IMAGE TRANSFORMER

Niki Parmar*, Ashish Vaswani*, Jakob Uszkoreit,
Noam Shazeer, Lukasz Kaiser, Alex Ku, Dustin Tran

Google AI

Probabilistic Image Generation

Model the joint distribution of pixels, factor into product of conditional distributions

Assigning probabilities allows measuring generalization

PixelRNNs, PixelCNN, PixelCNN++ are current state of the art,

RNNs are slow, require complex conditioning for best quality

CNNs are parallelizable but incorporate gating to match RNNs in quality

However …

Modeling long-range dependencies with CNNs requires either

Many layers likely making training harder

Large kernels at large parameter/computational cost

Self Attention

Self-attention worked great for language

'Sparsely parameterized', allowing larger receptive field

Trivial to parallelize per layer

However …

Quadratic complexity in structure size prohibitive for images

$O(n^2d)$ becomes expensive because n >> d

Loss Functions

Cross Entropy

Discretized Mixture of Logistics

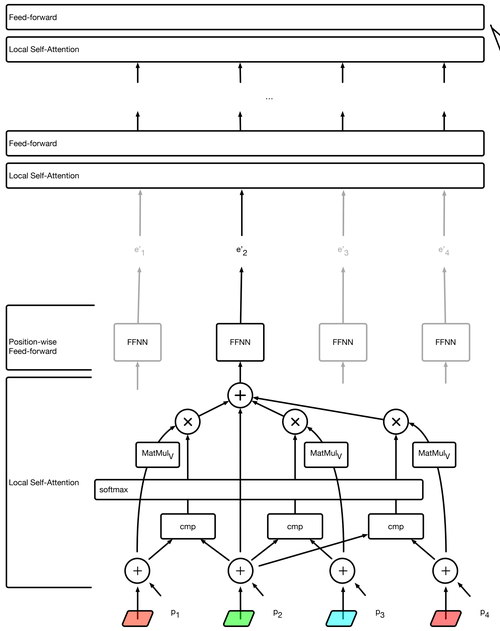
Position Embeddings

Sines and cosines

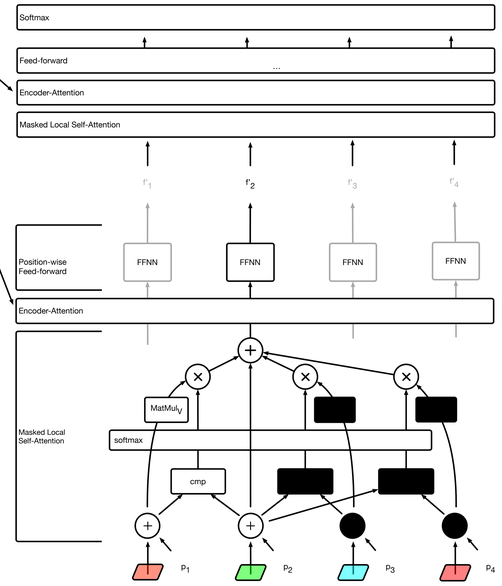
Learned

* Equal Contribution and random ordering

ENCODER



DECODER



Local Self-Attention

Constant 'path length' between pixels in receptive field

Gating/multiplication enables crisp error propagation

Local Attention

Restrict the attention windows to be local neighborhoods

Good assumption for images because of spatial locality

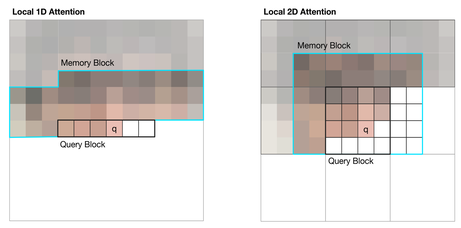


Image partitioned as non-overlapping query blocks and overlapping memory blocks. 1D local attention factorizes a sequence and 2D local attention balances number of pixels next to and above the query block.

CONDITIONAL IMAGE GENERATION

	Loss Type	NLL CIFAR-10 (Test)	NLL ImageNet (Validation)
Pixel CNN			
Row Pixel RNN		3.00	3.86
Gated Pixel CNN		3.03	3.83
Pixel CNN++		2.92	-
PixelSNAIL		**2.85**	3.8
Image Transformer, 1D local	cat	2.9	**3.78**
Image Transformer, 1D local	dmol	2.9	3.79

Negative log likelihood of various models on CIFAR-10 and ImageNet Datasets. Image Transformer performs significantly better on ImageNet.

SUPER-RESOLUTION HUMAN EVAL

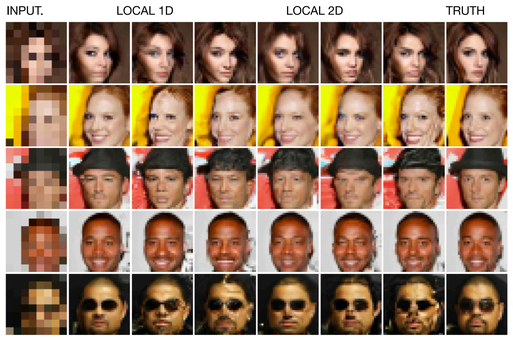
Model	% Fooled			
	Γ =n/a	Γ = 1.0	Γ = 0.9	Γ = 0.8
ResNet	4.0			
srez GAN	8.5			
Pixel Recursive CNN	-	11.0	10.4	10.25
Image Transformer, 1D local	-	**35.94 ± 3.0**	33.5 ± 3.5	29.6 ± 4.0
Image Transformer, 2D local	-	**36.11 ± 3.0**	34 ± 3.5	30.64 ± 4.0

Human Eval performance for the Image Transformer on CelebA. The fraction of humans fooled is significantly better than the previous state of art.

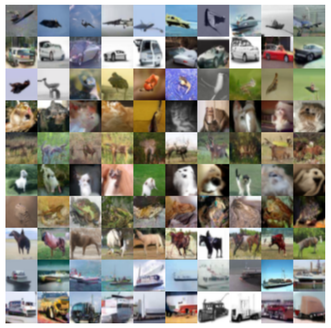
Code Link
https://github.com/tensorflow/tensor2tensor

Scan me

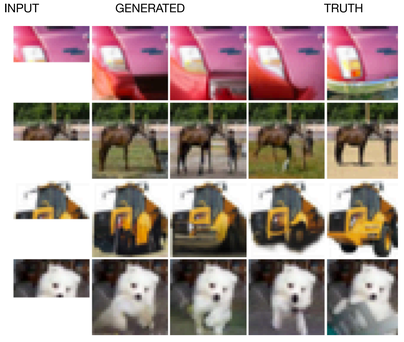
SUPER-RESOLUTION ON CELEBA



CIFAR-10 SAMPLES



CIFAR-10 COMPLETION



SUPER-RESOLUTION ON CIFAR-10