

GANchors: Realistic Image Perturbation Distributions for Anchors Using Generative Models

Kurtis Evan David¹ Harrison Keane² Jun Min Noh²

Abstract

We extend and improve the work of Model Agnostic Anchors for explanations on image classification through the use of generative adversarial networks (GANs). Using GANs, we generate samples from a more realistic perturbation distribution, by optimizing under a lower dimensional latent space. This increases the trust in an explanation, as results now come from images that are more likely to be found in the original training set of a classifier, rather than an overlay of random images. A large drawback to our method is the computational complexity of sampling through optimization; to address this, we implement more efficient algorithms, including a diverse encoder. Lastly, we share results from the MNIST and CelebA datasets, and note that our explanations can lead to smaller and higher precision anchors.*

1. Introduction

As use of machine learning models becomes increasingly regulated, explainability of models similarly becomes a pressing issue. For example, the EU's 2018 GDPR regulation stipulates that all individuals have a right to an explanation of the decisions reached using their data. Explainability of models can be approximated by global interpretability of sets of predictions, or local interpretation of individual predictions. (Ribeiro et al., 2018) proposed Model Agnostic Anchors as a method of achieving local interpretability. Their work formally defines an anchor A by the following:

$$\mathbb{E}_{\mathcal{D}(z|A)}[1_{f(x)=f(z)}] \geq \tau, A \subseteq x \quad (1)$$

i.e. a subset of features in a datapoint x where, if the other

¹Department of Computer Science, University of Texas at Austin ²Department of Electrical and Computer Engineering, University of Texas at Austin. Correspondence to: kurtis.e.david@gmail.com <
 >.

*The associated code and datasets can be found at: <https://github.com/kurtisdavid/ImageAnchors>



Figure 1. Comparison of samples generated using the original method with the samples generated with GANchors. Our samples are a more faithful representation of the true underlying data distribution.

features are perturbed in some distribution \mathcal{D} , then the prediction of the model remains the same on the new sampled data z . The threshold τ is a desired confidence, and for their tests, they use $\tau = 0.95$. However, since this expectation is intractable, so they provide a probabilistic definition for (1), given that (1) defines $prec(A)$:

$$P(prec(A) \geq \tau) \geq 1 - \delta \quad (2)$$

i.e. with δ high confidence, $prec(A) \geq \tau$. So long as this is achieved, then we can consider A as an anchor. However, to find the "best" anchor, they introduce their optimization on the coverage of an Anchor, where $cov(A) = \mathbb{E}[A \subseteq z]$:

$$\max_{A \text{ s.t. } P(prec(A) \geq \tau) \geq 1 - \delta} cov(A) \quad (3)$$

To find the anchors, they approximate the highest precision calculation by formulating the search as a *multi-armed bandit* problem. They utilize the KL-LUCB algorithm to solve this. Furthermore, because this is a purely greedy approach, to obtain a better possible solution for (3), they also implement beam search as a way to increase exploration of the solution space.



Figure 2. Original Anchor Method does not create images that are representative of the underlying distribution the model was trained for.

2. Motivation and Contributions

An important part of Anchors (Ribeiro et al., 2018) is the definition of $D(z|A)$, representing a perturbation distribution that lies in the original training distribution. For domains such as text, this was achieved via replacing non anchored words in sentences with new words of the same parts of speech and word embeddings of the original words. However, the authors did not use a realistic $D(z|A)$ for the image domain examples, instead producing them by overlaying segmentations onto random test images as can be seen in Figure 2.

The significant contributions of this paper are as follows:

1. We improve anchors for images by defining a more realistic $D(z|A)$ by performing stochastic gradient descent on a trained generative network.
2. We propose several methods of reducing computational complexity of optimization, including a *diverse encoder* to obtain initial projections to the latent space.

Our work improves anchors for images by generating samples from a perturbation distribution that is more consistent with the authors' original theory. Although the random image and the potential anchors come from the same underlying distribution, the combination of them both result in images that are not represented in the data distribution and are more representative of being an outlier. We do not believe an explanation based on samples generated from such a distribution are trustworthy. Results from this type of analysis can result from unexplained feature interactions, depending on the sensitivity of the desired classification model. To amend this, we propose using Generative Adversarial Networks (Goodfellow et al., 2014), trained to generate new samples that come from the original training distribution.

3. Proposed Method and Related Work

3.1. Generative Networks

Generative Models are a class of models that attempt to learn the distribution of the data and generate samples from a representative distribution. While this class includes various models such as bayesian networks and Markov Random Fields, for this work we focus on generative adversarial networks, or GANs. GANs typically consist of two networks: a generator and a discriminator that compete in a minimax objective. The goal of the discriminator is to classify images as real or fake, while the generators goal is the fool the discriminator by generating images indistinguishable from the real distribution. To sample from a generator, the generator is trained to map random noise vectors z in a lower dimensional latent space to the domain of the target distribution. If successfully trained, the generator is able to create new and diverse samples that mimic the original dataset. We will refer to this as the manifold of the GAN for the rest of the paper.

This ability to model underlying distributions combined with recent advancements in the field make GANs suitable for sampling perturbations of data points for the Anchor algorithm. In addition, researchers have been moving towards sharing trained models, making it easier to access high quality reconstructions without having to train complex networks ourselves. For testing, we explored several pretrained GANs, including DCGAN (Radford et al., 2016), BigGAN (Brock et al., 2019), and ProGAN (Karras et al., 2018).

3.2. Image In-painting

One key observation to be made with their current image framework is their use of superpixels as explainable features made from a segmentation map. This map contains labels for each pixel, thus to extract a certain segment i from the image x , one can apply an element wise product with a binary matrix A to extract a masked image $\hat{x} = A \circ x$ (we simplify this to Ax for the rest of the paper). Thus, our task is closely related to solving the following optimization:

$$\min_y \|\hat{x} - Ay\| \quad (4)$$

The result of equation (1) would reconstruct an image y that would match the masked areas of the image the most, under a norm such as $L2$. Because A describes a binary segmentation map, this problem can be viewed as image inpainting, i.e. the task of filling in holes of a masked image. There are multiple applications for image inpainting due to its versatility, but the most common use for it is to remove undesired part of an image and replace it with realistic pixels that blends in with the rest of the image. However, our purpose of inpainting is not to just fill in with

plausible pixels, but also to generate images that are realistic in terms of context. Another challenge that with inpainting is that the superpixels found during the Anchor algorithm are relatively small compared to the rest of the image. This means that the hole that needs to be filled in is larger than the part of the image that the inpainting method is optimizing over.

One state-of-the-art inpainting method, PatchMatch, fills in masked holes with smooth results, but uses image statistics of the unmasked image to fill in the missing portion (Barnes et al., 2009). This prevents the method from using a semantically-aware approach and often fills in the holes with undesired parts of superpixels. Additional limitations of other methods include strong assumptions of masked holes. For example, Pathak et al. assumes that the masked hole is a rectangle with dimensions of 64×64 that is centered in the middle of the image (Pathak et al., 2016). This would be unfeasible for our research since random segmentation can return scattered and irregular superpixels. Finally, Partial Convolution (Liu et al., 2018) is an inpainting method proposed by NVIDIA that uses deep learning to fill irregular holes by learning semantic priors and hidden representations. However, there are still limitations of this model in that it is deterministic and there is not a way to properly generate multiple realistic perturbations given an anchor.

3.3. Compressed Sensing with Generative Models

Image inpainting is one problem contained in the topic of Compressed Sensing. The goal is to reconstruct original measurements x^* given lossy measurements \hat{x} defined by a measurement matrix A .

$$\hat{x} = Ax^* + \eta$$

This equation is actually overdetermined, since \hat{x} actually lies in a lower dimensional subspace. To solve this and create natural solutions, a sparse prior is induced on possible reconstructions y . We can reframe Equation (4) to:

$$\min_y \|\hat{x} - Ay\| + \lambda \|y\|_1 \quad (5)$$

However, this can still be computationally heavy if the domain is in the image space. Bora et al. suggested alternative to (5) by restricting possible reconstructions y on the manifold of a trained GAN (Bora et al., 2017). This can be found in Equation (3). This enforces a natural image prior based off of the quality of the GAN, and reduces the optimization to a significantly lower-dimensional latent space. One of their applications was image inpainting, and we believe that there is promise in utilizing their framework due to the increasing capabilities of these generative networks, as well as being able to mimic the original training set – the exact problem we are trying to solve.

$$\min_z \|\hat{x} - AG(z)\| \quad (6)$$

3.4. Realistic Image Sampling

Under the optimization of (6), we observe that this is highly non-convex, due to the complex structure of generative networks. We can use this observation to apply a simple sampling algorithm; depending on the initialization of the latent vector $z \in \mathbb{R}^d$, we may obtain very different reconstructions from $G(z)$. Additionally, because it is unfeasible to repeatedly optimize (2) to small values, we introduce a reconstruction threshold ξ . We then define valid reconstructions y from the GAN that contain an anchor A as:

$$y = G(z) \text{ s.t. } \|\hat{x} - Ay\| < \xi \quad (7)$$

Reconstructions y can now be found by running gradient descent. However, to use these reconstructions in the Anchor algorithm, we must guarantee that \hat{x} is within the image. Thus, the final reconstruction uses the generated image to fill in the holes of \hat{x} found in (4) as well as Algorithm 1. Examples compared to original random stitching method can be seen in Figure 4.

$$y = (1 - A)G(z) + \hat{x} \text{ s.t. } \|\hat{x} - AG(z)\| < \xi \quad (8)$$

Algorithm 1 Sample $y \sim D(\cdot | A)$

Require: Trained GAN G , Threshold ξ , Anchor A , Target \hat{x} , Learning rate α

```

1: function SAMPLE( $G, \xi, A, \hat{x}, \alpha$ )
2:    $z \leftarrow \mathcal{N}(0, 1)$ 
3:   while  $\|\hat{x} - AG(z)\| > \xi$  do
4:      $z \leftarrow z - \alpha \nabla \|\hat{x} - AG(z)\|$ 
5:   end while
6:    $y \leftarrow (1 - A)G(z) + \hat{x}$ 
7:   return  $y$ 
8: end function

```

Following the midterm project, we realized that our original algorithm is biased towards generating samples immediately below the threshold ξ . There exist other possible images that could have lower anchor reconstructions, so we wanted to extend our work to include these possibilities. To implement this, we pass different thresholds $< \xi$ into our reconstruction algorithm, sampled according to a threshold distribution. However, this can become computationally inefficient if much smaller thresholds are consistently used.

Thus, we first apply a probabilistic prior on possible reconstructions. Specifically, we contend that the densities of these reconstructions decreases as ξ decreases. This is because at each possible level, there will be fewer local minima that can reach small reconstruction errors on the

manifold. Figure 3 visualizes this concept. To incorporate this prior, we can define a decreasing probability distribution to sample other thresholds $\xi' < \xi$ where ξ is the original max threshold. Algorithm 2 provides an example we use in our project on how to do this. We are essentially collecting points centered at ξ and throwing away any invalid thresholds. We also use a small standard deviation to mimic an exponential decrease of densities. Future work can make this more accurate or learn the true distribution of our algorithm.

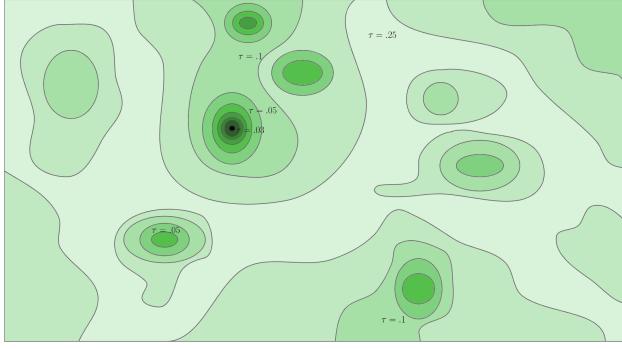


Figure 3. Non-convex solution space for Anchor Reconstruction. Darker contours imply a smaller reconstruction error.

Algorithm 2 Sample $\xi' < \xi$

Require: Max Threshold ξ

```

1: function THRESHOLD-SAMPLE( $\xi$ )
2:   while True do
3:      $\xi' \sim \mathcal{N}(\mu = \xi, \sigma = \xi/6)$ 
4:     if  $\xi' > 2\xi$  or  $\xi' < 0$  then
5:       continue
6:     end if
7:     if  $\xi' > \xi$  then
8:        $\xi' \leftarrow \xi - \xi'$ 
9:     end if
10:    return  $\xi'$ 
11:   end while
12: end function

```

4. Speed Improvements

When we implement Algorithm 1 to compute anchors, it quickly becomes intractable in that the *KL-LUCB* algorithm to estimate the precision will require many samples. In following sections, we will cover improved methods that increase the computational efficiency of GANchors.

4.1. Batch Sampling

One natural improvement to Algorithm 1 is to initialize a large batch of z and jointly optimize each under the mean

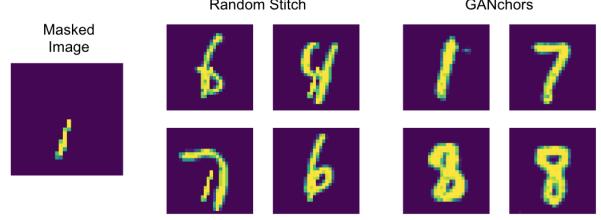


Figure 4. Comparison of Sampling using MNIST

loss. This is easily done in deep learning libraries such as PyTorch (Paszke et al., 2017). To make this as efficient as possible, we apply additional modifications. First, we only update latent vectors z if their reconstruction error of the anchor is above the desired threshold. If they are below the threshold, we instantly collect the reconstructions $G(z)$ and replace these latent vectors with new ones from the same initial distribution.

Secondly, if applying Algorithm 2 to obtain several possible ξ' , a naive matching algorithm would suffer from unlucky initialization. To remedy this, we sort latent vectors z by their anchor reconstruction loss, decreasing, and sort thresholds ξ' , increasing, and match based on the same indices. We chose to do this because it can be done easily using vectorized operations on tensors. This approach is summarized in Algorithm 3.

Algorithm 3 Batch Sample $y \sim D(\cdot|A)$

Require: Trained GAN G , Thresholds $\{\xi'\}$, Anchor A , Target \hat{x} , Learning rate α , Number of Samples N

```

1: function SAMPLE-BATCH( $G, \{\xi'\}, A, \hat{x}, \alpha, N$ )
2:   Fill set  $Z$  with  $N$  samples from  $z \sim \mathcal{N}(0, 1)$ 
3:    $Y \leftarrow \emptyset$ 
4:    $\Xi \leftarrow \{\xi'\}$ 
5:   while  $|Y| < N$  do
6:      $\mathcal{L} \leftarrow \{||\hat{x} - AG(z_i)||\}$ 
7:     Sort  $\mathcal{L}$  increasing. Use indices to resort  $Z$ .
8:     Sort  $\Xi$  decreasing
9:     Match  $Z$  and  $\Xi$ 
10:     $\hat{Z} \leftarrow \{z_i \text{ s.t. } ||\hat{x} - AG(z_i)|| \leq \xi'_i\}$ 
11:     $Z \leftarrow \{z_i \text{ s.t. } ||\hat{x} - AG(z_i)|| > \xi'_i\}$ 
12:     $Y \leftarrow Y \cup (1 - A)G(\hat{Z}) + \hat{x}$ 
13:    Remove  $\xi'_i$  used in  $\hat{Z}$  from  $\Xi$ 
14:   end while
15:   return  $Y$ 
16: end function

```

4.2. Diverse Encoder

When applying our algorithms to our datasets as seen in Section 5, we found that the speeds were still incomparable to the original method. To push this further, we make use of a technique used in (Zhu et al., 2016). Their objective was to project an input image x to the manifold of possible images generated on a trained GAN, by following a similar optimization to Algorithm 1, without any masking. They also noted that this is highly dependent on initializations and wanted to reduce time taken. As a solution, they incorporate another encoder network P that maps the input image to some vector in the d -dimensional latent space of the GAN G . This encoder is optimized with the following:

$$\min_{\theta_P} \mathbb{E}_x [\|x - G(P(x))\|] \quad (9)$$

Essentially, their goal is to have some deterministic network to immediately project the input image to a good reconstruction that lies on the manifold of G . This provided a significant speedup to their setup; however, this is not immediately usable in our case. Our goal is to replace the initial sampling of our latent vectors z with a similar network, but if they all begin at one point (since P is deterministic), then we will no longer get diverse samples.

To solve this problem, we instead create a *diverse encoder* \hat{P} , that outputs more than one vector z . Suppose that the latent space of G is in \mathbb{R}^{100} . An encoder from (Zhu et al., 2016) would output a 100-dimensional vector. Instead of doing this, \hat{P} will output an $N \times 100$ -dimensional vector where N denotes the number of possibly encoded vectors desired from the network. So, if $N = 8$, then it will output 8 possible latent vectors, from a single input image x : $\{z_i | i = 1, \dots, 8\}$ such that each $G(z_i)$ is a faithful reconstruction of x . However, we want to encourage diversity – trivially, \hat{P} can learn to output N copies of the same vector, and we reach a similar issue. Thus we add in an additional regularization term that computes pairwise distances between the different encodings:

$$\min_{\theta_{\hat{P}}} \mathbb{E}_x [\sum_{z_i \in \hat{P}(x)}^N \|x - G(z_i)\| - \lambda \sum_{i \neq j}^N \|z_i - z_j\|] \quad (10)$$

Notice that because we are trying to minimize (10), this would push the different outputs of \hat{P} towards different locations in the latent space. Because this is unconstrained (since the network could just output very large and different vectors but not any good reconstructions), we apply a maximum t on the norm of the λ term. Numerically it is modified to the following:

$$\min_{\theta_{\hat{P}}} \mathbb{E}_x [\sum_{z_i \in \hat{P}(x)}^N \|x - G(z_i)\| + \lambda \sum_{i \neq j}^N (\frac{\|z_i - z_j\| - t}{t})^2] \quad (11)$$

Lastly, because our sampling algorithm takes a masked image \hat{x} as input, we train \hat{P} to reconstruct masked images instead. The final version can be found in (12), where A is a randomized segmentation map. To augment Algorithm 3, we replace the initialization of z in Line 2 by first encoding the target \hat{x} and then adding random noise to help induce more diversity.

$$\min_{\theta_{\hat{P}}} \mathbb{E}_x [\sum_{z_i \in \hat{P}(x)}^N \|Ax - AG(z_i)\| + \lambda \sum_{i \neq j}^N (\frac{\|z_i - z_j\| - t}{t})^2] \quad (12)$$

5. Experiments and Results

5.1. MNIST

To initially validate our various algorithms during testing, we opted to use MNIST (LeCun & Cortes, 2010). We first trained a small convolutional neural network to predict the class and achieved a 99% accuracy on the test set. We also use a pretrained DCGAN implementation (Csinva, 2019) for all of our testing. To run explanations using our algorithm, we set a max anchor MSE threshold $\xi = 0.05$. To obtain samples for speed and precision analysis, we took a single test image from each class and ran three trials on each to compare the different methods in Section 4 as well as their baseline method. We limited testing to this scenario due to the large complexity of the naive methods. For segmentations, we use the quickshift algorithm from *scikit-image*. We wanted each example to produce exactly 15 segments each, so we modify the max distance parameter through a binary search to obtain the desired number of segments.

| Sampling Method | Time (min.) |
|--------------------|-------------|
| Random Stitching | 0.08 |
| Batch | 34.87 |
| Batch + Batch Norm | 4.46 |
| Diverse Encoder | 1.62 |

Table 1. Average Explanation Time

For batch processing, we use a batch size of 64 and apply random restarts every 1000 iterations to help escape sub-optimal local minima on the manifold. We use the Adam optimizer to run gradient descent. During experiments, we

noticed that batched processing took much longer than expected; upon inspection of the GAN’s architecture, we found batch normalization layers (Ioffe & Szegedy, 2015). We hypothesize that this created a significant bottleneck for our algorithm.

Batch normalization aims to solve the problem of internal covariate shift, or unscaled input distributions within each layer of the network. The hypothesis is that network performance, originally on classification tasks, can be improved significantly if inputs to each layer can be normalized, much like how input features are normalized during pre-processing stages. This has also extended to generative networks, as it can lead to more stable training. During training, the layers collects the minibatch mean (μ_B) and variance (σ_B^2) of its inputs and compute the new output y with the following:

$$\begin{aligned}\hat{x}_i &\leftarrow \frac{x - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}} \\ y_i &\leftarrow \gamma \hat{x}_i + \beta\end{aligned}$$

Figure 5. Batch Normalization Layer Operations

The learned parameters are γ and β . The issue arises during the layer’s behavior during training and testing mode. In training mode, it uses μ_B and σ_B^2 , but *also* computes a running mean and variance, μ_D and σ_D^2 , respectively. In testing mode, it uses these estimates μ_D and σ_D^2 to normalize, with the same scheme as in Figure 5. This is an issue, because these learned estimates from the pretrained GAN do not reflect the new batches during our sampling. The original training assumed an input distribution of $\mathcal{N}(0, 1)$, whereas our input distribution is constantly changing – it is only optimized so that every output from the GAN achieve an MSE reconstruction threshold of the desired anchor. When only using training mode, our time significantly decreases, as seen in Table 1.

Lastly, we implemented a diverse encoder for the MNIST dataset. We first run the quickshift algorithm on the dataset to obtain segmentations beforehand – this would speed up training time as they can be expensive to compute in real time. Then, we create a small convolutional neural network that takes a *masked image* as input and then outputs 8 possible encodings, by outputting an 800 dimensional vector, since the latent space of the GAN is in \mathbb{R}^{100} . To train this we use the loss function in Equation (8) of Section 4 with $\lambda = 1$, and $t = 10$. We also apply an upperbound on the norm of pairwise errors to be 10. The input segmentations are randomized, by selecting each segment label with a probability of 0.5. We apply early stopping after a single iteration through the training set. We show examples of both

good and bad results in Figure 6 and 7.

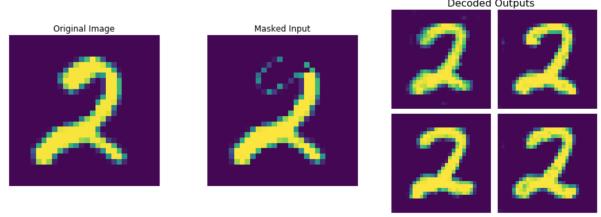


Figure 6. Good Diverse Encoder Outputs

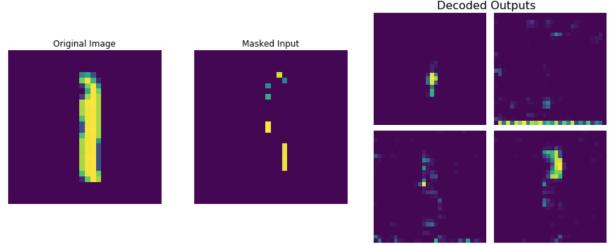


Figure 7. Bad Diverse Encoder Outputs

First note that this diverse encoder is only trying to produce encoded vectors that when decoded and masked, match well the the masked input. The reason we distinguish between good/bad in these figures are the whole images produced by the network. In Figure 6, it actually produces realistic two’s similar to the original image, but Figure 7 shows noisy and nonsensical images. This is most likely due to a small amount of disconnected signals that are used as inputs to the encoder. Nevertheless, we are only looking to use these encoded vectors as *starting areas* for optimization, so they do not have to be perfect at reconstruction.

To test the effectiveness of these methods, we first measure the average time to produce a single explanation with different sampling methods for $\mathcal{D}(z|A)$. These can be found in Table 1. We see that improvement to sampling changes the average time by an order of magnitude, however cannot reach the times produced by the original random sampling method. We believe that if the encoder was more finely trained, faster times can be achieved, but that tradeoff is acceptable to a certain extent for more realistic sampling.

Additionally, we measure the precision of our explanations on the same set as another metric of comparison. These can be found in Table 2. Immediately we can see that our explanations achieve a higher precision, and smaller sizes of anchors. We hypothesize that this happens due to the inherent structure of natural images, especially those that

are learned by a trained GAN. Given a subset of pixels that must be reconstructed, there exist a smaller amount of possible digits. For example, in Figure 6, we can see that the structure of the masked input contains the curves one would only see in handwritten 2s. There exist correlations between the grouping of pixels that actually provide more information for the classification, since the GAN would only create perturbation that look like 2s. Thus, as the anchor size increases, the classifier sees a more and more restricted perturbation distribution. We assert that this is expected behavior, as we are now only testing the classifier on true digits, rather than a random mixture of different parts. This however weakens the explanation, as the segments shown now describe *pixels that result in the same prediction, or are correlated to other pixels that result in the same prediction*. Further work must still be done to strengthen these conclusions. Nevertheless, the explanations are more faithful to the true distribution of data, as the explanations from the original random stitching method can also be a result of correlations between outlier signals of the random images. We have also included side by side comparisons in Figure 8.

| Sampling Method | Precision | Size of Anchor |
|------------------|-----------|----------------|
| Random Stitching | 0.982 | 4.2 |
| Batch + BN | 0.987 | 1.87 |
| Diverse Encoder | 0.993 | 1.63 |

Table 2. MNIST Anchor Precisions. We omit non Batch Norm results as samples were initialized the same way. Size of Anchor defined as the average number of segments.

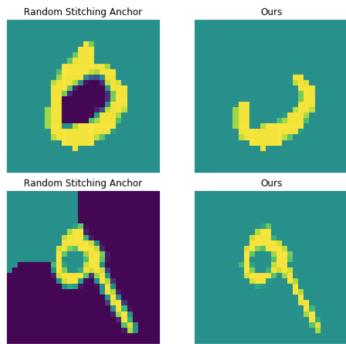


Figure 8. Computed Anchor Comparisons Left: Original Random Sampling. Right: Sampling from a trained GAN

5.2. CelebA-HQ

While the original authors’ explored anchors on ImageNet, state of the art GANs for the dataset were mainly class-

condition, including BigGAN. It proved to be a challenge to generate samples below a desired threshold ξ while selecting random classes. Due to this, we opted to use CelebA-HQ (Tero Karras, 2018), a dataset of 1024x1024x3 images of celebrity faces created by passing CelebA through a super-resolution network. The generative model associated with this dataset is ProGAN (Karras et al., 2018), a progressively grown generator which outputs high resolution, realistic 1024x1024x3 images of faces. For the purposes of computation time and memory, we scaled down both the dataset and generator output to 256x256x3.

For the purposes of testing, we trained a convolutional neural network using ResNet18 to predict whether the image was smiling or not. The classifier we attempted to explain achieved a validation accuracy of 93%. The units for CelebA anchors were segments achieved using SLIC (numsegments = 10, compactness = 20). Finally, we had to qualitatively select an anchor MSE threshold at which to collect the samples based on running time, quality of the match, and variation of results. After testing in the range of [0.01, .15], we found that .075 would be sufficient for our experiments. The results of this threshold testing can be seen in Figure 9. The CelebA experiments were ran on a single GTX1080 with an i5-4790k at 3.5Ghz.

The anchors our method found were consistent with intuition in that the anchor was the lower face, mouth, or combination of the two. Naturally, these are important segments for smiling and thus result in a valid explanation of the image. All of the anchors seen in Figure 10 generated by our algorithm achieved a precision of 1.0.

We also try to train a diverse encoder on the CelebAHQ dataset. To do this, we obtain masks from CelebAHQ-Mask (Liu, 2019) and create an encoder that outputs $N = 2$ encodings. This is limited due to the large memory consumption needed to use ProGAN. We trained on random segmentations, much like in MNIST, with a batch size of 8. To stabilize training, we found that we also had to apply L_2 regularization on the encoder’s outputs, as they tend to diverge towards large vectors. We use $\lambda = 1e-3$ and L_2 regularization parameter of $1e-6$. We have included results of an example encoding and patched up results in Figure 11. Unlike MNIST, our encoder was unable to achieve both decent reconstructions as well as diversity. The reconstructions themselves look very blurry, as if they are average images in the dataset; it proved difficult to train due to the sheer size of the input images. In addition, when we used these as starting points for our batch search, this resulted in all very similar images, with less quality than compared to Figure 9. For this reason, we stuck to randomly initializing from $\mathcal{N}(0, 1)$.

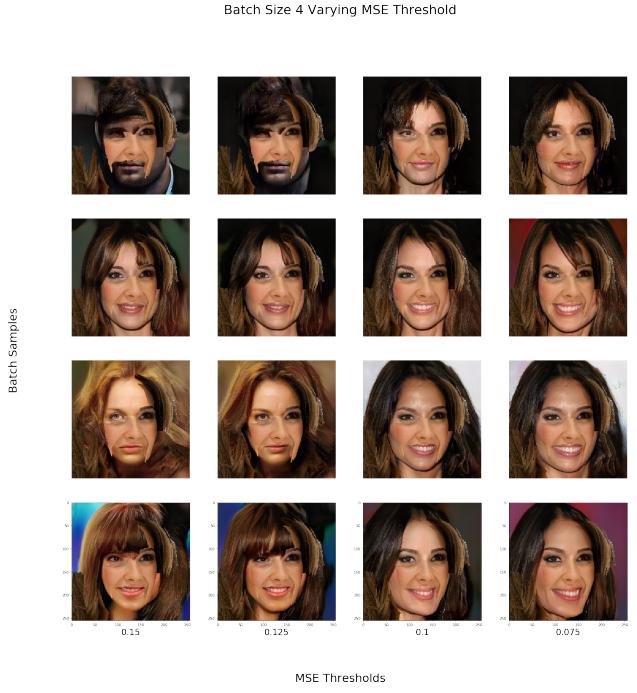


Figure 9. Batch samples generated at various MSEs. As the threshold decreases, the samples have lower variance, but higher running time. Additionally for anchor purposes, a baseline variance is necessary to actually test the potential anchor. We found .075 balanced the these tradeoffs appropriately.

6. Discussion

6.1. Limitations

A considerable drawback of this approach is the computational complexity of having to run a GAN. Even though we were able to dramatically increase the average speed of generating anchors for images from hours to a few minutes, the time to train a GAN as well as sample enough coverage samples for a robust anchor is still considerably high, for example if one wanted to generate explanations for a large dataset. Furthermore, running ProGAN with a batch size of 4 inside our anchor sample generation algorithm required at least 8GB of GPU memory on a single GTX1080. The high memory consumption limits the speedup gained through batch, and could potentially be optimized for a larger speedup. This bottleneck also makes it difficult to train a successful diverse encoder on large datasets, but we hope that the idea can be explored more.

Our work also relied heavily on qualitatively searching for optimal hyperparameters for generating realistic images. We considered various boundary error methods to measure how well a sampled background fit the anchor, but in cases with a clean segmentation cut there would be higher acceptable



Figure 10. Anchors for the prediction of smiling or not smiling. For most cases, the anchor involved the lower face and mouth, intuitively sufficient for smiles. All anchors seen above achieved a precision of 1.0. Compared with the bottom row from the original method, our anchors tend to be smaller with better precision.

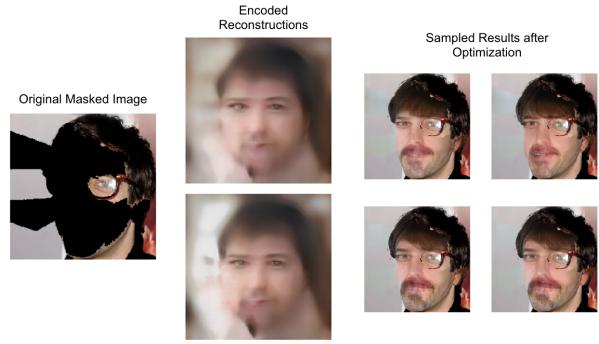


Figure 11. Diverse Encoder on CelebA-HQ. Left: Original Masked Image Middle: Reconstructions from Diverse Encoder Outputs Right: Results from applying our Sampling Algorithms

boundary differences than an anchor in the center of a homogeneous area. Without making considerable assumptions, we were unable to define a contextual realism metric of how well an anchor maps onto a sample without extending the anchor outside of its bounds or using information from the original image not found in the anchor.

Finally, our work assumes that an unconditional GAN exists for the data distribution that the classifier is trained on. Thus, the challenges of training a GAN are transferred to our work. This difficulty can be relieved through sharing of models and public access to datasets and generators.

6.2. Future Work

While our project was limited to the image domain, the use of GANchors could be applied to other domains, be that text, tabular or time series. As GANs model a mapping from a latent space to the underlying data distribution, they could theoretically be used to sample from any distribution in most

domains. By incorporating GANs into the anchor pipeline, this technique could be increasingly model-agnostic, simply accepting a generator and explainable feature extractor specific to the model/domain. In addition, we would also like to supplement our hypotheses of more realistic perturbations by developing a new metric that could reflect the similarity to a reference set. Something along the lines of a BLEU score but for images is what we have in mind.

References

- Barnes, C., Shechtman, E., Finkelstein, A., and Goldman, D. B. PatchMatch: A randomized correspondence algorithm for structural image editing. *ACM Transactions on Graphics (Proc. SIGGRAPH)*, 28(3), August 2009.
- Bora, A., Jalal, A., Price, E., and Dimakis, A. G. Compressed sensing using generative models. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 537–546, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. URL <http://proceedings.mlr.press/v70/bora17a.html>.
- Brock, A., Donahue, J., and Simonyan, K. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019. URL <https://openreview.net/forum?id=B1xsqj09Fm>.
- Csinva. Pretrained gans in pytorch for mnist/cifar. https://github.com/csinva/pytorch_gan_pretrained, 2019.
- Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative Adversarial Networks. *arXiv e-prints*, art. arXiv:1406.2661, Jun 2014.
- Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ICML’15, pp. 448–456. JMLR.org, 2015. URL <http://dl.acm.org/citation.cfm?id=3045118.3045167>.
- Karras, T., Aila, T., Laine, S., and Lehtinen, J. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018. URL <https://openreview.net/forum?id=Hk99zCeAb>.
- LeCun, Y. and Cortes, C. MNIST handwritten digit database. 2010. URL <http://yann.lecun.com/exdb/mnist/>.
- Liu, G., Reda, F. A., Shih, K. J., Wang, T.-C., Tao, A., and Catanzaro, B. Image inpainting for irregular holes using partial convolutions. In *The European Conference on Computer Vision (ECCV)*, September 2018.
- Liu, Z. A large-scale face dataset for face parsing, recognition, generation and editing. <https://github.com/switchablenorms/CelebAMask-HQs>, 2019.
- Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. Automatic differentiation in pytorch. 2017.
- Pathak, D., Krähenbühl, P., Donahue, J., Darrell, T., and Efros, A. Context encoders: Feature learning by inpainting. In *CVPR*, 2016.
- Radford, A., Metz, L., and Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. In *ICLR*, 2016.
- Ribeiro, M. T., Singh, S., and Guestrin, C. Anchors: High-precision model-agnostic explanations. In McIlraith, S. A. and Weinberger, K. Q. (eds.), *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pp. 1527–1535. AAAI Press, 2018. URL <https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16982>.
- Tero Karras, Timo Aila, S. L. J. L. Progressive growing of gans for improved quality, stability, and variation. https://github.com/tkarras/progressive_growing_of_gans, 2018.
- Zhu, J.-Y., Krähenbühl, P., Shechtman, E., and Efros, A. A. Generative visual manipulation on the natural image manifold. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2016.