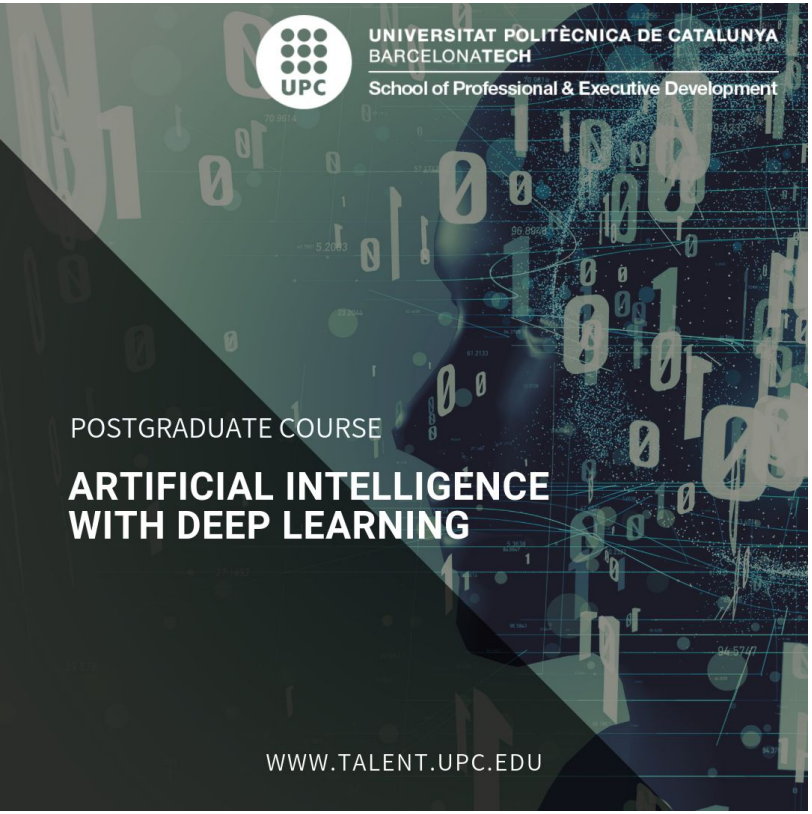


#DLUPC

Deep Learning 15

Interpretability (lab)

Daniel Fojo
dani.fojo@gmail.com

Research Engineer
Disney Research

Janna Escur
janna.eg@gmail.com

Engineer
Crisalix

Acknowledgements

Amaia Salvador
amaia.salvador@upc.edu

PhD Candidate
Universitat Politècnica de Catalunya

UNIVERSITAT POLITÈCNICA DE CATALUNYA

Eva Mohedano
eva.mohedano@insight-centre.org

Postdoctoral Researcher
Insight-centre for Data Analytics
Dublin City University

DCU

Insight
Centre for Data Analytics

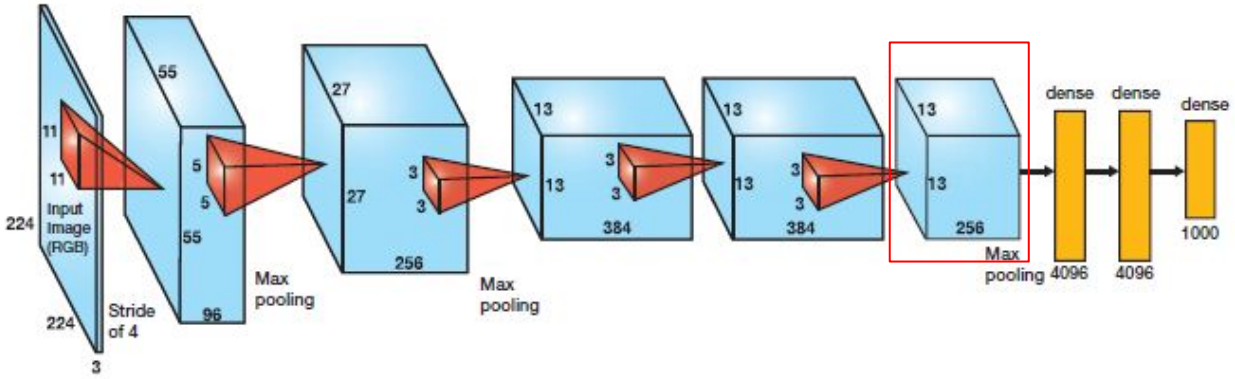
Interpretability

- Learned weights
- Activations from data
- Gradient-based
- Activation Maximization (AM)
- **Extra: Artistic variations**

Artistic variations: DeepDream

https://github.com/google/deepdream

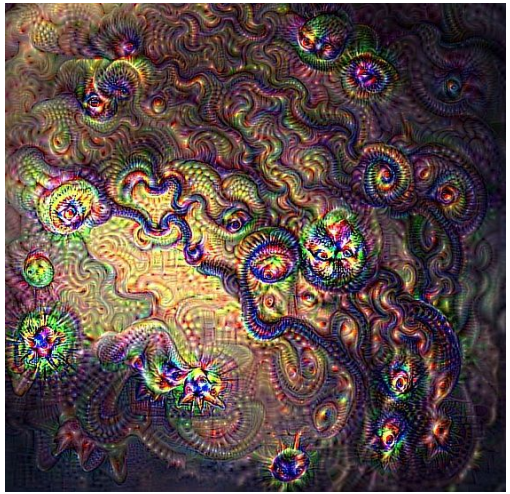
Artistic variations: DeepDream



1. Forward image up to some layer (e.g. conv5)
2. Set the gradients to equal the layer activations
3. Backprop to get gradient on the image
4. Update image (small step in the gradient direction)
5. Repeat

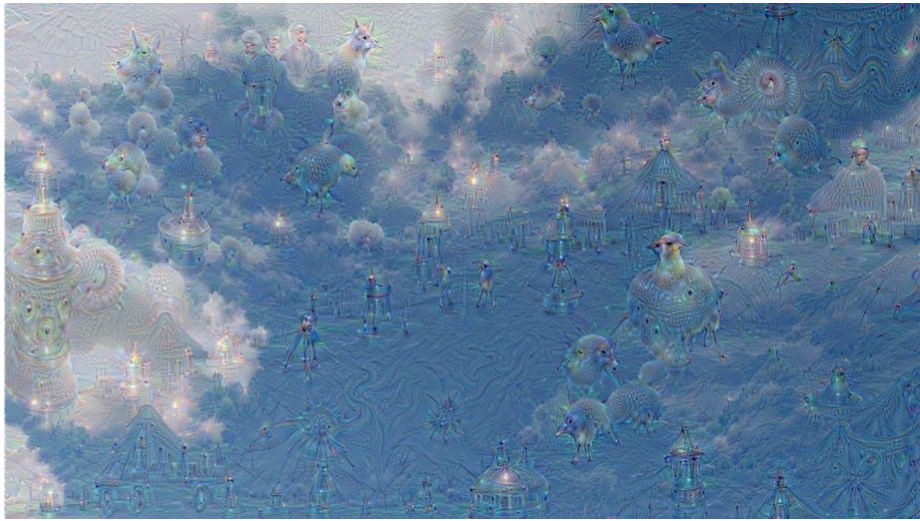
Artistic variations: DeepDream

1. Forward image up to some layer (e.g. conv5)
2. **Set the gradients to equal the layer activations**
3. Backprop to get gradient on the image
4. Update image (small step in the gradient direction)
5. Repeat

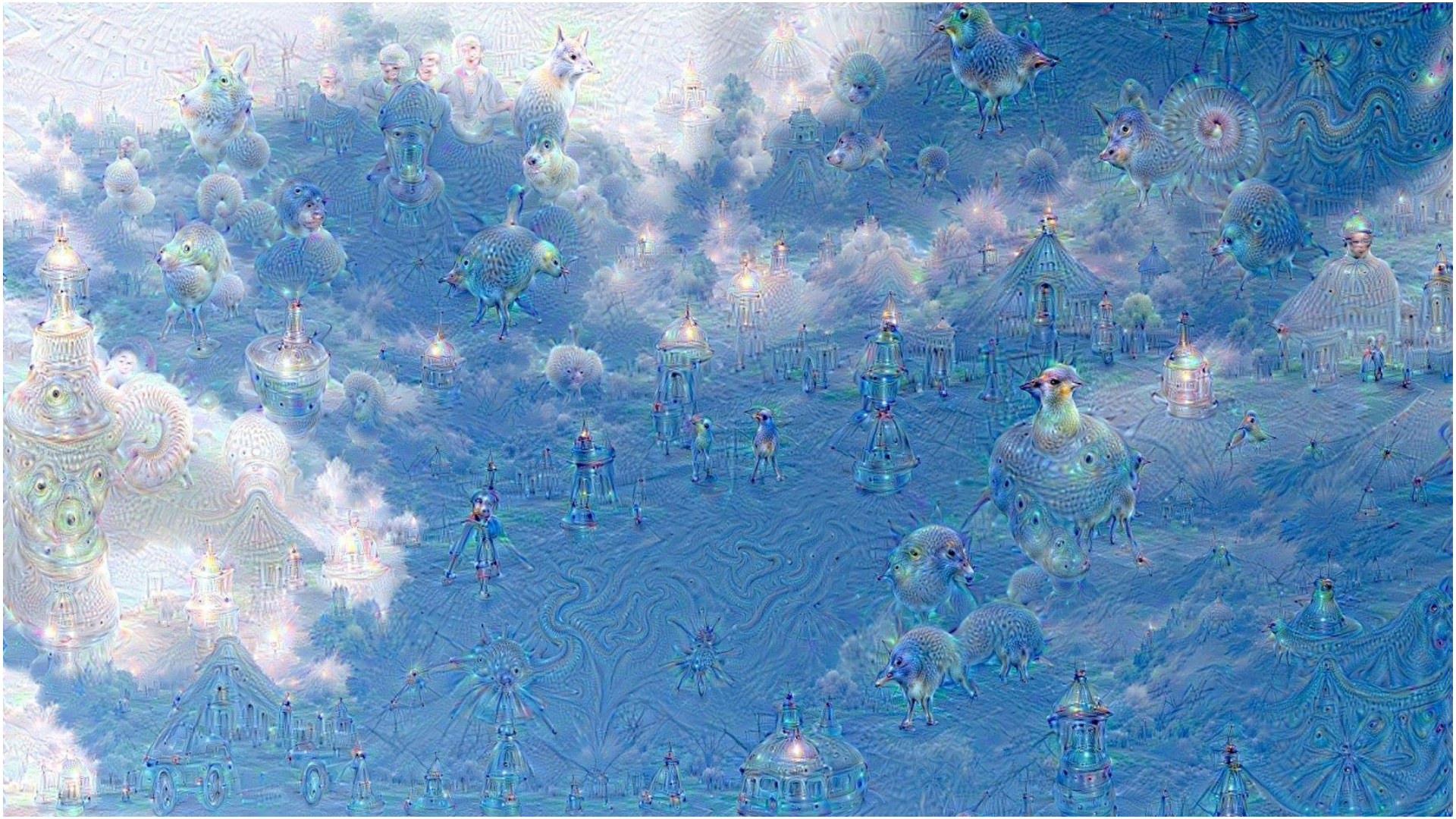


At each iteration, the image is updated to **boost all features** that activated in that layer in the forward pass.

Artistic variations: DeepDream



More examples here



Artistic variations: DeepDream

The network tends to generate animals because it was trained with many pictures of animals.

"Admiral Dog!"

"The Pig-Snail"

"The Camel-Bird"

"The Dog-Fish"

Google Research, “Going deeper into neural networks” - DeepDream (2015)

Artistic variations: DeepDream

If we apply the algorithm iteratively on its own outputs and apply some zooming after each iteration, we get an endless stream of new impressions, exploring the set of things the network knows about.

Google Research, “Going deeper into neural networks” - DeepDream (2015)

Video: Jonah Nordberg. http://johan-nordberg.com/

Neural Style

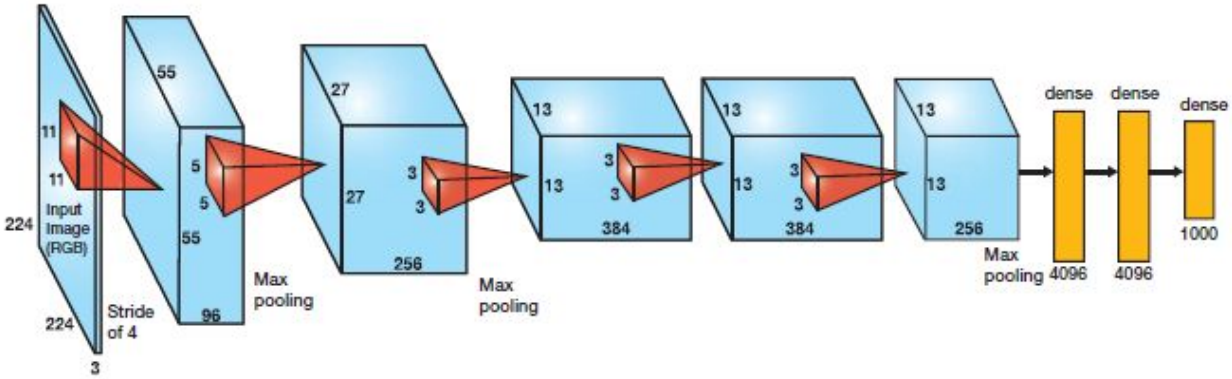
Style image

Content image

Result

Gatys, Leon A., Alexander S. Ecker, and Matthias Bethge. "Image style transfer using convolutional neural networks.", CVPR 2016. [video]

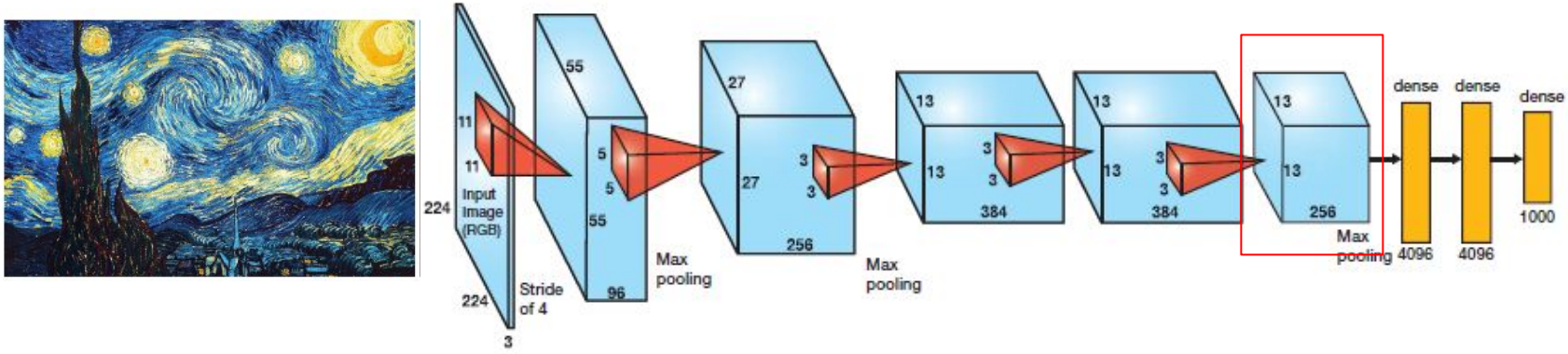
Neural Style



Extract raw activations in all layers. These activations will represent the contents of the image.

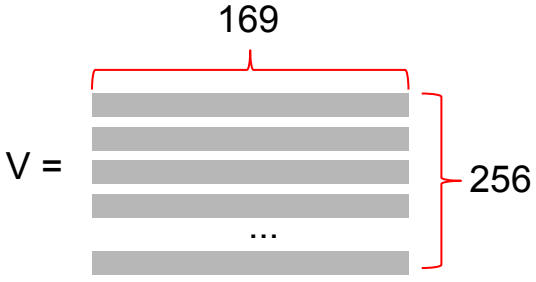
Gatys, Leon A., Alexander S. Ecker, and Matthias Bethge. "Image style transfer using convolutional neural networks.", CVPR 2016. [video]

Neural Style



- Activations are also extracted from the style image for all layers.
- Instead of the raw activations, **gram matrices (G)** are computed at each layer to represent the style.

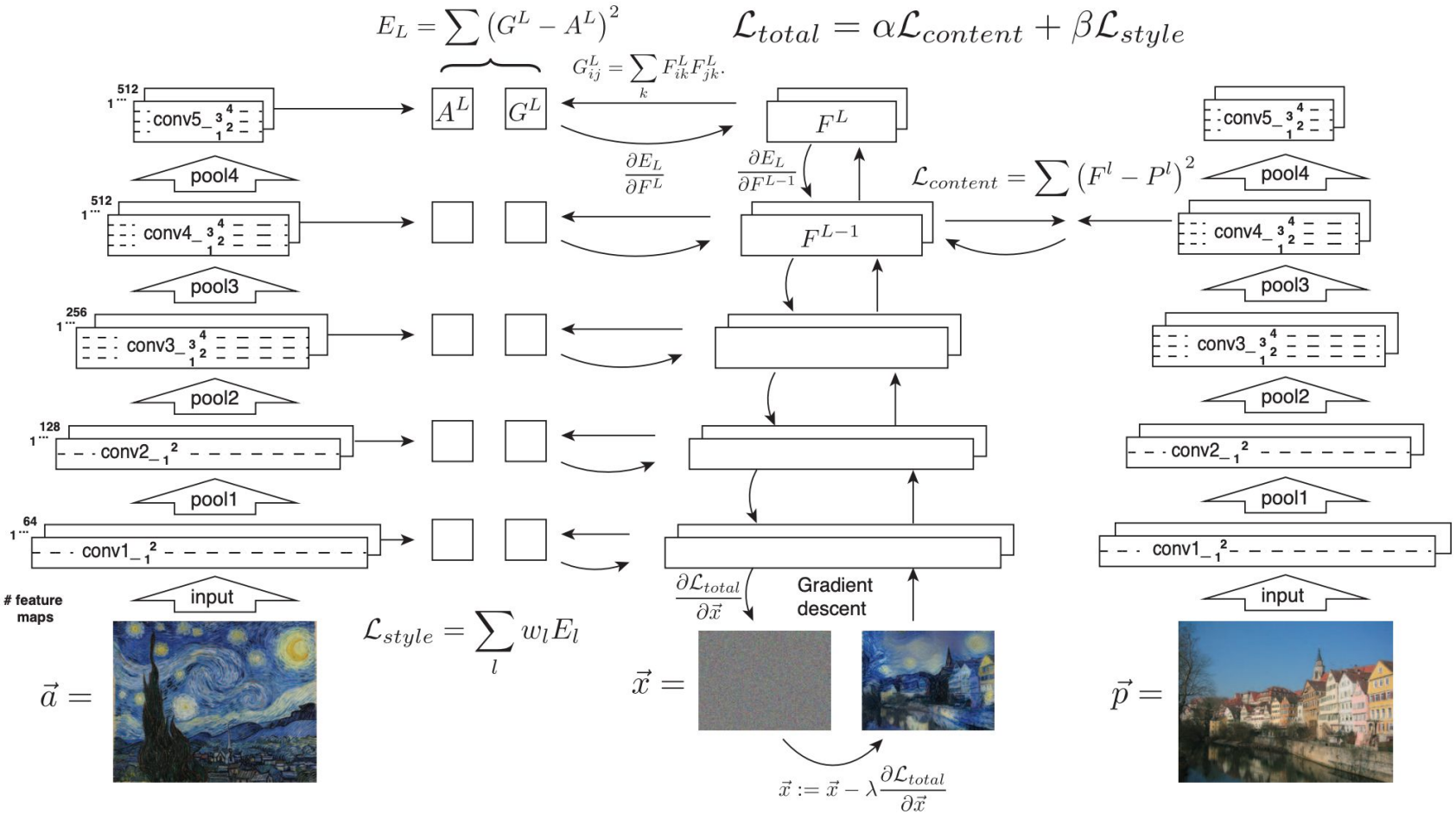
E.g. at conv5 [13x13x256], reshape to:



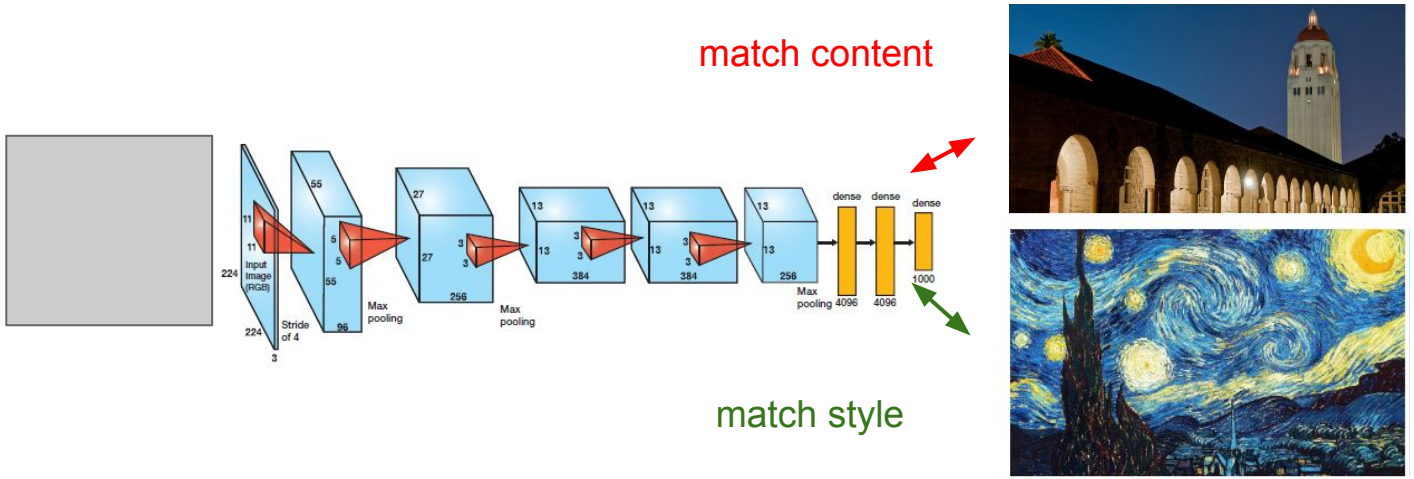
$G = V^T V$

The Gram matrix G gives the correlations between filter responses.

Neural Style



Neural Style



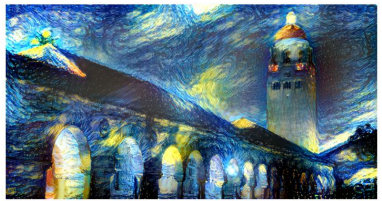
$$\mathcal{L}_{total}(\vec{p}, \vec{a}, \vec{x}) = \alpha \mathcal{L}_{content}(\vec{p}, \vec{x}) + \beta \mathcal{L}_{style}(\vec{a}, \vec{x})$$

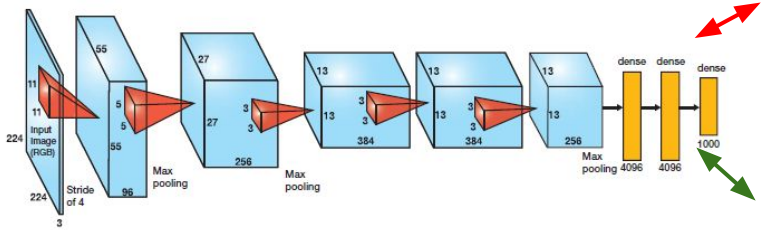
Match activations from content image

Match gram matrices from style image

Gatys, Leon A., Alexander S. Ecker, and Matthias Bethge. "Image style transfer using convolutional neural networks.", CVPR 2016. [video]

Neural Style





match content

match style

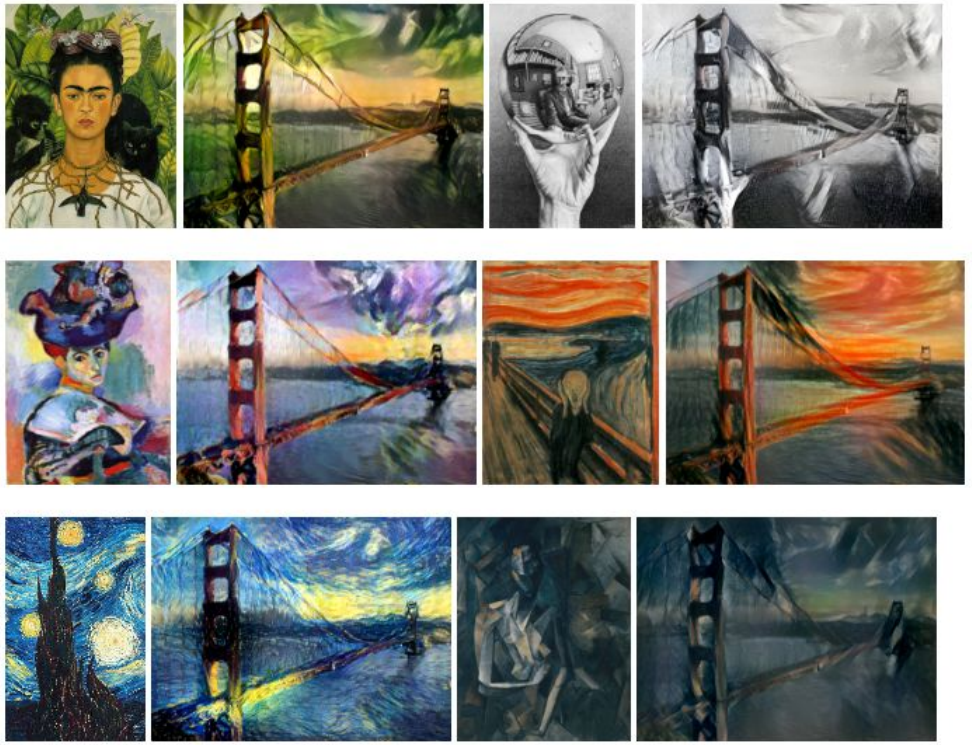
$$\mathcal{L}_{total}(\vec{p}, \vec{a}, \vec{x}) = \alpha \mathcal{L}_{content}(\vec{p}, \vec{x}) + \beta \mathcal{L}_{style}(\vec{a}, \vec{x})$$

Match activations from content image

Match gram matrices from style image

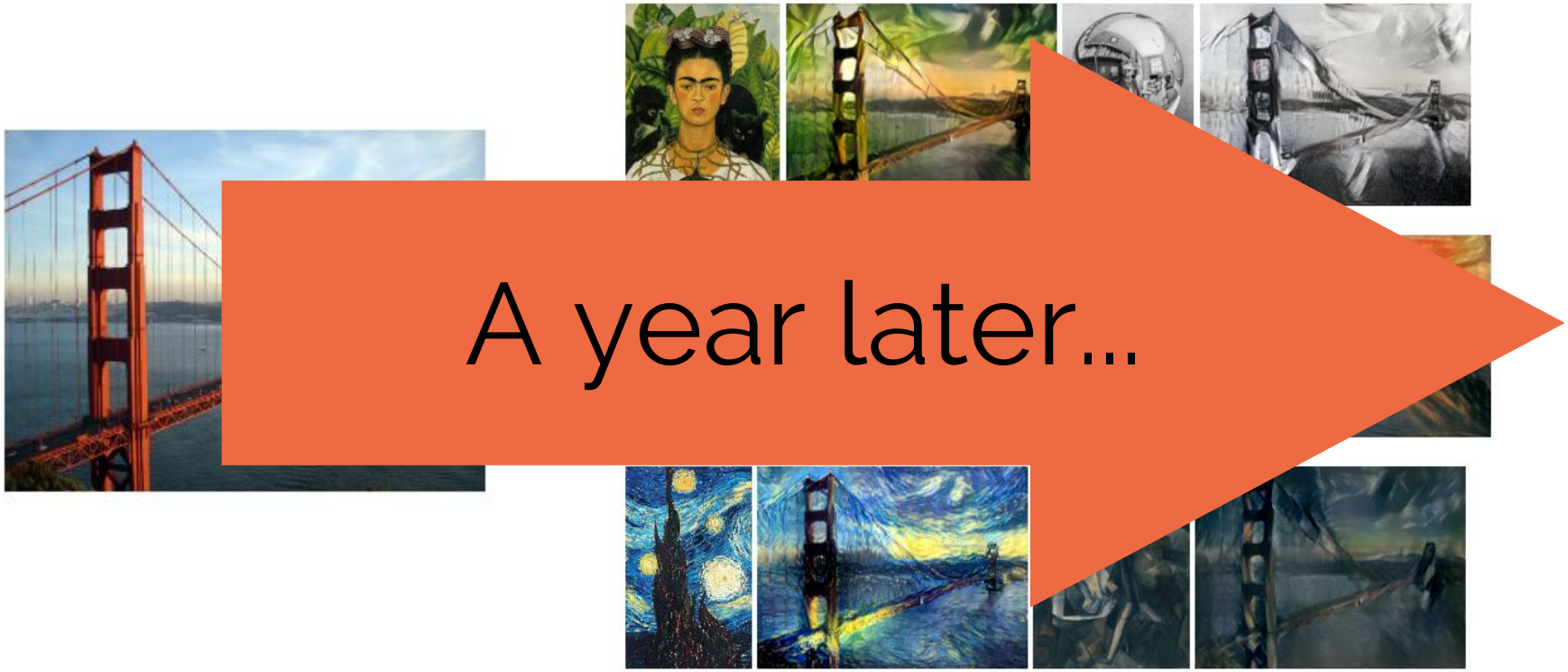
Gatys, Leon A., Alexander S. Ecker, and Matthias Bethge. "Image style transfer using convolutional neural networks.", CVPR 2016. [video]

Neural Style



Gatys, Leon A., Alexander S. Ecker, and Matthias Bethge. "Image style transfer using convolutional neural networks.", CVPR 2016. [video]

Neural Style



Gatys, Leon A., Alexander S. Ecker, and Matthias Bethge. "Image style transfer using convolutional neural networks.", CVPR 2016. [video]






Content Image | Style Image | Result

Luan, Fujun, Sylvain Paris, Eli Shechtman, and Kavita Bala. "Deep photo style transfer." CVPR 2017.






Content Image | Style Image | Result

Luan, Fujun, Sylvain Paris, Eli Shechtman, and Kavita Bala. "Deep photo style transfer." CVPR 2017.

Lab

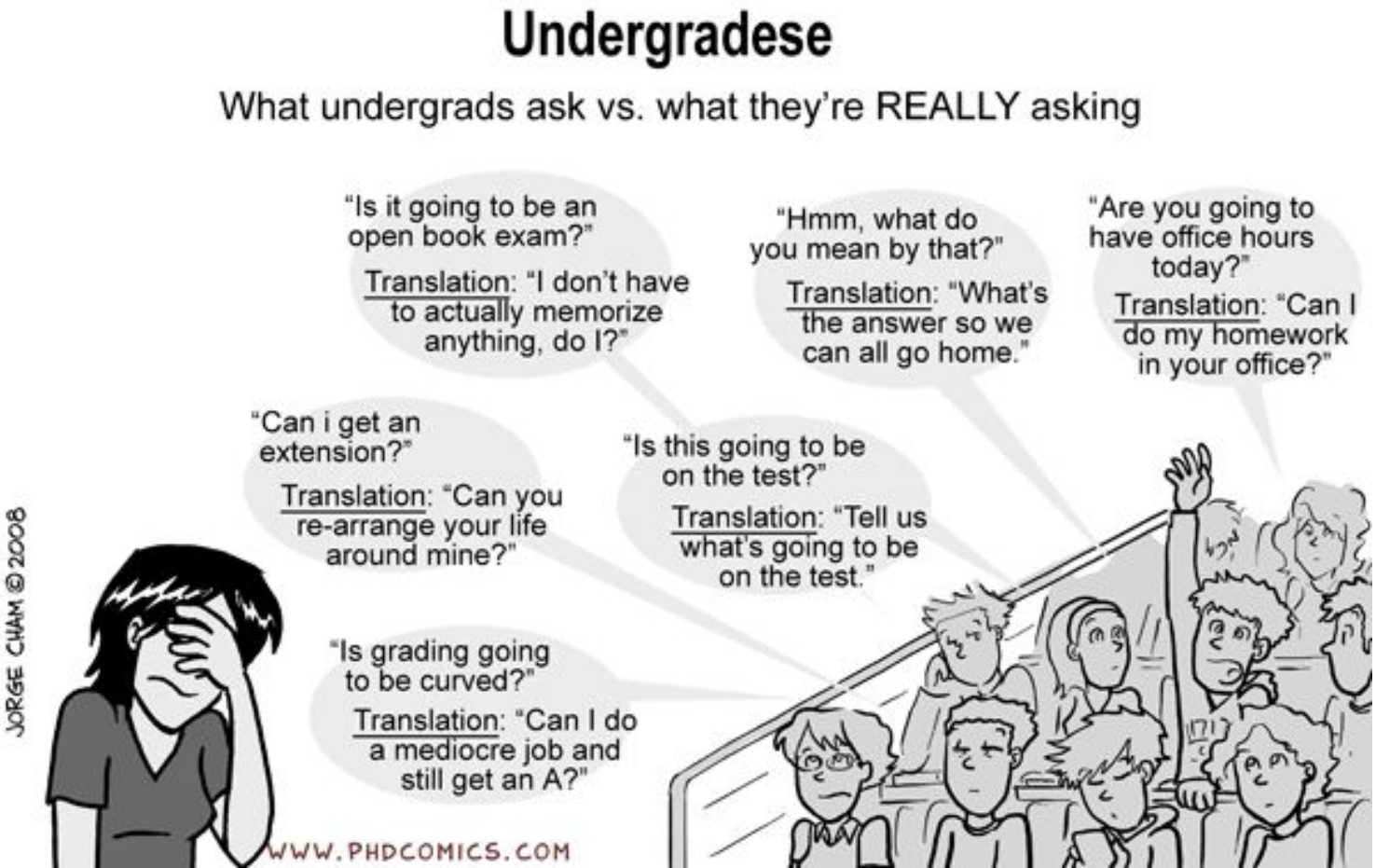
Amaia Salvador
amaia.salvador@upc.edu

PhD Candidate
Universitat Politècnica de Catalunya

UNIVERSITAT POLITÈCNICA DE CATALUNYA

- Deep Dream (optional)
- Style transfer

Questions ?


Undergradese
What undergrads ask vs. what they're REALLY asking
"Is it going to be an open book exam?"
Translation: "I don't have to actually memorize anything, do I?"
"Hmm, what do you mean by that?"
Translation: "What's the answer so we can all go home."
"Are you going to have office hours today?"
Translation: "Can I do my homework in your office?"
"Can i get an extension?"
Translation: "Can you re-arrange your life around mine?"
"Is this going to be on the test?"
Translation: "Tell us what's going to be on the test."
"Is grading going to be curved?"
Translation: "Can I do a mediocre job and still get an A?"
JORGE CHAM © 2008
WWW.PHDCOMICS.COM