

Going Beyond Explaining CNNs

Vineeth N Balasubramanian

Department of Computer Science and Engineering
Indian Institute of Technology, Hyderabad



Review: Axioms of Attribution¹

Completeness

For any input x , the sum of the feature attributions equals $F(x) = \sum_i A_i^F(x)$

Sensitivity

If x has only one non-zero feature and $F(x) \neq 0$, then the attribution to that feature should be non-zero

Implementation Invariance

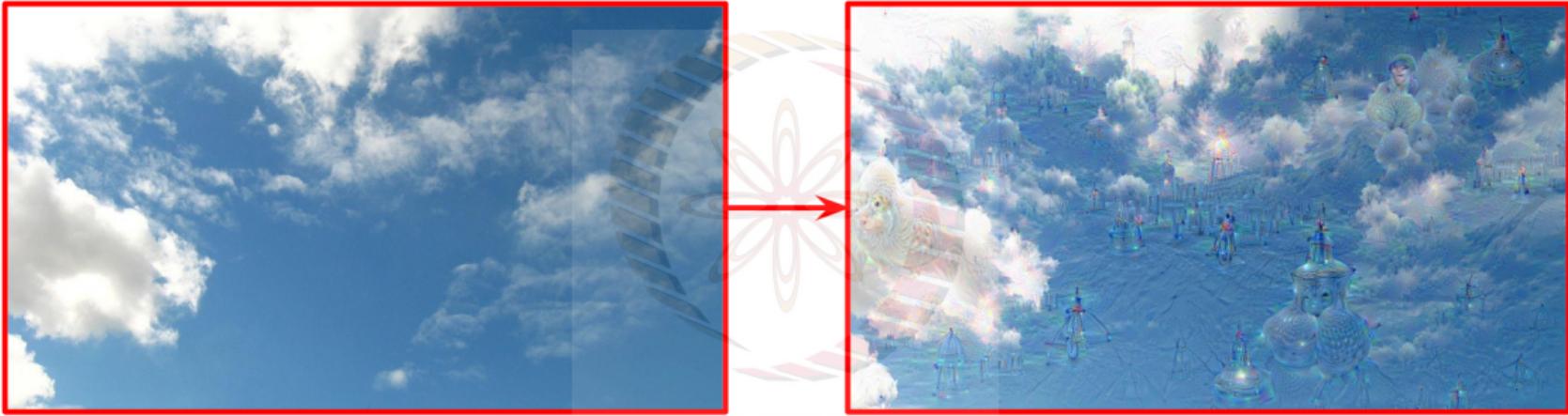
When two neural networks compute the same mathematical function $F(x)$, regardless of how differently they are implemented, the attributions to all features should always be identical.

Symmetry-Preserving

For any input x where the two values of two symmetric features are the same, their attributions should be identical as well.

¹Sundararajan et al, Axiomatic Attribution for Deep Networks, ICML 2017

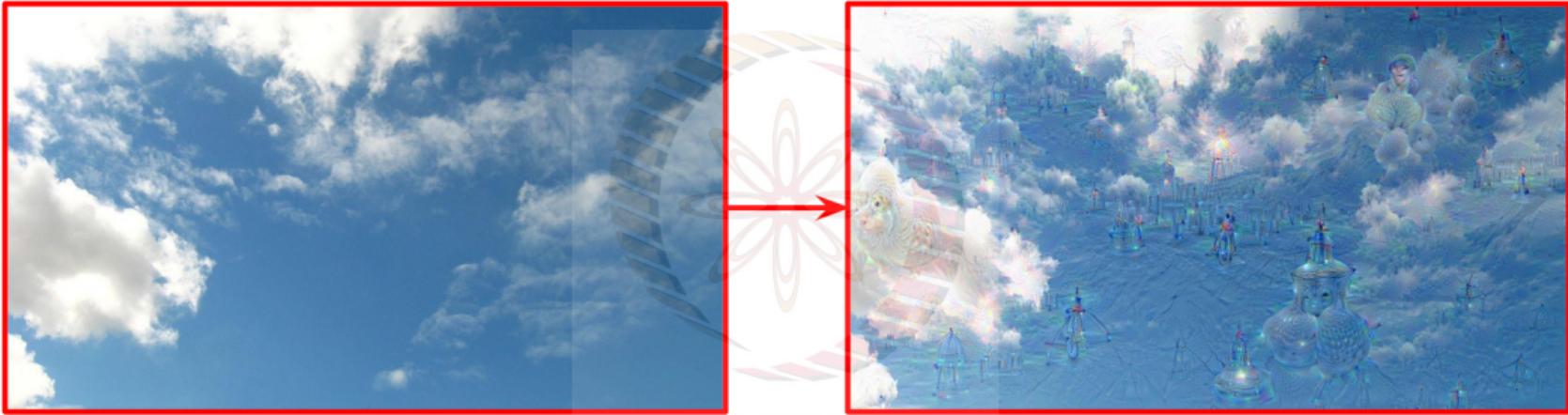
DeepDream²



- Modifies a given image in a way that **boosts** all activations at any layer, creating a feedback loop

²Mordvintsev et al, Deepdream - a code example for visualizing neural networks, 2015

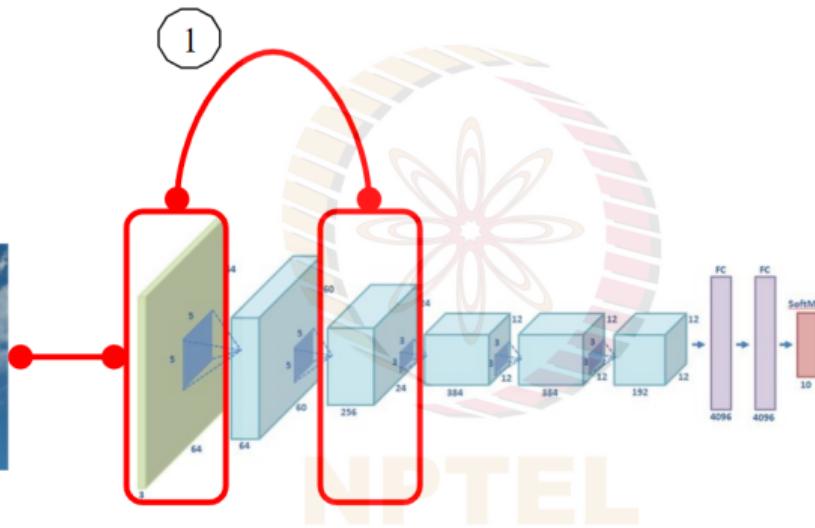
DeepDream²



- Modifies a given image in a way that **boosts** all activations at any layer, creating a feedback loop
- Any slightly detected dog face will be made more and more dog-like over time

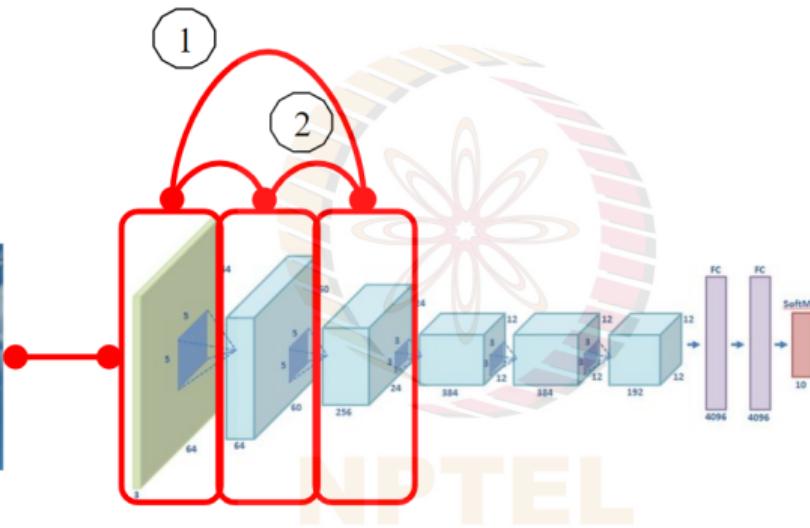
²Mordvintsev et al, Deepdream - a code example for visualizing neural networks, 2015

DeepDream



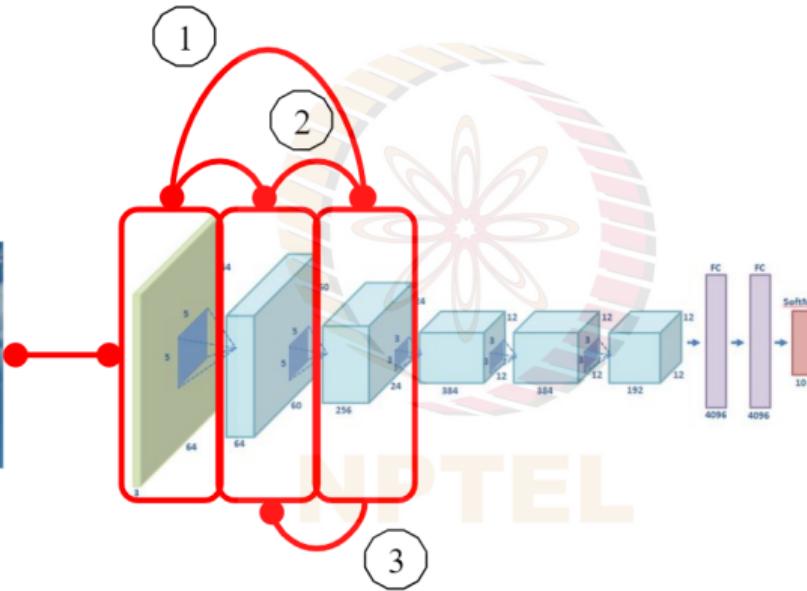
1) Choose a layer/filter.

DeepDream



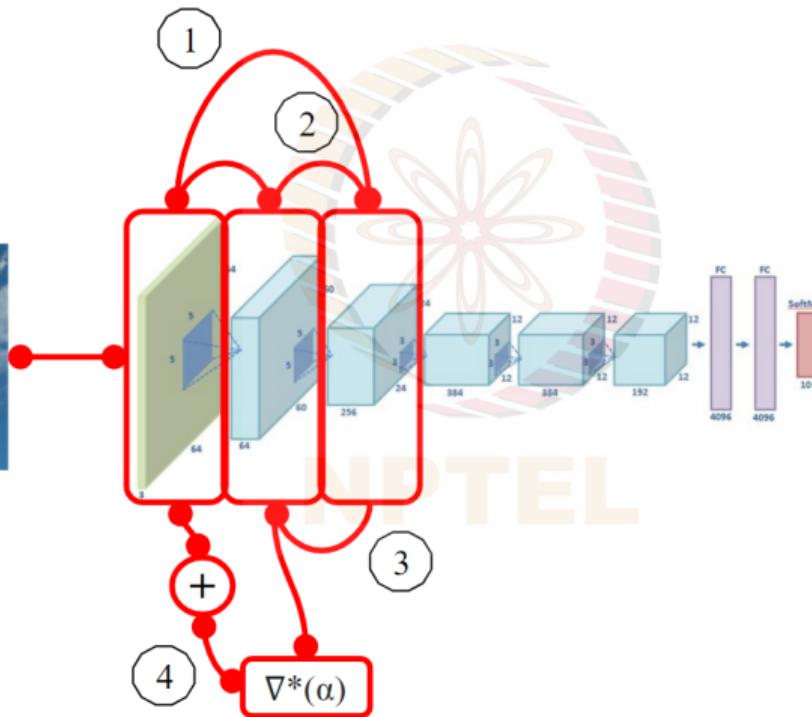
- 1) Choose a layer/filter.
- 2) Compute the activations for your image up to that layer.

DeepDream



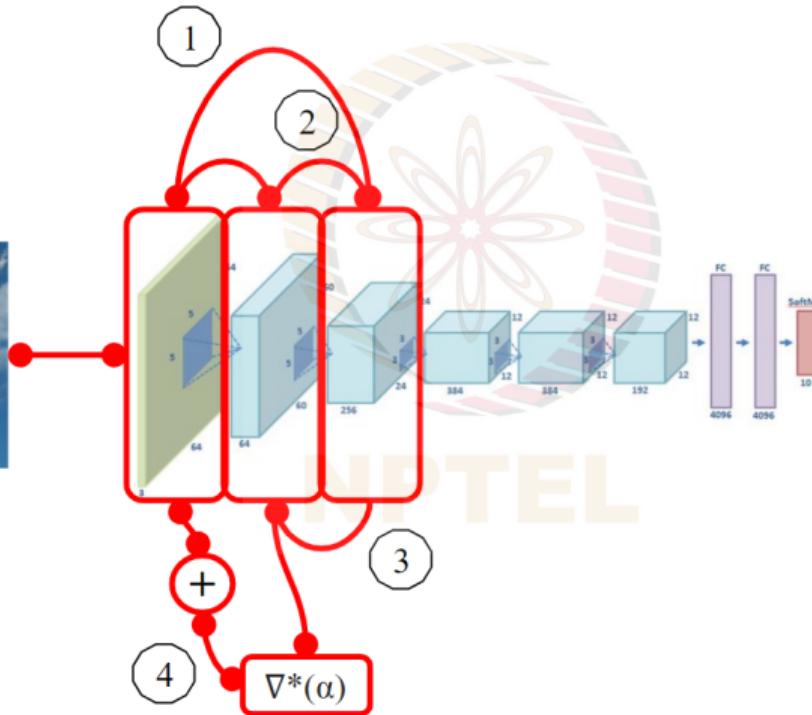
- 1) Choose a layer/filter.
- 2) Compute the activations for your image up to that layer.
- 3) Backpropagate the activations of your filter back to the input image.

DeepDream



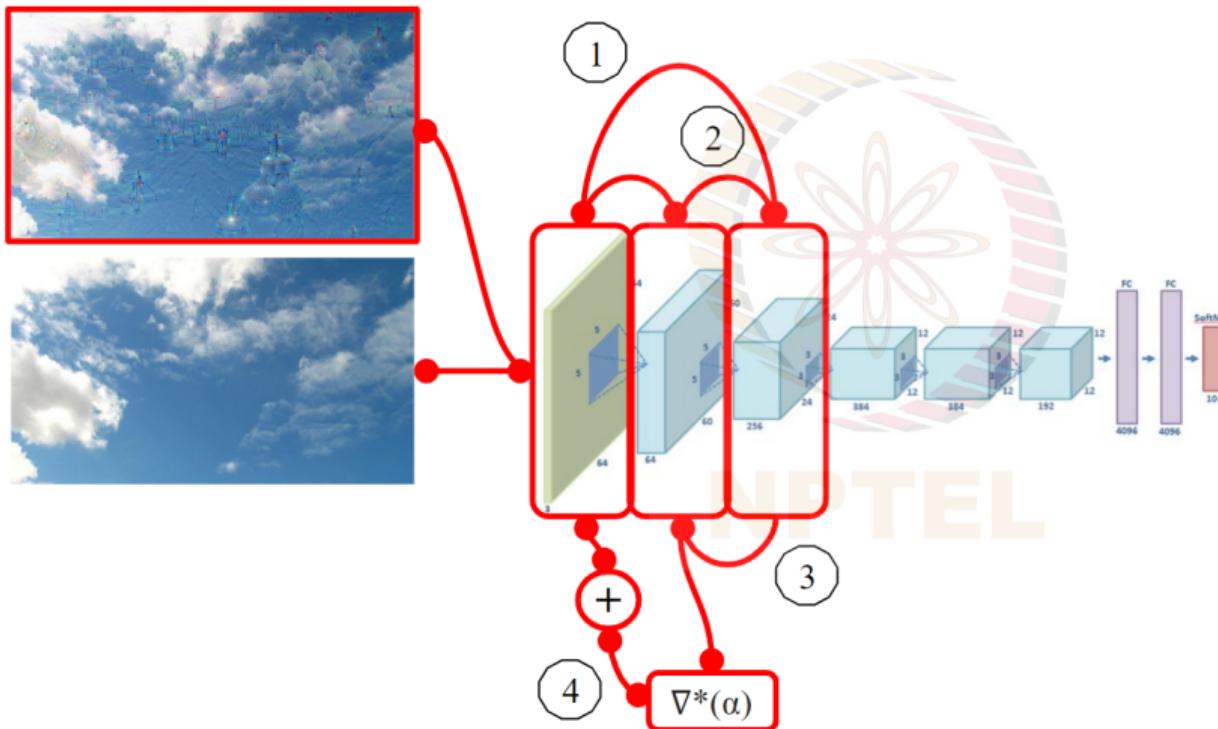
- 1) Choose a layer/filter.
- 2) Compute the activations for your image up to that layer.
- 3) Backpropagate the activations of your filter back to the input image.
- 4) Multiply the gradients (∇) with your learning rate (α) and add them to your input image

DeepDream



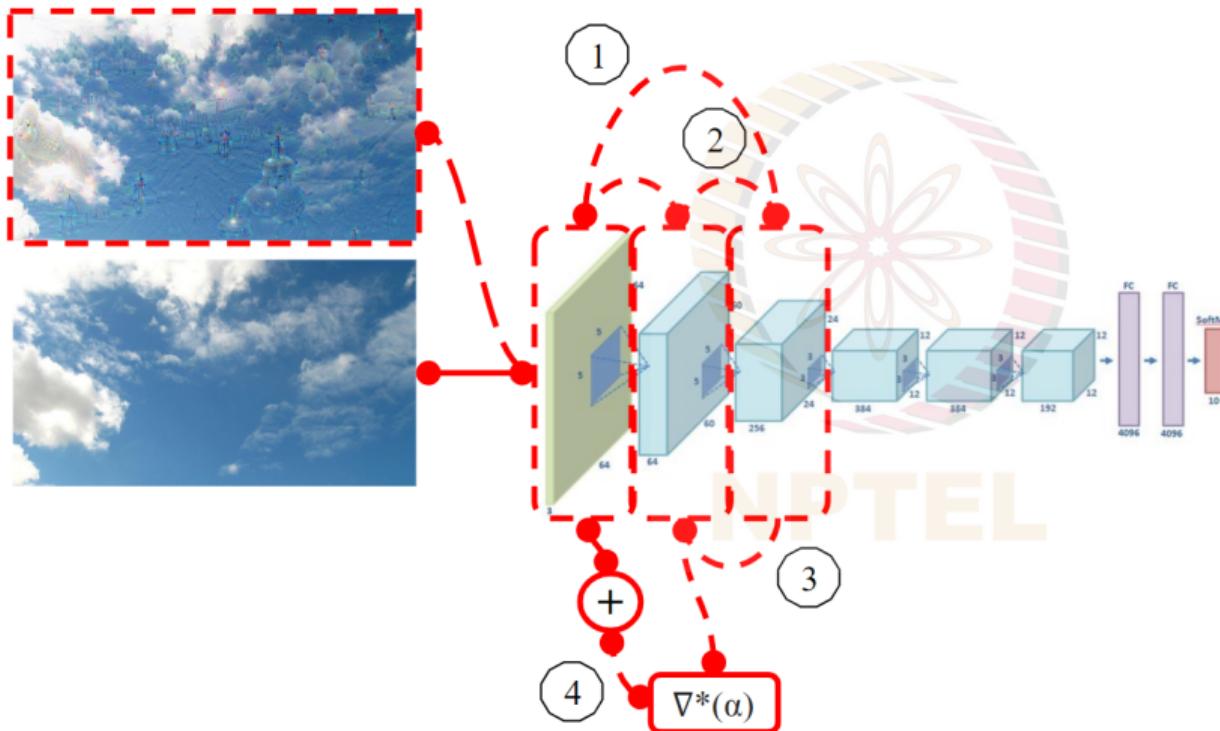
- 1) Choose a layer/filter.
- 2) Compute the activations for your image up to that layer.
- 3) Backpropagate the activations of your filter back to the input image.
- 4) Multiply the gradients (∇) with your learning rate (α) and add them to your input image
- 5) Go back to 2.

DeepDream



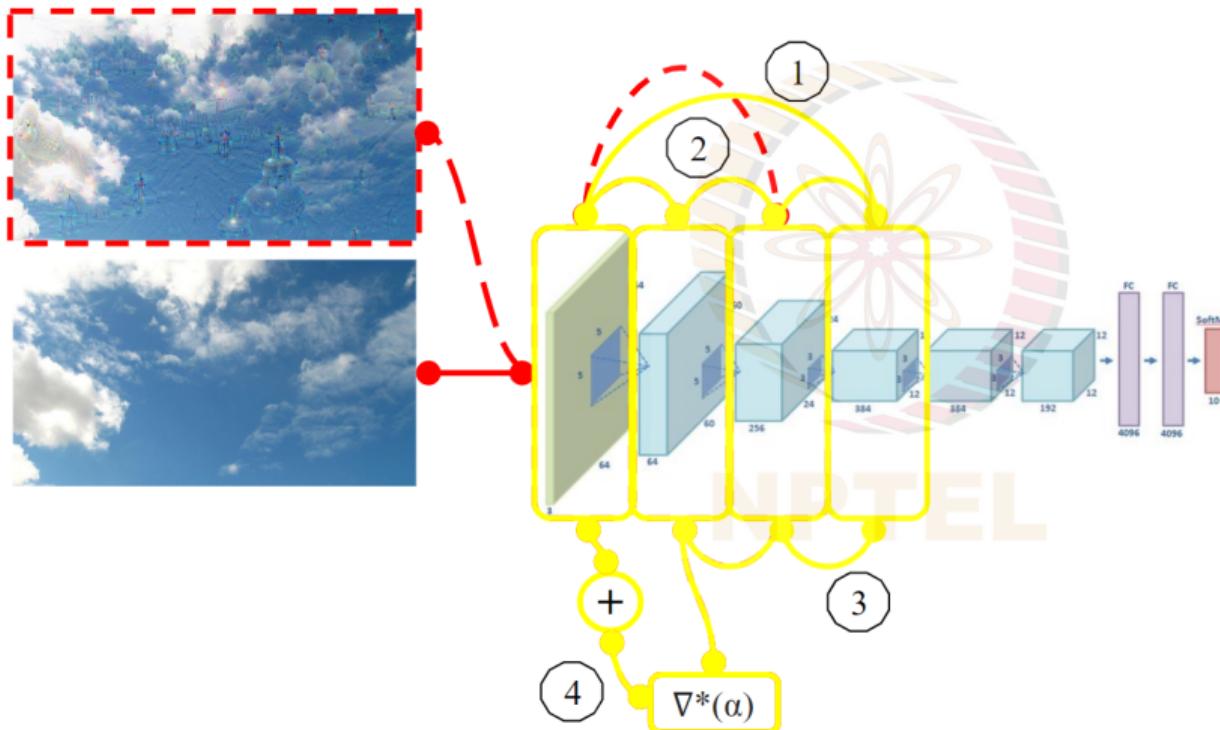
- 1) Choose a layer/filter.
- 2) Compute the activations for your image up to that layer.
- 3) Backpropagate the activations of your filter back to the input image.
- 4) Multiply the gradients (∇) with your learning rate (α) and add them to your input image
- 5) Go back to 2.

DeepDream



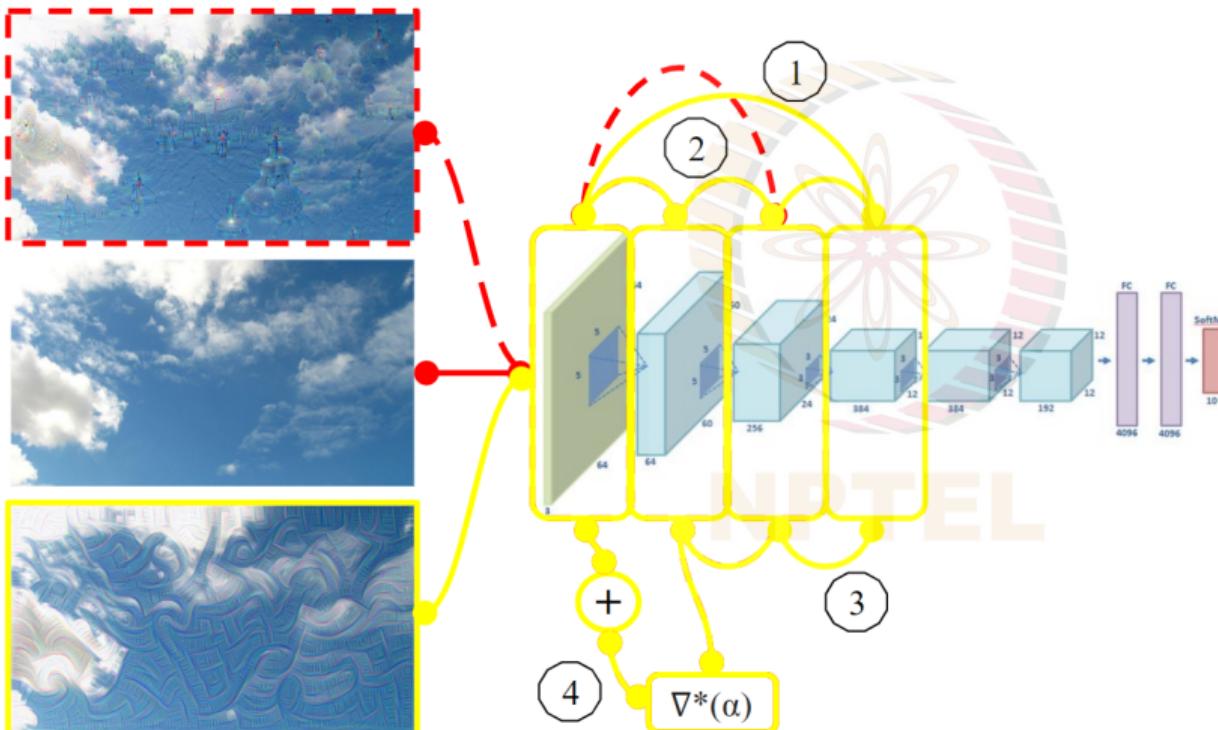
- 1) Choose a layer/filter.
- 2) Compute the activations for your image up to that layer.
- 3) Backpropagate the activations of your filter back to the input image.
- 4) Multiply the gradients (∇) with your learning rate (α) and add them to your input image
- 5) Go back to 2.

DeepDream



- 1) Choose a layer/filter.
- 2) Compute the activations for your image up to that layer.
- 3) Backpropagate the activations of your filter back to the input image.
- 4) Multiply the gradients (∇) with your learning rate (α) and add them to your input image
- 5) Go back to 2.

DeepDream



Higher layers produce complex features, while lower ones enhance edges and textures.

DeepDream: Examples



Horizon



Tower



NPTEL

DeepDream: Examples



Horizon



Tower



Trees



Building

DeepDream: Examples



Horizon



Tower



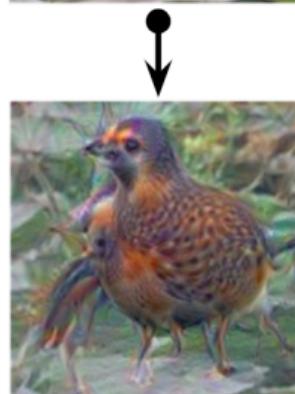
Trees



Building



Leaves

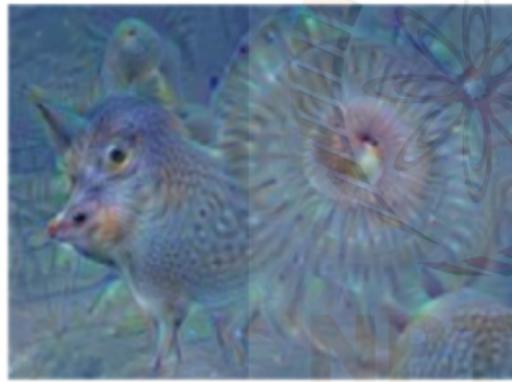


Birds & Insects

DeepDream: Examples



Admiral-Dog



Pig-Snail



Camel-Bird



Dog-Fish

Credit: Fei-Fei Li and Andrej Karpathy, CS231n course, Stanford, Winter 2016

Neural Style



Neural Style



Neural Style



+



Overlay

=



Neural Style



+



=



Neural Style



+



Human

Neural Style



⊕





+



Orsay

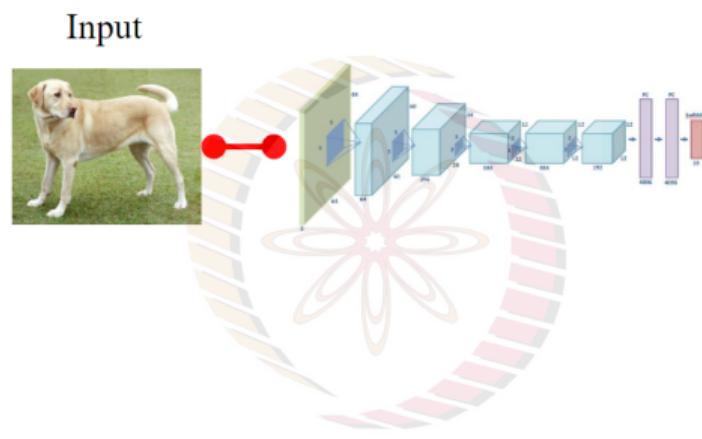
Hvan

Neural Style



³Gatys et al, A Neural Algorithm of Artistic Style, 2015

Neural Style⁴

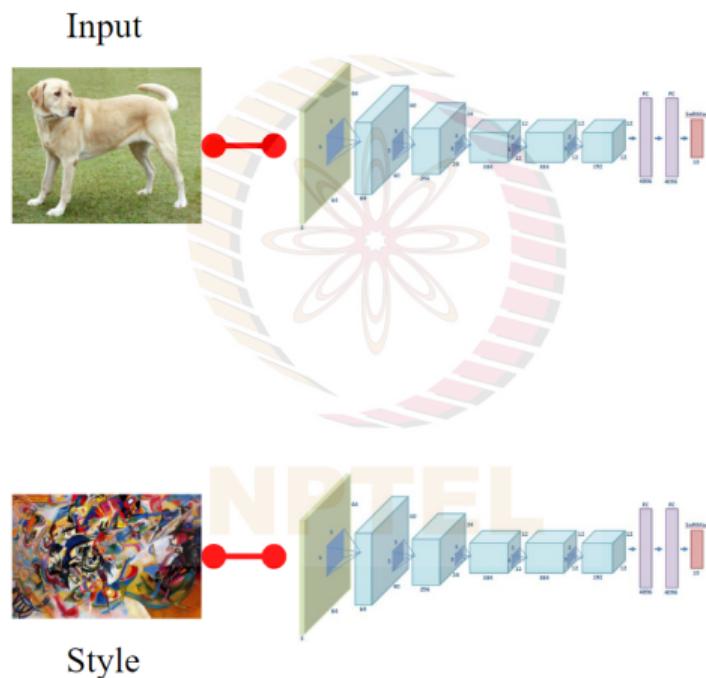


1) Extract **input targets** :
ConvNet activations of all layers
for the given input image.

NPTEL

⁴Gatys et al, A Neural Algorithm of Artistic Style, 2015

Neural Style⁴

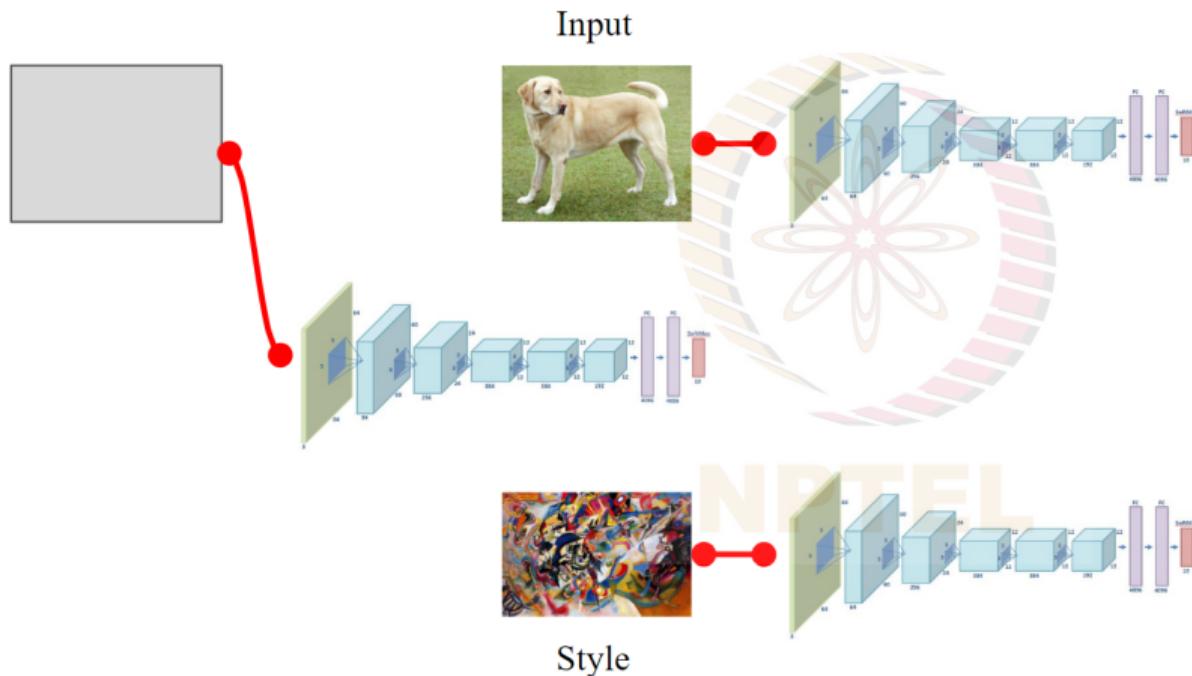


1) Extract **input targets** :
ConvNet activations of all layers
for the given input image.

2) Extract **style targets** : Gram
matrix of ConvNet activations
of all layers for the given style
image.

⁴Gatys et al, A Neural Algorithm of Artistic Style, 2015

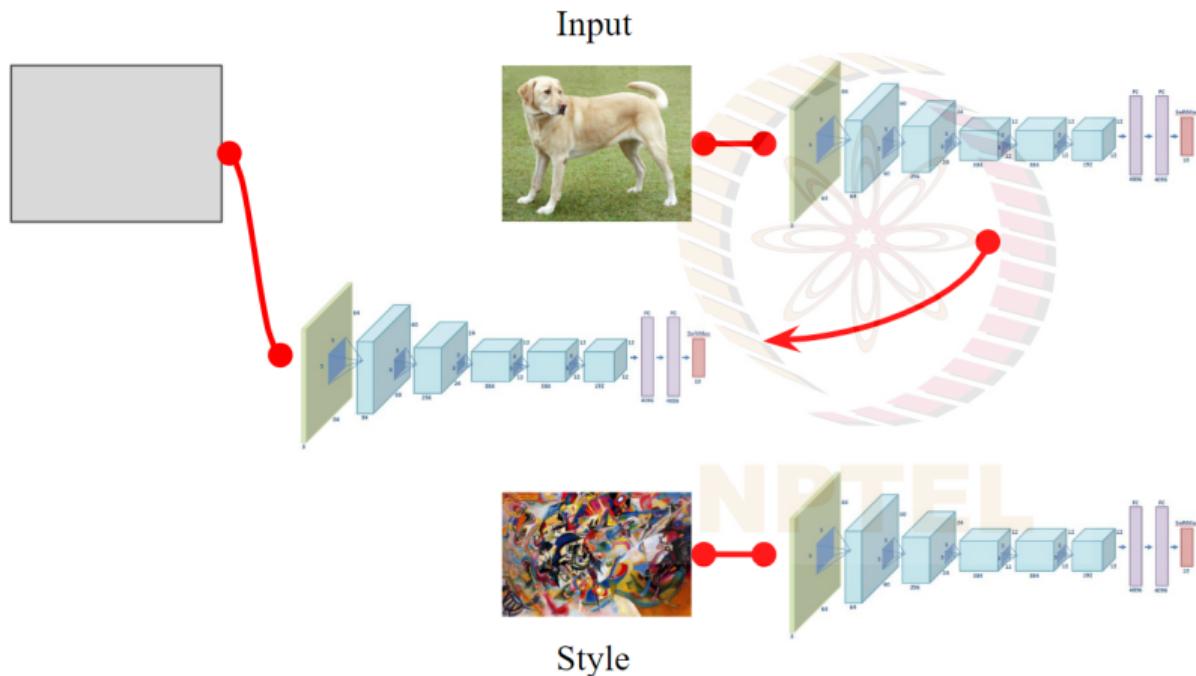
Neural Style⁴



- 1) Extract **input targets** : ConvNet activations of all layers for the given input image.
- 2) Extract **style targets** : Gram matrix of ConvNet activations of all layers for the given style image.
- 3) Initialize a new network.

⁴Gatys et al, A Neural Algorithm of Artistic Style, 2015

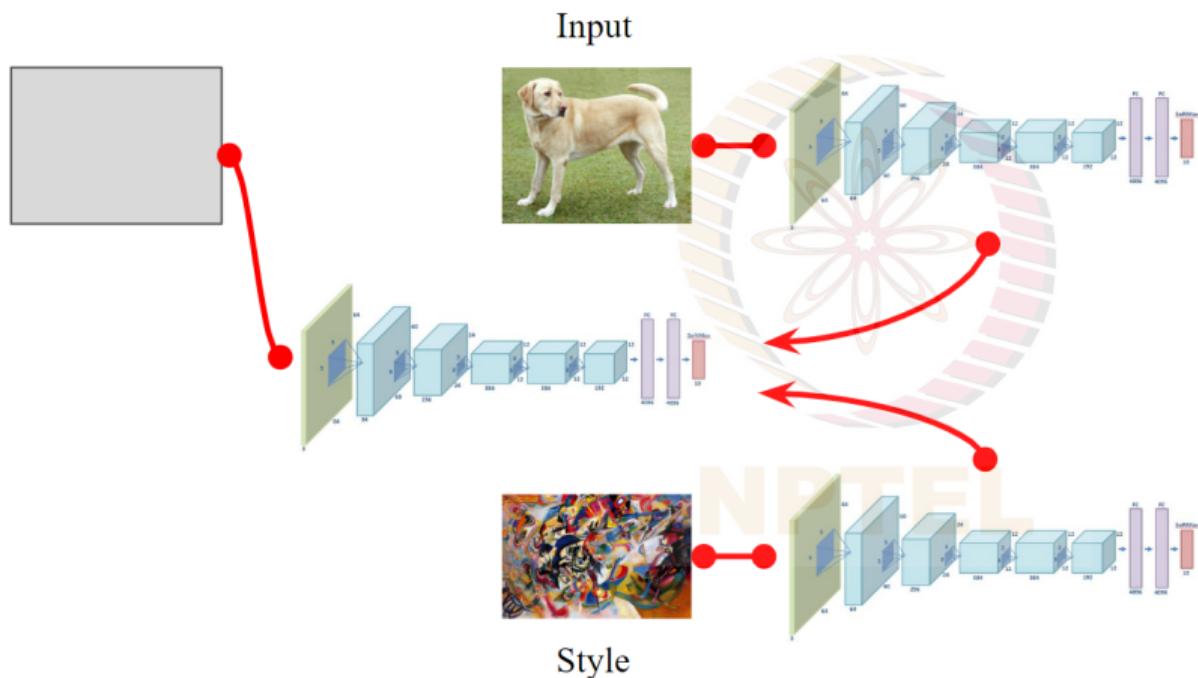
Neural Style⁴



- 1) Extract **input targets** :
ConvNet activations of all layers
for the given input image.
- 2) Extract **style targets** : Gram
matrix of ConvNet activations
of all layers for the given style
image.
- 3) Initialize a new network.
- 4) Optimize over image to match:
 - Activations of **input**.

⁴Gatys et al, A Neural Algorithm of Artistic Style, 2015

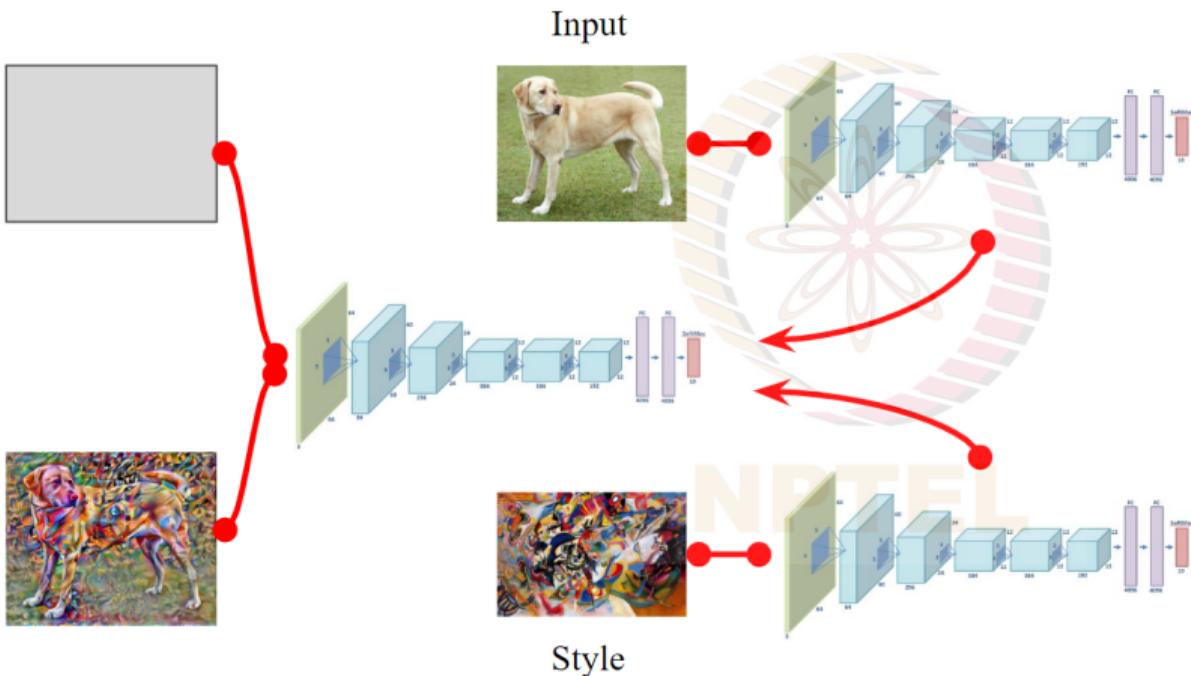
Neural Style⁴



- 1) Extract **input targets** : ConvNet activations of all layers for the given input image.
- 2) Extract **style targets** : Gram matrix of ConvNet activations of all layers for the given style image.
- 3) Initialize a new network.
- 4) Optimize over image to match:
 - Activations of **input**.
 - Gram matrix of activations of **style**.

⁴Gatys et al, A Neural Algorithm of Artistic Style, 2015

Neural Style⁴



- 1) Extract **input targets** : ConvNet activations of all layers for the given input image.
- 2) Extract **style targets** : Gram matrix of ConvNet activations of all layers for the given style image.
- 3) Initialize a new network.
- 4) Optimize over image to match:
 - Activations of **input**.
 - Gram matrix of activations of **style**.

⁴Gatys et al, A Neural Algorithm of Artistic Style, 2015

Neural Style: Examples



Credit: [Thushan Ganegedara, Intuitive Guide to Neural Style Transfer, TowardsDataScience](#)

Neural Style: Examples

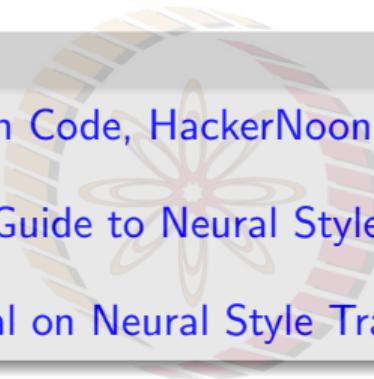


Credit: [Artistic Style Transfer with TensorFlow Lite](#)

Homework

Readings

- Sarthak Gupta, DeepDream with Code, HackerNoon
- Thushan Ganegedara, Intuitive Guide to Neural Style Transfer, Towards Data Science
- (Optional) Another good tutorial on Neural Style Transfer on Towards Data Science



Exercises

- Watch this fun video on YouTube of using DeepDream on videos, try to figure out how this was done!

NPTEL