

Your Local GAN: Designing Two Dimensional Local Attention Mechanisms for Generative Models (arXiv :1911.12287)

Giannis Daras
National Technical
University of Athens

Augustus Odena
Google Brain

Han Zhang
Google Brain

Alexandros G. Dimakis
UT Austin

Abstract

We introduce a new local sparse attention layer that preserves two-dimensional geometry and locality. We show that by just replacing the dense attention layer of SAGAN with our construction, we obtain very significant FID, Inception score and pure visual improvements. FID score is improved from 18.65 to 15.94 on ImageNet, keeping all other parameters the same. The sparse attention patterns that we propose for our new layer are designed using a novel information theoretic criterion that uses information flow graphs.

We also present a novel way to invert Generative Adversarial Networks with attention. Our method uses the attention layer of the discriminator to create an innovative loss function. This allows us to visualize the newly introduced attention heads and show that they indeed capture interesting aspects of two-dimensional geometry of real images.

Attention

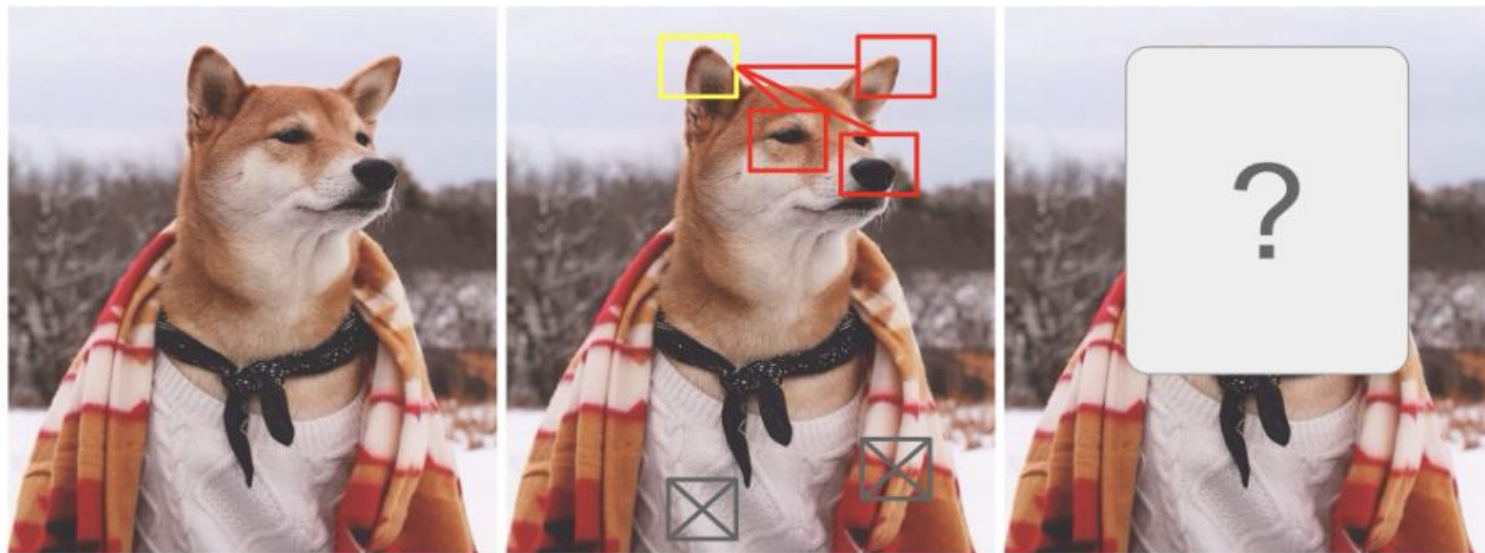


Fig. 1. A Shiba Inu in a men's outfit. The credit of the original photo goes to Instagram [@mensweardog](#).

Attention

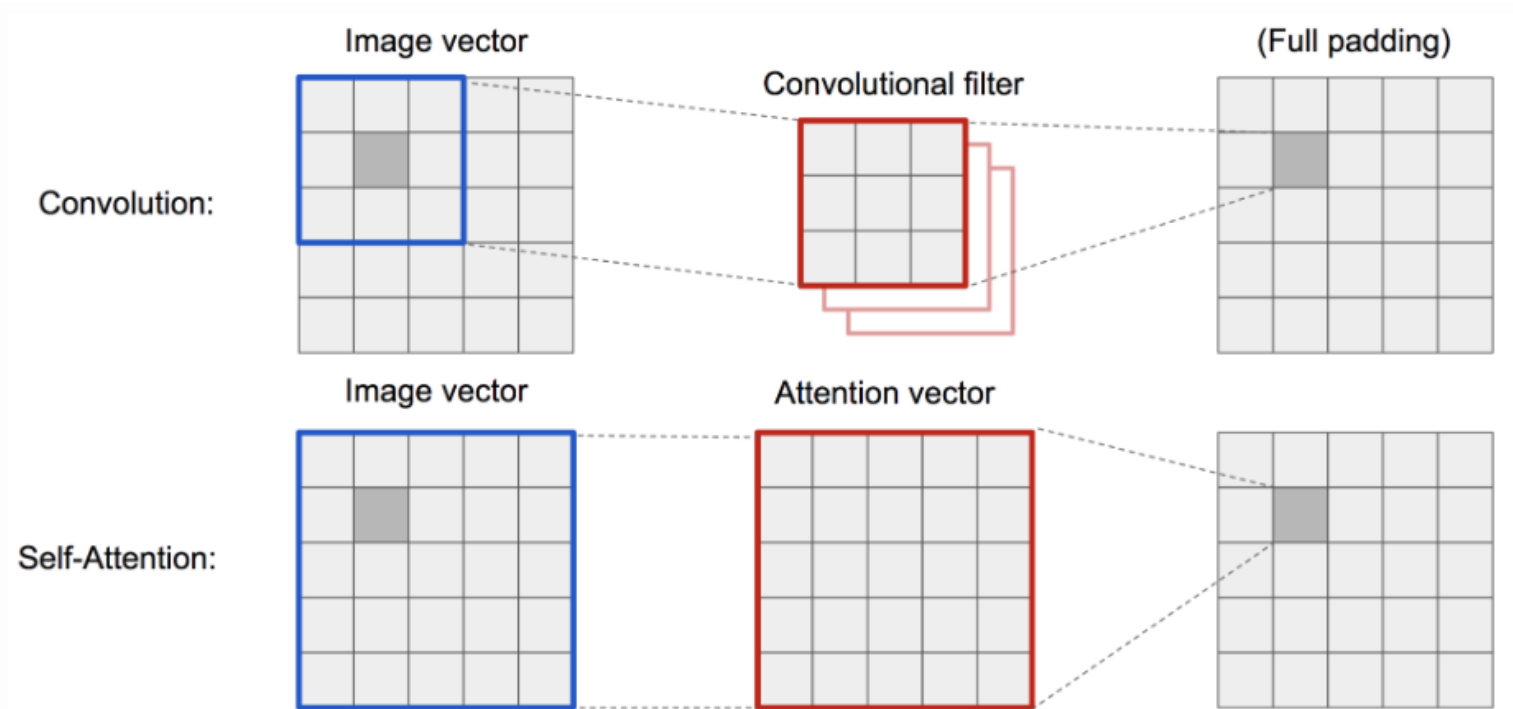


Fig. 19. Convolution operation and self-attention have access to regions of very different sizes.

Attention

Then we apply the dot-product attention to output the self-attention feature maps:

$$\alpha_{i,j} = \text{softmax}(f(\mathbf{x}_i)^\top g(\mathbf{x}_j))$$

$$\mathbf{o}_j = \sum_{i=1}^N \alpha_{i,j} h(\mathbf{x}_i)$$

Key: $f(\mathbf{x}) = \mathbf{W}_f \mathbf{x}$

Query: $g(\mathbf{x}) = \mathbf{W}_g \mathbf{x}$

Value: $h(\mathbf{x}) = \mathbf{W}_h \mathbf{x}$

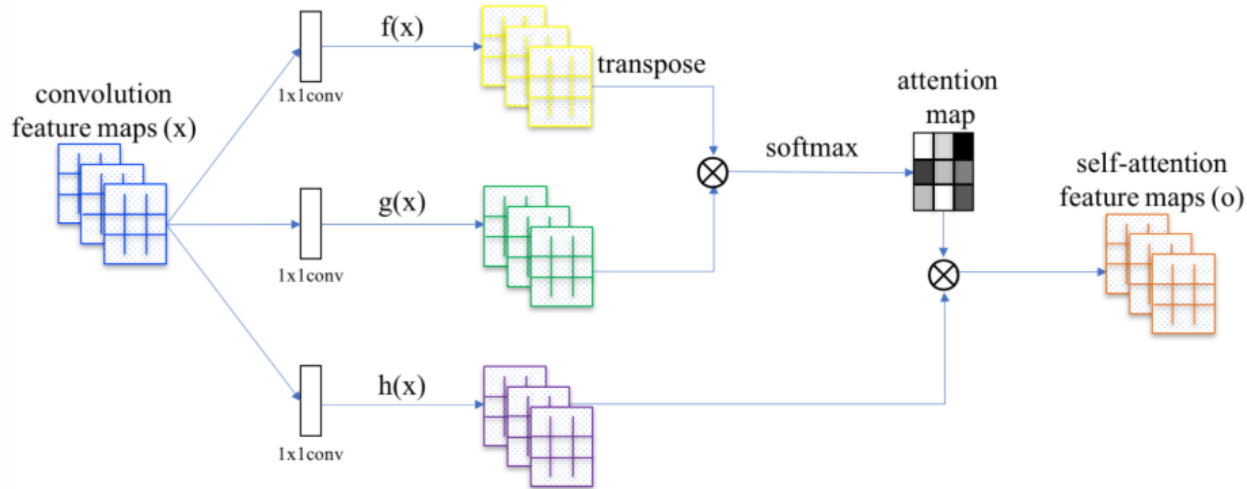


Fig. 20. The self-attention mechanism in SAGAN. (Image source: Fig. 2 in [Zhang et al., 2018](#))

Limitation

Inefficiency Memory and time complexity - $O(N^2)$

Need $N \times N$ attention matrix for every layer and attention head,

Dense attention does not benefit from locality

most dependencies in images relate to nearby neighborhoods of pixels

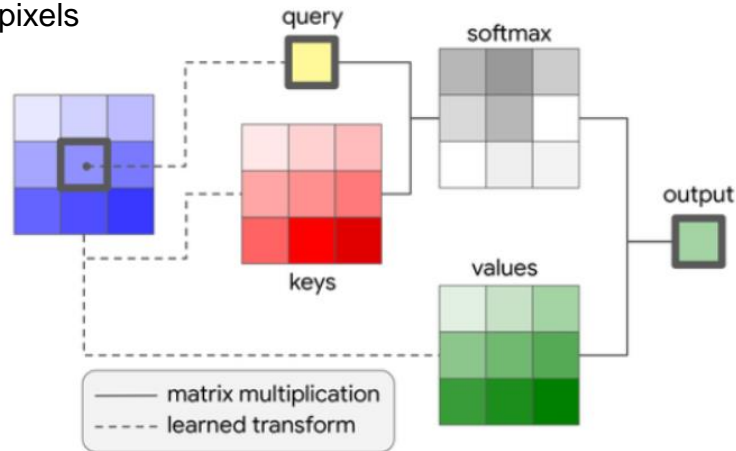


Figure 6: Local attention layer with spatial extent $k=3$ in [4].

Sparse attention

Generating Long Sequences with Sparse Transformers , OpenAI(2019)

Reduce complexity to $O(N\sqrt{N})$

Split attention in multiple steps

an attention sparsification in p steps is described by binary masks $\{M1, M2, M3, M4, \dots, Mp\}$

$$A_{X,Y} = X_Q \cdot Y_K^T, \in \mathbb{R}^{N_X \times N_Y}$$

$$A_{X,Y}^i[a, b] = \begin{cases} A_{X,Y}[a, b], & M^i[a, b] = 1 \\ -\infty, & M^i[a, b] = 0 \end{cases}$$

*‘any pair of input nodes is **connected** to any pair of output nodes with **two edge-disjoint paths**’*

Your Local GAN - *Two Dimensional Local Attention*

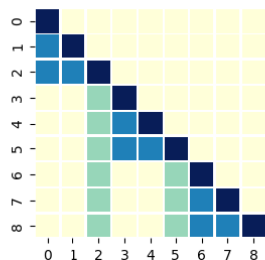
Information Flow Graphs (IFGs)

- How to distributed storage (Node) in Network systems
- Directed acyclic graphs
- Each storage node is connected by a directed edge with capacity equal to the amount of information that can be stored into that node

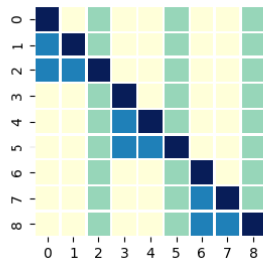
Full Information Attention Sparsification

- **every token** of every stage of an attention layer is represented by a **storage node**
- all the tokens have the same size, we can eliminate vertex splitting and compactly represent each storage node by a single vertex

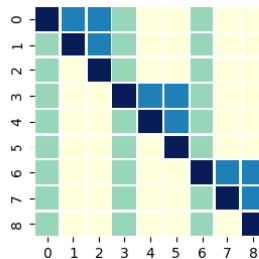
Your Local GAN



(a) Attention masks for Fixed Pattern [6].



(b) Attention masks for Left To Right (LTR) pattern.

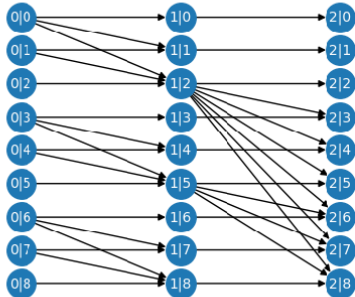


(c) Attention masks for Right To Left (RTL) pattern.

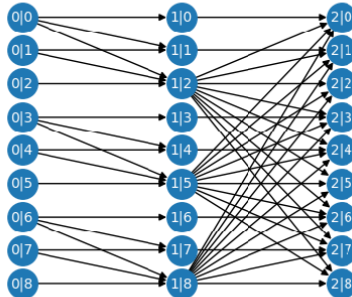
Left to Right (LTR) / Right to Left (RTL)

9x 9 masks associated Information Flow Graphs

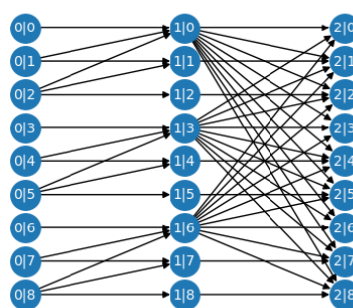
Attention only to $O(N\sqrt{N})$ positions



(d) Information Flow Graph associated with Fixed Pattern. This pattern *does not have Full Information*, i.e. there are dependencies between nodes that the attention layer cannot model. For example, there is no path from node 0 of V^0 to node 1 of V^2 .

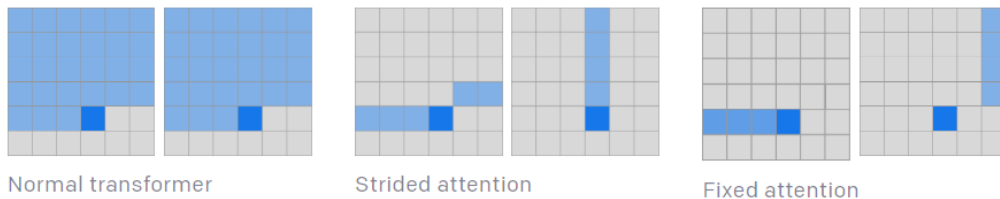


(e) Information Flow Graph associated with LTR. This pattern has **Full Information**, i.e. there is a path between any node of V^0 and any node of V^2 . Note that the number of edges is only increased by a constant compared to the Fixed Attention Pattern [6], illustrated in 2d.

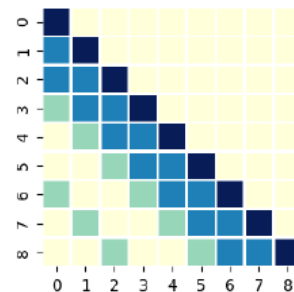


(f) Information Flow Graph associated with RTL. This pattern also has **Full Information**. RTL is a "transposed" version of LTR, so that local context at the right of each node is attended at the first step.

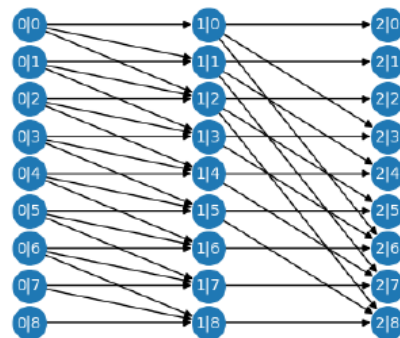
Your Local GAN



The first version, *strided* attention, is roughly equivalent to each position attending to its row and its column, and is similar to the attention pattern learned by the network above. (Note that the column attention can be equivalently formulated as attending to the row of the transposed matrix). The second version, *fixed* attention, attends to a fixed column and the elements after the latest column element, a pattern we found useful for when the data didn't fit into a two-dimensional structure (like text). For more details, we refer readers to our paper.



(a) Attention masks for Strided Pattern [6].



(c) Information Flow Graph associated with Strided Pattern. This pattern *does not have Full Information*, i.e. there are dependencies between nodes that the attention layer cannot model. For example, there is no path from node 2 of V^0 to node 1 of V^2 .

ESA (Enumerate, Shift, Apply)

Two-Dimensional Locality - ESA (Enumerate, Shift, Apply)

- Make 1-D sparsifications to become aware of 2-D locality
- **enumerate** pixels of the image based on their Manhattan distance from the pixel at location (0, 0)
- **Shift** the indices of any given one-dimensional sparsification to match the **Manhattan distance** enumeration instead of the reshape enumeration
- **apply** this new one dimensional sparsification pattern, that respects two-dimensional locality, to the one-dimensional reshaped version of the image

ESA (Enumerate, Shift, Apply)

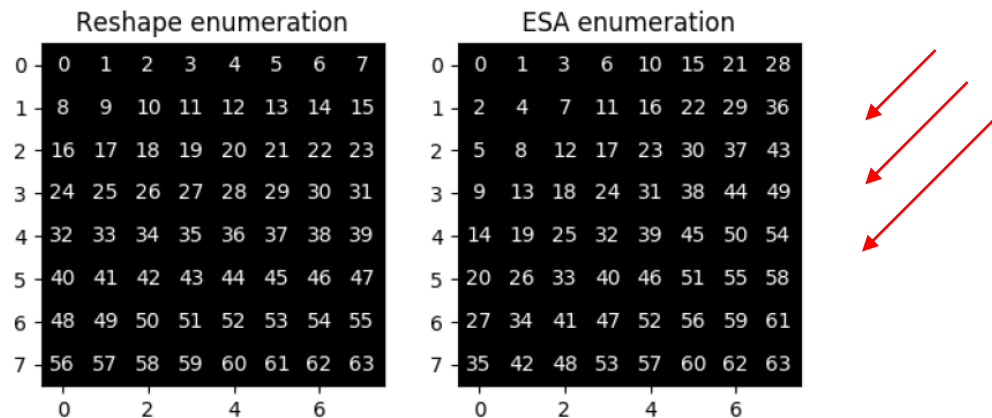


Figure 3: Reshape and ESA enumerations of the cells of an image grid that show how image grid is projected into a line. (Left) Enumeration of pixels of an 8×8 image using a standard reshape. This projection maintains locality only in rows. (Right) Enumeration of pixels of an 8×8 image, using the ESA framework. We use the Manhattan distance from the start $(0,0)$ as a criterion for enumeration. Although there is some distortion due to the projection into 1-D, locality is mostly maintained.

Inverting Generative Models with Attention

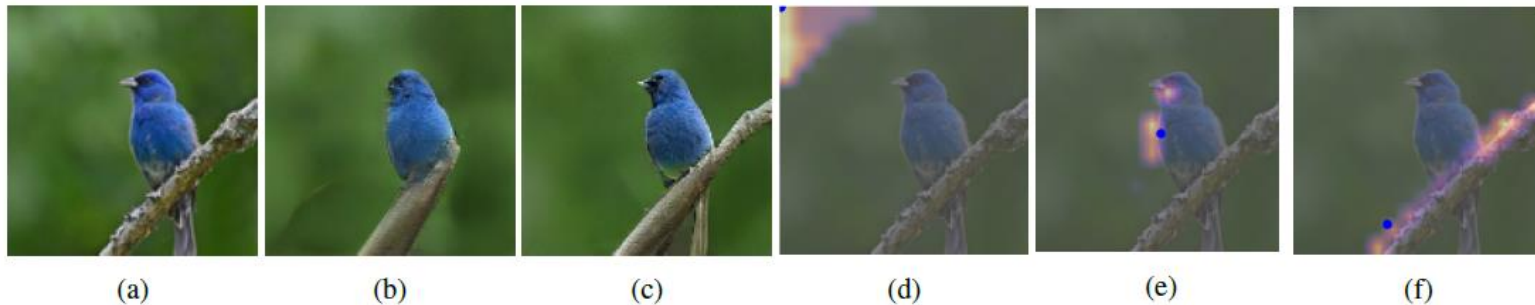


Figure 6: Inverted image of an indigo bird and visualization of the attention maps for specific query points. (a) The original image. Again, this was obtained with a Google image search and was not in the training set. (b) Shows how previous inversion methods fail to reconstruct the head of the bird and the branch. (c) A successful inversion using our method. (d) Specifically, [6d](#) shows how attention uses our ESA trick to model background, homogeneous areas. (e) Attention applied to the bird. (f) Attention applied with a query on the branch. Notice how attention is non-local and captures the full branch.

Inverting Generative Models with Attention

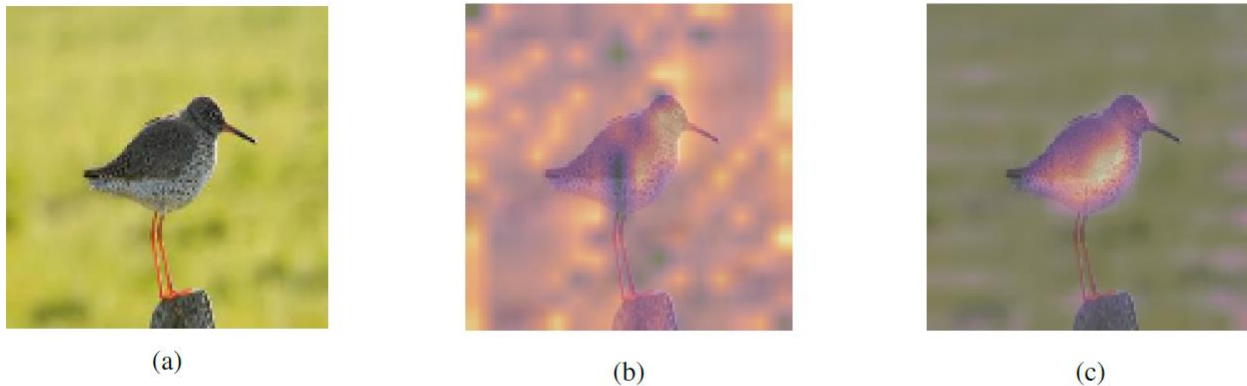


Figure 8: (a) Real image of a redshank. (b) Saliency map extracted from **all** heads of the Discriminator. (c) Saliency map extracted from a **single** head of the Discriminator. Weighting our loss function with (b) does not have a huge impact, as the attention weights are almost uniform. Saliency map from (c) is more likely to help correct inversion of the bird. We can use saliency maps from other heads to invert the background as well.

Experimental Validation

- Change only the attention layer of SAGAN
- Trained all models for up to 1,500,000 steps on individual
- Cloud TPU v3 devices (v3-8)

Experimental Validation

	# Heads	FID	Inception
SAGAN	1	18.65	52.52
SAGAN	8	20.09	46.01
YLG-SAGAN	8	15.94	57.22
YLG - No ESA	8	17.47	51.09
YLG - Strided	8	16.64	55.21

Table 1: ImageNet Results: Table of results after training SAGAN and YLG-SAGAN on ImageNet. Table also includes Ablation Studies (SAGAN 8 heads, YLG - No ESA, YLG - Strided). Our best model, **YLG**, achieves **15.94** FID and **57.22** Inception score. Our scores correspond to **14.53%** and **8.95%** improvement to FID and Inception respectively. We emphasize that these benefits are obtained by only one layer change to SAGAN, replacing dense attention with the local sparse attention layer that we introduce.

Experimental Validation

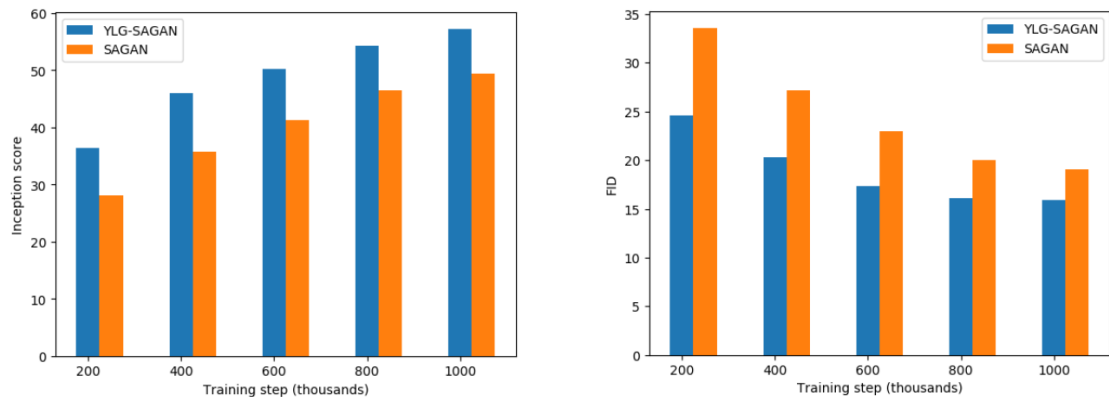
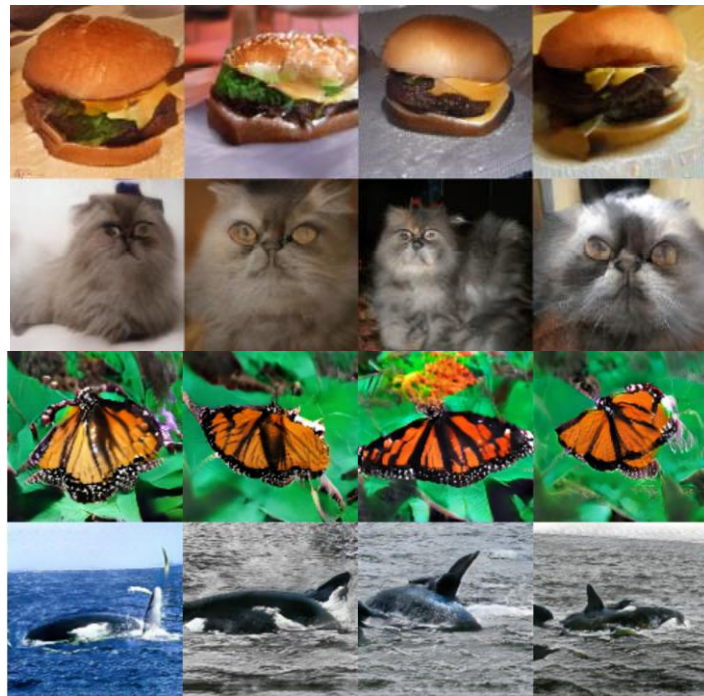


Figure 4: Training comparison for YLG-SAGAN and SAGAN. We plot every 200k steps the Inception score (a) and the FID (b) of both YLG-SAGAN and SAGAN, up to 1M training steps on ImageNet. As it can be seen, YLG-SAGAN converges much faster compared to the baseline. Specifically, we obtain our best FID at step 865k, while SAGAN requires over 1.3M steps to reach its FID performance peak. Comparing peak performance for both models, we obtain an improvement from 18.65 to **15.94** FID, by only changing the attention layer.

Experimental Validation

Generated images from YLG SAGAN divided by ImageNet category



Conclusion

- introduce a new local sparse attention layer preserves two-dimensional image locality and can support good information flow through attention steps
- To visualize our attention maps on natural images
- Achieve 14.53% improvement FID score of SAGAN and 8.95% improvement in Inception score

Question ?