Seminar Explainable AI

Module 6

Selected Methods Part 2

Deconv-Inverting-Guided-Backprop-DGN-TCAV

Andreas Holzinger

Human-Centered AI Lab (Holzinger Group)

Institute for Medical Informatics/Statistics, Medical University Graz, Austria

and

Explainable AI-Lab, Alberta Machine Intelligence Institute, Edmonton, Canada

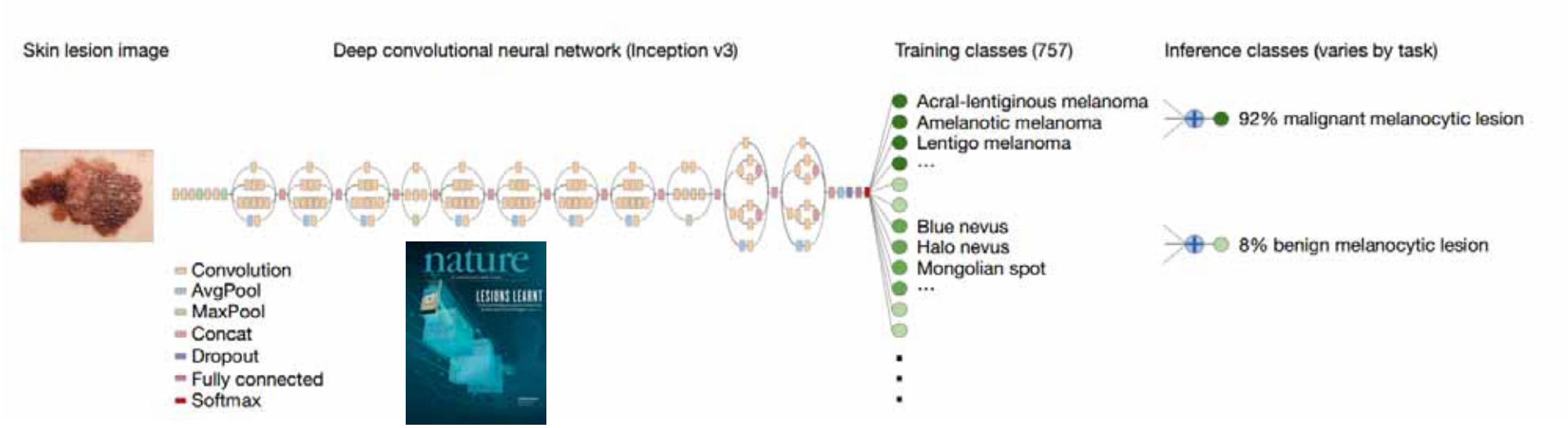
This is the version for printing and reading. The lecture version is didactically different.

AGENDA

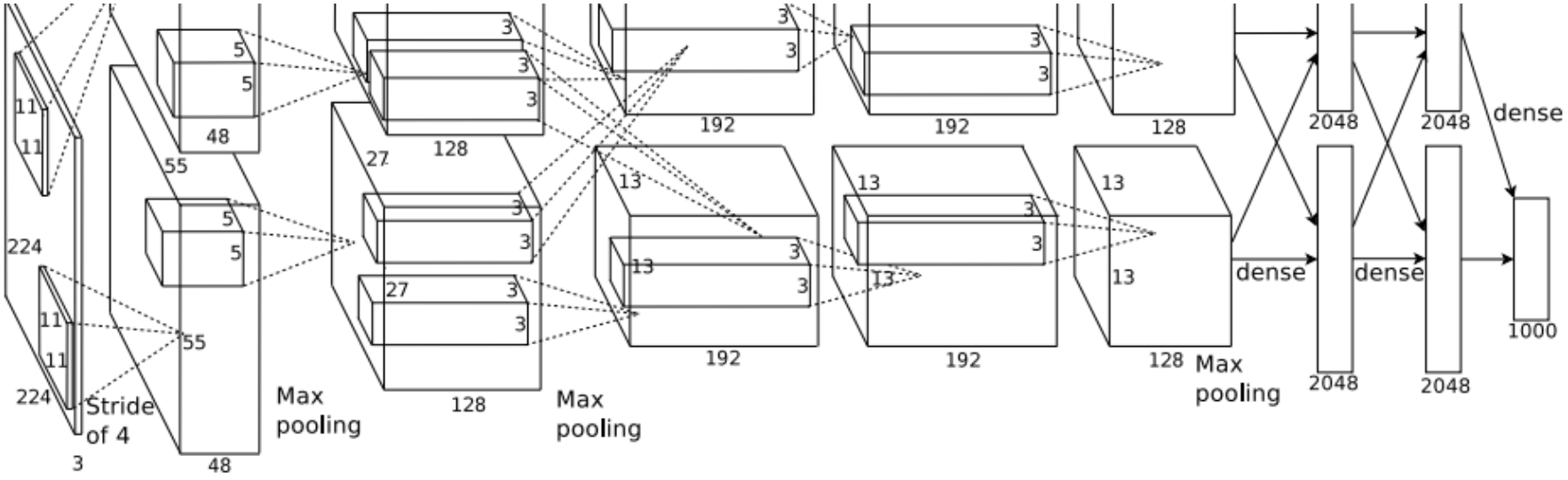
- 00 Reflection
- 01 Visualizing CNN with Deconvolution
- 02 Inverting CNN
- 03 Guided Backpropagation
- 04 Deep Generator Networks
- 05 Testing with Concept Activation Vectors

01 Visualizing CNN with Deconvolution

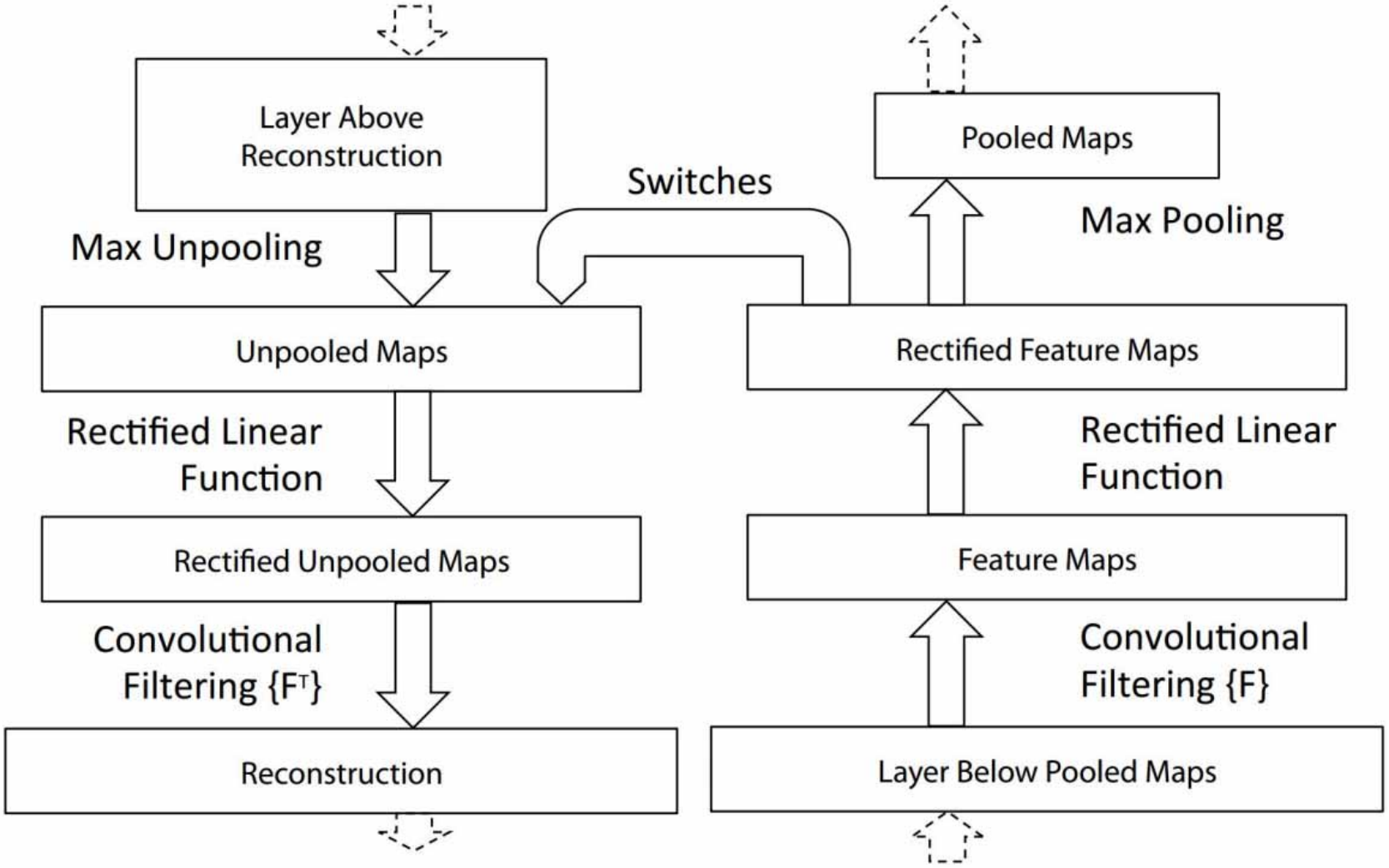
Our typical example to answer the question of why



Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M. & Thrun, S. 2017. Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542, (7639), 115-118, doi:10.1038/nature21056.

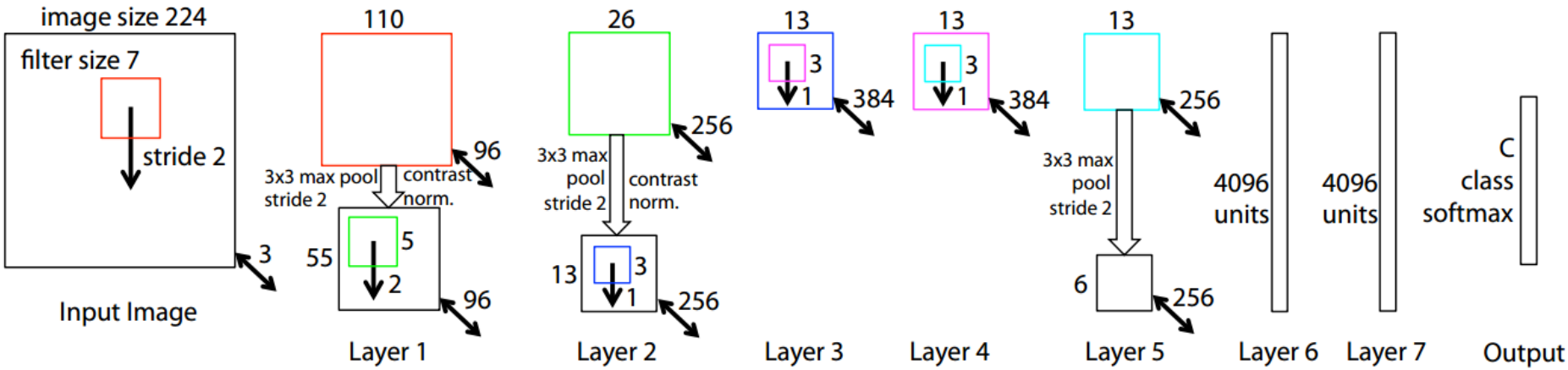


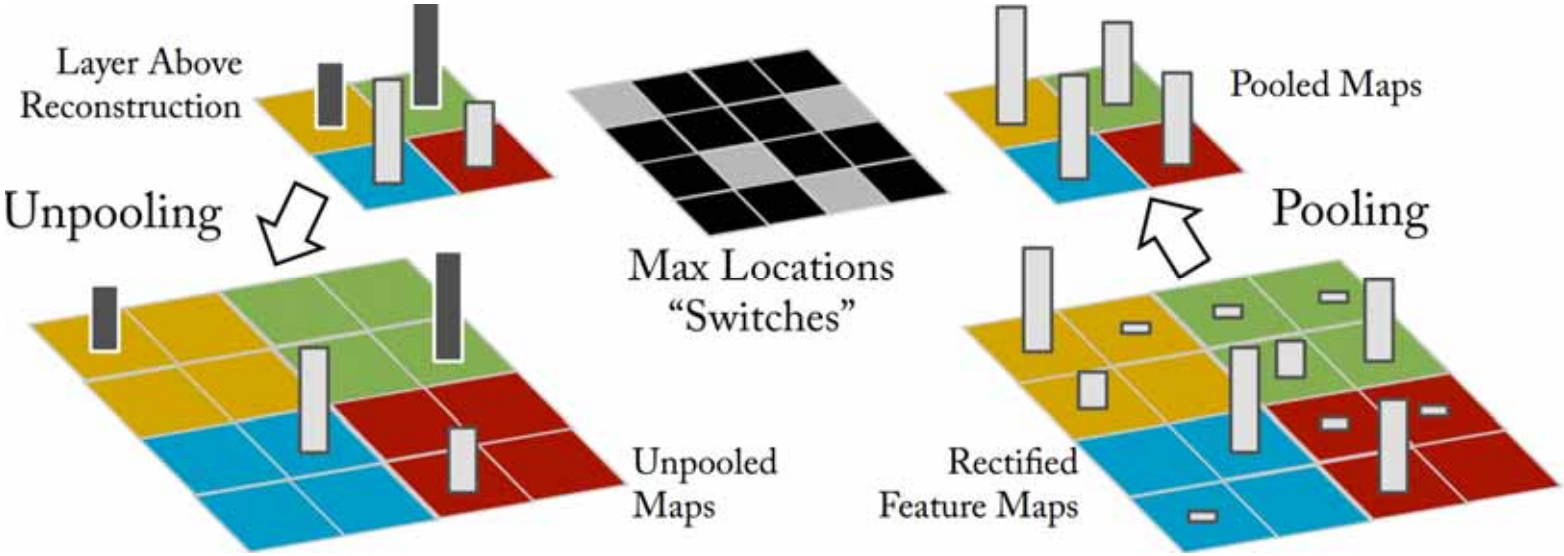
Example: Interpretable Deep Learning Model



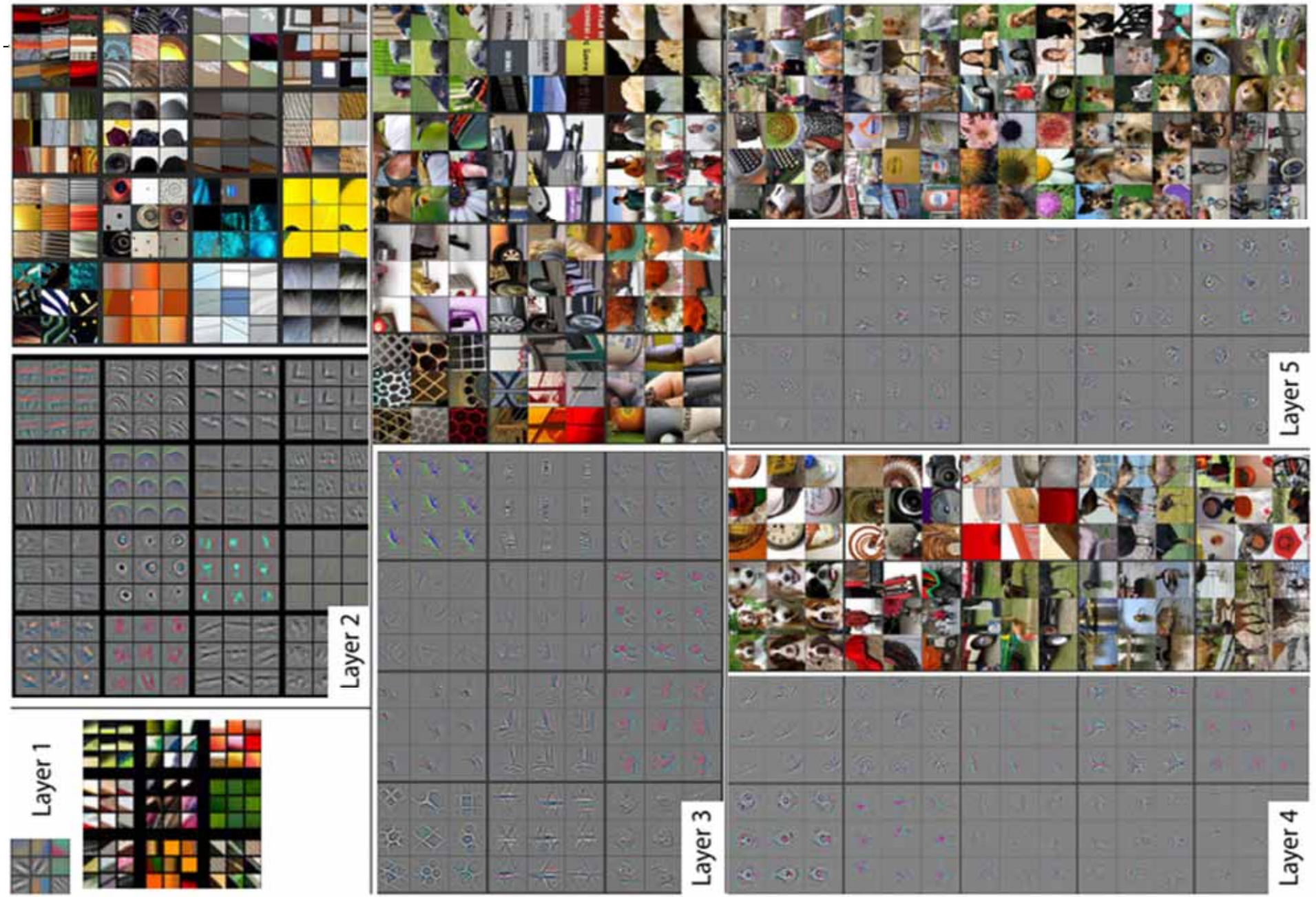
Matthew D. Zeiler & Rob Fergus 2013. Visualizing and Understanding Convolutional Networks. arXiv:1311.2901.

Visualizing a Conv Net with a De-Conv Net



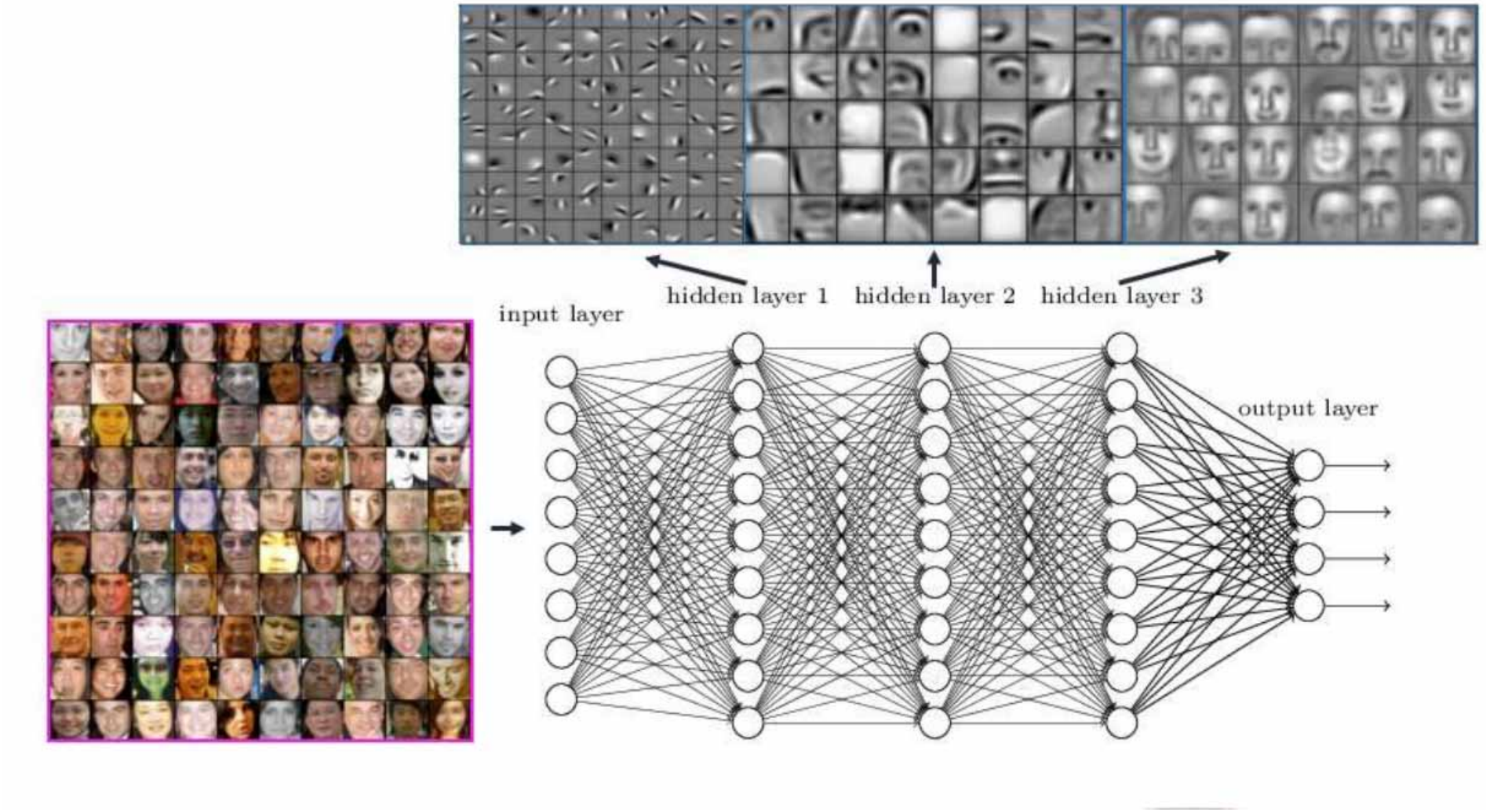


Matthew D. Zeiler & Rob Fergus 2014. Visualizing and understanding convolutional networks. In: D., Fleet, T., Pajdla, B., Schiele & T., Tuytelaars (eds.) ECCV, Lecture Notes in Computer Science LNCS 8689. Cham: Springer, pp. 818-833, doi:10.1007/978-3-319-10590-1_53.



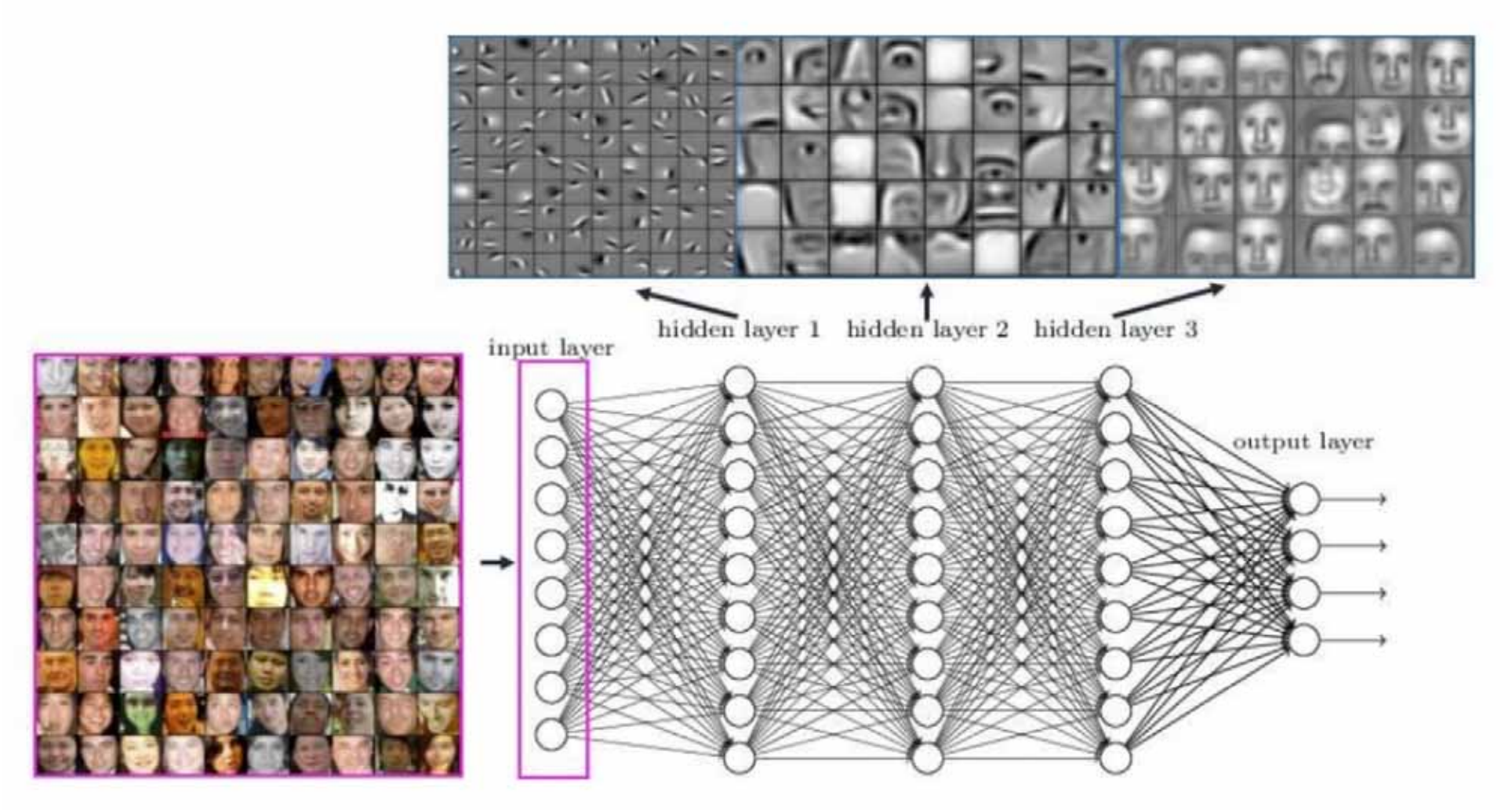
Matthew D. Zeiler & Rob Fergus 2013. Visualizing and Understanding Convolutional Networks. arXiv:1311.2901.

The world is compositional (Yann LeCun)

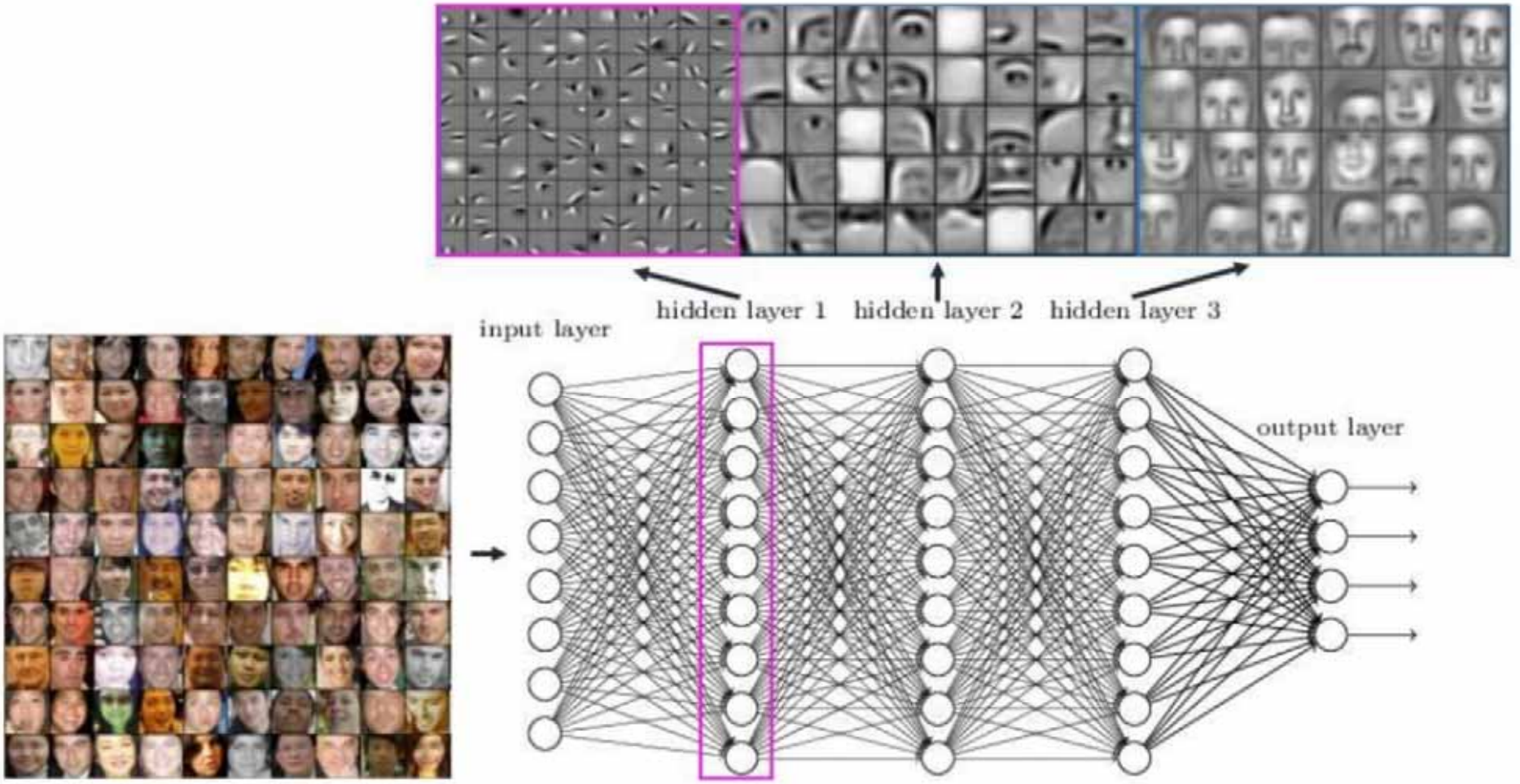


Matthew D. Zeiler & Rob Fergus 2013. Visualizing and Understanding Convolutional Networks. arXiv:1311.2901

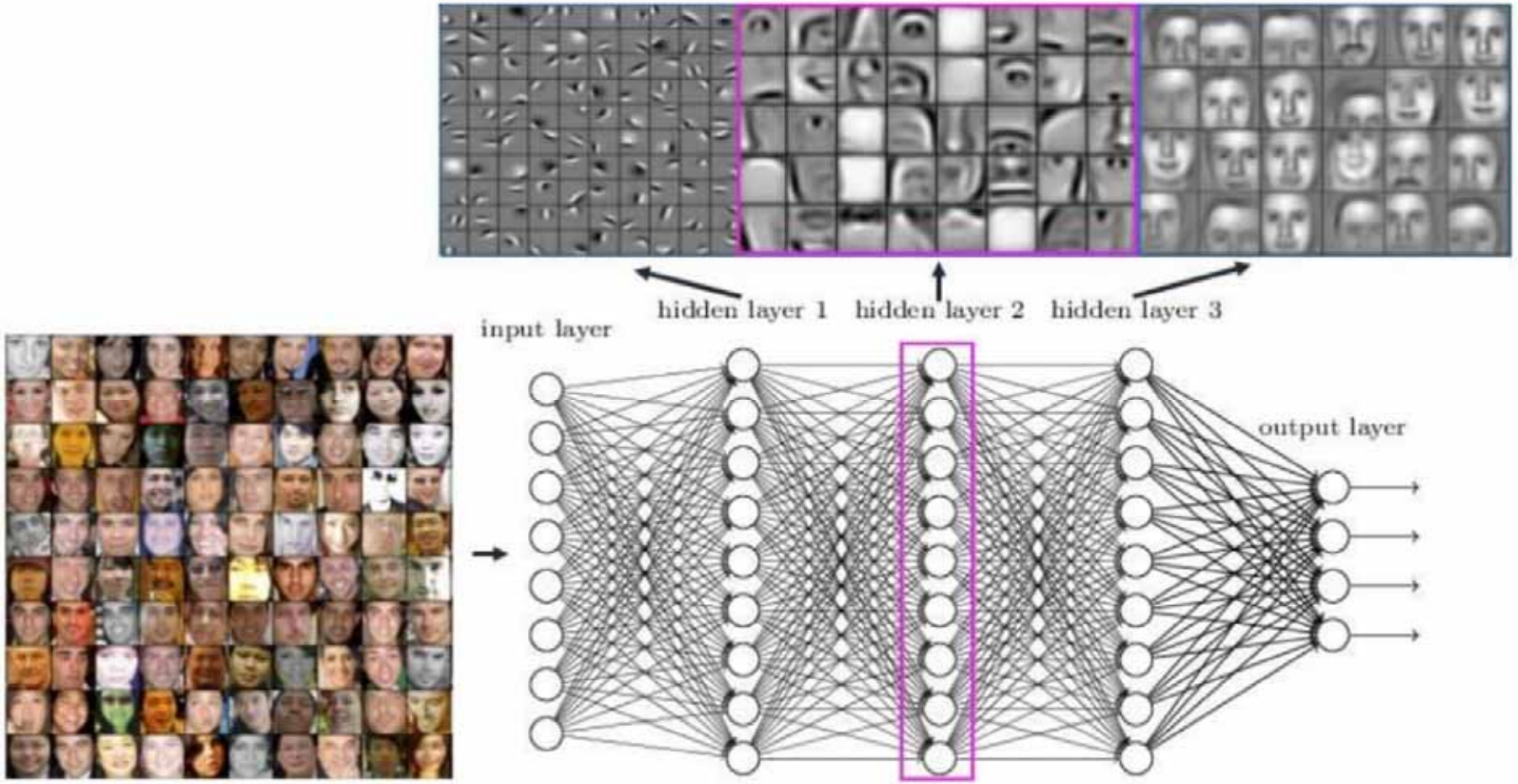
The world is compositional (Yann LeCun)



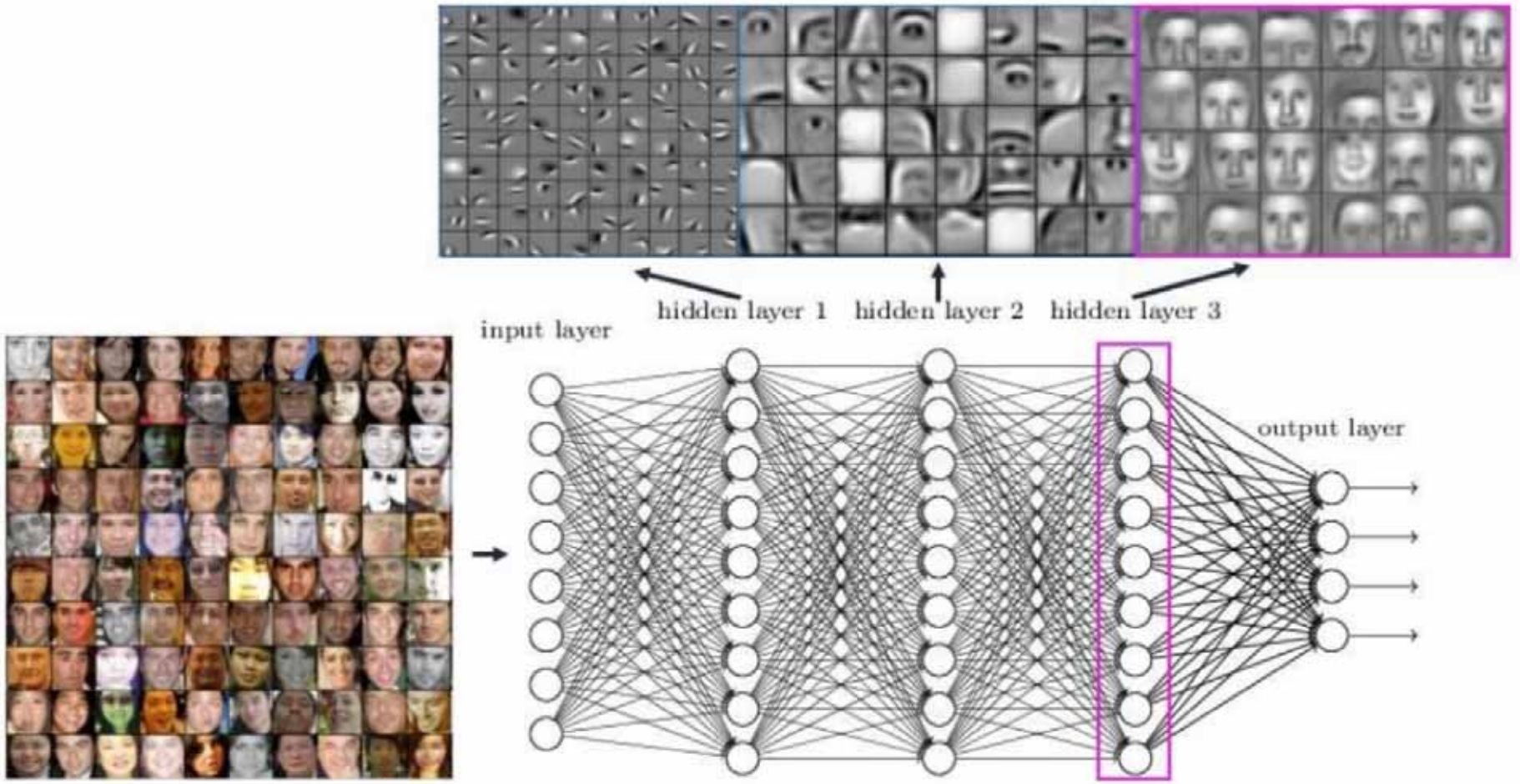
Matthew D. Zeiler & Rob Fergus 2013. Visualizing and Understanding Convolutional Networks. arXiv:1311.2901



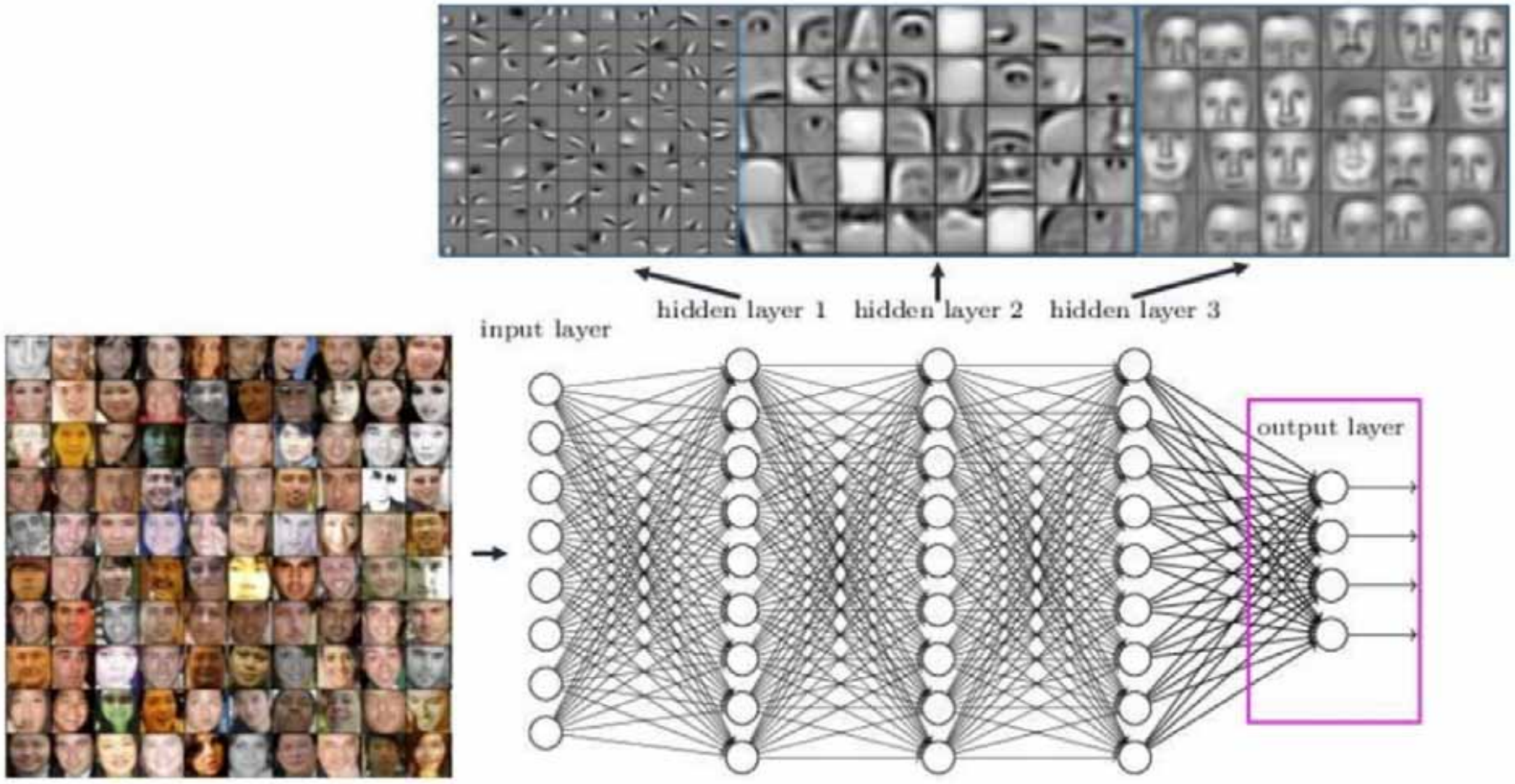
Matthew D. Zeiler & Rob Fergus 2013. Visualizing and Understanding Convolutional Networks. arXiv:1311.2901



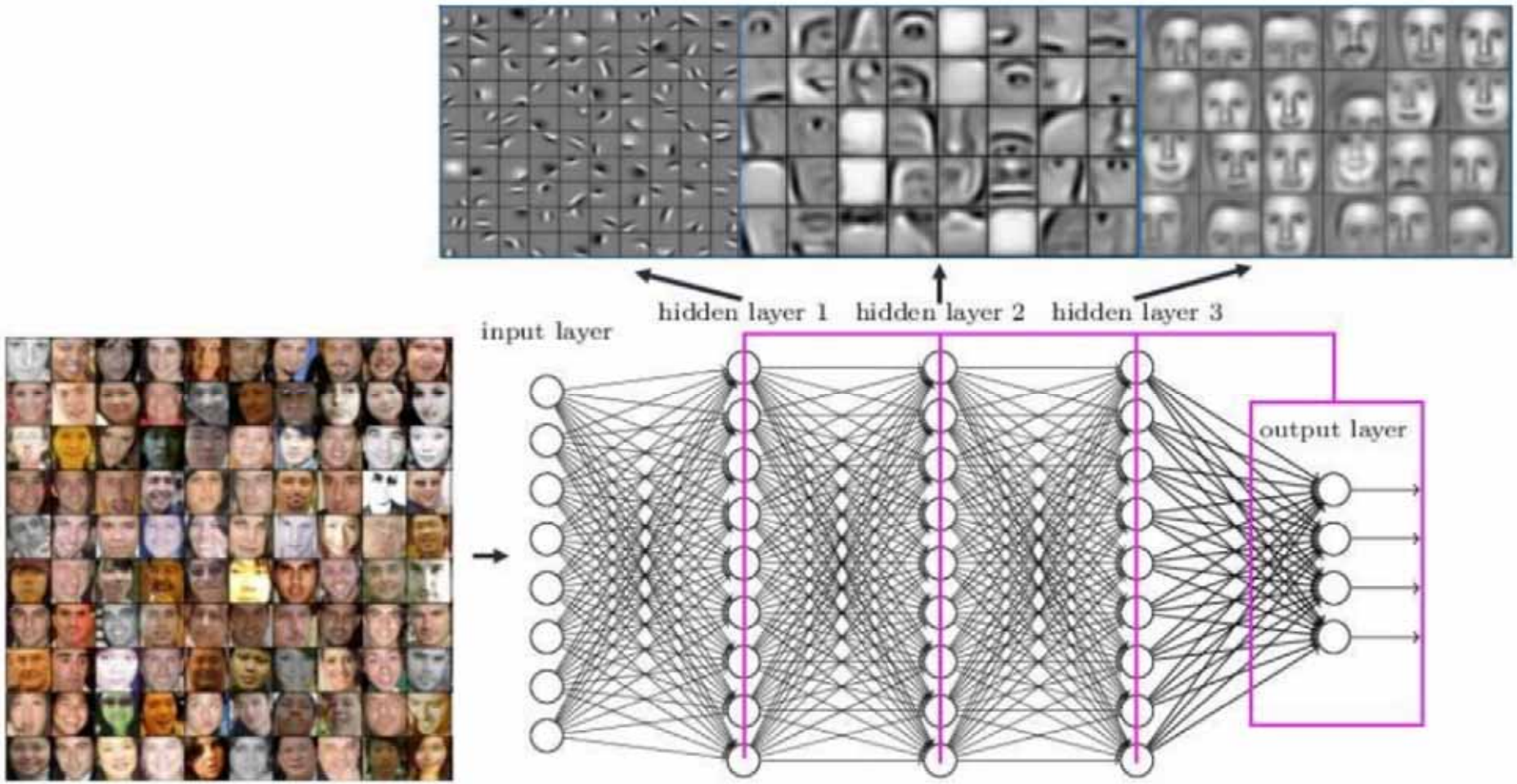
Matthew D. Zeiler & Rob Fergus 2013. Visualizing and Understanding Convolutional Networks. arXiv:1311.2901



Matthew D. Zeiler & Rob Fergus 2013. Visualizing and Understanding Convolutional Networks. arXiv:1311.2901



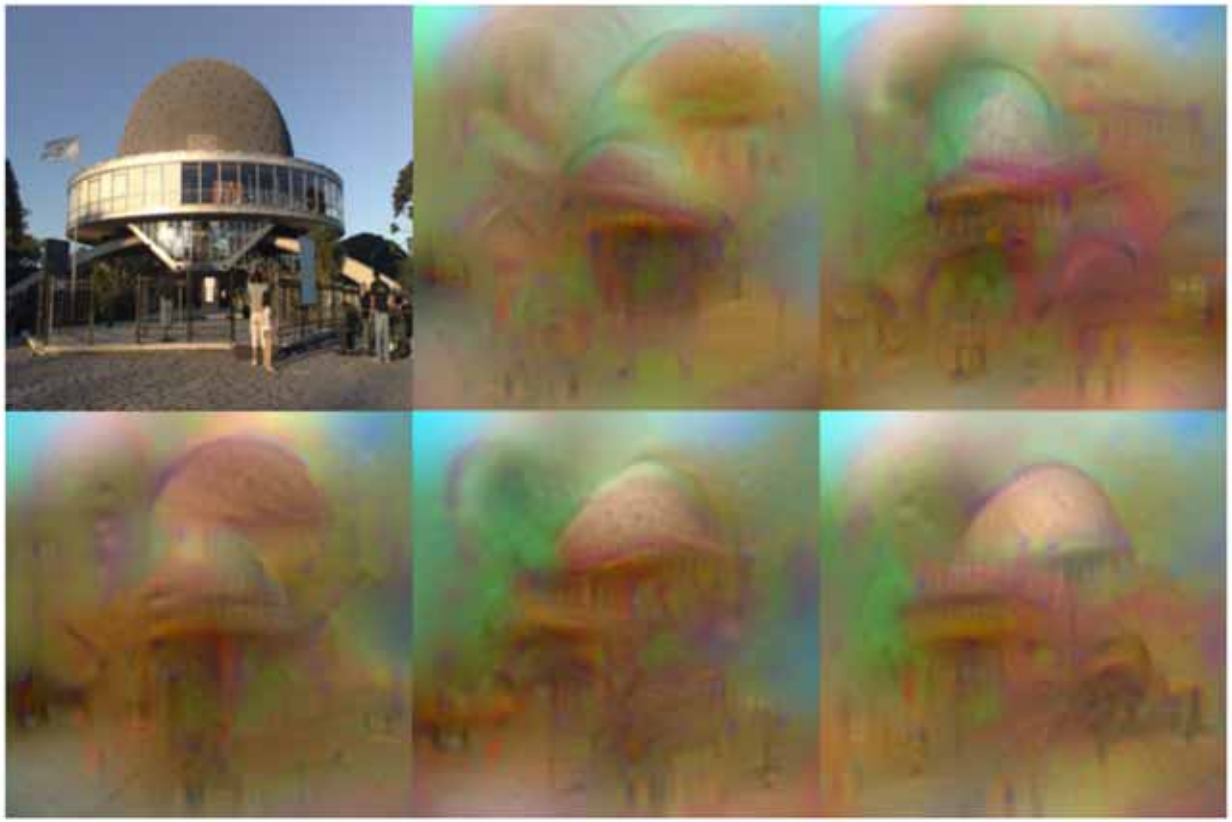
Matthew D. Zeiler & Rob Fergus 2013. Visualizing and Understanding Convolutional Networks. arXiv:1311.2901



Matthew D. Zeiler & Rob Fergus 2013. Visualizing and Understanding Convolutional Networks. arXiv:1311.2901

02 Inverting Convolutional Neural Networks

What is encoded by a CNN ? (before softmax)



$$\mathbf{x}^* = \underset{\mathbf{x} \in \mathbb{R}^{H \times W \times C}}{\operatorname{argmin}} \ \ell(\Phi(\mathbf{x}), \Phi_0) + \lambda \mathcal{R}(\mathbf{x})$$

Aravindh Mahendran & Andrea Vedaldi. Understanding deep image representations by inverting them. Proceedings of the IEEE conference on computer vision and pattern recognition, 2015. 5188-5196, doi:10.1109/CVPR.2015.7299155.

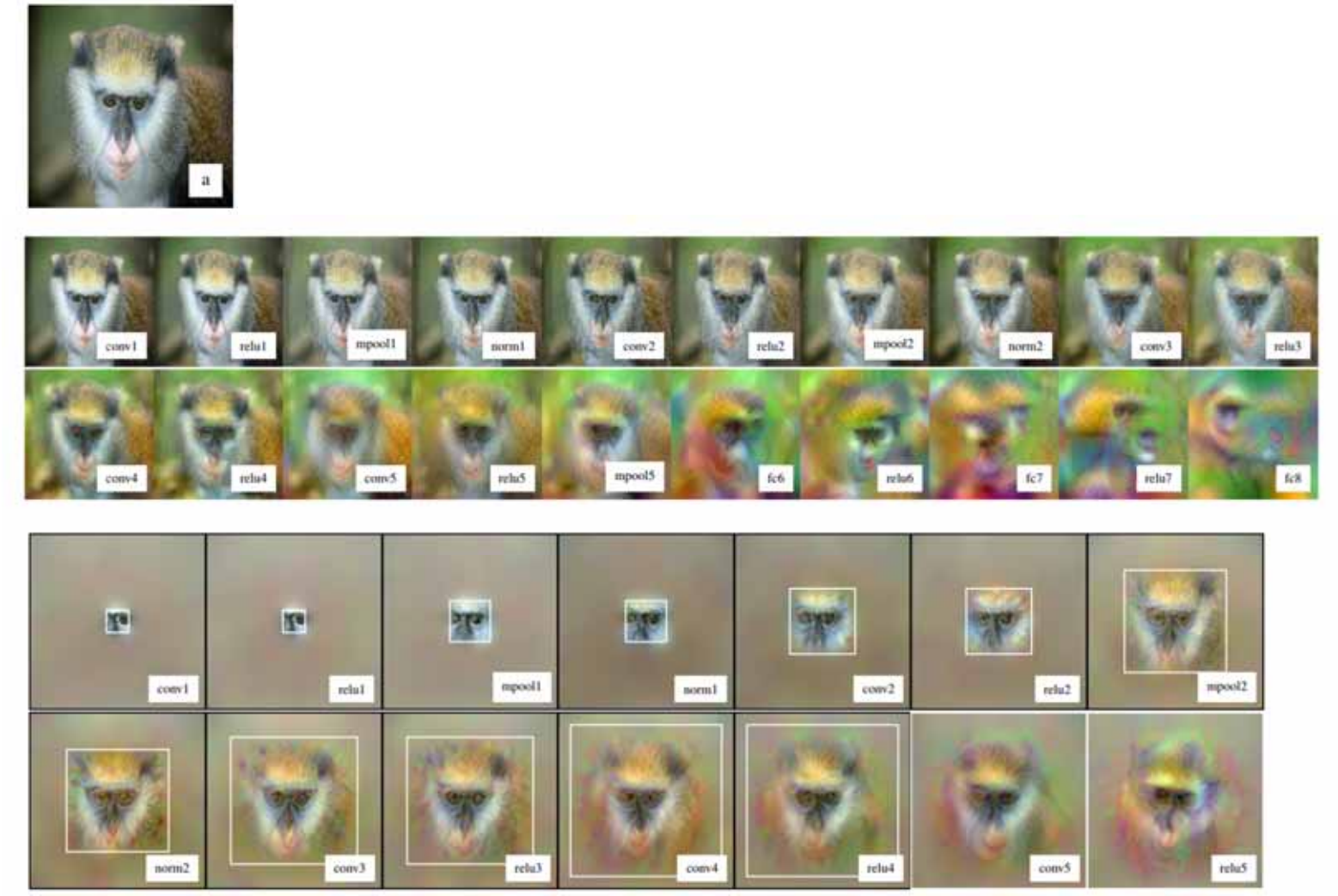
Loss function and total variation regularizer

$$\ell(\Phi(\mathbf{x}), \Phi_0) = \|\Phi(\mathbf{x}) - \Phi_0\|^2$$

$$\mathcal{R}_{V^\beta}(f) = \int_\Omega \left(\left(\frac{\partial f}{\partial u}(u,v) \right)^2 + \left(\frac{\partial f}{\partial v}(u,v) \right)^2 \right)^{\frac{\beta}{2}} du\, dv$$

$$\mathcal{R}_{V^\beta}(\mathbf{x}) = \sum_{i,j} \left((x_{i,j+1} - x_{ij})^2 + (x_{i+1,j} - x_{ij})^2 \right)^{\frac{\beta}{2}}$$

Aravindh Mahendran & Andrea Vedaldi. Understanding deep image representations by inverting them. Proceedings of the IEEE conference on computer vision and pattern recognition, 2015. 5188-5196, doi:10.1109/CVPR.2015.7299155.



Aravindh Mahendran & Andrea Vedaldi. Understanding deep image representations by inverting them. Proceedings of the IEEE conference on computer vision and pattern recognition, 2015. 5188-5196, doi:10.1109/CVPR.2015.7299155.

03 Guided Backpropagation

CIFAR-10 classification error (without data augmentation)

Model	Error (%)	# parameters
without data augmentation		
Model A	12.47%	≈ 0.9 M
Strided-CNN-A	13.46%	≈ 0.9 M
ConvPool-CNN-A	**10.21%**	≈ 1.28 M
ALL-CNN-A	10.30%	≈ 1.28 M
Model B	10.20%	≈ 1 M
Strided-CNN-B	10.98%	≈ 1 M
ConvPool-CNN-B	9.33%	≈ 1.35 M
ALL-CNN-B	**9.10%**	≈ 1.35 M
Model C	9.74%	≈ 1.3 M
Strided-CNN-C	10.19%	≈ 1.3 M
ConvPool-CNN-C	9.31%	≈ 1.4 M
ALL-CNN-C	**9.08%**	≈ 1.4 M

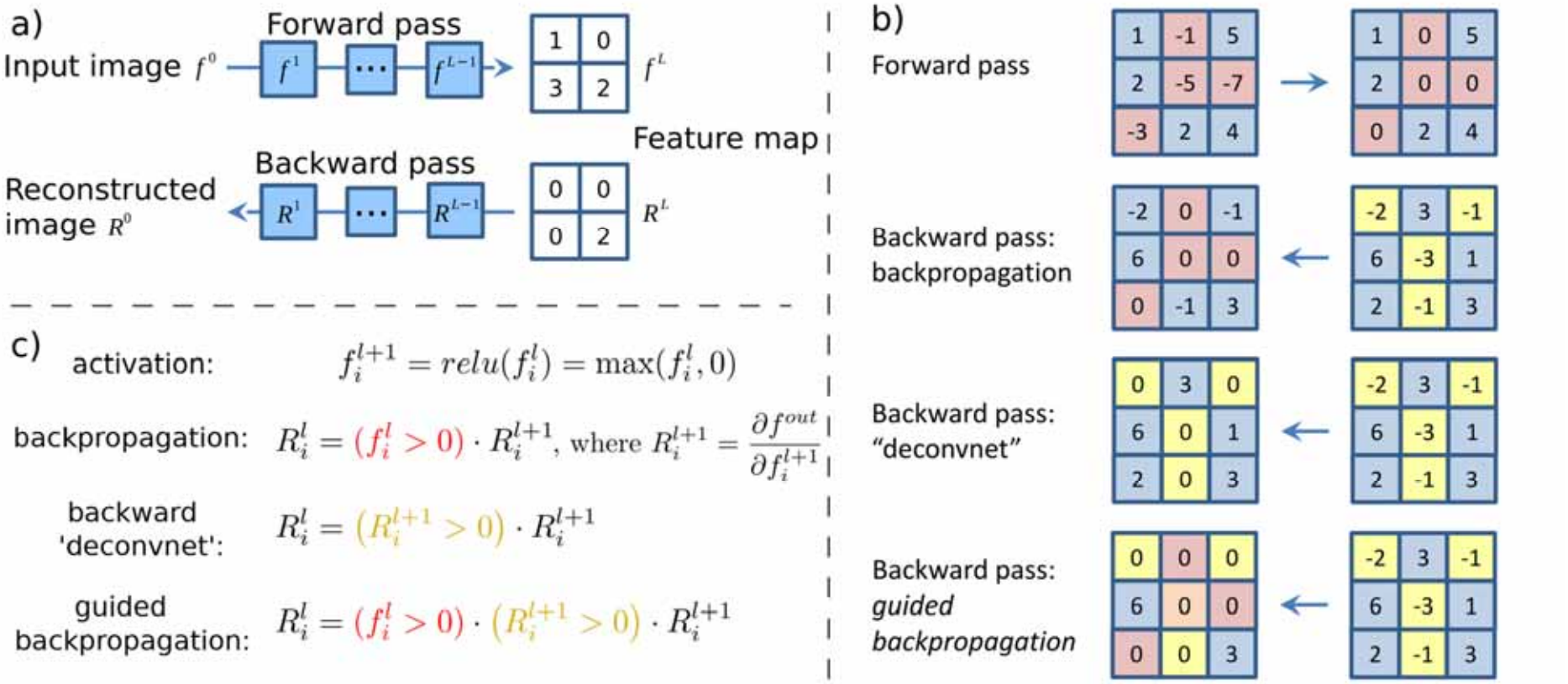
Method	Error (%)	# params
without data augmentation		
Maxout [1]	11.68%	> 6 M
Network in Network [2]	10.41%	≈ 1 M
Deeply Supervised [3]	9.69%	≈ 1 M
ALL-CNN (Ours)	**9.08%**	≈ 1.3 M
with data augmentation		
Maxout [1]	9.38%	> 6 M
DropConnect [2]	9.32%	-
dasNet [4]	9.22%	> 6 M
Network in Network [2]	8.81%	≈ 1 M
Deeply Supervised [3]	7.97%	≈ 1 M
ALL-CNN (Ours)	**7.25%**	≈ 1.3 M

Method	Error (%)
CNN + tree prior [5]	36.85%
Network in Network [2]	35.68%
Deeply Supervised [3]	34.57%
Maxout (larger) [4]	34.54%
dasNet [4]	33.78%
ALL-CNN (Ours)	33.71%
Fractional Pooling (1 test) [6]	**31.45%**
Fractional Pooling (12 tests) [6]	**26.39%**

Method	Error (%)
with large data augmentation	
Spatially Sparse CNN [6]	4.47%
Large ALL-CNN (Ours)	4.41%
Fractional Pooling (1 test) [6]	4.50%
Fractional Pooling (100 tests) [6]	**3.47%**

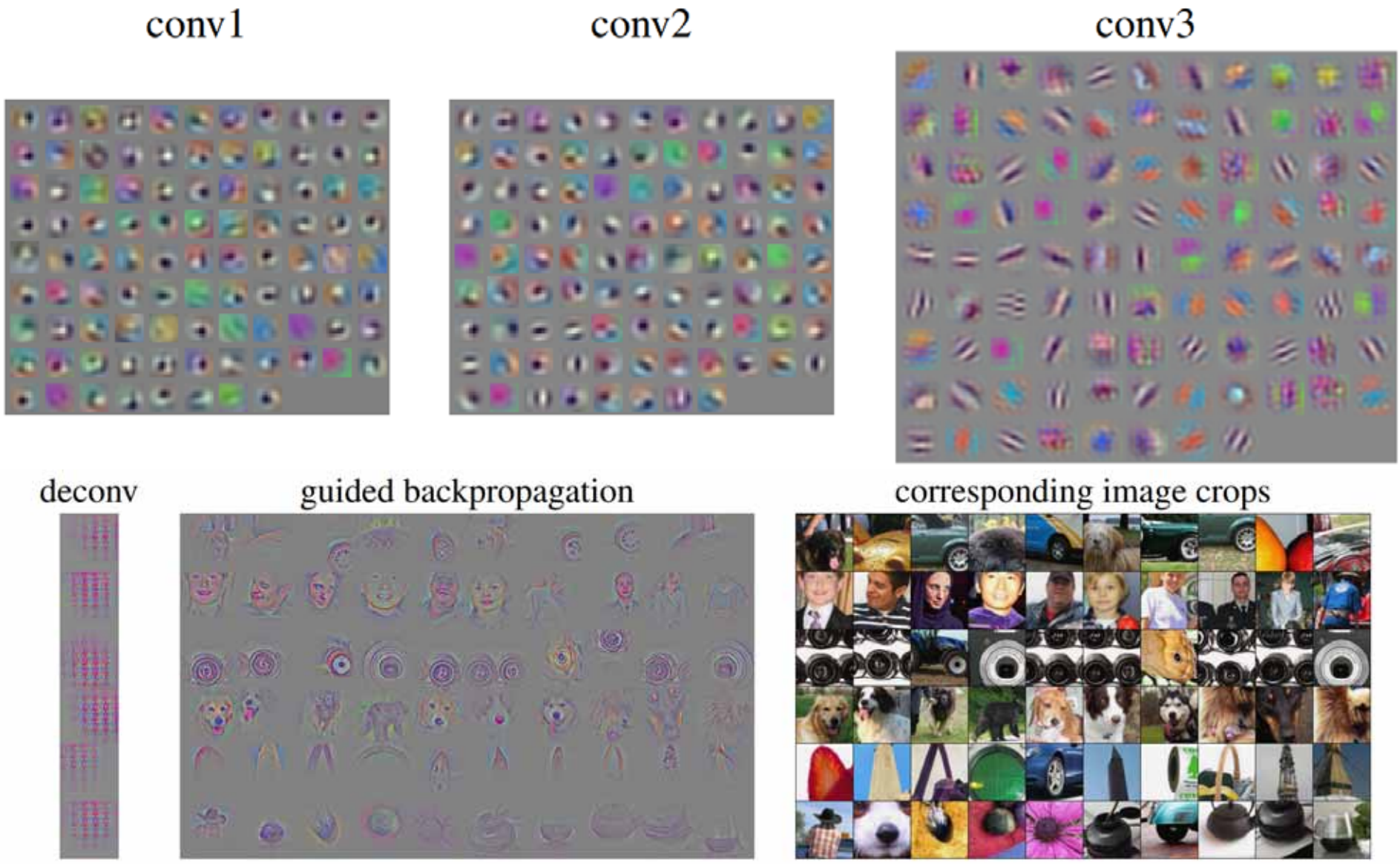
Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox & Martin Riedmiller 2014. Striving for simplicity: The all convolutional net. arXiv:1412.6806.

Visualising the activations of high layer neurons



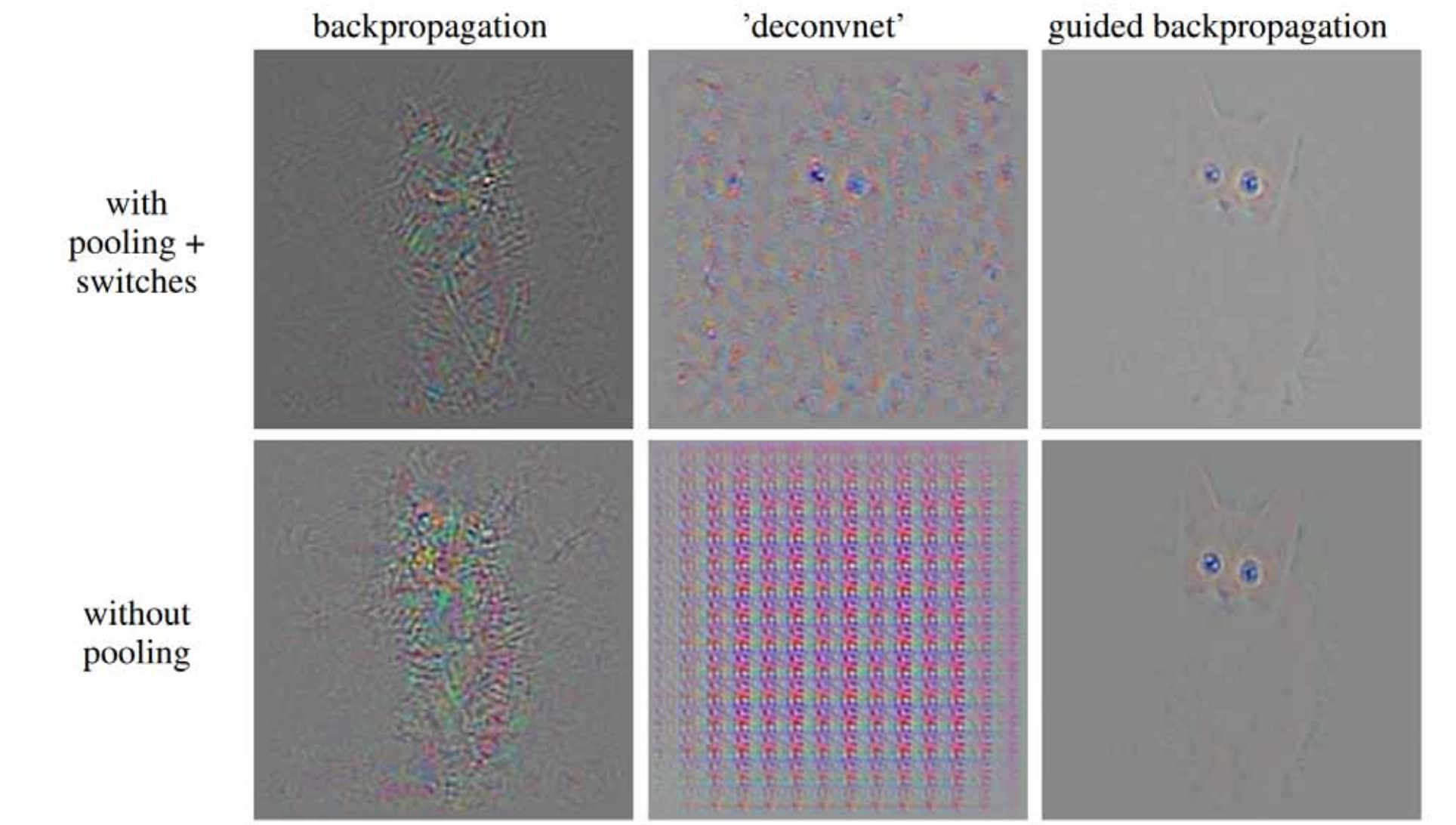
Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox & Martin Riedmiller 2014. Striving for simplicity: The all convolutional net. arXiv:1412.6806.

Guided Backpropagation



Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox & Martin Riedmiller 2014. Striving for simplicity: The all convolutional net. arXiv:1412.6806.

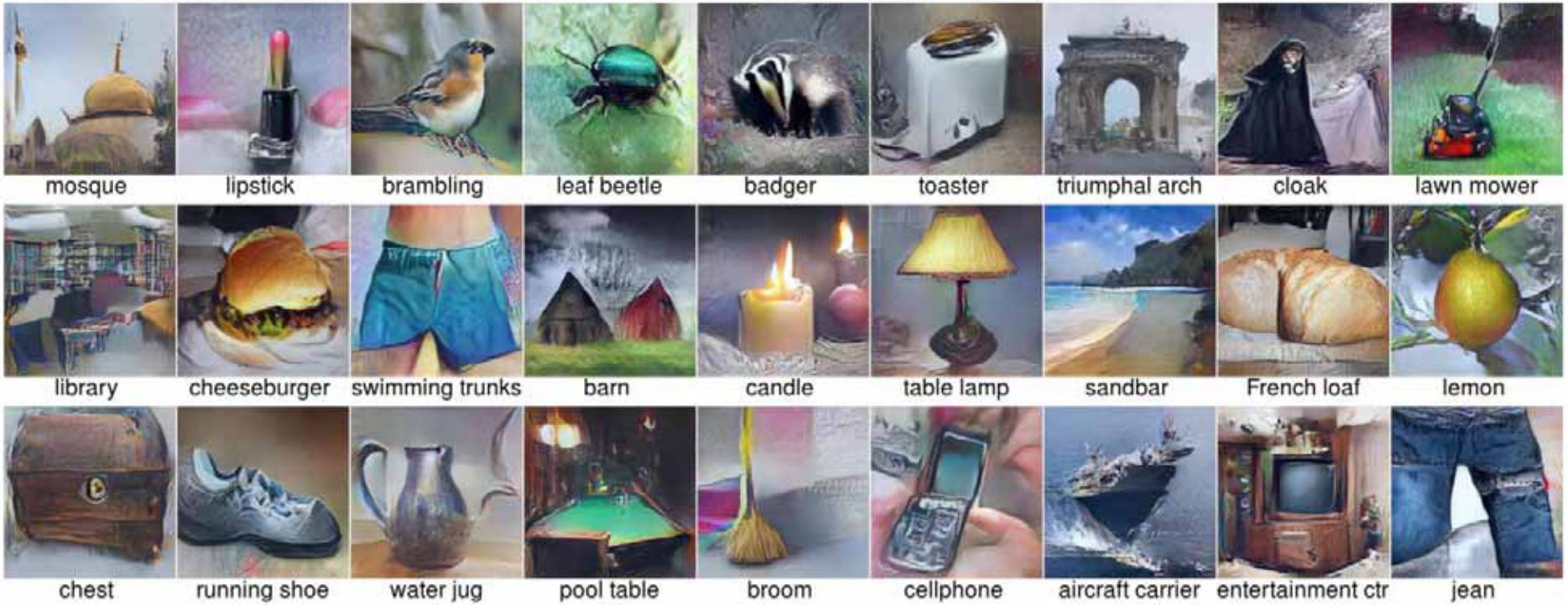
Vis of descriptive image regions



Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox & Martin Riedmiller 2014. Striving for simplicity: The all convolutional net. arXiv:1412.6806.

04 Deep Generator Networks (DGN)

Learning generative models of natural images

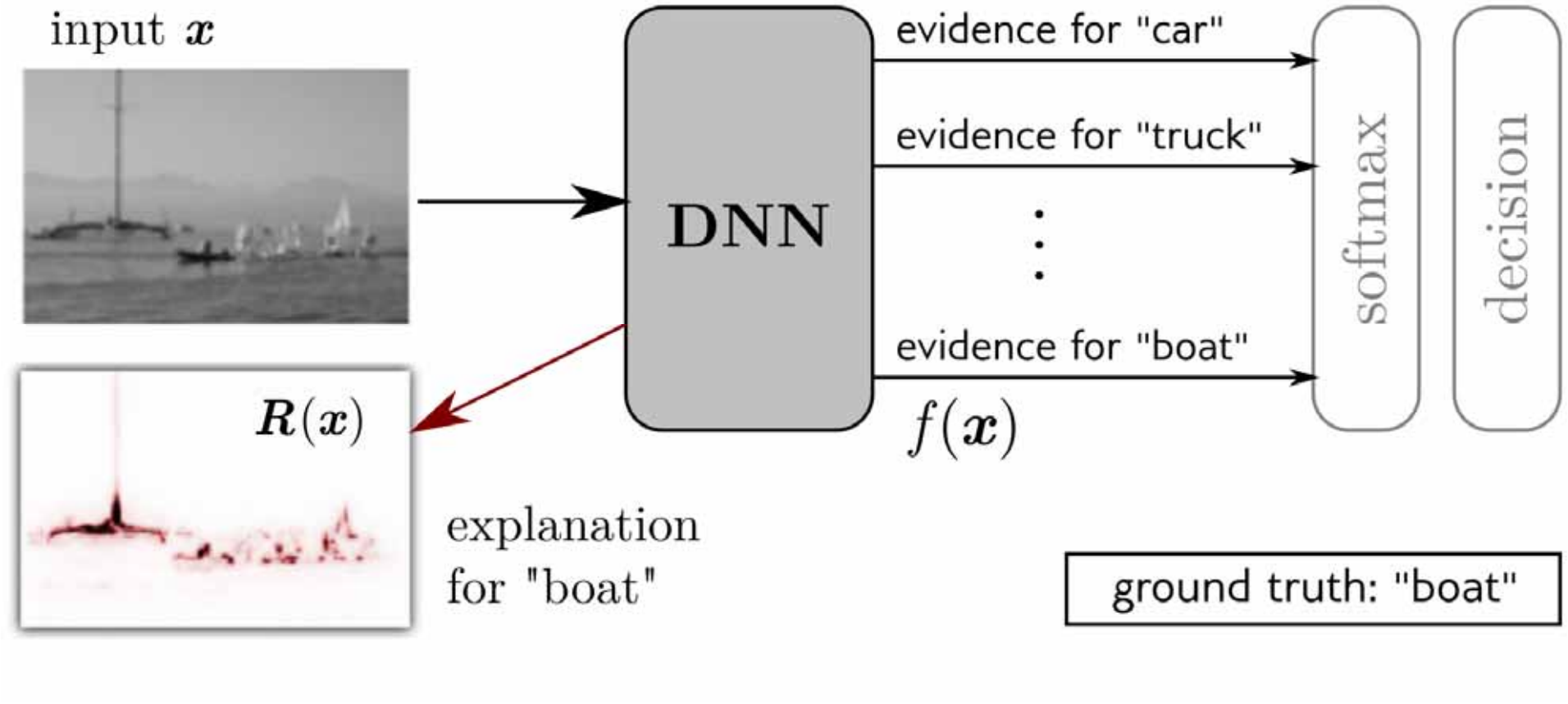


- Idea: incorporating a generative model in the activation maximization framework and redefine the optimization problem as:

$$\max_{z \in \mathcal{Z}} \log p(\omega_c \mid g(\boldsymbol{z})) - \lambda \|\boldsymbol{z}\|^2$$

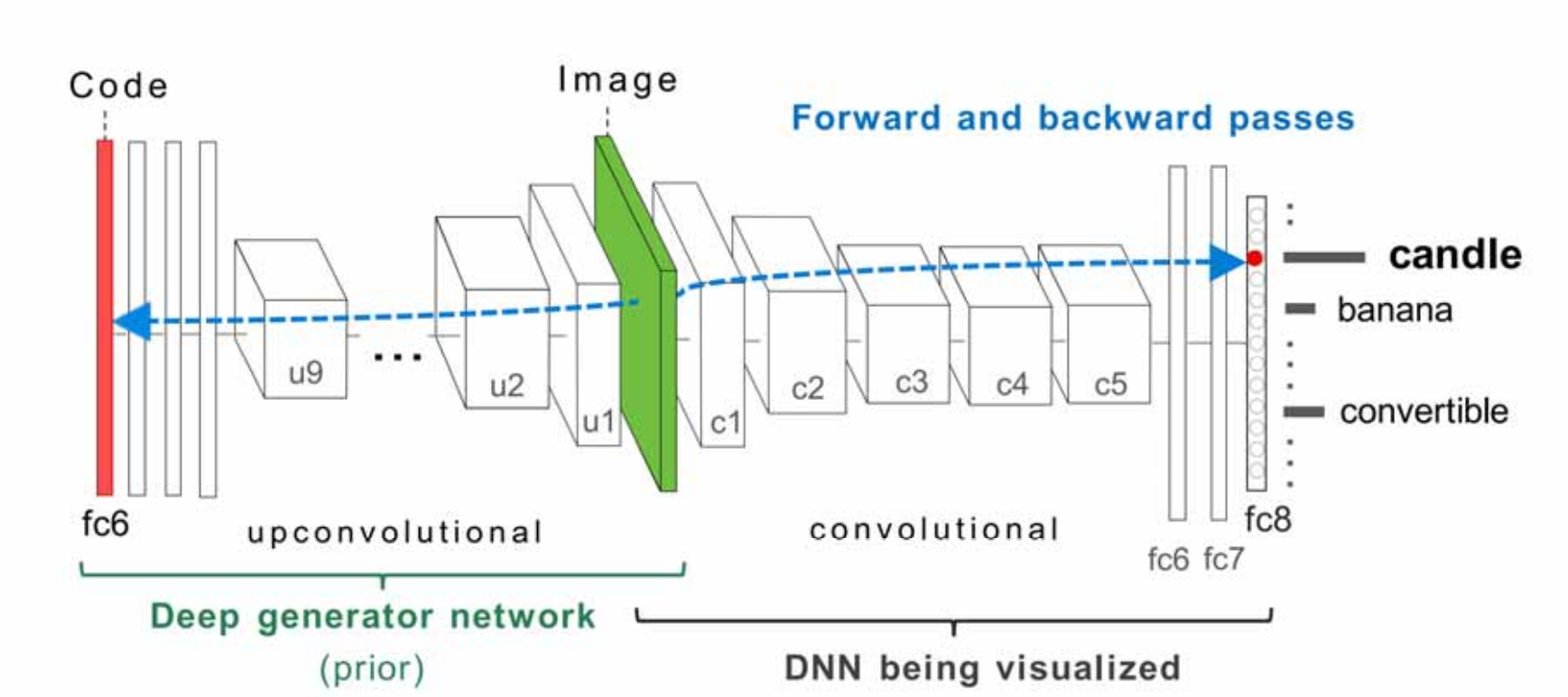
Anh Nguyen, Alexey Dosovitskiy, Jason Yosinski, Thomas Brox & Jeff Clune. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. Advances in Neural Information Processing Systems (NIPS 2016), 2016 Barcelona. 3387-3395.

Basic principle of explanation of prediction f(x)



Grégoire Montavon, Wojciech Samek & Klaus-Robert Müller 2018. Methods for interpreting and understanding deep neural networks. Digital Signal Processing, 73, 1-15, doi:10.1016/j.dsp.2017.10.011.

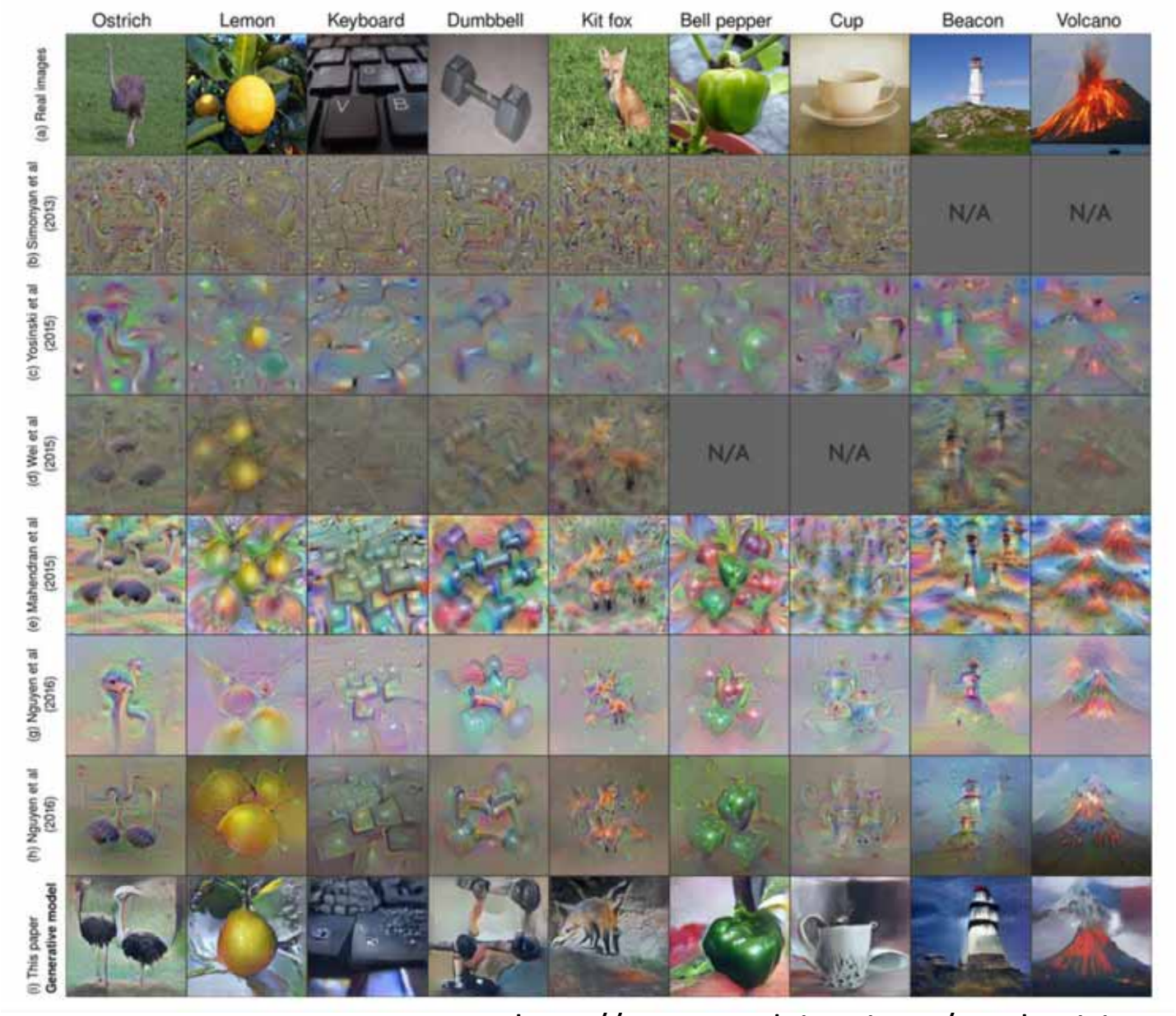
Learning generative models of natural images



$$\widehat{\mathbf{y}^l} = \arg\max_{\mathbf{y}^l} (\Phi_h(G_l(\mathbf{y}^l)) - \lambda \|\mathbf{y}^l\|)$$

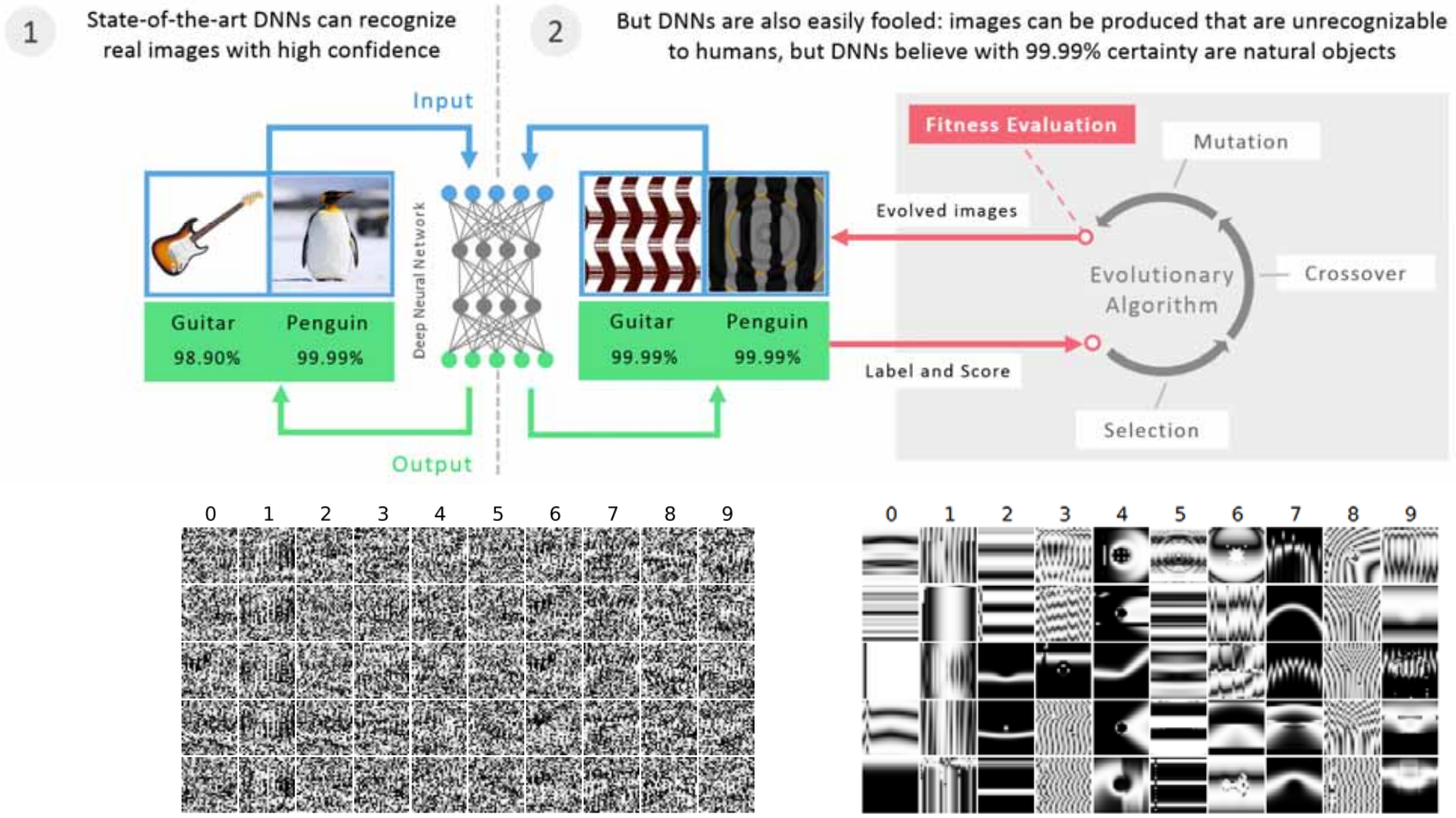
Anh Nguyen, Alexey Dosovitskiy, Jason Yosinski, Thomas Brox & Jeff Clune. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. Advances in Neural Information Processing Systems (NIPS 2016), 2016 Barcelona. 3387-3395.

Comparison of methods to DGN



http://www.evolvingai.org/synthesizing

Example for Explanation Needed: adversarials



Anh Nguyen, Jason Yosinski & Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015 Boston (MA). IEEE, 427-436.

http://www.evolvingai.org/fooling

05 Testing with Concept Activation Vectors

- "It's not enough to know if a model works, we need to know how it works"

… if Sundar Pichai is saying this …

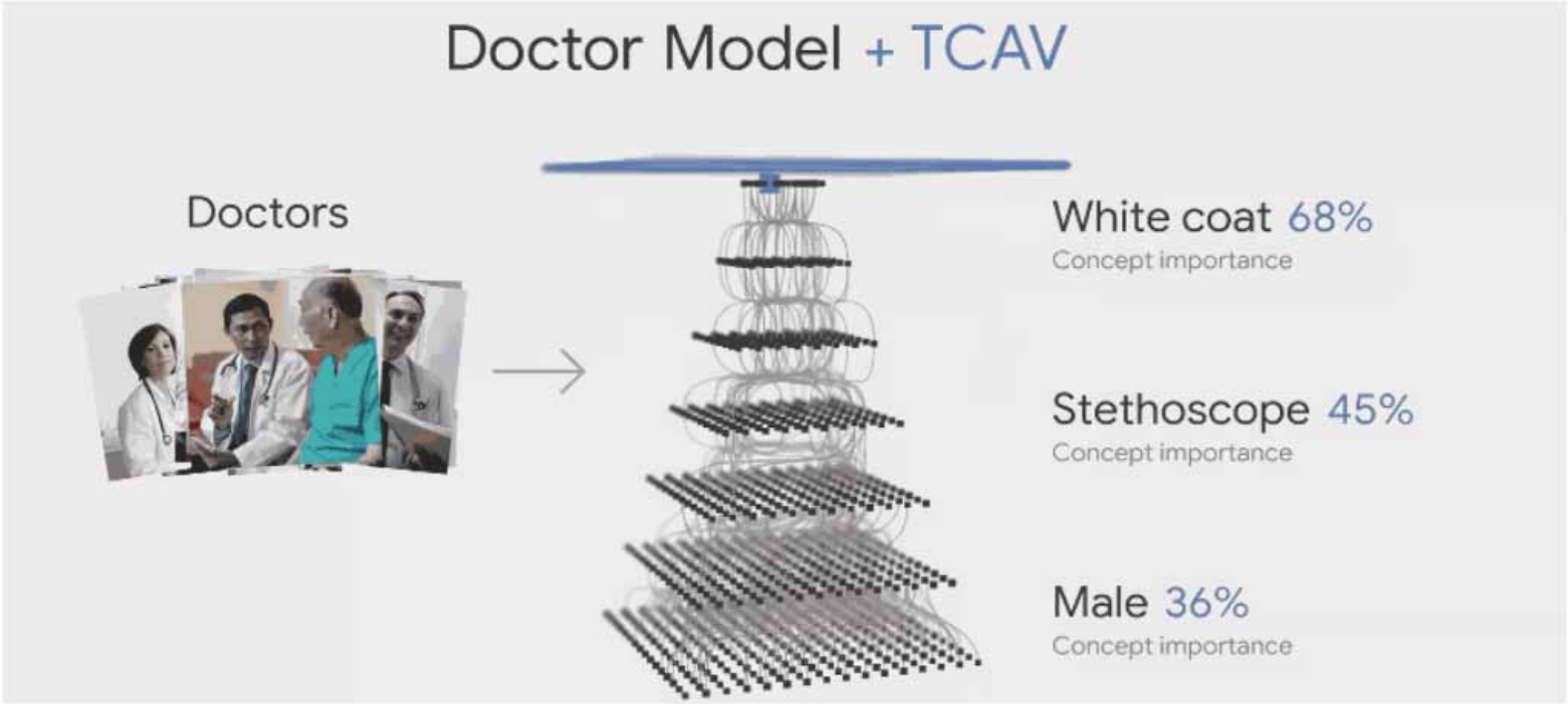
Sundar Pichai

Born	Pichai Sundararajan June 10, 1972 (age 47) Madurai, Tamil Nadu, India
Alma mater	Indian Institute of Technology Kharagpur (BTech) Stanford University (MS) The Wharton School (MBA)
Salary	US$1,881,066 (2018) US$1,333,557 (2017)[1] US$199.7 million[2] (2016)
Title	CEO of Google and Alphabet
Board member of	Alphabet Inc.[3] Magic Leap[4]

TCAV at a glance

- ML models work on **low-level features** (edges, lines, pixel, …)
- Humans are working on **high-level concepts** (shape, size, …)
- Every pixel of an image is a input feature and are numbers, which do not make sense to humans.
- TCAV enables to provide an explanation that is generally true for a class of interest, beyond one image (global explanation).
- TCAV learns ‘concepts’ from examples. For instance, TCAV needs a couple of examples of ‘female’, and something ‘not female’ to learn a “gender” concept. The goal of TCAV is to determine how much a concept (e.g., gender, race) was important for a prediction in a trained model even if the concept was not part of the training.

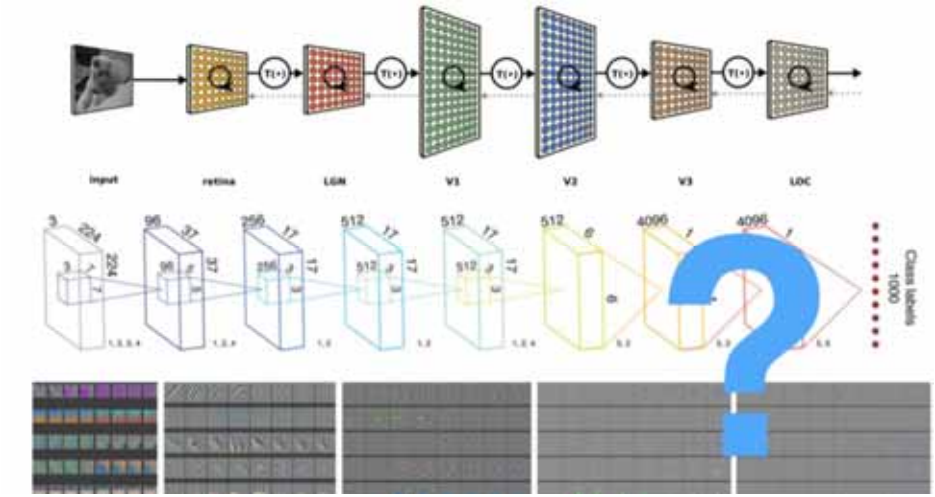
Top-Level intro



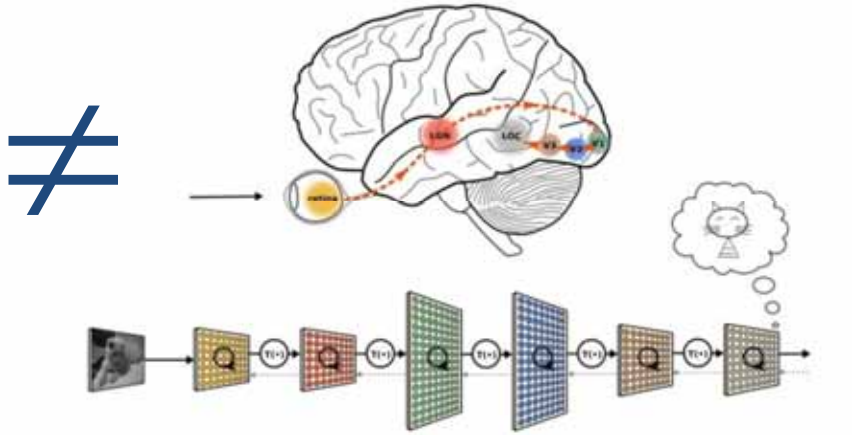
TCAV shows that being male is an important 'concept' when deciding if an image belongs to a doctor or not

https://towardsdatascience.com/tcav-interpretability-beyond-feature-attribution-79b4d3610b4d

Example: Concept Activation Vector (CAV)



Yann Lecun, Yoshua Bengio & Geoffrey Hinton 2015. Deep learning. Nature, 521, (7553), 436-444, doi:10.1038/nature14539.



$$\frac{\partial h_k(x)}{\partial x_{a,b}}$$

Humans work in another vector space which is spanned by **implicit knowledge** vectors corresponding to an unknown set of human interpretable concepts.

$$S_{C,k,l}(x) = \lim_{\epsilon \to 0} \frac{h_{l,k}\left(f_l(x) + \epsilon v_C^l\right) - h_{l,k}\left(f_l(x)\right)}{\epsilon} = \nabla h_{l,k}\left(f_l(x)\right) \cdot v_C^l$$

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler & Fernanda Viegas. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors, ICML, 2018. 2673-2682.

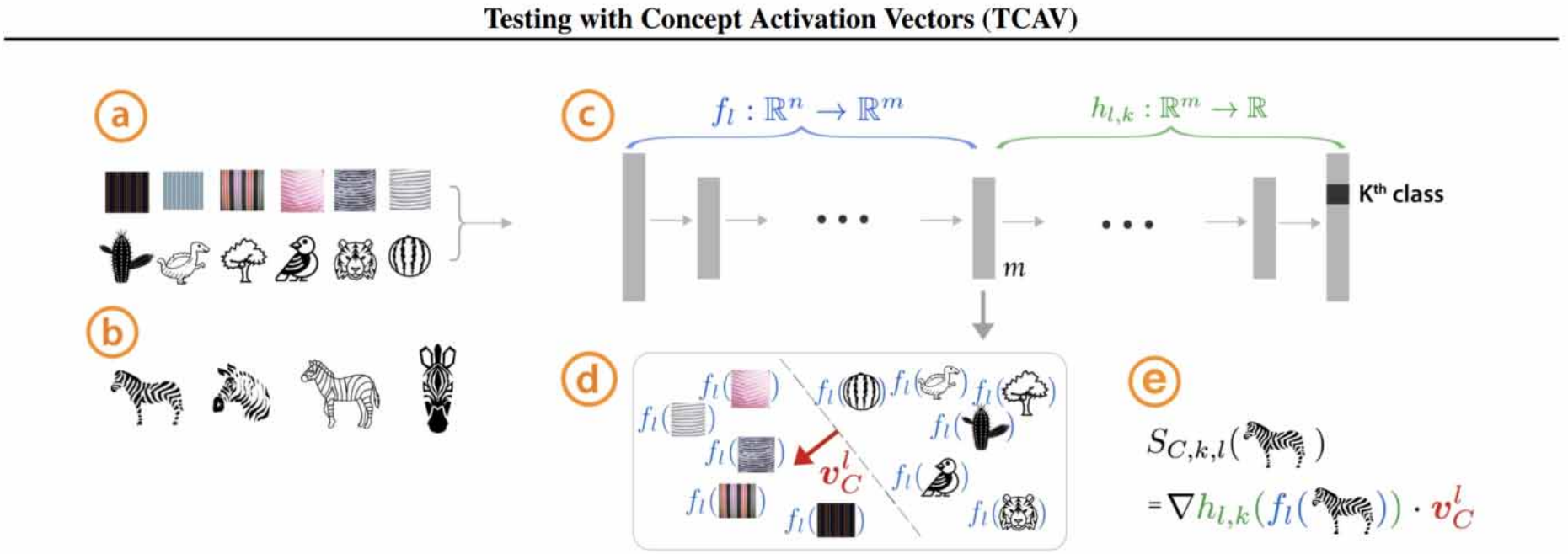


Figure 1. **Testing with Concept Activation Vectors:** Given a user-defined set of examples for a concept (e.g., 'striped'), and random examples ⓐ, labeled training-data examples for the studied class (zebras) ⓑ, and a trained network ⓒ, TCAV can quantify the model's sensitivity to the concept for that class. CAVs are learned by training a linear classifier to distinguish between the activations produced by a concept's examples and examples in any layer ⓓ. The CAV is the vector orthogonal to the classification boundary (v_C^l, red arrow). For the class of interest (zebras), TCAV uses the directional derivative $S_{C,k,l}(\boldsymbol{x})$ to quantify conceptual sensitivity ⓔ.

https://github.com/tensorflow/tcav

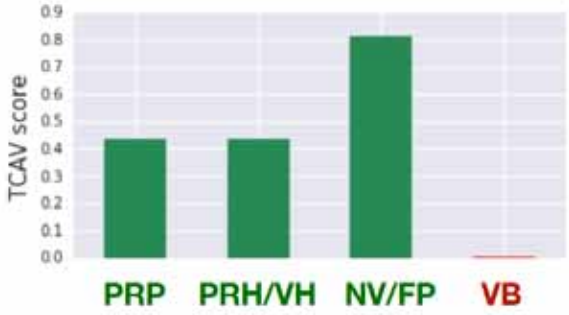
Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler & Fernanda Viegas. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). International Conference on Machine Learning (ICML), 2018. 2673-2682.

TCAV Application example: Diabetic retinopathy

Prediction class	Prediction accuracy
DR level 4	High

Example

TCAV scores



TCAV shows the model is **consistent** with doctor's knowledge when model is **accurate**

DR level 1	Med

TCAV for DR level 1

TCAV shows the model is **inconsistent** with doctor's knowledge for classes when model is less accurate

Green: domain expert's label on concepts belong to the level
Red: domain expert's label on concepts does not belong to the level

A controlled experiment with ground truth

Construction of 2 TCAVS:

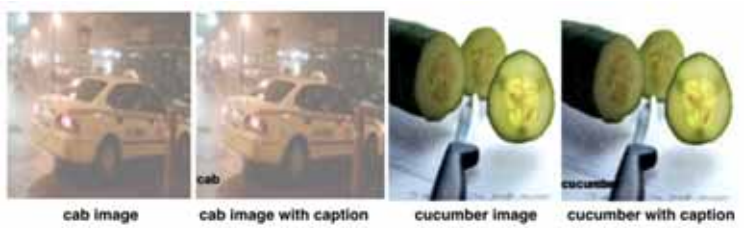
- Image concept
- Caption concept

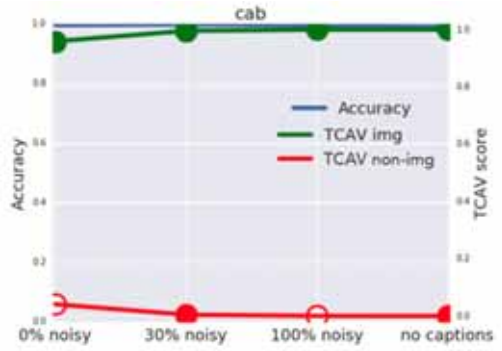
Caption is not always correct:

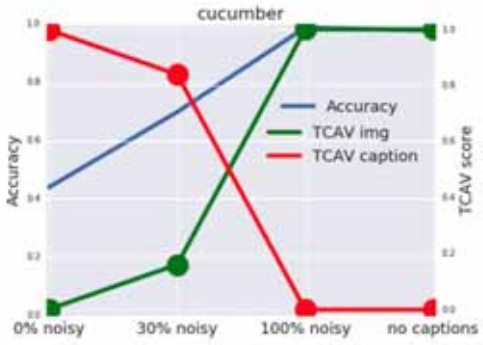
- Model with 0% noise: correct caption
- Model with 100% noise: always erroneous caption

Which concept is the network using?

- TCAVQ closely matches ground truth.
- For cab images, the network used image concept more than the caption concept, regardless of % of noise.



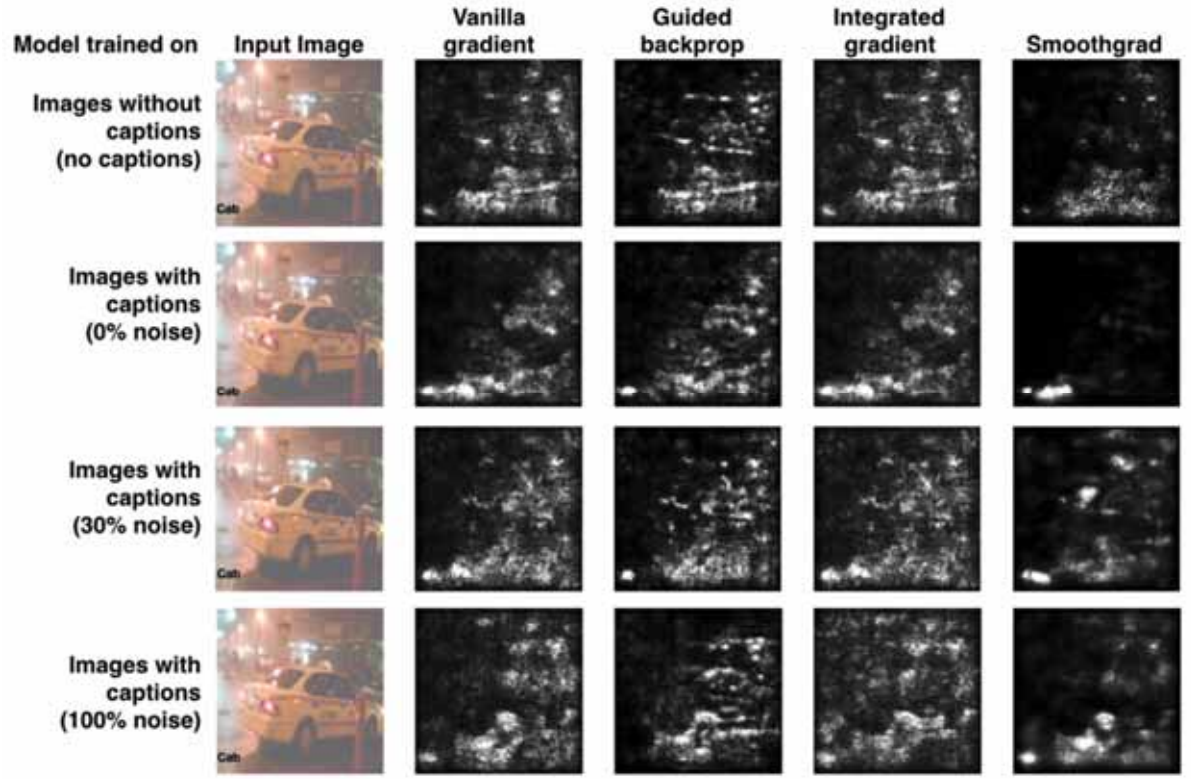


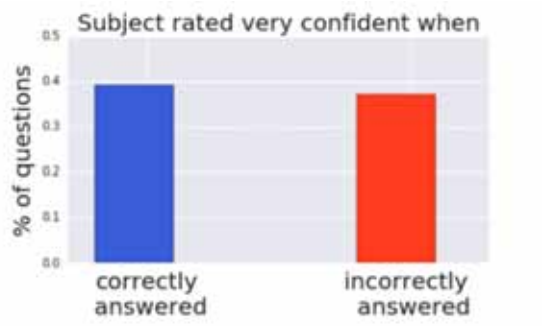


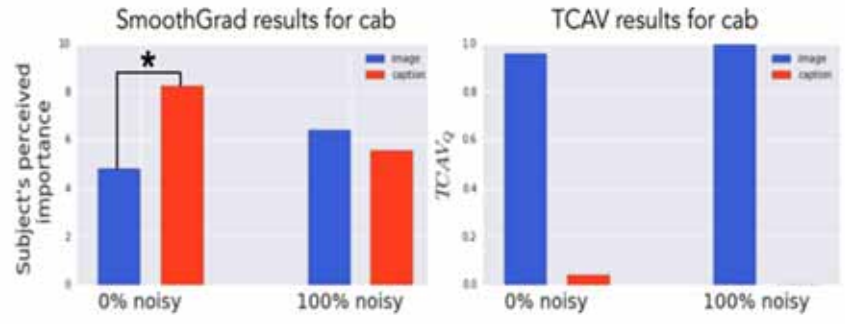
Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler & Fernanda Viegas. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). International Conference on Machine Learning (ICML), 2018. 2673-2682.

A controlled experiment with ground truth

- Saliency map can not focus on caption ->
- Saliency map should focus on caption ->
- Saliency map possibly can focus on caption ->
- Saliency map should not focus on caption ->







Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler & Fernanda Viegas. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). International Conference on Machine Learning (ICML), 2018. 2673-2682.

- TCAV correct – image concept always important

- Note: Humans are very confident – even when they are wrong!
- Conclusion: Saliency maps may be misleading!

- Thank you – see you in the next lecture!