

UNSUPERVISED REPRESENTATION LEARNING WITH DEEP CONVOLUTIONAL GENERATIVE ADVERSARIAL NETWORKS

Alec Radford & Luke Metz

indico Research

Boston, MA

{alec, luke}@indico.io

Soumith Chintala

Facebook AI Research

New York, NY

soumith@fb.com

ABSTRACT

In recent years, supervised learning with convolutional networks (CNNs) has seen huge adoption in computer vision applications. Comparatively, unsupervised learning with CNNs has received less attention. In this work we hope to help bridge the gap between the success of CNNs for supervised learning and unsupervised learning. We introduce a class of CNNs called deep convolutional generative adversarial networks (DCGANs), that have certain architectural constraints, and demonstrate that they are a strong candidate for unsupervised learning. Training on various image datasets, we show convincing evidence that our deep convolutional adversarial pair learns a hierarchy of representations from object parts to scenes in both the generator and discriminator. Additionally, we use the learned features for novel tasks - demonstrating their applicability as general image representations.

1 INTRODUCTION

Learning reusable feature representations from large unlabeled datasets has been an area of active research. In the context of computer vision, one can leverage the practically unlimited amount of unlabeled images and videos to learn good intermediate representations, which can then be used on a variety of supervised learning tasks such as image classification. We propose that one way to build good image representations is by training Generative Adversarial Networks (GANs) (Goodfellow et al., 2014), and later reusing parts of the generator and discriminator networks as feature extractors for supervised tasks. GANs provide an attractive alternative to maximum likelihood techniques. One can additionally argue that their learning process and the lack of a heuristic cost function (such as pixel-wise independent mean-square error) are attractive to representation learning. GANs have been known to be unstable to train, often resulting in generators that produce nonsensical outputs. There has been very limited published research in trying to understand and visualize what GANs learn, and the intermediate representations of multi-layer GANs.

In this paper, we make the following contributions

- We propose and evaluate a set of constraints on the architectural topology of Convolutional GANs that make them stable to train in most settings. We name these class of architectures Deep Convolutional GANs (DCGAN)
- We use the trained discriminators for image classification tasks, showing competitive performance with other unsupervised algorithms.
- We visualize the filters learnt by GANs and empirically show that specific filters have learned to draw specific objects.

- We show that the generators have interesting vector arithmetic properties allowing for easy manipulation of many semantic qualities of generated samples.

2 RELATED WORK

2.1 REPRESENTATION LEARNING FROM UNLABELED DATA

Unsupervised representation learning is a fairly well studied problem in general computer vision research, as well as in the context of images. A classic approach to unsupervised representation learning is to do clustering on the data (for example using K-means), and leverage the clusters for improved classification scores. In the context of images, one can do hierarchical clustering of image patches (Coates & Ng, 2012) to learn powerful image representations. Another popular method is to train auto-encoders (convolutionally, stacked (Vincent et al., 2010), separating the what and where components of the code (Zhao et al., 2015), ladder structures (Rasmus et al., 2015)) that encode an image into a compact code, and decode the code to reconstruct the image as accurately as possible. These methods have also been shown to learn good feature representations from image pixels. Deep belief networks (Lee et al., 2009) have also been shown to work well in learning hierarchical representations.

2.2 GENERATING NATURAL IMAGES

Generative image models are well studied and fall into two categories: parametric and non-parametric.

The non-parametric models often do matching from a database of existing images, often matching patches of images, and have been shown to be used in texture synthesis (Efros et al., 1999), super-resolution (Freeman et al., 2002) and in-painting (Hays & Efros, 2007).

Parametric models for generating images has been explored extensively (for example on MNIST digits or for texture synthesis (Portilla & Simoncelli, 2000)). However, generating natural images of the real world have had not much success until recently. A variational sampling approach to generating images (Kingma & Welling, 2013) has had some success, but the samples often suffer from being blurry. Another approach generates images using an iterative forward diffusion process (Sohl-Dickstein et al., 2015). Generative Adversarial Networks (Goodfellow et al., 2014) generated images suffering from being noisy and incomprehensible. A laplacian pyramid extension to this approach (Denton et al., 2015) showed higher quality images, but they still suffered from the objects looking wobbly because of noise introduced in chaining multiple models. A recurrent network approach (Gregor et al., 2015) and a deconvolution network approach (Dosovitskiy et al., 2014) have also recently had some success with generating natural images. However, they have not leveraged the generators for supervised tasks.

2.3 VISUALIZING THE INTERNALS OF CNNS

One constant criticism of using neural networks has been that they are black-box methods, with little understanding of what the networks do in the form of a simple human-consumable algorithm. In the context of CNNs, Zeiler et. al. (Zeiler & Fergus, 2014) showed that by using deconvolutions and filtering the maximal activations, one can find the approximate purpose of each convolution filter in the network. Similarly, using a gradient descent on the inputs lets us inspect the ideal image that activates certain subsets of filters (Mordvintsev et al.).

3 APPROACH AND MODEL ARCHITECTURE

Historical attempts to scale up GANs using CNNs to model images have been unsuccessful. This motivated the authors of LAPGAN (Denton et al., 2015) to develop an alternative approach to iteratively upscale low resolution generated images which can be modeled more reliably. We also encountered difficulties attempting to scale GANs using CNN architectures commonly used in the supervised literature. However, after extensive model exploration we identified a family of archi-

tures that resulted in stable training across a range of datasets and allowed for training higher resolution and deeper generative models.

Core to our approach is adopting and modifying three recently demonstrated changes to CNN architectures.

The first is the all convolutional net (Springenberg et al., 2014) which replaces deterministic spatial pooling functions (such as maxpooling) with strided convolutions, allowing the network to learn its own spatial downsampling. We use this approach in our generator, allowing it to learn its own spatial upsampling, and discriminator.

Second is the trend towards eliminating fully connected layers on top of convolutional features. The strongest example of this is global average pooling which has been utilized in state of the art image classification models (Mordvintsev et al.). We found global average pooling increased model stability but hurt convergence speed. A middle ground of directly connecting the highest convolutional features to the input and output respectively of the generator and discriminator worked well. The first layer of the GAN, which takes a uniform noise distribution Z as input, could be called fully connected as it is just a matrix multiplication, but the result is reshaped into a 4-dimensional tensor and used as the start of the convolution stack. For the discriminator, the last convolution layer is flattened and then fed into a single sigmoid output. See Fig. 1 for a visualization of an example model architecture.

Third is Batch Normalization (Ioffe & Szegedy, 2015) which stabilizes learning by normalizing the input to each unit to have zero mean and unit variance. This helps deal with training problems that arise due to poor initialization and helps gradient flow in deeper models. This proved critical to get deep generators to begin learning, preventing the generator from collapsing all samples to a single point which is a common failure mode observed in GANs. Directly applying batchnorm to all layers however, resulted in sample oscillation and model instability. This was avoided by not applying batchnorm to the generator output layer and the discriminator input layer.

The ReLU activation (Nair & Hinton, 2010) is used in the generator with the exception of the output layer which uses the Tanh function. We observed that using a bounded activation allowed the model to learn more quickly to saturate and cover the color space of the training distribution. Within the discriminator we found the leaky rectified activation (Maas et al., 2013) (Xu et al., 2015) to work well, especially for higher resolution modeling. This is in contrast to the original GAN paper, which used the maxout activation (Goodfellow et al., 2013).

Architecture guidelines for stable Deep Convolutional GANs

- Replace any pooling layers with strided convolutions (discriminator) and fractional-strided convolutions (generator).
- Use batchnorm in both the generator and the discriminator.
- Remove fully connected hidden layers for deeper architectures.
- Use ReLU activation in generator for all layers except for the output, which uses Tanh.
- Use LeakyReLU activation in the discriminator for all layers.

4 DETAILS OF ADVERSARIAL TRAINING

We trained DCGANs on three datasets, Large-scale Scene Understanding (LSUN) (Yu et al., 2015), Imagenet-1k and a newly assembled Faces dataset. Details on the usage of each of these datasets are given below.

No pre-processing was applied to training images besides scaling to the range of the tanh activation function [-1, 1]. All models were trained with mini-batch stochastic gradient descent (SGD) with a mini-batch size of 128. All weights were initialized from a zero-centered Normal distribution with standard deviation 0.02. In the LeakyReLU, the slope of the leak was set to 0.2 in all models. While previous GAN work has used momentum to accelerate training, we used the Adam optimizer (Kingma & Ba, 2014) with tuned hyperparameters. We found the suggested learning rate of 0.001, to be too high, using 0.0002 instead. Additionally, we found leaving the momentum term β_1 at the



Figure 1: DCGAN generator used for LSUN scene modeling. A 100 dimensional uniform distribution Z is projected to a small spatial extent convolutional representation with many feature maps. A series of four fractionally-strided convolutions (in some recent papers, these are wrongly called deconvolutions) then convert this high level representation into a 64×64 pixel image. Notably, no fully connected or pooling layers are used.

suggested value of 0.9 resulted in training oscillation and instability while reducing it to 0.5 helped stabilize training.

4.1 LSUN

As visual quality of samples from generative image models has improved, concerns of over-fitting and memorization of training samples have risen. To demonstrate how our model scales with more data and higher resolution generation, we train a model on the LSUN bedrooms dataset containing a little over 3 million training examples. Recent analysis has shown that there is a direct link between how fast models learn and their generalization performance (Hardt et al., 2015). We show samples from one epoch of training (Fig.2), mimicking online learning, in addition to samples after convergence (Fig.3), as an opportunity to demonstrate that our model is not producing high quality samples via simply overfitting/memorizing training examples. No data augmentation was applied to the images.

4.1.1 DEDUPLICATION

To further decrease the likelihood of the generator memorizing input examples (Fig.2) we perform a simple image de-duplication process. We fit a 3072-128-3072 de-noising dropout regularized RELU autoencoder on 32x32 downsampled center-crops of training examples. The resulting code layer activations are then binarized via thresholding the ReLU activation which has been shown to be an effective information preserving technique (Srivastava et al., 2014) and provides a convenient form of semantic-hashing, allowing for linear time de-duplication. Visual inspection of hash collisions showed high precision with an estimated false positive rate of less than 1 in 100. Additionally, the technique detected and removed approximately 275,000 near duplicates, suggesting a high recall.

4.2 FACES

We scraped images containing human faces from random web image queries of peoples names. The people names were acquired from dbpedia, with a criterion that they were born in the modern era. This dataset has 3M images from 10K people. We run an OpenCV face detector on these images, keeping the detections that are sufficiently high resolution, which gives us approximately 350,000 face boxes. We use these face boxes for training. No data augmentation was applied to the images.



Figure 2: Generated bedrooms after one training pass through the dataset. Theoretically, the model could learn to memorize training examples, but this is experimentally unlikely as we train with a **small learning rate and minibatch SGD**. We are aware of no prior empirical evidence demonstrating memorization with SGD and a small learning rate in only one epoch.



Figure 3: Generated bedrooms after five epochs of training. There appears to be evidence of visual under-fitting via repeated textures across multiple samples.

4.3 IMAGENET-1K

We use Imagenet-1k (Deng et al., 2009) as a source of **natural images for unsupervised training**. We train on 32×32 min-resized center crops. No data augmentation was applied to the images.

5 EMPIRICAL VALIDATION OF DCGANS CAPABILITIES

5.1 CLASSIFYING CIFAR-10 USING GANS AS A FEATURE EXTRACTOR

One common technique for evaluating the quality of unsupervised representation learning algorithms is to apply them as a feature extractor on supervised datasets and evaluate the performance of linear models fitted on top of these features.

On the CIFAR-10 dataset, a very strong baseline performance has been demonstrated from a well tuned single layer feature extraction pipeline utilizing K-means as a feature learning algorithm. When using a very large amount of feature maps (4800) this technique achieves 80.6% accuracy. An unsupervised multi-layered extension of the base algorithm reaches 82.0% accuracy (Coates & Ng, 2011). To evaluate the quality of the representations learned by DCGANs for supervised tasks, we train on Imagenet-1k and then use the discriminator’s convolutional features from all layers, maxpooling each layers representation to produce a 4×4 spatial grid. These features are then flattened and concatenated to form a 28672 dimensional vector and a regularized linear L2-SVM classifier is trained on top of them. This achieves 82.8% accuracy, out performing all K-means based approaches. Notably, the discriminator has many less feature maps (512 in the highest layer) compared to K-means based techniques, but does result in a larger total feature vector size due to the many layers of 4×4 spatial locations. The performance of DCGANs is still less than that of Exemplar CNNs (Dosovitskiy et al., 2015), a technique which trains normal discriminative CNNs in an unsupervised fashion to differentiate between specifically chosen, aggressively augmented, exemplar samples from the source dataset. Further improvements could be made by finetuning the discriminator’s representations, but we leave this for future work. Additionally, since our DCGAN was never trained on CIFAR-10 this experiment also demonstrates the domain robustness of the learned features.

Table 1: CIFAR-10 classification results using our pre-trained model. Our DCGAN is not pre-trained on CIFAR-10, but on Imagenet-1k, and the features are used to classify CIFAR-10 images.

Model	Accuracy	Accuracy (400 per class)	max # of features units
1 Layer K-means	80.6%	63.7% ($\pm 0.7\%$)	4800
3 Layer K-means Learned RF	82.0%	70.7% ($\pm 0.7\%$)	3200
View Invariant K-means	81.9%	72.6% ($\pm 0.7\%$)	6400
Exemplar CNN	84.3%	77.4% ($\pm 0.2\%$)	1024
DCGAN (ours) + L2-SVM	82.8%	73.8% ($\pm 0.4\%$)	512

6 INVESTIGATING AND VISUALIZING THE INTERNALS OF THE NETWORKS

We investigate the trained generators and discriminators in a variety of ways. We do not do any kind of nearest neighbor search on the training set. Nearest neighbors in pixel or feature space are trivially fooled (Theis et al., 2015) by small image transforms. We also do not use log-likelihood metrics to quantitatively assess the model, as it is a poor (Theis et al., 2015) metric.

6.1 WALKING IN THE LATENT SPACE

The first experiment we did was to understand the landscape of the latent space. Walking on the manifold that is learnt can usually tell us about signs of memorization (if there are sharp transitions) and about the way in which the space is hierarchically collapsed. If walking in this latent space results in semantic changes to the image generations (such as objects being added and removed), we can reason that the model has learned relevant and interesting representations. The results are shown in Fig.4.

6.2 VISUALIZING THE DISCRIMINATOR FEATURES

Previous work has demonstrated that supervised training of CNNs on large image datasets results in very powerful learned features (Zeiler & Fergus, 2014). Additionally, supervised CNNs trained on

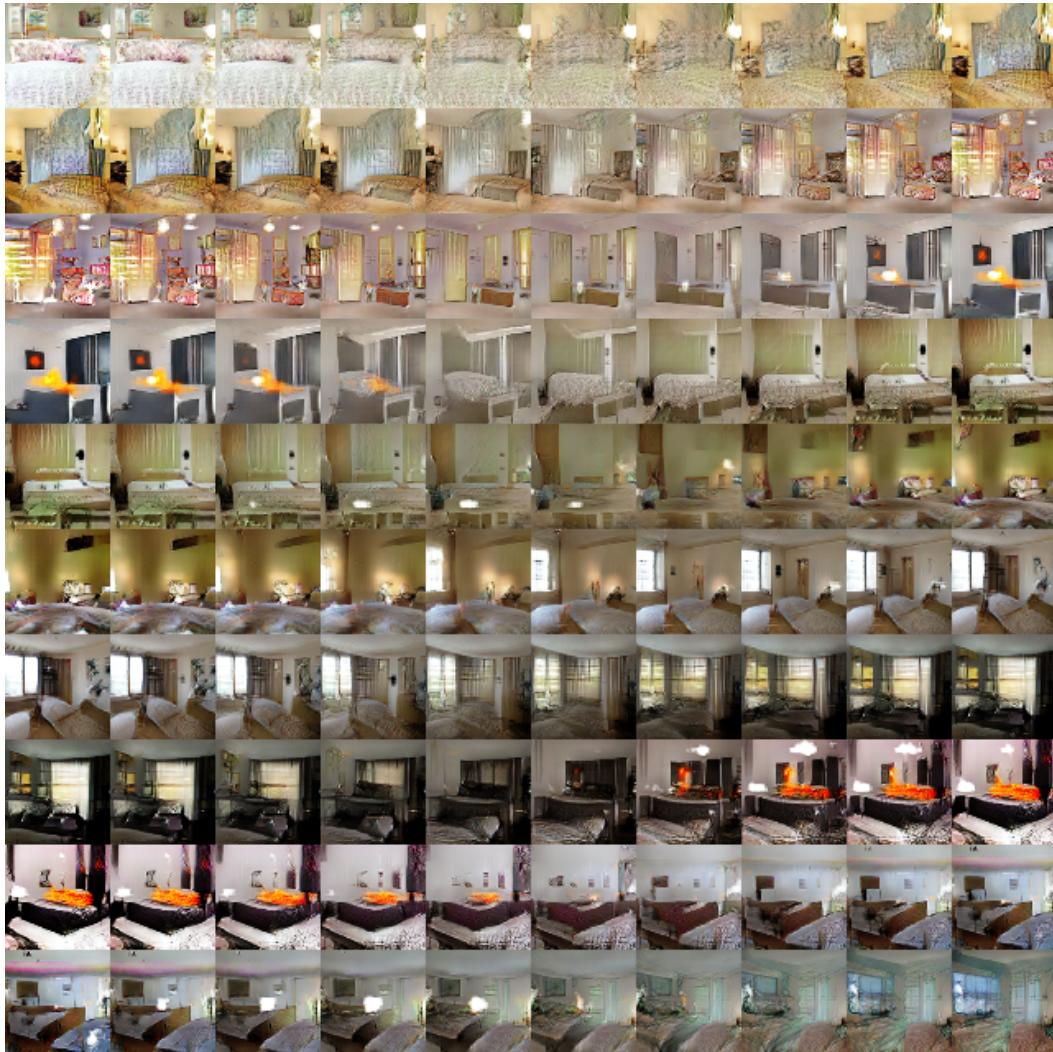


Figure 4: Top rows: Interpolation between a series of 9 random points in Z show that the space learned has smooth transitions, with every image in the space plausibly looking like a bedroom. In the 6th row, you see a room without a window slowly transforming into a room with a giant window. In the 10th row, you see what appears to be a TV slowly being transformed into a window.

scene classification learn object detectors (Oquab et al., 2014). We demonstrate that an unsupervised DCGAN trained on a large image dataset can also learn a hierarchy of features that are interesting. Using guided backpropagation as proposed by (Springenberg et al., 2014), we show in Fig.5 that the features learnt by the discriminator activate on typical parts of a bedroom, like beds and windows. For comparison, in the same figure, we give a baseline for randomly initialized features that are not activated on anything that is semantically relevant or interesting.

6.3 MANIPULATING THE GENERATOR REPRESENTATION

6.3.1 FORGETTING TO DRAW CERTAIN OBJECTS

In addition to the representations learnt by a discriminator, there is the question of what representations the generator learns. The quality of samples suggest that the generator learns specific object representations for major scene components such as beds, windows, lamps, doors, and miscellaneous furniture. In order to explore the form that these representations take, we conducted an experiment to attempt to remove windows from the generator completely.



Figure 5: On the right, guided backpropagation visualizations of maximal axis-aligned responses for the first 6 learned convolutional features from the last convolution layer in the discriminator. Notice a significant minority of features respond to beds - the central object in the LSUN bedrooms dataset. On the left is a random filter baseline. Comparing to the previous responses there is little to no discrimination and random structure.



Figure 6: Top row: un-modified samples from model. Bottom row: the same samples generated with dropping out "window" filters. Some windows are removed, others are transformed into objects with similar visual appearance such as doors and mirrors. Although visual quality decreased, overall scene composition stayed similar, suggesting the generator has done a good job disentangling scene representation from object representation. Extended experiments could be done to remove other objects from the image and modify the objects the generator draws.

On 150 samples, 52 window bounding boxes were drawn manually. On the second highest convolution layer features, logistic regression was fit to predict whether a feature activation was on a window (or not), by using the criterion that activations inside the drawn bounding boxes are positives and random samples from the same images are negatives. Using this simple model, all feature maps with weights greater than zero (200 in total) were dropped from all spatial locations. Then, random new samples were generated with and without the feature map removal.

The generated images with and without the window dropout are shown in Fig.6, and interestingly, the network mostly forgets to draw windows in the bedrooms, replacing them with other objects.

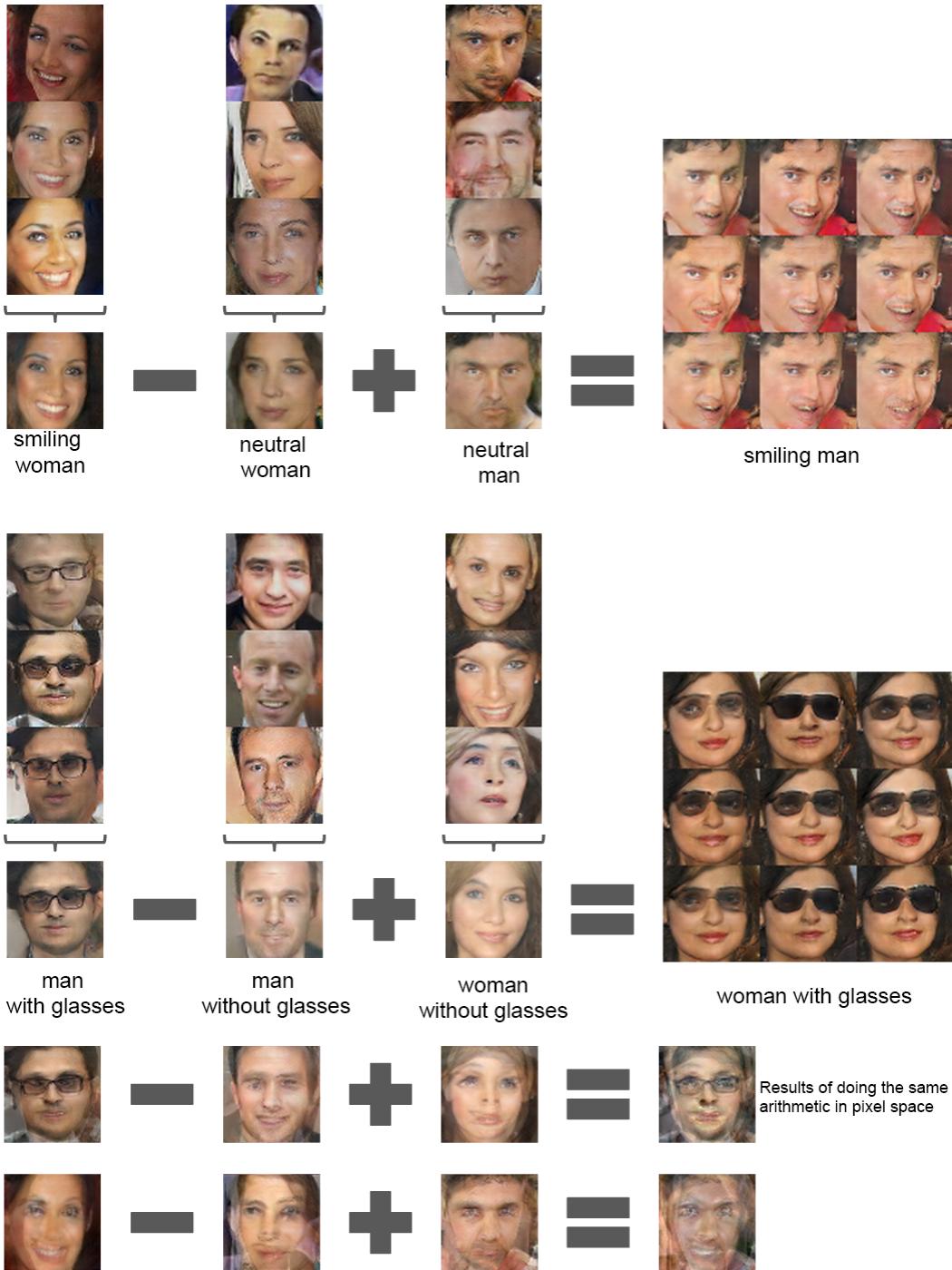


Figure 7: Vector arithmetic for visual concepts. For each column, the Z vectors of samples are averaged. Arithmetic was then performed on the mean vectors creating a new vector Y . The center sample on the right hand side is produced by feeding Y as input to the generator. To demonstrate the interpolation capabilities of the generator, uniform noise sampled with scale ± 0.25 was added to Y to produce the 8 other samples. Applying arithmetic in the input space (bottom two examples) results in noisy overlap due to misalignment.



Figure 8: A “turn” vector was created from four averaged samples of faces looking left vs looking right. By adding interpolations along this axis to random samples we were able to reliably transform their pose.

6.3.2 VECTOR ARITHMETIC ON FACE SAMPLES

In the context of evaluating learned representations of words (Mikolov et al., 2013) demonstrated that simple arithmetic operations revealed rich linear structure in representation space. One canonical example demonstrated that the vector(“King”) - vector(“Man”) + vector(“Woman”) resulted in a vector whose nearest neighbor was the vector for Queen. We investigated whether similar structure emerges in the Z representation of our generators. We performed similar arithmetic on the Z vectors of sets of exemplar samples for visual concepts. Experiments working on only single samples per concept were unstable, but averaging the Z vector for three exemplars showed consistent and stable generations that semantically obeyed the arithmetic. In addition to the object manipulation shown in (Fig. 7), we demonstrate that face pose is also modeled linearly in Z space (Fig. 8).

These demonstrations suggest interesting applications can be developed using Z representations learned by our models. It has been previously demonstrated that conditional generative models can learn to convincingly model object attributes like scale, rotation, and position (Dosovitskiy et al., 2014). This is to our knowledge the first demonstration of this occurring in purely unsupervised models. Further exploring and developing the above mentioned vector arithmetic could dramatically reduce the amount of data needed for conditional generative modeling of complex image distributions.

7 CONCLUSION AND FUTURE WORK

We propose a more stable set of architectures for training generative adversarial networks and we give evidence that adversarial networks learn good representations of images for supervised learning and generative modeling. There are still some forms of model instability remaining - we noticed as models are trained longer they sometimes collapse a subset of filters to a single oscillating mode. Further work is needed to tackle this from of instability. We think that extending this framework to other domains such as video (for frame prediction) and audio (pre-trained features for speech synthesis) should be very interesting. Further investigations into the properties of the learnt latent space would be interesting as well.

ACKNOWLEDGMENTS

We are fortunate and thankful for all the advice and guidance we have received during this work, especially that of Ian Goodfellow, Tobias Springenberg, Arthur Szlam and Durk Kingma. Additionally we'd like to thank all of the folks at indico for providing support, resources, and conversations, especially the two other members of the indico research team, Dan Kuster and Nathan Lintz. Finally, we'd like to thank Nvidia for donating a Titan-X GPU used in this work.

REFERENCES

- Coates, Adam and Ng, Andrew. Selecting receptive fields in deep networks. *NIPS*, 2011.
- Coates, Adam and Ng, Andrew Y. Learning feature representations with k-means. In *Neural Networks: Tricks of the Trade*, pp. 561–580. Springer, 2012.
- Deng, Jia, Dong, Wei, Socher, Richard, Li, Li-Jia, Li, Kai, and Fei-Fei, Li. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 248–255. IEEE, 2009.
- Denton, Emily, Chintala, Soumith, Szlam, Arthur, and Fergus, Rob. Deep generative image models using a laplacian pyramid of adversarial networks. *arXiv preprint arXiv:1506.05751*, 2015.
- Dosovitskiy, Alexey, Springenberg, Jost Tobias, and Brox, Thomas. Learning to generate chairs with convolutional neural networks. *arXiv preprint arXiv:1411.5928*, 2014.
- Dosovitskiy, Alexey, Fischer, Philipp, Springenberg, Jost Tobias, Riedmiller, Martin, and Brox, Thomas. Discriminative unsupervised feature learning with exemplar convolutional neural networks. In *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, volume 99. IEEE, 2015.
- Efros, Alexei, Leung, Thomas K, et al. Texture synthesis by non-parametric sampling. In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, volume 2, pp. 1033–1038. IEEE, 1999.
- Freeman, William T, Jones, Thouis R, and Pasztor, Egon C. Example-based super-resolution. *Computer Graphics and Applications, IEEE*, 22(2):56–65, 2002.
- Goodfellow, Ian J, Warde-Farley, David, Mirza, Mehdi, Courville, Aaron, and Bengio, Yoshua. Maxout networks. *arXiv preprint arXiv:1302.4389*, 2013.
- Goodfellow, Ian J., Pouget-Abadie, Jean, Mirza, Mehdi, Xu, Bing, Warde-Farley, David, Ozair, Sherjil, Courville, Aaron C., and Bengio, Yoshua. Generative adversarial nets. *NIPS*, 2014.
- Gregor, Karol, Danihelka, Ivo, Graves, Alex, and Wierstra, Daan. Draw: A recurrent neural network for image generation. *arXiv preprint arXiv:1502.04623*, 2015.
- Hardt, Moritz, Recht, Benjamin, and Singer, Yoram. Train faster, generalize better: Stability of stochastic gradient descent. *arXiv preprint arXiv:1509.01240*, 2015.
- Hauberg, Sren, Freifeld, Oren, Larsen, Anders Boesen Lindbo, Fisher III, John W., and Hansen, Lars Kair. Dreaming more data: Class-dependent distributions over diffeomorphisms for learned data augmentation. *arXiv preprint arXiv:1510.02795*, 2015.
- Hays, James and Efros, Alexei A. Scene completion using millions of photographs. *ACM Transactions on Graphics (TOG)*, 26(3):4, 2007.
- Ioffe, Sergey and Szegedy, Christian. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.
- Kingma, Diederik P and Ba, Jimmy Lei. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- Kingma, Diederik P and Welling, Max. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

- Lee, Honglak, Grosse, Roger, Ranganath, Rajesh, and Ng, Andrew Y. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 609–616. ACM, 2009.
- Loosli, Gaëlle, Canu, Stéphane, and Bottou, Léon. Training invariant support vector machines using selective sampling. In Bottou, Léon, Chapelle, Olivier, DeCoste, Dennis, and Weston, Jason (eds.), *Large Scale Kernel Machines*, pp. 301–320. MIT Press, Cambridge, MA., 2007. URL <http://leon.bottou.org/papers/loosli-canu-bottou-2006>.
- Maas, Andrew L, Hannun, Awni Y, and Ng, Andrew Y. Rectifier nonlinearities improve neural network acoustic models. In *Proc. ICML*, volume 30, 2013.
- Mikolov, Tomas, Sutskever, Ilya, Chen, Kai, Corrado, Greg S, and Dean, Jeff. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pp. 3111–3119, 2013.
- Mordvintsev, Alexander, Olah, Christopher, and Tyka, Mike. Inceptionism : Going deeper into neural networks. <http://googleresearch.blogspot.com/2015/06/inceptionism-going-deeper-into-neural.html>. Accessed: 2015-06-17.
- Nair, Vinod and Hinton, Geoffrey E. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pp. 807–814, 2010.
- Oquab, M., Bottou, L., Laptev, I., and Sivic, J. Learning and transferring mid-level image representations using convolutional neural networks. In *CVPR*, 2014.
- Portilla, Javier and Simoncelli, Eero P. A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision*, 40(1):49–70, 2000.
- Rasmus, Antti, Valpola, Harri, Honkala, Mikko, Berglund, Mathias, and Raiko, Tapani. Semi-supervised learning with ladder network. *arXiv preprint arXiv:1507.02672*, 2015.
- Sohl-Dickstein, Jascha, Weiss, Eric A, Maheswaranathan, Niru, and Ganguli, Surya. Deep unsupervised learning using nonequilibrium thermodynamics. *arXiv preprint arXiv:1503.03585*, 2015.
- Springenberg, Jost Tobias, Dosovitskiy, Alexey, Brox, Thomas, and Riedmiller, Martin. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.
- Srivastava, Rupesh Kumar, Masci, Jonathan, Gomez, Faustino, and Schmidhuber, Jürgen. Understanding locally competitive networks. *arXiv preprint arXiv:1410.1165*, 2014.
- Theis, L., van den Oord, A., and Bethge, M. A note on the evaluation of generative models. *arXiv:1511.01844*, Nov 2015. URL <http://arxiv.org/abs/1511.01844>.
- Vincent, Pascal, Larochelle, Hugo, Lajoie, Isabelle, Bengio, Yoshua, and Manzagol, Pierre-Antoine. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *The Journal of Machine Learning Research*, 11:3371–3408, 2010.
- Xu, Bing, Wang, Naiyan, Chen, Tianqi, and Li, Mu. Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853*, 2015.
- Yu, Fisher, Zhang, Yinda, Song, Shuran, Seff, Ari, and Xiao, Jianxiong. Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.
- Zeiler, Matthew D and Fergus, Rob. Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014*, pp. 818–833. Springer, 2014.
- Zhao, Junbo, Mathieu, Michael, Goroshin, Ross, and Lecun, Yann. Stacked what-where auto-encoders. *arXiv preprint arXiv:1506.02351*, 2015.

8 SUPPLEMENTARY MATERIAL

8.1 EVALUATING DCGANS CAPABILITY TO CAPTURE DATA DISTRIBUTIONS

We propose to apply standard classification metrics to a conditional version of our model, evaluating the conditional distributions learned. We trained a DCGAN on MNIST (splitting off a 10K validation set) as well as a permutation invariant GAN baseline and evaluated the models using a nearest neighbor classifier comparing real data to a set of generated conditional samples. We found that removing the scale and bias parameters from batchnorm produced better results for both models. We speculate that the noise introduced by batchnorm helps the generative models to better explore and generate from the underlying data distribution. The results are shown in Table 2 which compares our models with other techniques. The DCGAN model achieves the same test error as a nearest neighbor classifier fitted on the training dataset - suggesting the DCGAN model has done a superb job at modeling the conditional distributions of this dataset. At one million samples per class, the DCGAN model outperforms InfiMNIST (Loosli et al., 2007), a hand developed data augmentation pipeline which uses translations and elastic deformations of training examples. The DCGAN is competitive with a probabilistic generative data augmentation technique utilizing learned per class transformations (Hauberg et al., 2015) while being more general as it directly models the data instead of transformations of the data.

Table 2: Nearest neighbor classification results.

Model	Test Error @50K samples	Test Error @10M samples
AlignMNIST	-	1.4%
InfiMNIST	-	2.6%
Real Data	3.1%	-
GAN	6.28%	5.65%
DCGAN (ours)	2.98%	1.48%

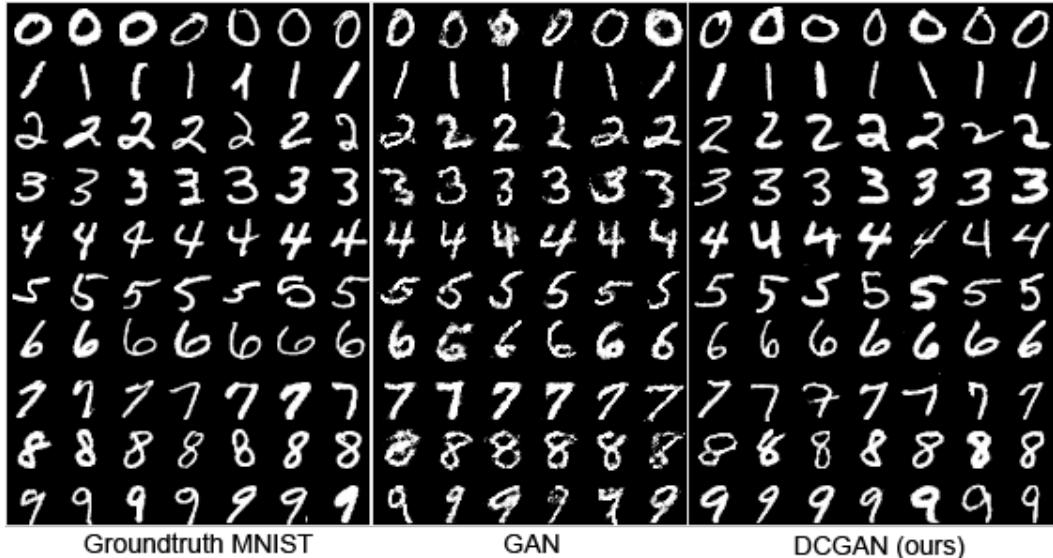


Figure 9: Side-by-side illustration of (from left-to-right) the MNIST dataset, generations from a baseline GAN, and generations from our DCGAN .



Figure 10: More face generations from our Face DCGAN.

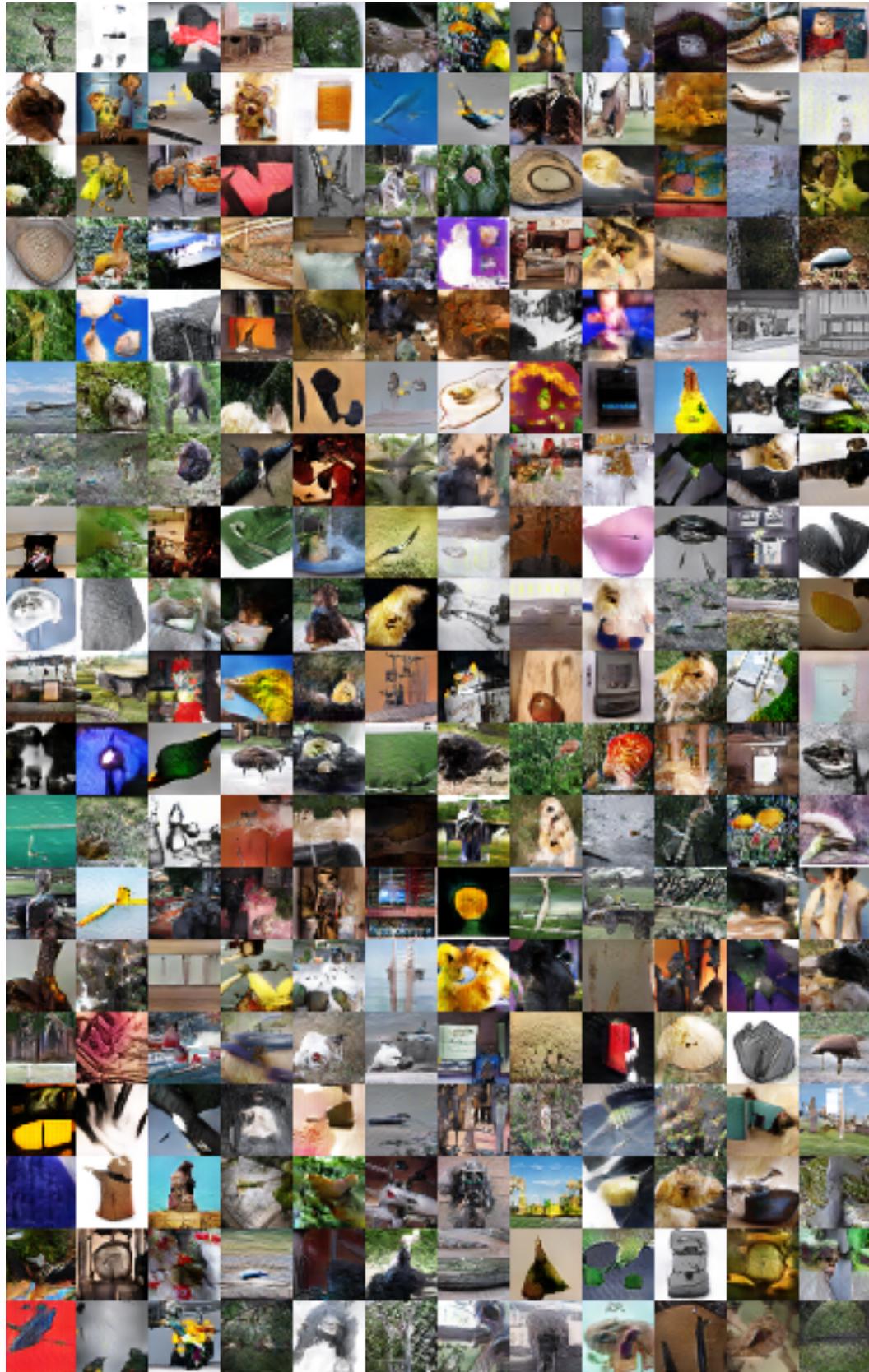


Figure 11: Generations of a DCGAN that was trained on the Imagenet-1k dataset.