WSAM: Visual Explanations from Style Augmentation as Adversarial Attacker

Felipe Moreno-Vera, Edgar Medina, and Jorge Poco

FAPERJ
Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro
FAPESP
FUNDAÇÃO DE AMPARO À PESQUISA DO ESTADO DE SÃO PAULO
FGV
FUNDAÇÃO GETULIO VARGAS
EMAp
ESCOLA DE MATEMÁTICA APLICADA
V-DS

Content

- Motivation

- Methodology
  - Styler Network
  - Style Activation Map
  - Weighted Style Activation Map

- Experiments
  - Train-test with Style Augmentation
  - SAM
  - WSAM
  - WSAM variance

- Conclusions
  - Main Contributions

Motivation

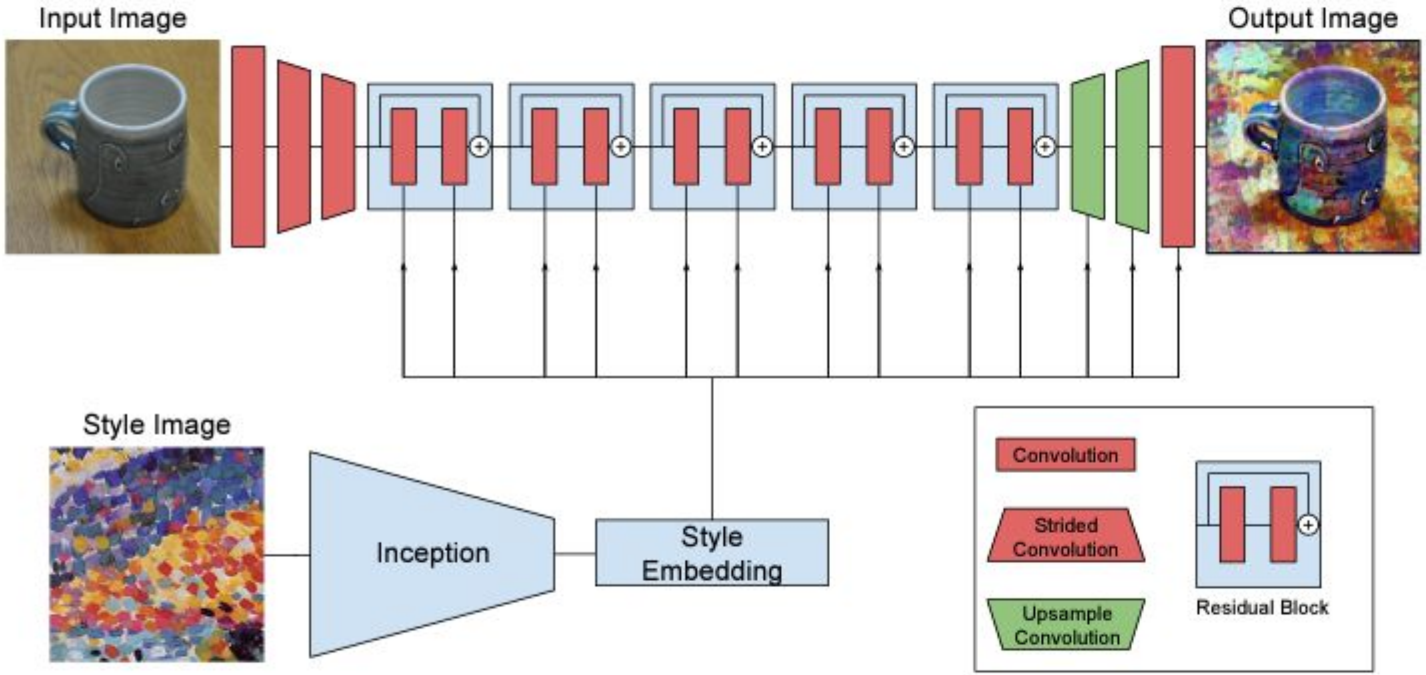
Data Augmentation: Style Augmentation

- Traditional methods on images are cutout, flip, crop, rotation, scaling, etc.

- ~80K styles
- Different values of intensity
- VGG-based CNN styler
- Shape is preserved but the style, including texture, color and contrast are randomized

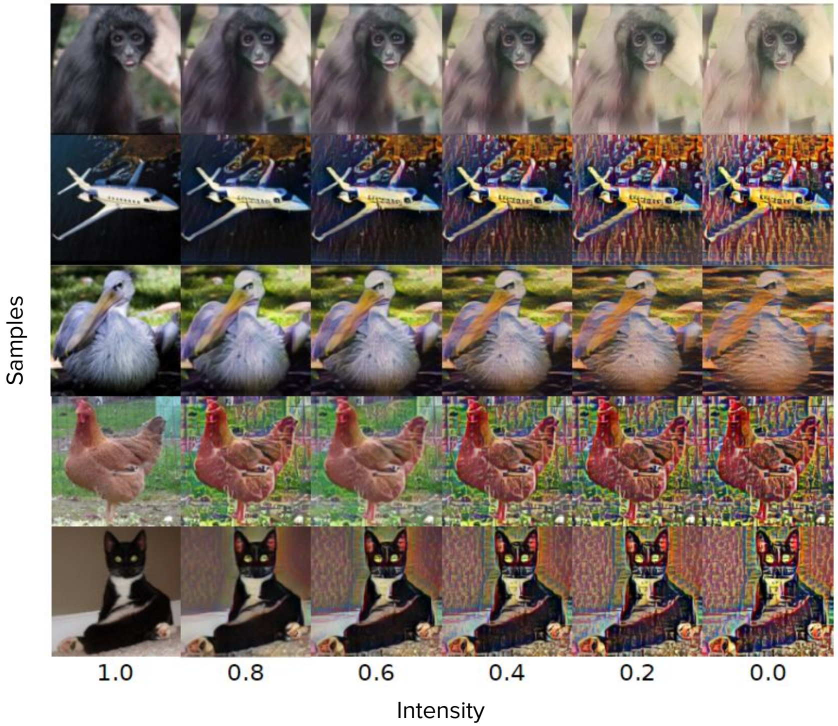
* Style Augmentation: Data Augmentation via Style Randomization, Jackson et al., 2019

Style Augmentation: by Style transfer

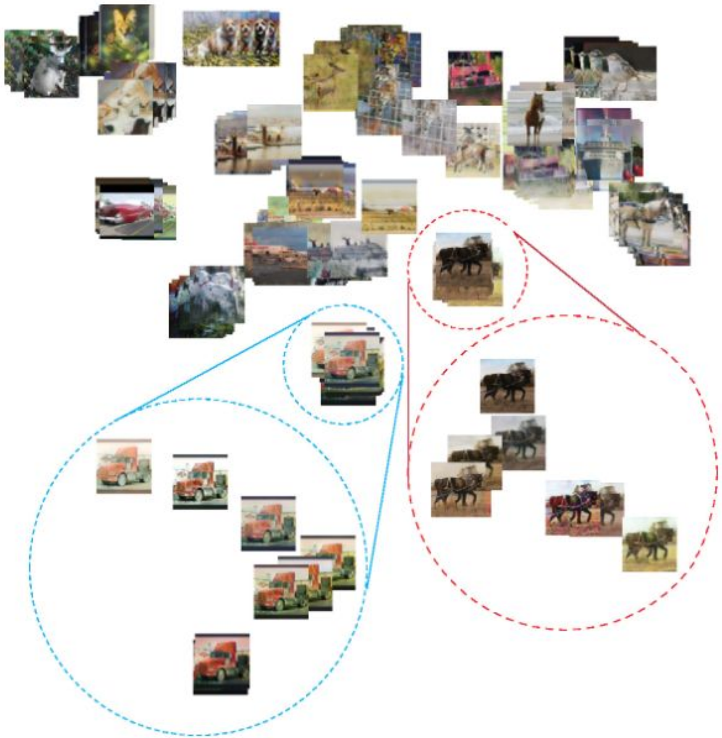


* Style Augmentation: Data Augmentation via Style Randomization, Jackson et al., 2019

Style Augmentation



The less intensity, the more style



Styled images visualization using T-sne

Limitations

- Not all styles provide good information.
- CNN-based Networks fall into adversarial attack tests.
- Some styles can distort input samples.

Proposal

- When Style Augmentation performs well as a data augmentation technique?
- Can we explain the impact of the stylization technique?
- Can we measure the impact of the stylization technique?

Methodology

Styler Network

$$\phi_c = \phi_1(VGG(\bar{c}_j))$$
$$\phi_s = \phi_2(VGG(\bar{s}_i))$$

* $T = \phi_c \phi_c^T \alpha \phi_s \phi_s^T$
* $o_i = T c_i$

we add some noise

$$\hat{z}_i \sim \bar{z}_i + \mathcal{N}(\mu_i, \sigma_i^2)$$

$$T = \phi_c \phi_c^T (\alpha \phi_c \phi_c^T + (1-\alpha)\hat{z}_i)$$
$$o_i = U(T\ C(c_j)) + (\alpha)\mu_{c_i} + (1-\alpha)\mu_{z_i}$$

- $\bar{s}_i$: zero-mean vectors styles feature map
- $\bar{c}_j$: zero-mean vectors of image sample
- o_i : styled image output
- ϕ_1 : image to vector
- ϕ_2 : Styler Net
- α : hyper-parameter of style

- $\bar{z}_i$: precomputed **s**tyle vectors
- $U(.)$: Linear Encoder Net**
- $C(.)$: Linear Decoder Net**
- μ_{c_i} : mean of images vectors
- μ_{z_i} : mean of embedded vectors

*Style Augmentation: Data Augmentation via Style Randomization, Jackson et al., 2019.

**Learning Linear Transformations for Fast Image and Video Style Transfer, Li et al., 2020.

Style Activation Maps (SAM)

$$\overset{*}{\delta_k^c} = \overbrace{\frac{1}{Z}\sum_i \sum_j}^{\text{GAP}} \underbrace{\frac{\partial y^c}{\partial A_{ij}^k}}_{\text{grad-backprop}}$$

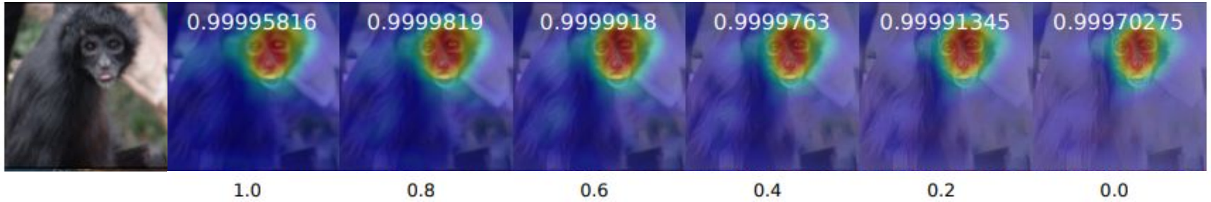
- δ :: Neuron Importance
- $A_{i,j}$: feature map
- **Strong against adversarial attacks

We multiply by styled feature maps

We take ReLU to keep only **positive influence**

$$SAM_{\alpha,\sigma}^c = ReLU\left(\sum_k \delta_k^c A_{\alpha,\sigma}^k\right)$$

- σ : style
- α : intensity
- k: k-th feature activation map
- c: class
- $A_{\alpha,\sigma}$: feature styled map



* Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization, Selvaraju et al., 2016

** Sanity Checks for Saliency Maps, Adebayo et al., 2021

Weight Style Activation Maps (WSAM)

$$WSAM^c = \frac{1}{\Omega}\sum_{\alpha}\sum_{\sigma} y^c_{\alpha,\sigma} \times SAM^c_{\alpha,\sigma}$$

- Ω : styles x intensities
- y: prediction probability
- SAM: Styled Activation Map

We saw the regions where all styles combined generate a high variance in sample

WSAM: Variance

$$WSAM^c_{variance} = \frac{1}{Z \times m} \sum_i^m (WSAM^c_i - y^c_i \times I^c_i)^2$$

- i: i-th sample
- m: number of samples
- c: class
- Z: image size (width x height)
- I^c_i: image with no style (alpha=1.0)
- $WSAM^c_i$: WSAM calculate for that sample i.
- y^c_i : probability of prediction

We calculate the total variance of all styles and intensities for all samples per each category.

Results

Evaluations: Performed on STL-10

Comparison between applying style augmentation to previous works

Network	Extra	Trad	Style	Acc
SWWAE	✓	✓		74.33
Exempla Conv	✓	✓		75.40
IIC	✓	✓		**88.80**
Baseline		✓		75.67
Ensemble		✓		77.62
STADA*		✓	✓	75.31
InceptionV3-299*		✓	✓	**80.80**
Xception-96*		✓	✓	**82.67**
Xception-128*		✓	✓	**85.11**

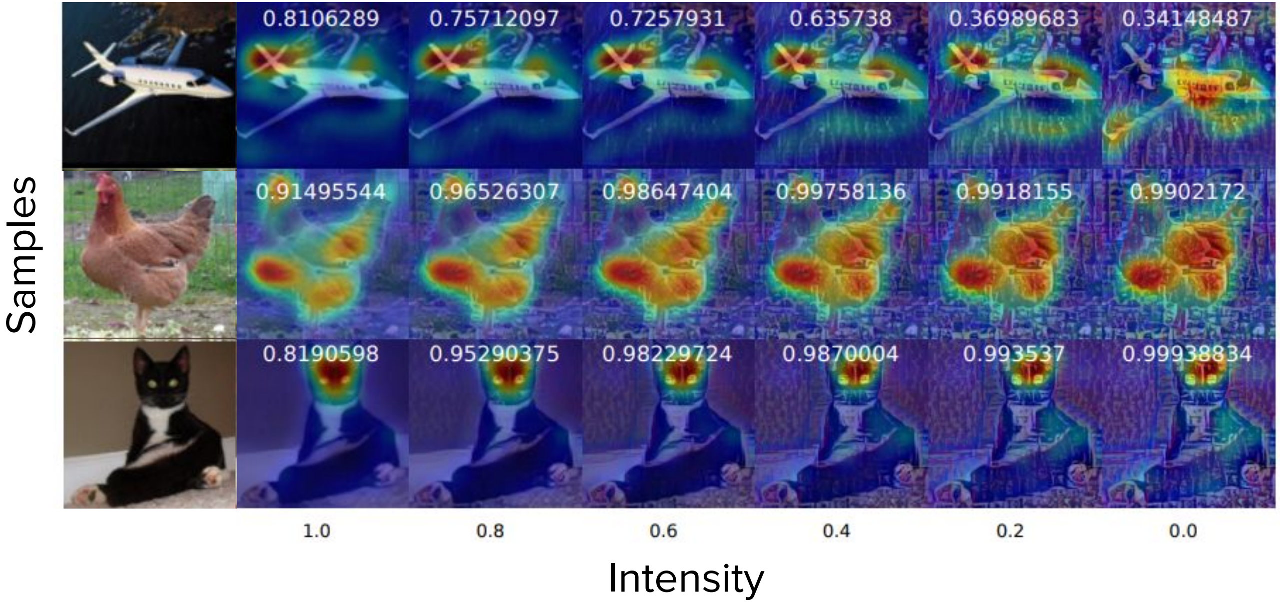
"Extra" column means additional data used in training.

WideResNet-101 has a 32x32 input shape.

Comparison between traditional and style augmentation with different input shapes

Network	Extra	Trad	Style	Acc
Xception-256*				73.37
		✓		86.19
			✓	74.89
		✓	✓	**86.85**
InceptionV4-299*				79.17
		✓		86.49
			✓	80.52
		✓	✓	**88.18**
WideResNet-96* (WRN)				77.28
		✓		87.26
			✓	83.58
		✓	✓	**88.83**
WideResNet-101* (WRN)				87.83
		✓		88.23
			✓	92.23
		✓	✓	**94.67**

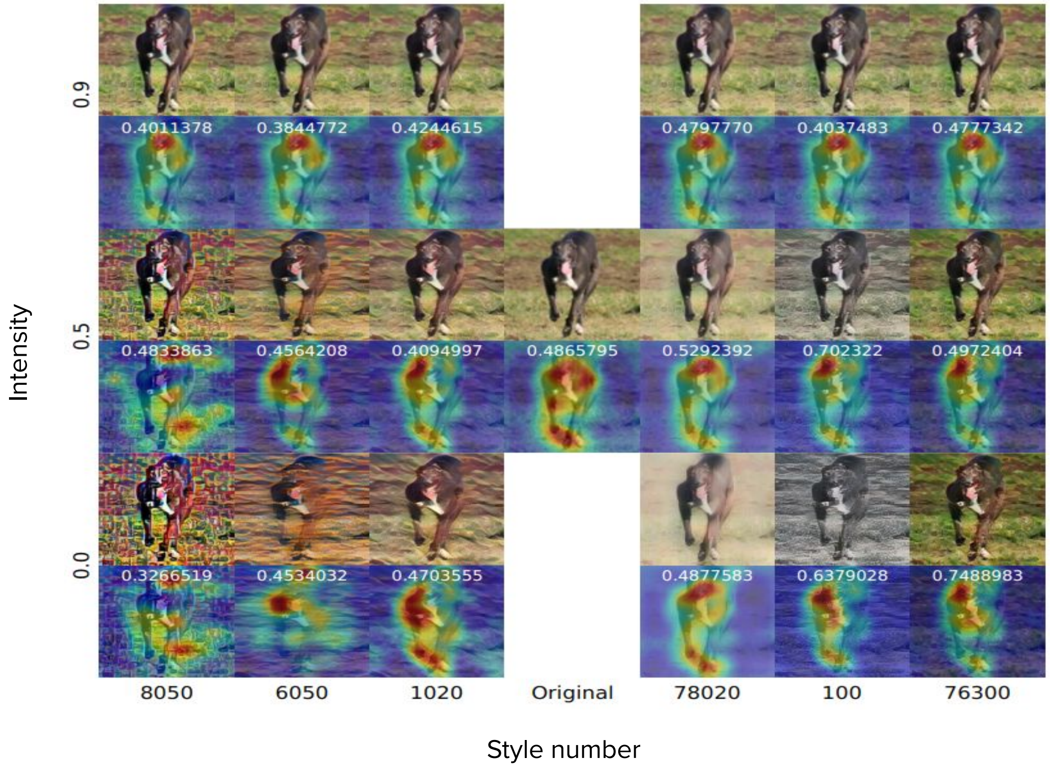
SAM: Impact of stylizing



The same style is applied to different samples.

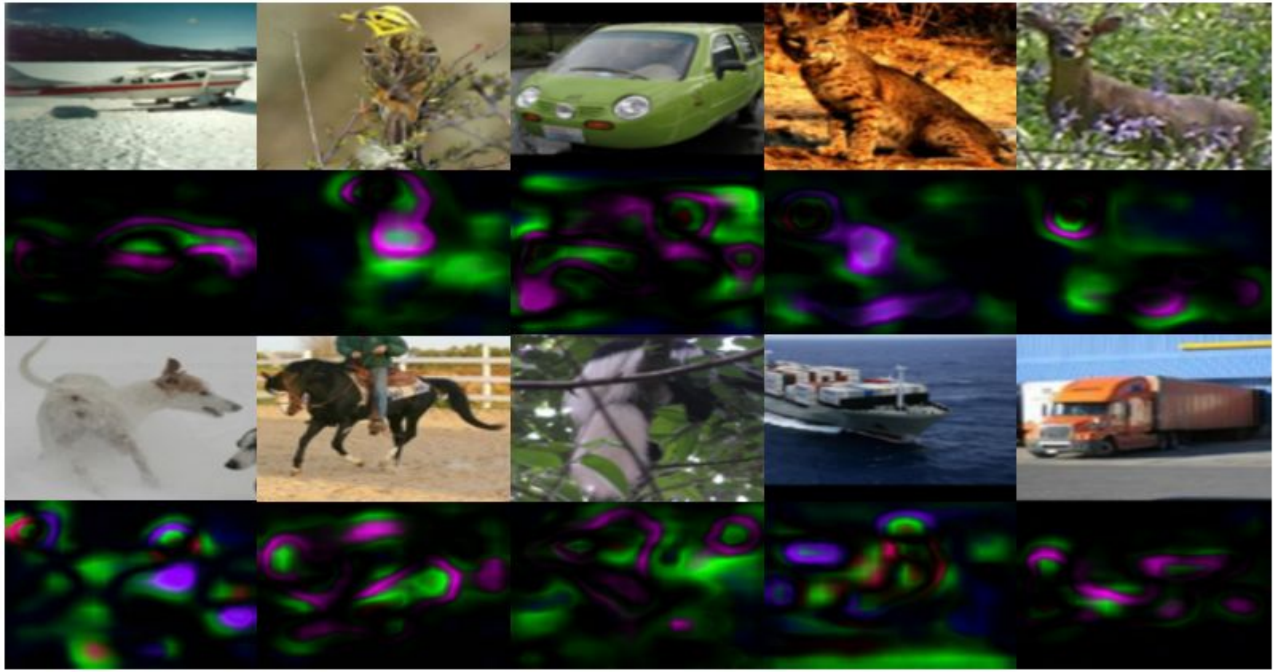
We note that for some categories**, coarse styles make a prediction difficult.**

SAM: Impact of stylizing in predictions



The more texture, the worse the result.

WSAM: Impact of styling as “noise” in samples

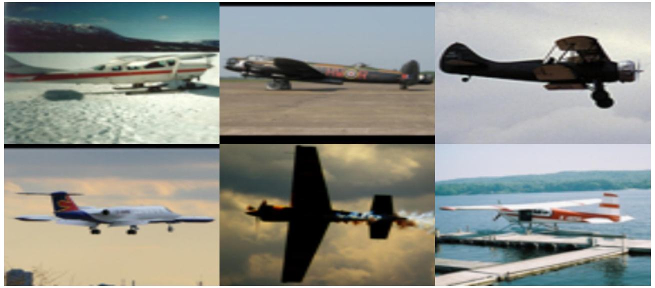


The more intensity used in stylization, the more noise (variance) is added.

WSAM: Measuring impact of styling in STL-10

WSAM variance calculated for each category in STL-10 dataset

$WSAM_{variance}$	Category	$WSAM_{variance}$	Category
airplane	0.107	horse	0.269
truck	0.129	bird	0.316
deer	0.175	dog	0.338
cat	0.193	monkey	0.380
car	0.228	ship	0.456



We note that high variance corresponds to the **ship** category, which has more PoV in samples.

Conclusions

.

- We found that the **best alpha** values are between 0.3 and 0.8
- Style Augmentation **works better** in **lower** image quality.
- **Style Activation Maps (SAM):** highlight impact of stylization in predictions
- **Weighted Style Activation Maps (WSAM)**: remark the total influence of styles in samples
- **WSAM variance**: measure the impact of stylized samples
- **Code:** https://github.com/fmorenovr/WSAM_Style/

Questions?