Apprentissage, réseaux de neurones et modèles graphiques (RCP209)
Deep Learning: open questions and perspectives

Nicolas Thome
Prenom.Nom@cnam.fr
`http://cedric.cnam.fr/vertigo/Cours/ml2/`

Département Informatique
Conservatoire Nationnal des Arts et Métiers (Cnam)

Outline

CNN and invariance

CNN and invariance

- Standard ConvNets: limited invariance capacity (small shifts)
- ImageNet: single centered object ≠ other datasets (VOC, MS COCO)

⇒ **Learn shift invariance: region alignment !**

⇒ **Deep learning + structured prediction !**

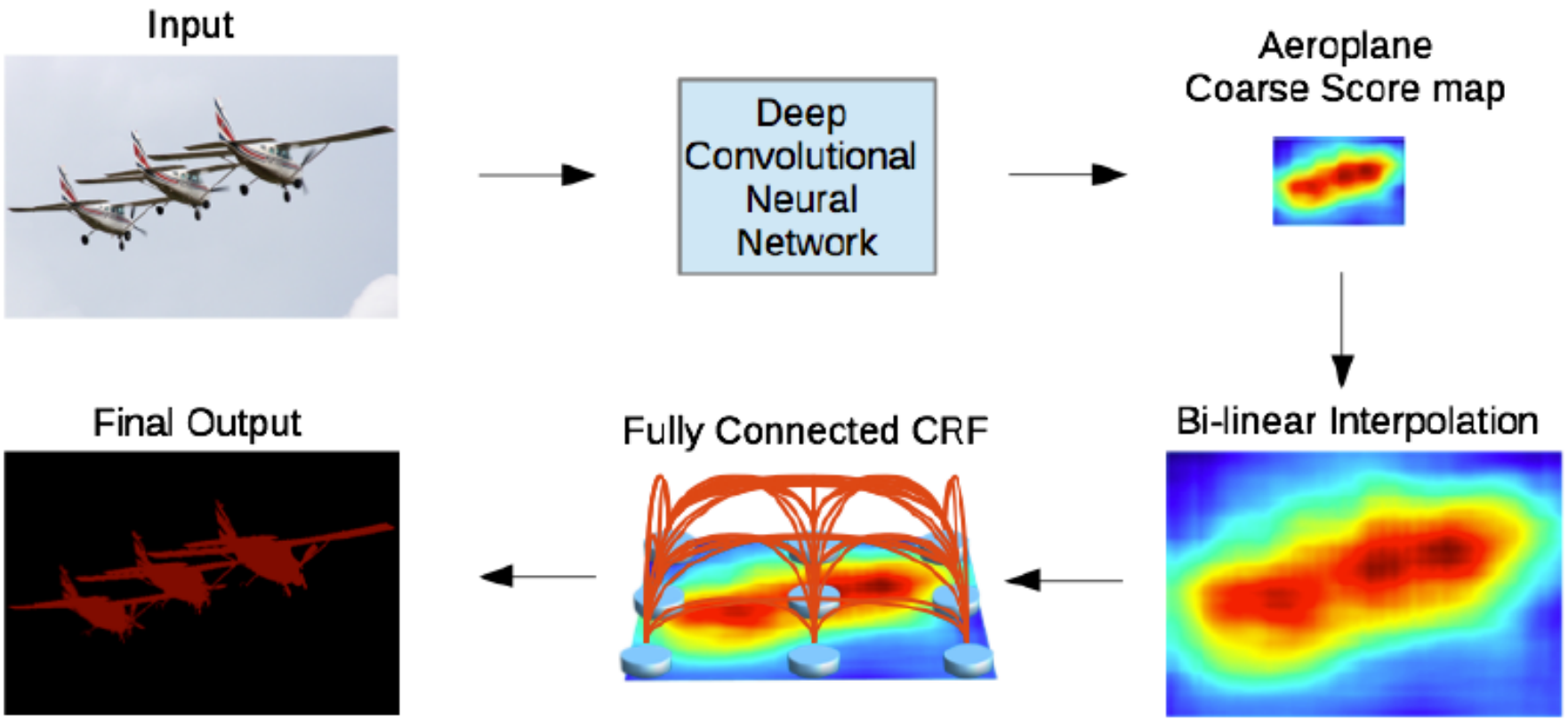
Deep Learning and Structured Prediction

- Structured prediction: graphical model in general (*e.g.* SSVM a specific model)
- Structured Prediction models (previous weeks): limited to log linear models with handcrafted features
- Combining Deep Learning & Structured Prediction
  - Solution: add a structured layer on top of your favorite deep model (*e.g.* ConvNet)
  - Issue : computational issue with Inference (and LAI for SSVM)
    - Methods for discrete outputs [CSYU15]
    - Recent models for continuous outputs [BM16, WFU16]
    - Approches to unroll inference: forward and backward passes of these deep structured models expressed as a set of standard layers [ZJRP$^+$15, BM16, WFU16]
      $\Rightarrow$ fast end-to-end training on GPUs.

Deep Learning and Structured Prediction

Ex: Semantic Segmentation

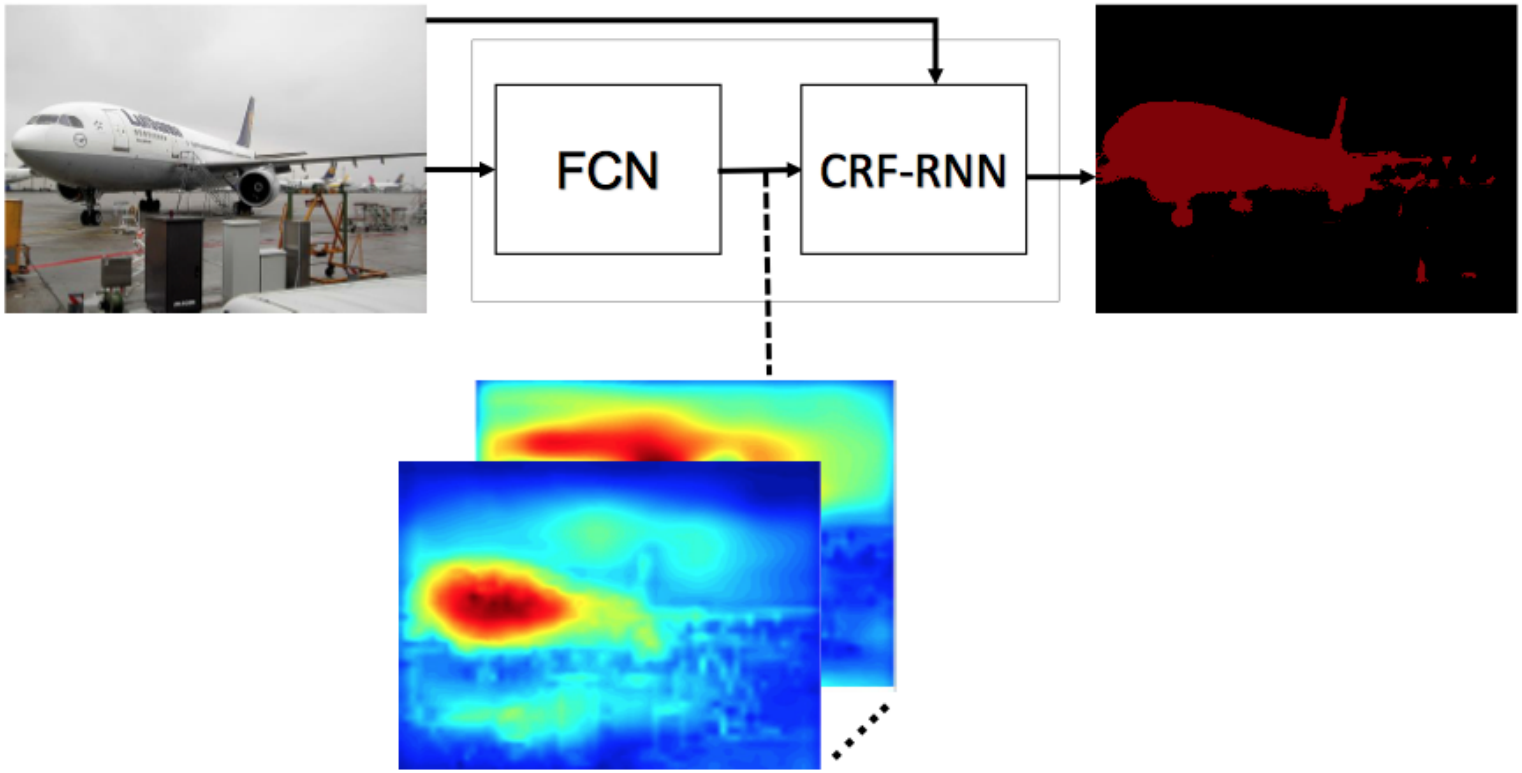
- Baseline model: DeepLab [CPK+15]: use Fully Convolutional Network (FCN) (fc layers cast as 1x1 conv + transfer layer)
- Per-pixel cross entropy loss
- CRF: post-processing to model correlation between outputs (pixel classes)



Deep Learning and Structured Prediction

Ex: Semantic Segmentation

- Extension: incorporate the CRF during training
  - Pair-wise term modeling correlation
  - End-to-end training with backprop
- CRF as RNN [ZJRP+15]: mean filed inference in CRF written as RNN



Recap: CNN and invariance

- Standard ConvNets: limited invariance capacity (small shifts)
- ImageNet: single centered object ≠ other datasets (VOC, MS COCO)

ImageNet

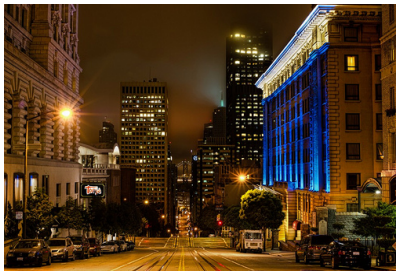



VOC 2007



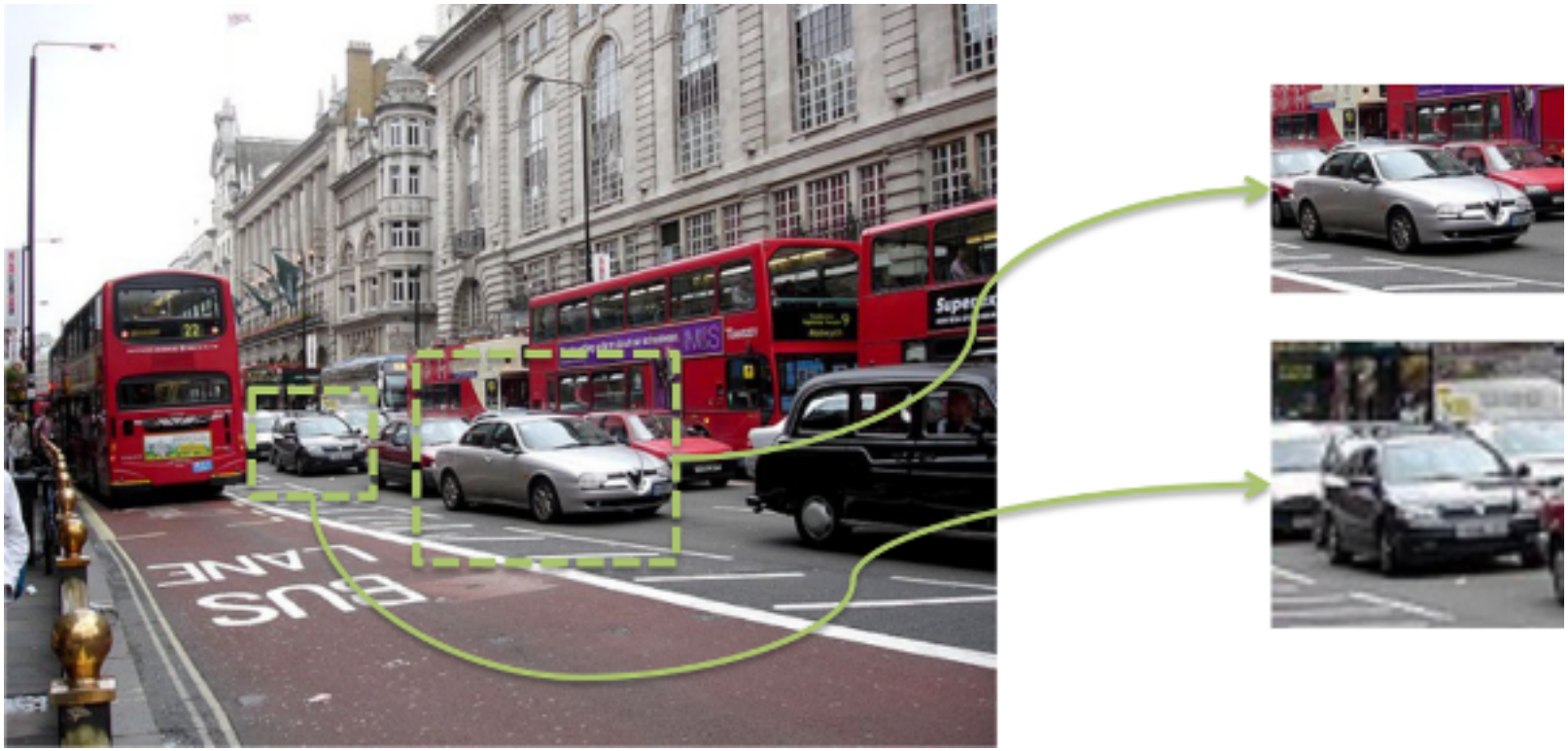

MS COCO






CNN and invariance

- Use regions to have images that look like ImageNet
- Using bounding box annotations [OBL+14]



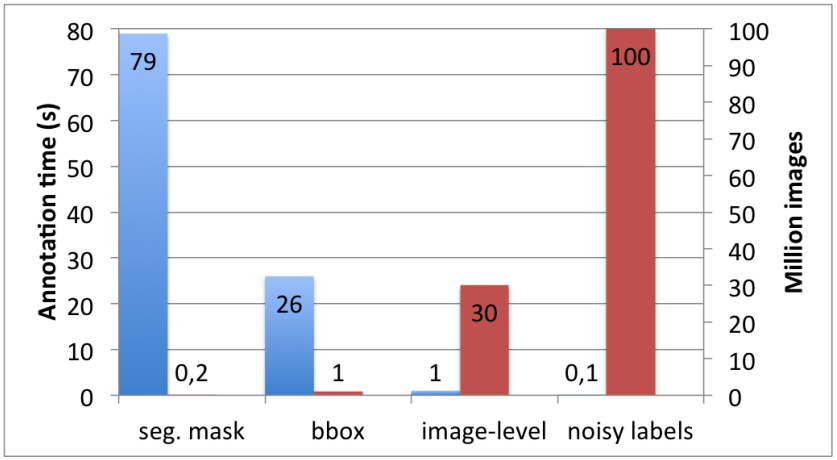
	Naive	Region
PASCAL VOC 2012	70.9%	78.7%

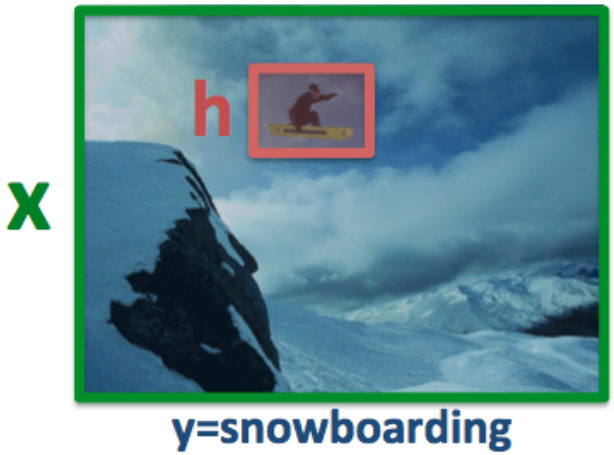
- Regions ⇒ better prediction

CNN and invariance

Weakly Supervised Learning

- Full annotations expensive ⇒ training with weak supervision [BRFFF16]



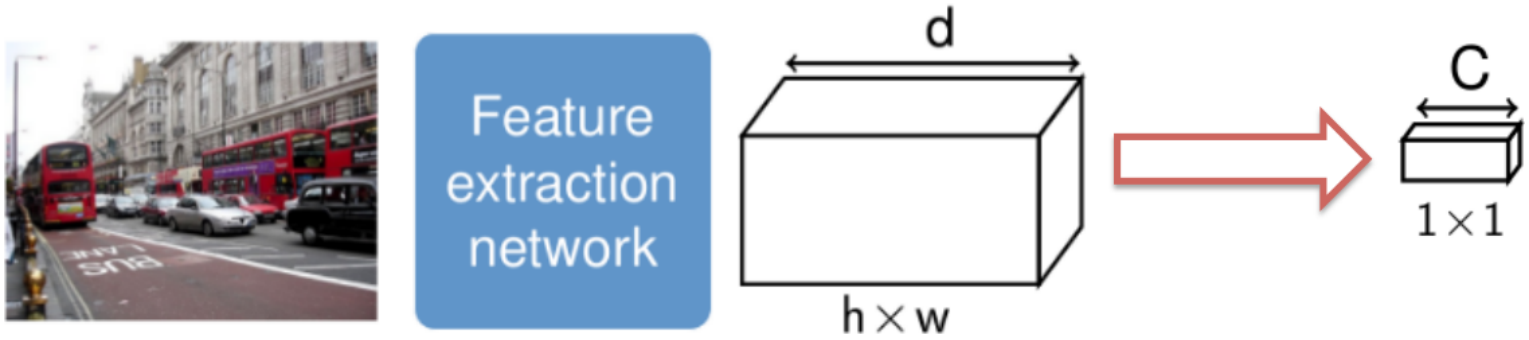


- **Incorporating latent variables h $\in \mathcal{H}$, *e.g.* training object detector from global image labels**

Variable	Notation	Space	Train	Test
Input	**x**	$\mathcal{X}$	observed	observed
Output	**y**	$\mathcal{Y}$	observed	unobserved
Latent	**h**	$\mathcal{H}$	unobserved	unobserved

Where to pool?

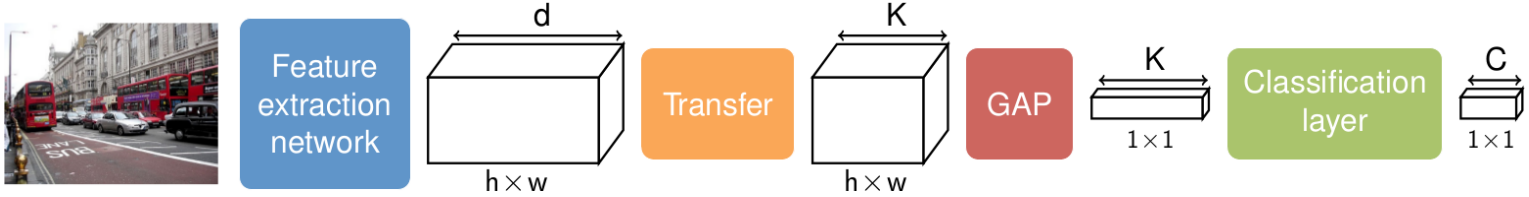
Transfer – Pooling – Classification



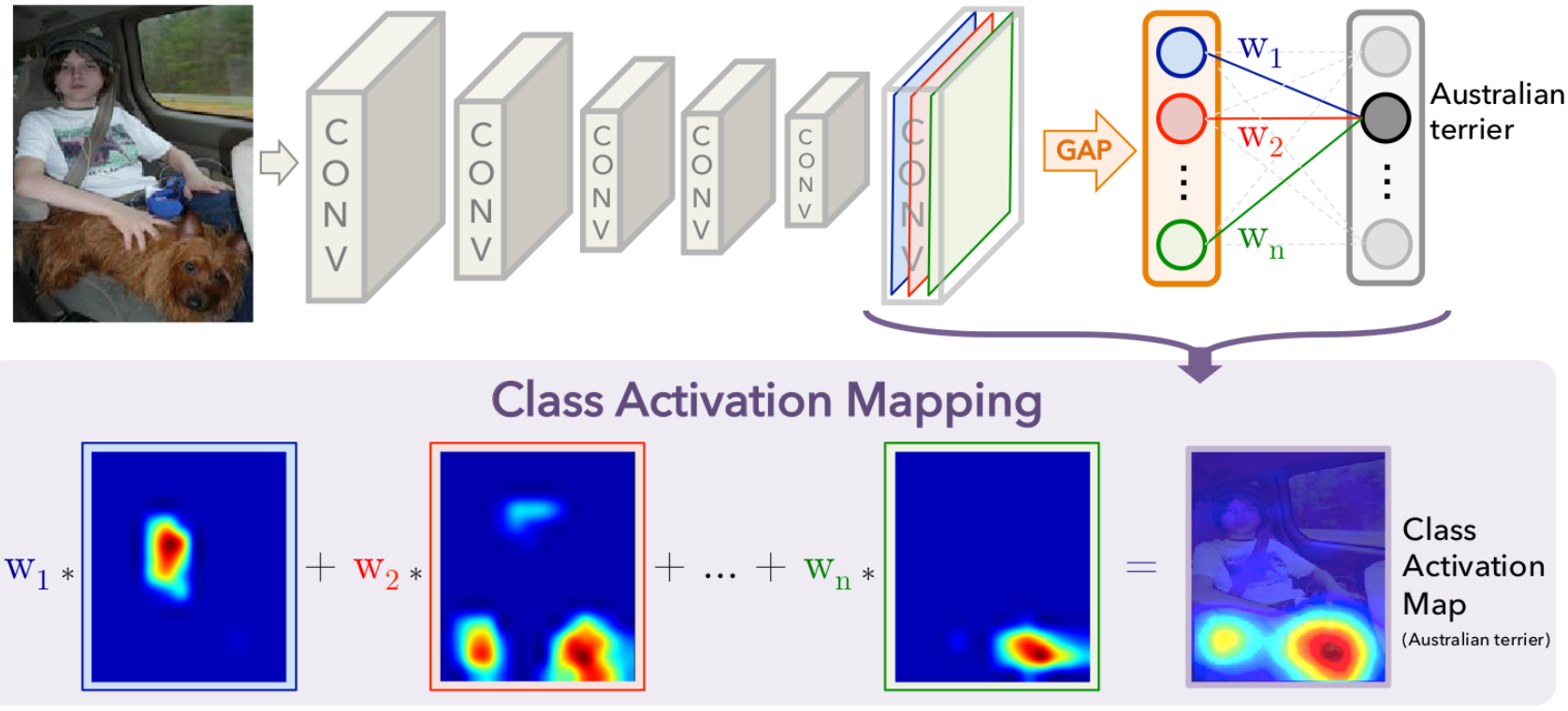
- Image-based strategy
- Region-based strategy

Image-based strategy

- Global Average Pooling [ZKA+16] (GoogLeNet, ResNet)

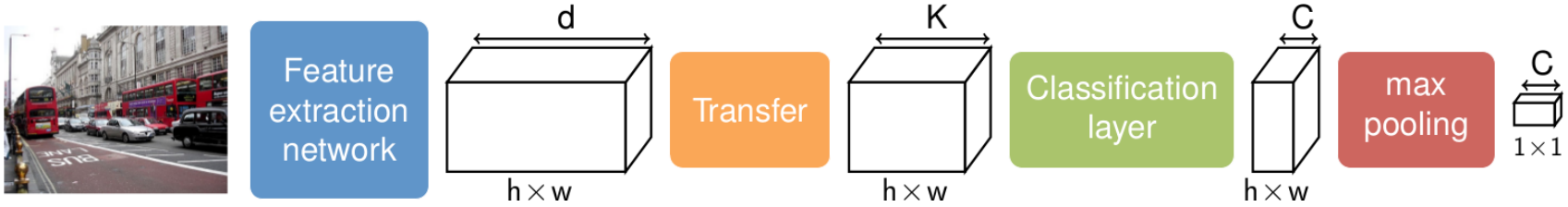


- Class Activation Mapping (CAM) for GAP

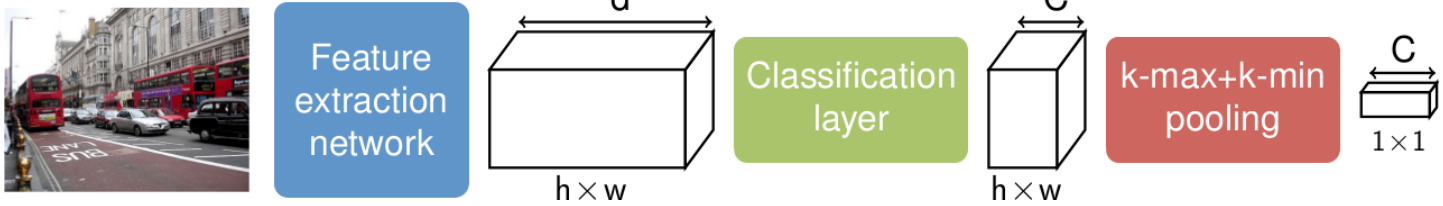


Region-based strategy

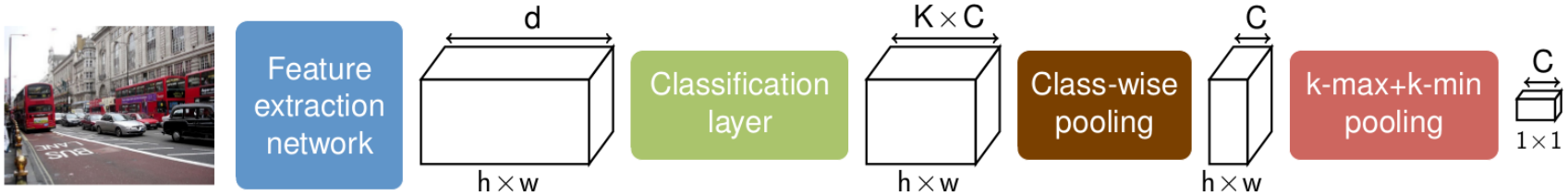
- Deep MIL [Oquab, CVPR15] [OBLS15]



- WELDON [Durand, CVPR16] [DTC16] (ProNet [Sun, CVPR16] [SPC+16])

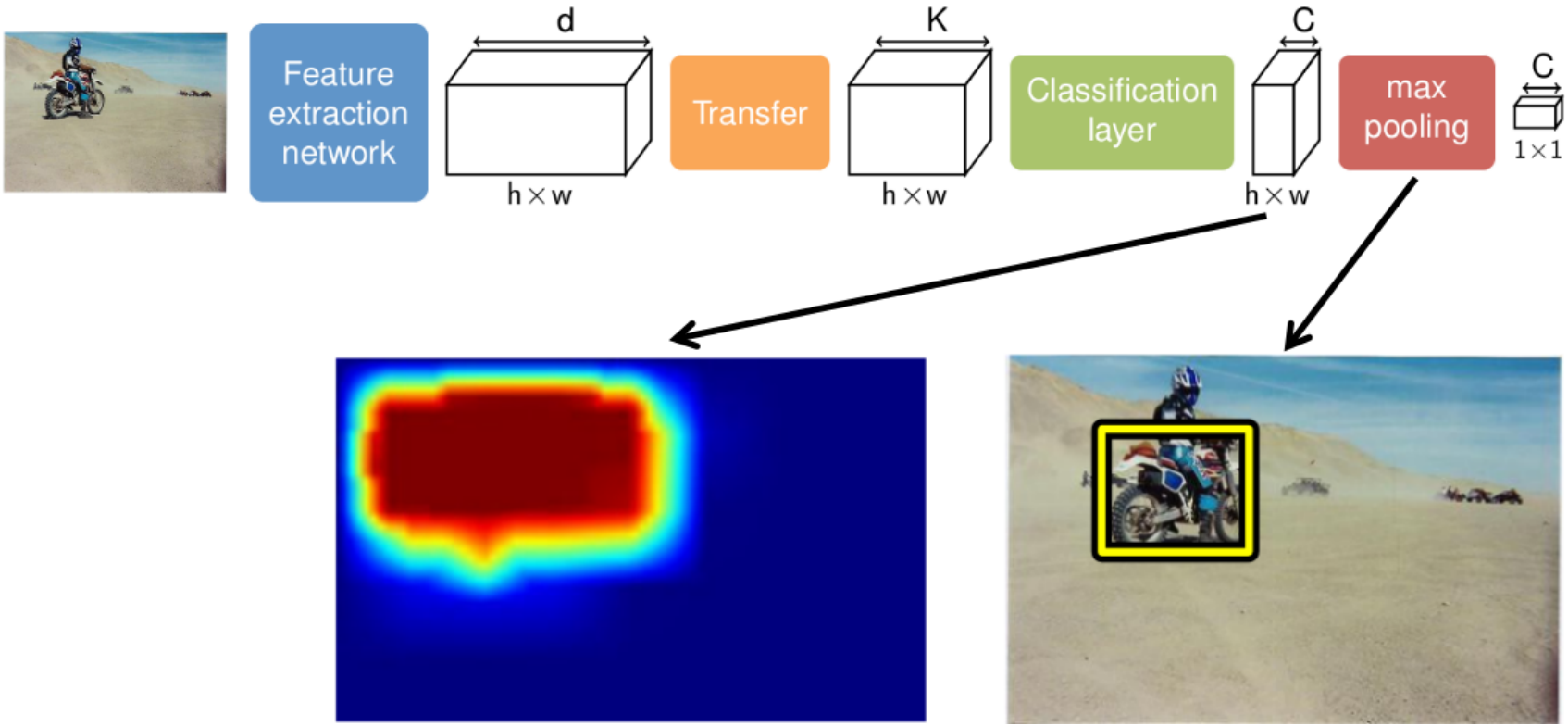


- WLIDCAT [Durand, CVPR17] [MDTC17]



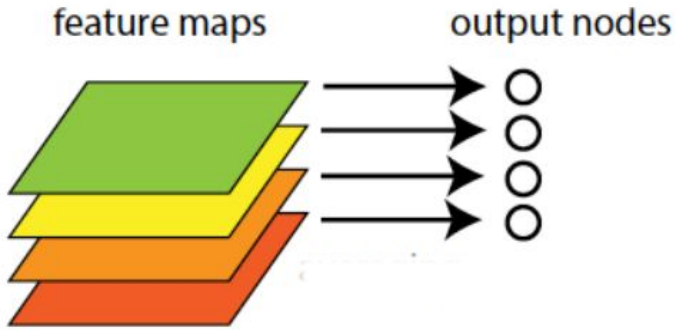
Region-based strategy

- CAM: direct in the map of depth C, *e.g.* for Deep MIL [OBLS15]



How to pool?

Pooling schemes



- Max [Oquab, CVPR15] [OBLS15]

$$y^c = \max_{i,j} z_{ij}^c \tag{1}$$

- GAP [Zhou, CVPR16] [ZKA+16]

$$y^c = \frac{1}{N} \sum_{i,j} z_{ij}^c \tag{2}$$

- LSE [Pinheiro, CVPR15] [PC15] / SPLeap [Kulkarni, ECCV16] [KJZ+16]

$$y^c = \frac{1}{\beta} \log \left(\frac{1}{N} \sum_{i,j} \exp(\beta \cdot z_{ij}^c) \right) \tag{3}$$

Max pooling limitation

- Classifying only with the max scoring region

- Loss of contextual information

Max pooling limitation

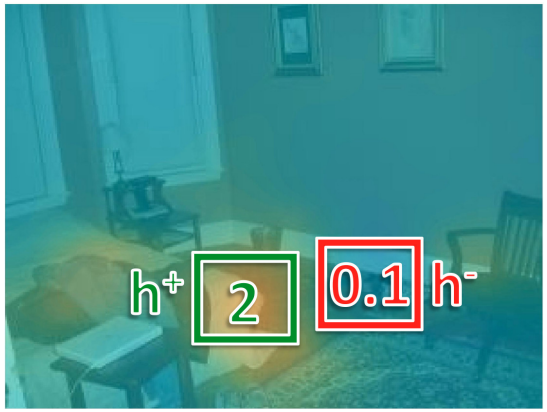
- Classifying only with the max scoring region

- Loss of contextual information

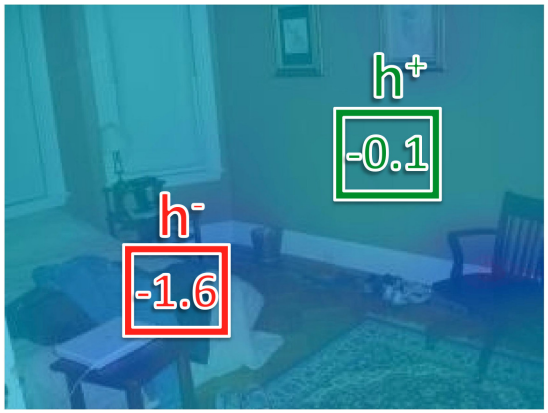
MANTRA [DTC15]: max+min pooling

- $\mathbf{h}^+$: presence of the class → high $\mathbf{h}^+$
- $\mathbf{h}^-$: localized evidence of the absence of class

original image



bedroom

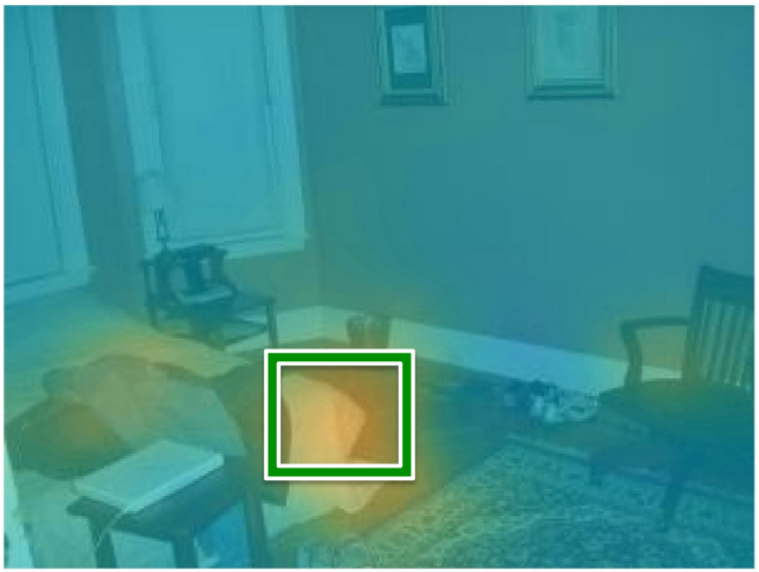


airport inside

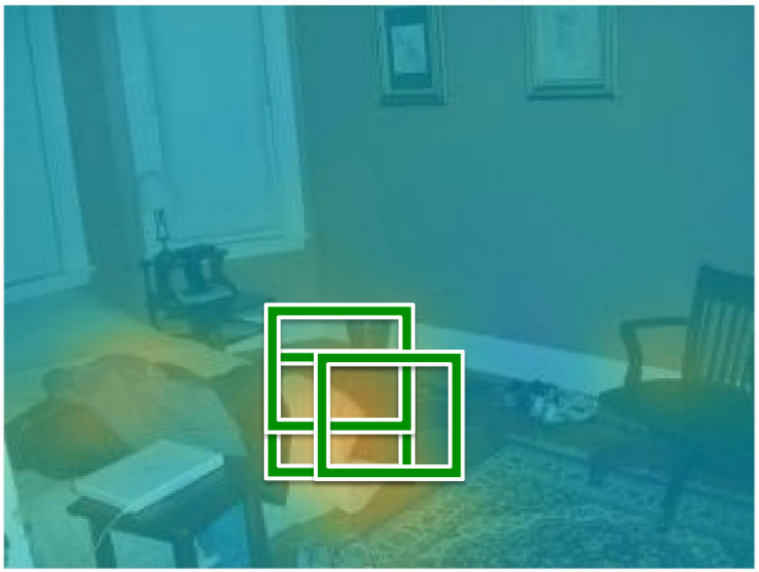
dining room

WELDON [DTC16] Pooling

- `max+min` strategy
- Top instances: using several regions, more robust region selection [Vasconcelos, CVPR15]



k=1



k=3

WELDON [DTC16] Pooling

- `max+min` strategy
- Top instances: using several regions, more robust region selection [Vasconcelos, CVPR15]

$$y^c = s^{top}_{k^+}(z^c) + s^{low}_{k^-}(z^c) \tag{4}$$

$$s^{top}_{k^+}(z^c) = \frac{1}{k^+} \sum_{i=1}^{k^+} i\text{-th-max}(z^c) \tag{5}$$

$$s^{low}_{k^-}(z^c) = \frac{1}{k^-} \sum_{i=1}^{k^-} i\text{-th-min}(z^c) \tag{6}$$

WILDCAT [MDTC17] Pooling

- `max+min`: complementary information
- Different kind of information

$$y^c = s^{top}_{k^+}(z^c) + \alpha \cdot s^{low}_{k^-}(z^c) \tag{7}$$

- α: trade off parameter.

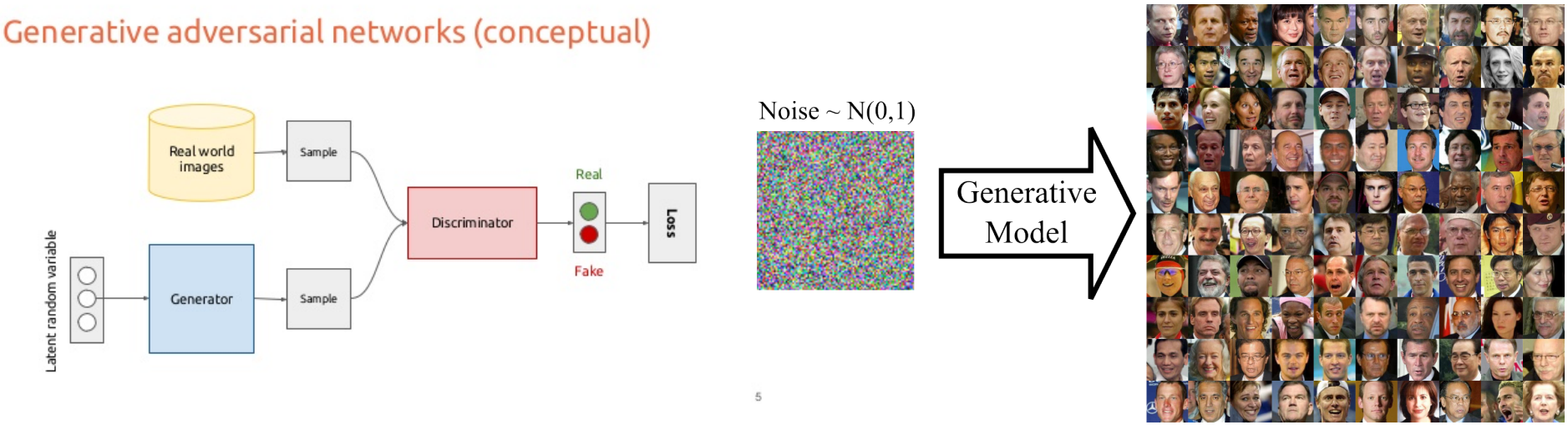
Pooling	k^+	k^-	α
Maximum	1	0	0
GAP	n	0	0
WELDON	k	k	1

Outline

Ongoing Issues in Deep Learning

Unsupervised Training

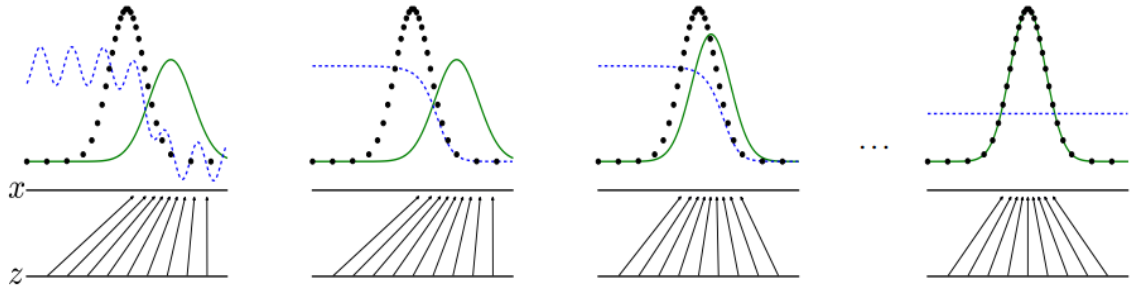
- Standard ways to perform unsupervised: learning representations fitting data well, *e.g.* Maximum likelihood, reconstruction error, *etc*
- Success of deep learning essentially for supervised problem
- Solution: cast unsupervised problem as a supervised one
  $\Rightarrow$ **auto-supervision**
  - Trendy example: Generative Adversarial Networks (GAN) [GPAM+14]



Ongoing Issues in Deep Learning

Unsupervised Training: GAN

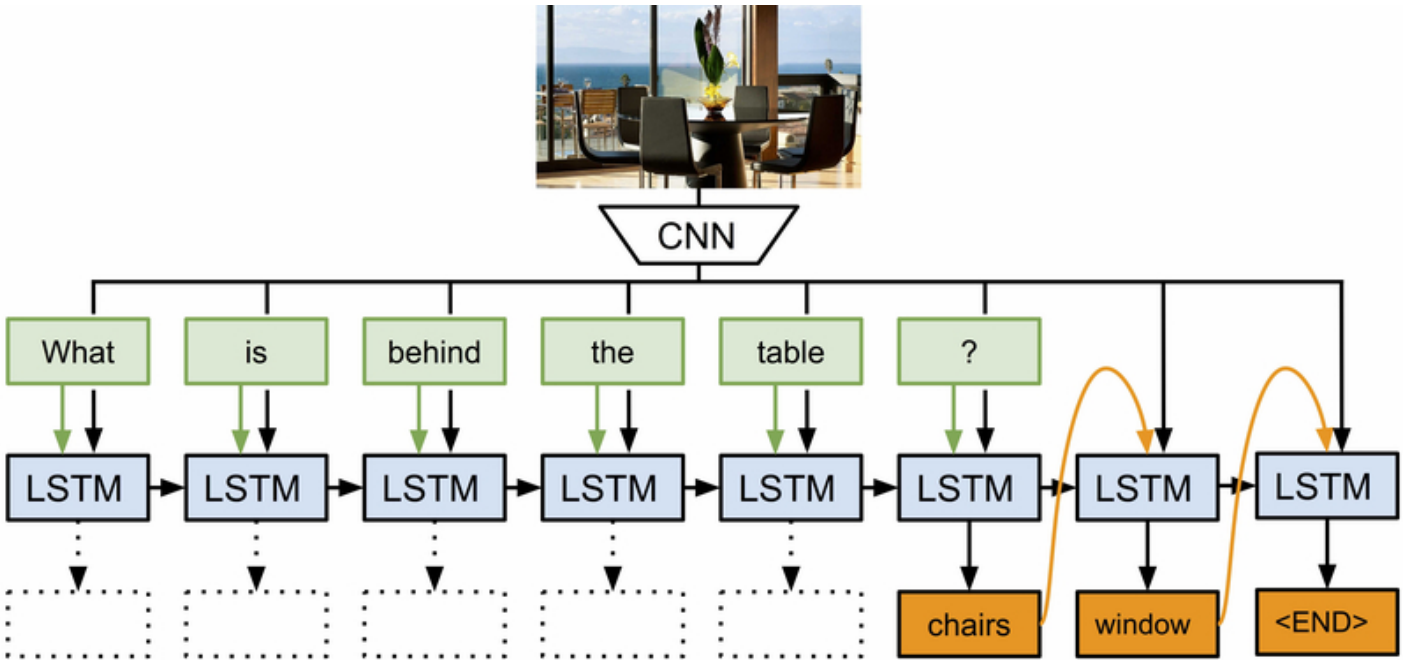
- Unsupervised problem ⇒ 2-player game theory problem
- Interesting results: optimal generator learns data distribution



Ongoing Issues in Deep Learning

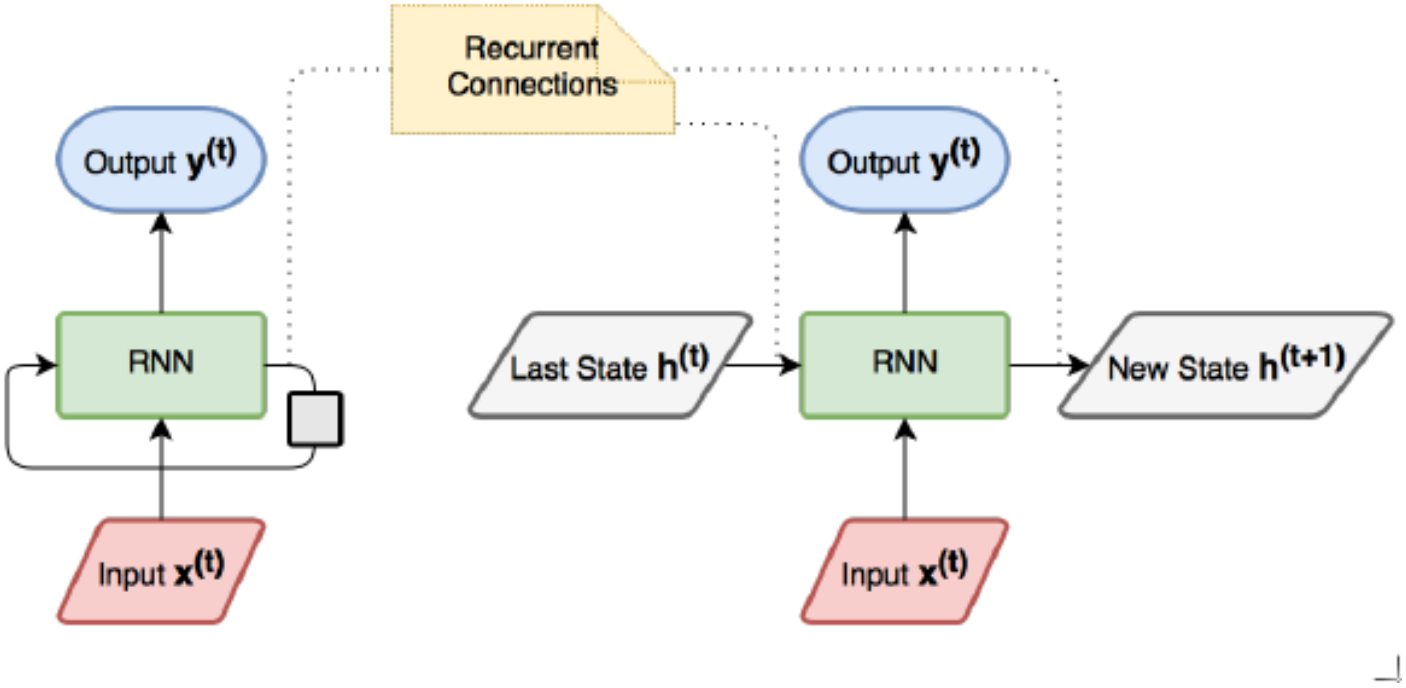
New Tasks in Artificial Intelligence

- Vision and language, Visual Question Answering (VQA), or image captioning
- Language: use of Reurrent Neural Networks (RNNs)



Credit: M. Malinowski [MRF15]

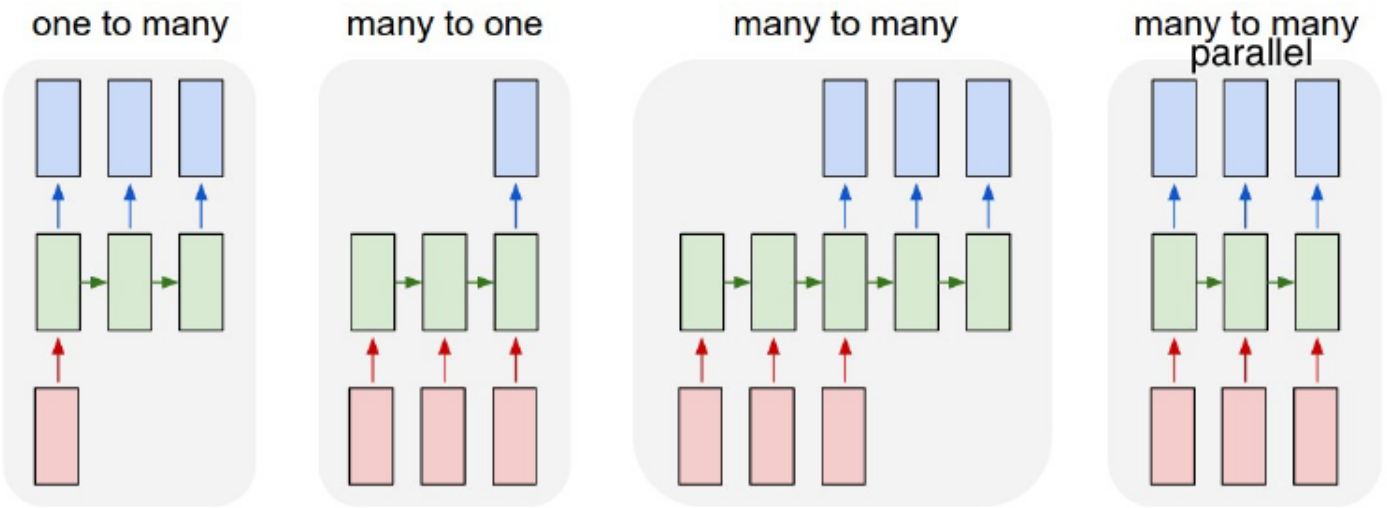
Reurrent Neural Networks (RNNs)



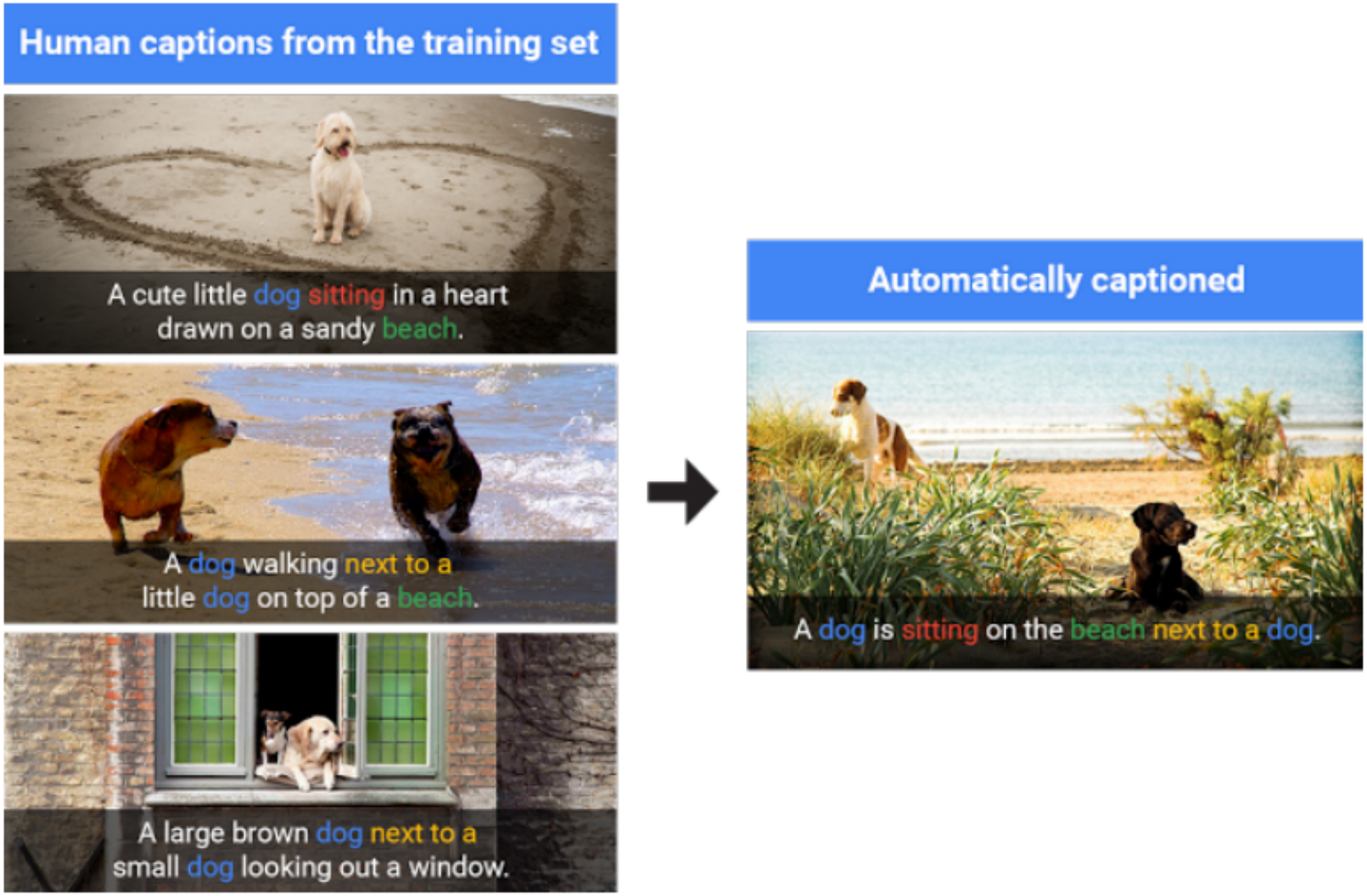
- Input vector $x(t)$, *e.g.* word (text) or image representation (CNN).
- Input/Output $h(t)$: vector representing model "short-term memory"
- Output vector $y(t)$: task dependent
- All parameters trained with backpropagation through time.

Reurrent Neural Networks (RNNs)

Sequence modeling with RNNs

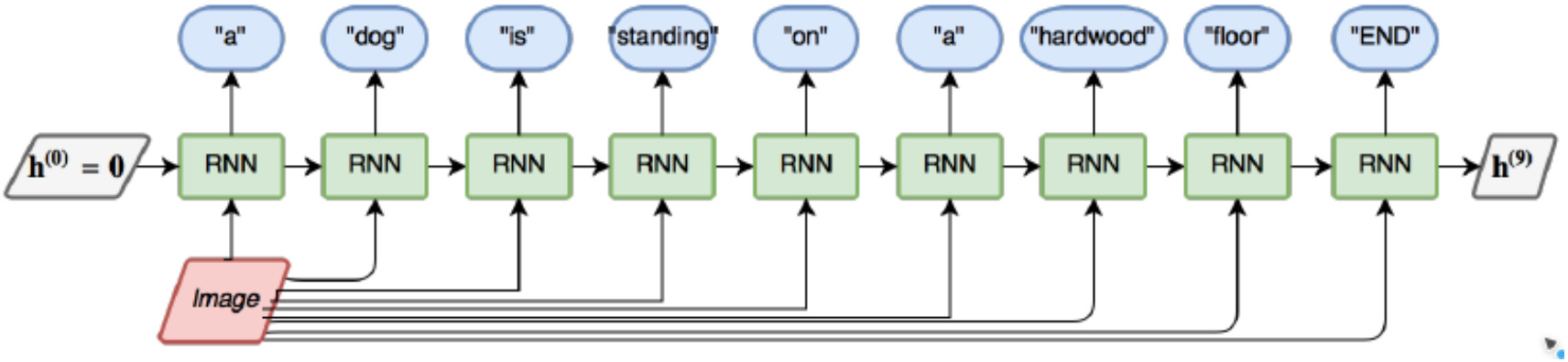


One to Many - Image captioning

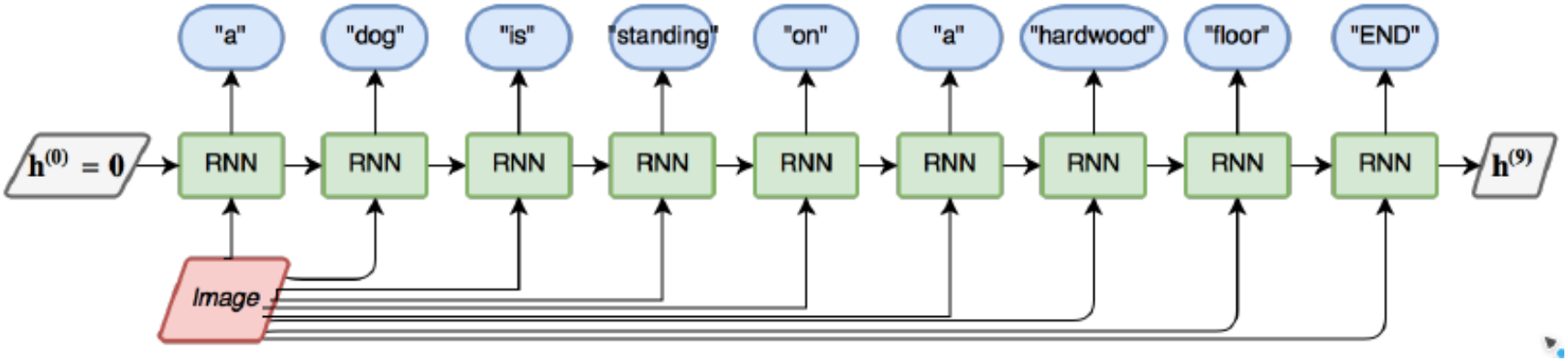


One to Many - Image captioning

- Show, Attend and Tell, Xu *et. al.*, ICML15 [XBK+15]



- Karpathy, CVPR15 [KL15]



Many to One - Visual Question Answering (VQA)

Goal : build a system that can answer questions about images

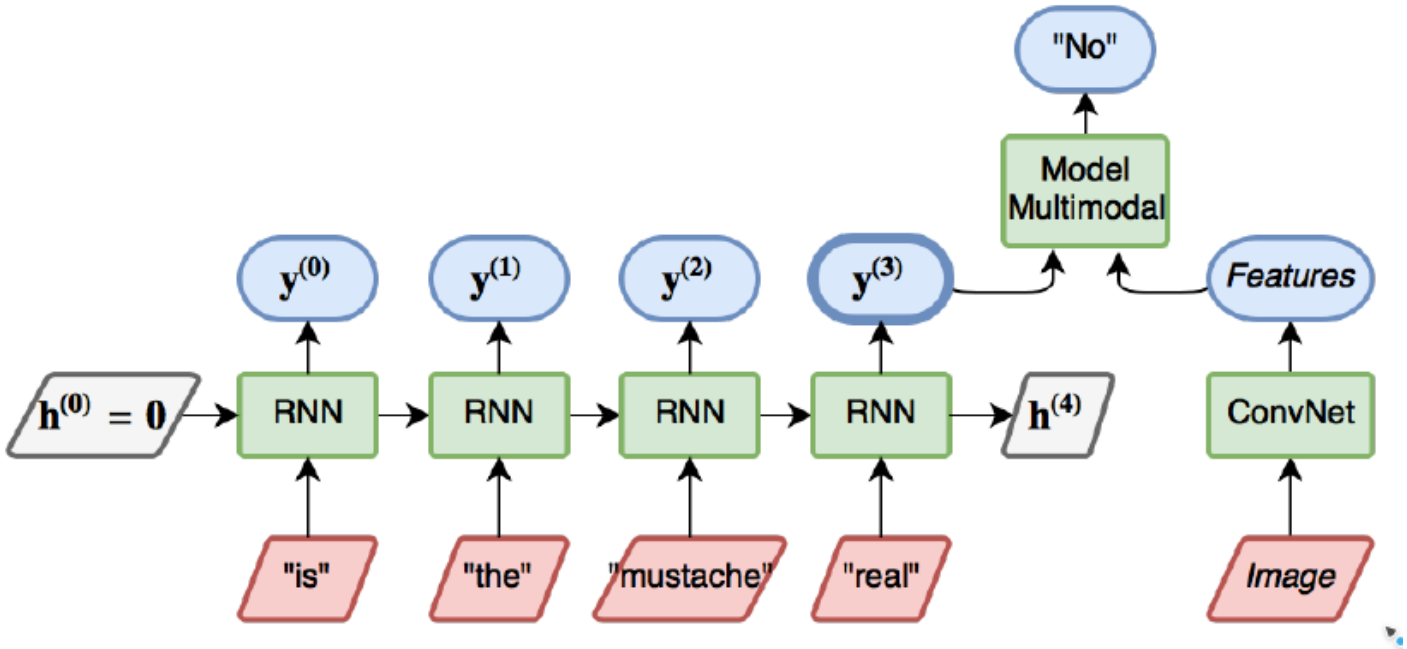
How many slices of pizza are there?
Is this a vegetarian pizza?

Does it appear to be rainy?
Does this person have 20/20 vision?

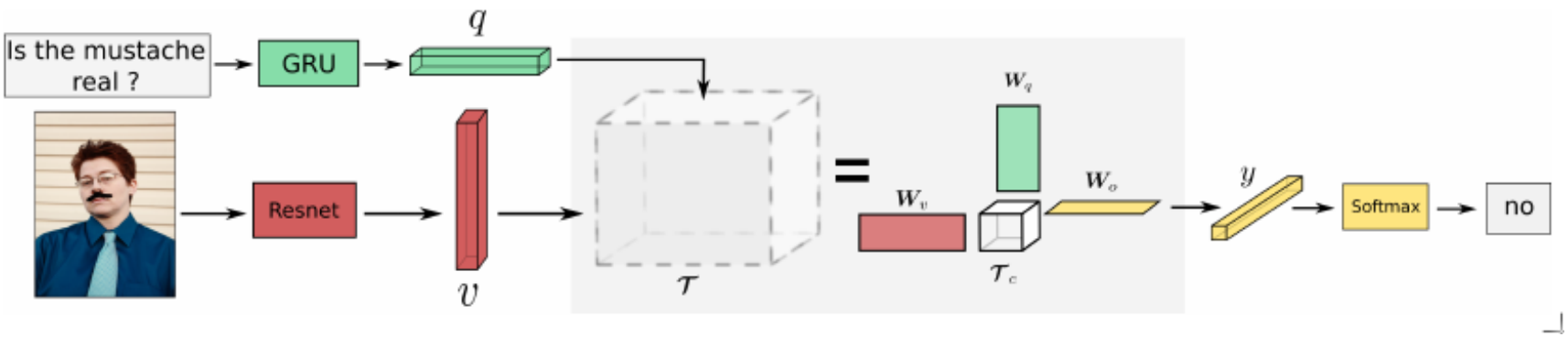
What color are her eyes?
What is the mustache made of?

Many to One - Visual Question Answering (VQA)

- Input: question & image
- Output: answer



VQA : Multi-modal Fusion

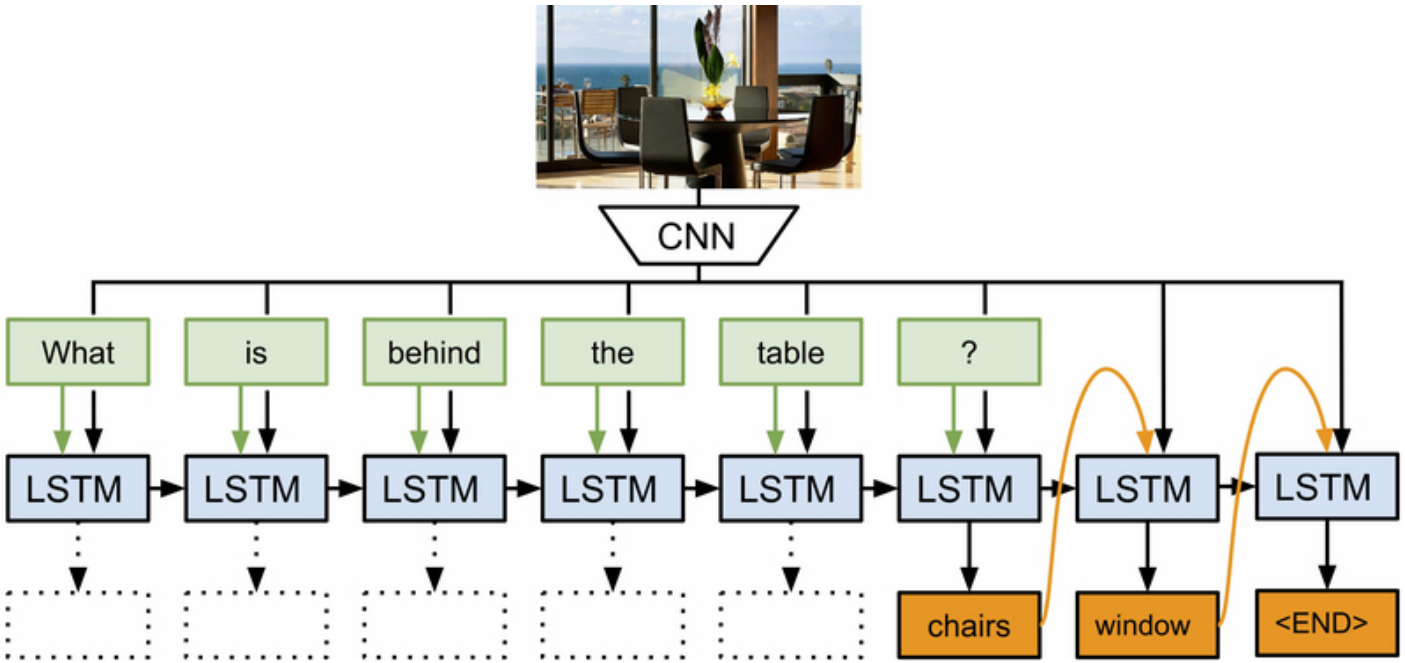


- Mono-modal representations:
  - Visual representation: ResNet-152
  - Question representation: pre-trained GRU
- How to perform multi-modal fusion ⇒ Tucker decomposition [BCCT17]

Ongoing Issues in Deep Learning

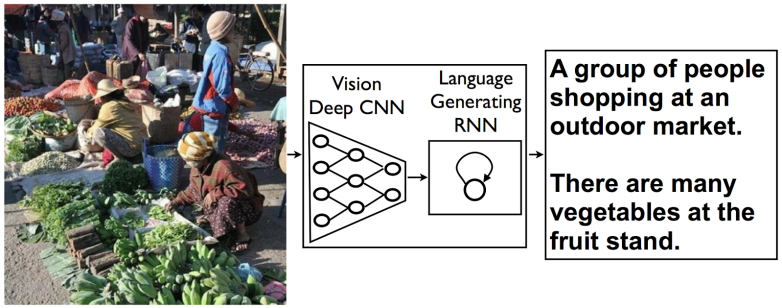
New Tasks in Artificial Intelligence

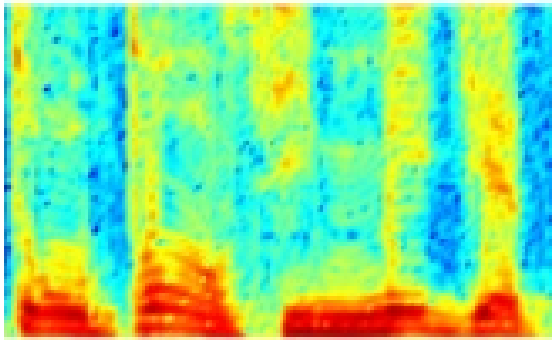
- But still a long way to go toward real AI ...



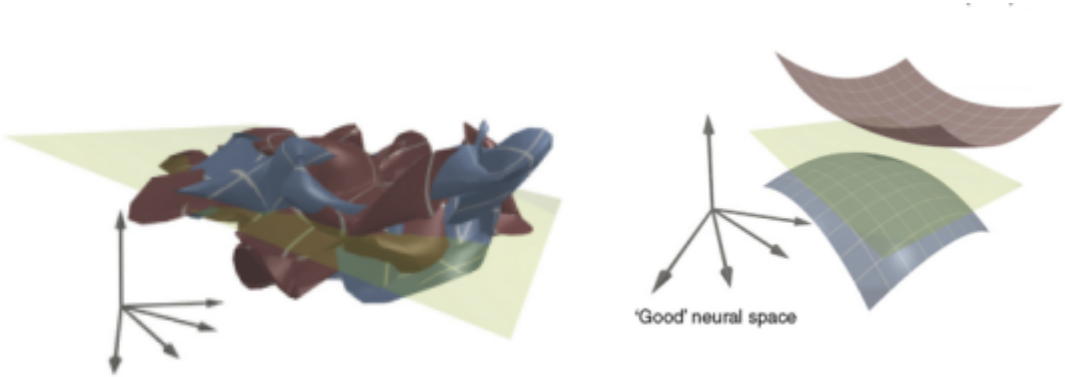
Credit: M. Malinowski [MRF15]

Conclusion





- Deep Learning: huge impact in terms of experimental results
- BUT: formal understanding still limited,
  - Optimization: non-convex problem
  - Model: ability to untangle manifold
  - Robustness to over-fitting & generalization

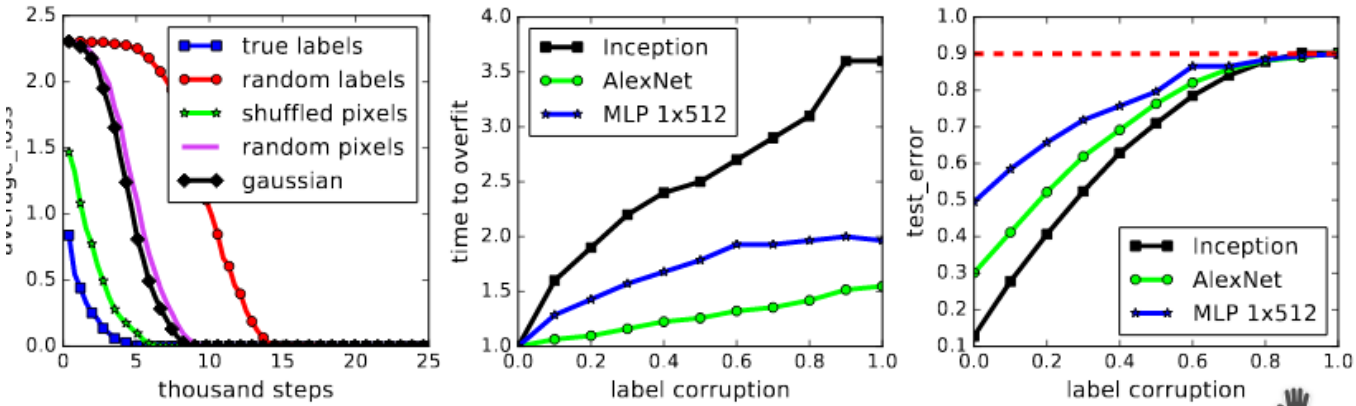


Deep Learning and generalization

- Rademacher complexity: capacity of a model to fit random label :

$$\mathcal{R}_n(\mathcal{H}) = E_\sigma \left[sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i h(x_i) \right]$$

- Rethinking generalization: Zhang *et. al.* ICLR17 [ZBH+17]



(a) learning curves (b) convergence slowdown (c) generalization error growth

  - Deep models easily fits random labels !!
  - $\mathcal{R}_n(\mathcal{H}) \approx 1 \Rightarrow$ no theoretical guarantee on generalization performances
- Classical learning theory insufficient to explain the good generalization behavior of deep models

References I


Hedi Ben-younes, Rémi Cadène, Matthieu Cord, and Nicolas Thome, *MUTAN: multimodal tucker fusion for visual question answering*, CoRR abs/1705.06676 (2017).


David Belanger and Andrew McCallum, *Structured prediction energy networks*, ICML, JMLR Workshop and Conference Proceedings, vol. 48, JMLR.org, 2016, pp. 983–992.


Bearman, Russakovsky, Ferrari, and Fei-Fei, *What's the Point: Semantic Segmentation with Point Supervision*, ECCV (2016).


Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille, *Semantic image segmentation with deep convolutional nets and fully connected crfs*, ICLR, 2015.


Liang-Chieh Chen, Alexander G. Schwing, Alan L. Yuille, and Raquel Urtasun, *Learning deep structured models*, Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015, 2015, pp. 1785–1794.


Thibaut Durand, Nicolas Thome, and Matthieu Cord, *MANTRA: Minimum Maximum Latent Structural SVM for Image Classification and Ranking*, International Conference on Computer Vision (ICCV), 2015.


_____, *WELDON: Weakly Supervised Learning of Deep Convolutional Neural Networks*, Computer Vision and Pattern Recognition (CVPR), 2016.


Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, *Generative adversarial nets*, Advances in Neural Information Processing Systems 27 (Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, eds.), Curran Associates, Inc., 2014, pp. 2672–2680.


Praveen Kulkarni, Frédéric Jurie, Joaquin Zepeda, Patrick Pérez, and Louis Chevallier, *Spleap: Soft pooling of learned parts for image classification*, ECCV, 2016.

References II


Andrej Karpathy and Fei-Fei Li, *Deep visual-semantic alignments for generating image descriptions*, IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015, 2015, pp. 3128–3137.


Taylor Mordan, Thibaut Durand, Nicolas Thome, and Matthieu Cord, *WILDCAT: Weakly Supervised Learning of Deep ConvNets for Image Classification, Localization and Segmentation*, Computer Vision and Pattern Recognition (CVPR), 2017.


Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz, *Ask your neurons: A neural-based approach to answering questions about images*, 2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015, 2015, pp. 1–9.


Maxime Oquab, Leon Bottou, Ivan Laptev, Josef Sivic, et al., *Learning and transferring mid-level image representations using convolutional neural networks*, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014.


M. Oquab, L. Bottou, I. Laptev, and J. Sivic, *Is object localization for free? âĂŞ weakly-supervised learning with convolutional neural networks*, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015.


Pedro O. Pinheiro and Ronan Collobert, *From image-level to pixel-level labeling with convolutional networks*, CVPR, 2015.


Chen Sun, Manohar Paluri, Ronan Collobert, Ram Nevatia, and Lubomir Bourdev, *ProNet: Learning to Propose Object-Specific Boxes for Cascaded Neural Networks*, CVPR, 2016.


Shenlong Wang, Sanja Fidler, and Raquel Urtasun, *Proximal deep structured models*, NIPS, 2016, pp. 865–873.

References III


Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio, *Show, attend and tell: Neural image caption generation with visual attention*, Proceedings of the 32nd International Conference on Machine Learning (ICML-15) (David Blei and Francis Bach, eds.), JMLR Workshop and Conference Proceedings, 2015, pp. 2048–2057.


Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals, *Understanding deep learning requires rethinking generalization*, 2017.


Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip H. S. Torr, *Conditional random fields as recurrent neural networks*, Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV) (Washington, DC, USA), ICCV '15, IEEE Computer Society, 2015, pp. 1529–1537.


B. Zhou, A. Khosla, Lapedriza. A., A. Oliva, and A. Torralba, *Learning Deep Features for Discriminative Localization.*, CVPR (2016).