} \leftarrow G(s, \varphi(t)) \]

\[ L_{style} = \mathbb{E}_{t, z \sim \mathcal{N}(0,1)} \| z - S(G(z, \varphi(t))) \|_2^2 \]
Experiments

• Datasets:

  • CUB birds (11788 images, 200 classes, 5 captions per image)
    • Split to disjoint classes: 150 train+val, 50 test
  • Oxford-102 flowers (8189 images, 102 categories, 5 captions per image)
    • 82 train+val, 20 test
Experiments

• Text features

• Pre-training on deep deep convolutional-recurrent text encoder (char level) with Google LeNet image embeddings
Qualitative Results

an all black bird with a distinct thick, rounded bill.

this small bird has a yellow breast, brown crown, and black superciliary

a tiny bird, with a tiny beak, tarsus and feet, a blue crown, blue coverts, and black cheek patch

this bird is different shades of brown all over with white and black spots on its head and back

the grey bird has a light grey head and grey webbed feet
Qualitative Results

GT
this flower is white and pink in color, with petals that have veins.

GAN
these flowers have petals that start off white in color and end in a dark purple towards the tips.

GAN - CLS
bright droopy yellow petals with burgundy streaks, and a yellow stigma.

GAN - INT
a flower with long pink petals and raised orange stamen.

GAN - INT - CLS
the flower shown has a blue petals with a white pistil in the center.
Disentangling Style and Content

- Quantification of success is based on pose verification and background verification

- Similar pairs of images constructed for each task via K-means:
  - Avg. RGB for background color
  - Keypoint coordinates for pose
Disentangling Style and Content

- Similar and different images are fed into Style network, then cosine similarity is calculated based on the resulting encodings:
Pose and Background Style Transfer

Text descriptions (content)  Images (style)

The bird has a yellow breast with grey features and a small beak.

This is a large white bird with black wings and a red head.

A small bird with a black head and wings and features grey wings.

This bird has a white breast, brown and white coloring on its head and wings, and a thin pointy beak.

A small bird with white base and black stripes throughout its belly, head, and feathers.

A small sized bird that has a cream belly and a short pointed bill.

This bird is completely red.

This bird is completely white.

This is a yellow bird. The wings are bright blue.
Sentence Interpolation

Figure 8. Left: Generated bird images by interpolating between two sentences (within a row the noise is fixed). Right: Interpolating between two randomly-sampled noise vectors.
GAN CLS on MS COCO

GT

Ours

a group of people on skis stand on the snow.

a man in a wet suit riding a surfboard on a wave.

two giraffe standing next to each other in a forest.

two plates of food that include beans, guacamole and rice.

a green plant that is growing out of the ground.

there is only one horse in the grassy field.

a pitcher is about to throw the ball to the batter.

a picture of a very clean living room.

a sheep standing in an open grass field.

a toilet in a small room with a window and unfinished walls.
THANKS FOR YOUR ATTENTION