Master TRIED
Reconnaissance des formes et méthodes neuronales (US330X) - Neural Networks and Deep Learning

Nicolas Thome
Conservatoire Nationnal des Arts et Métiers (Cnam)
Laboratoire CEDRIC - équipe Vertigo

Outline

Deep Features: Domain Adaptation for Localized Tasks

From [Noh et al., 2017]

From [Cao et al., 2017]

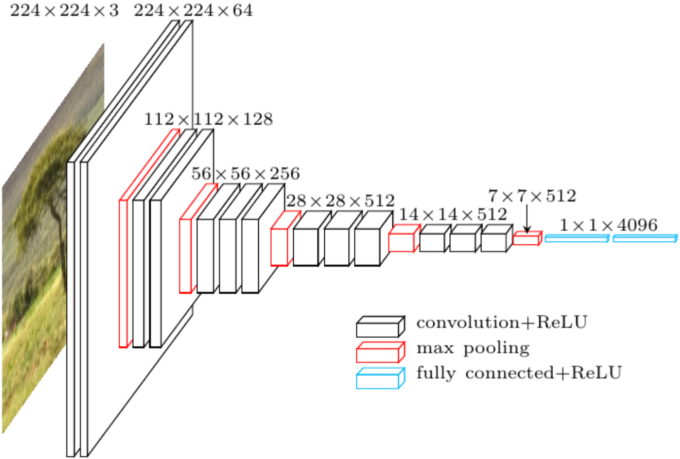
- Local information needed: various applications, *e.g.* localization, segmentation, retrieval, pose estimation, *etc*

Deep Features for Localized Tasks

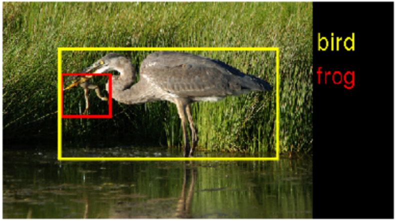
- Core (simple) idea: deep features for local information in image regions
  - Crop given image sub-area
  - Rescale → ImageNet input size, *e.g.* 224×224

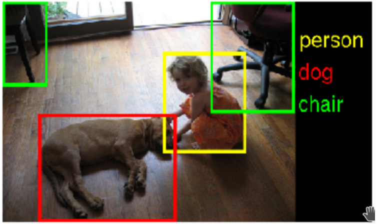
Deep Features for Localized Tasks

- Core idea: deep features for local information in image regions
  - Extract Deep Features with ConvNet pre-trained on ImageNet



Example: Object Localization





- **Object Localization**: rectangular Bounding Box (BB) aroud each object in the image
- **Localization as classification**: classify each region into K+1 (background) classes

Localization with Region-CNN [Girshick et al., 2014]

1. R-CNN, 1^{st} step: extract a set of region proposal candidates
   - Goal: pre-select candidates based on their "objectness"
   - Low-level, unsupervised
   - Many approaches,*e.g.* selective search [Uijlings et al., 2013]

Localization with Region-CNN [Girshick et al., 2014]

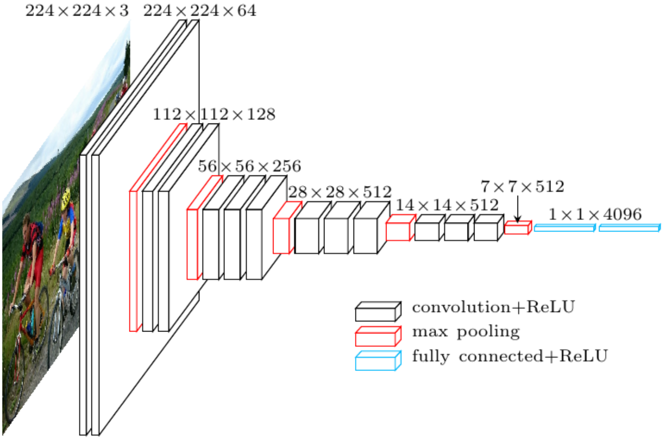
2. R-CNN, 2^{nd} step: classifiy each regions proposal
   - Rescale proposal & extract deep feature
   - Add transfer layer with $K+1$ classes
     - +BB regression, *i.e.* remap proposal (red) $\rightarrow$ GT BB (green)

Semantic Image Segmentation

- **Label each image pixel into $K + 1$ (background) classes**
- Extract deep features on regions centered at each pixel (cf localization)?
  - Naive solution very inefficient , does not scale!
  - Ex: 500×500 image $\Rightarrow$ 25000 regions with a single scale!

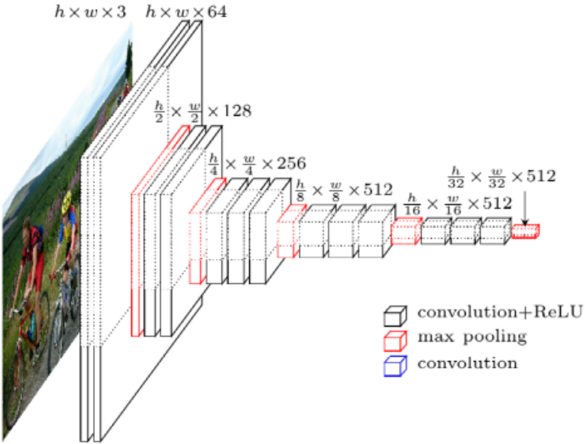
Semantic Segmentation with Fully Convolutionnal Networks

- 224×224 input image: apply [Conv-FC], *e.g.* VGG



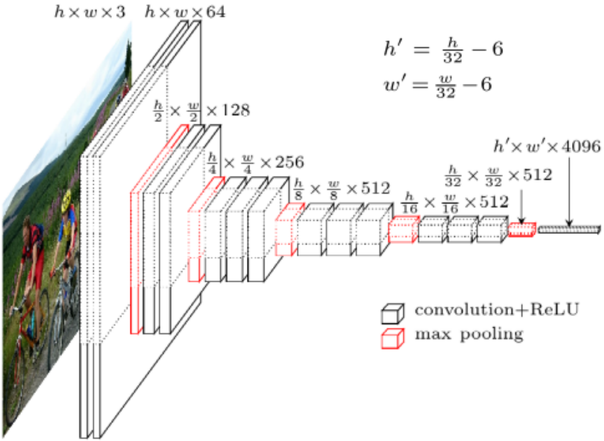
Semantic Segmentation with Fully Convolutionnal Networks

- Conv layer directly applicable to bigger image, size $w \times h$
- How to transfer FC layers? (direct with base FCN, *e.g.* ResNet)



Semantic Segmentation with Fully Convolutionnal Networks

- FC $\Leftrightarrow$ conv with $7 \times 7 \times 512$ filters
- Ex: input image $= 512^2$, $w' = 10$, $h' = 10$



Semantic Segmentation with Fully Convolutionnal Networks

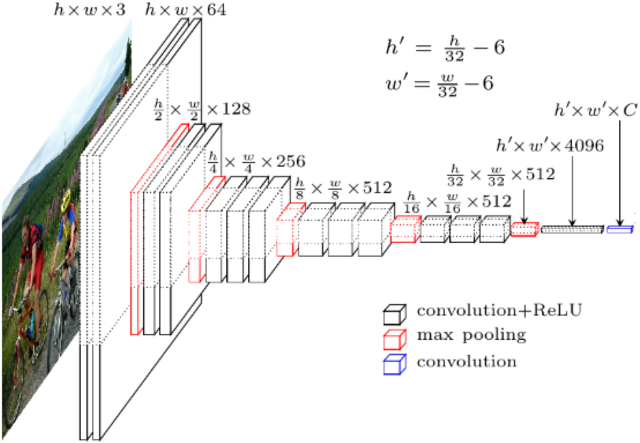
- Ex: input image = 512×512, $w' = 10, h' = 10$

Semantic Segmentation with Fully Convolutionnal Networks

- Ex: input image = 512×512, $w' = 10, h' = 10$
- Receptive field, features extracted $\approx$ rescaled region and apply ConvNet

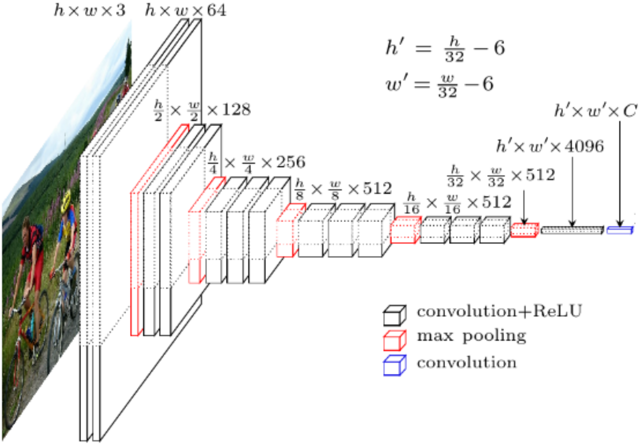
Semantic Segmentation with Fully Convolutionnal Networks

- Add transfer layer ($C = K + 1$ classes) to classify each of the $w' \times h'$ regions
- Fully connected layer on each region: 1×1 convolution + softmax



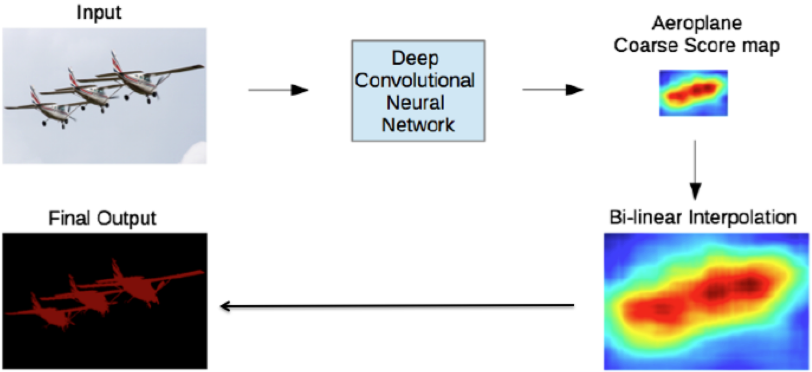
Semantic Segmentation: DeepLab [Chen et al., 2015b]

- Fully Convolutional Network outputs $w' \times h' \times C$ tensor
- How to train it from $w \times h \times C$ annotations?



Semantic Segmentation: DeepLab [Chen et al., 2015b]

- DeepLab: simply interpolate maps → $w \times h \times C$
- Cross-entroy loss for each pixel



Deep Learning and Structured Prediction

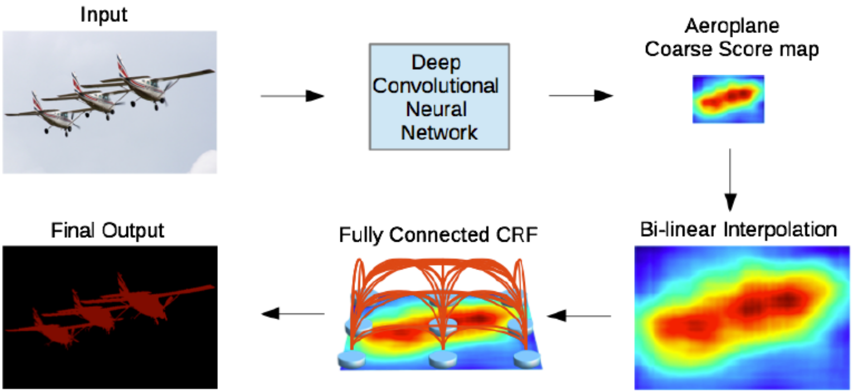
- Structured prediction: graphical model in general (*e.g.* SSVM or CRF specific models)
  - Model correlation between output variables
- Structured Prediction models (previous weeks): limited to log linear models with handcrafted features
- Combining Deep Learning & Structured Prediction
  - Solution: 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 [Chen et al., 2015a]
    - Recent models for continuous outputs [Belanger and McCallum, 2016, Wang et al., 2016]
    - Approches to unroll inference: forward and backward passes of these deep structured models expressed as a set of standard layers [Zheng et al., 2015, Belanger and McCallum, 2016, Wang et al., 2016]

      $\Rightarrow$ fast end-to-end training on GPUs.

DL & Structured Prediction: Semantic Segmentation

DeepLab [Chen et al., 2015b]

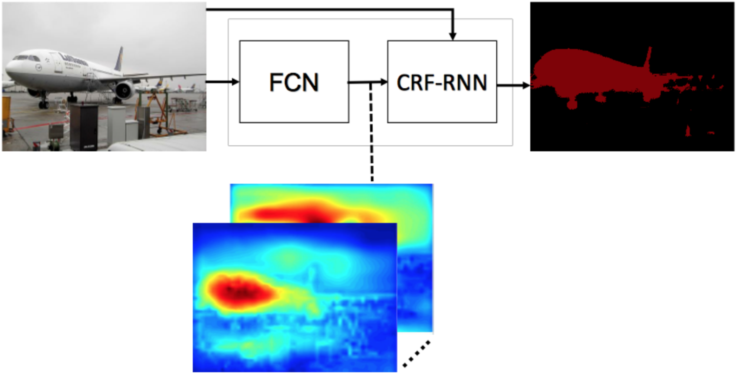
- Per-pixel cross entropy loss ⇒ classify each pixel independently
- CRF: post-processing to model correlation between outputs



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 [Zheng et al., 2015]: mean filed inference in CRF written as RNN



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 !**

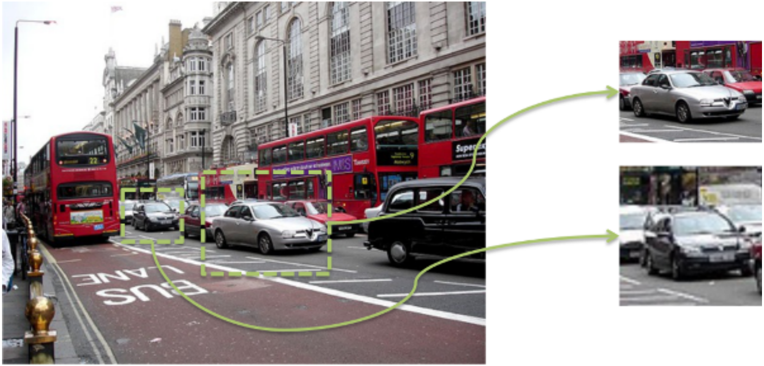
ImageNet

VOC 2007

MS COCO

CNN and invariance

- Use regions to have images that look like ImageNet
- Using bounding box annotations [Oquab et al., 2014]



	Naive	Region
PASCAL VOC 2012	70.9%	78.7%

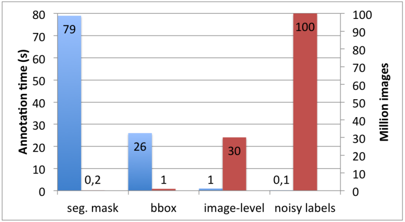
- **Regions ⇒ better prediction**

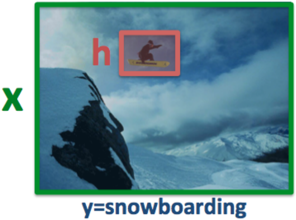
CNN and invariance

Weakly Supervised Learning

- Full annotations expensive ⇒ training with weak supervision

  [Bearman et al., 2016]



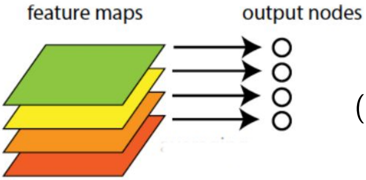


- **Incorporating latent variables h ∈ $\mathcal{H}$, *e.g.* training object detector from global image labels**

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

Weakly Supervised Learning

How to pool? Pooling schemes



- Max [Oquab et al., 2015]

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

- GAP [Zhou et al., 2016]

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

- Soft-max [Pinheiro and Collobert, 2015, Kulkarni et al., 2016]

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

Average pooling limitation

- Classifying with all regions
- Not efficient for small objects: lots of "noisy" regions

Max pooling limitation

- Classifying only with the max scoring region

- Loss of contextual information

Max pooling limitation

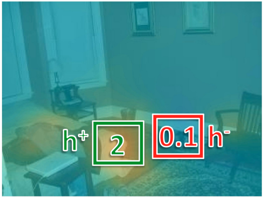
- Classifying only with the max scoring region

- Loss of contextual information

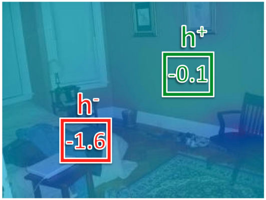
MANTRA [Durand et al., 2015]: max+min pooling

- h^+: presence of the class → high h^+
- h^-: localized evidence of the absence of class

original image



bedroom



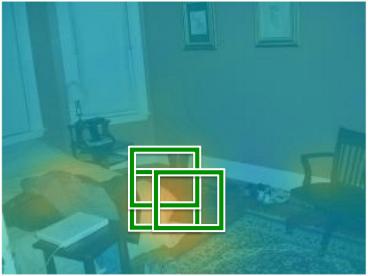
airport inside

dining room

WELDON [Durand et al., 2016] Pooling

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

k=1



k=3

WELDON [Durand et al., 2016] Pooling

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

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

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

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

WILDCAT [Mordan et al., 2017] Pooling

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

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

- α: trade off parameter.

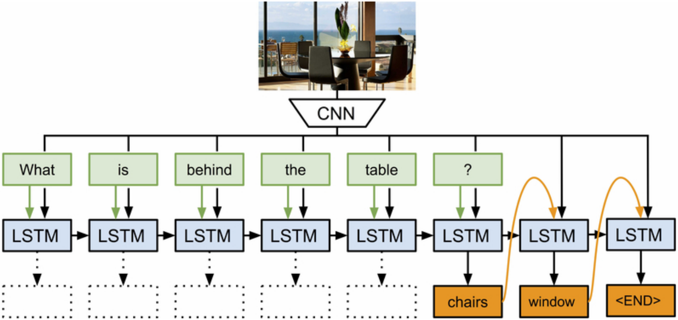
Pooling	k^+	k^-	α
Maximum	1	0	0
GAP	n	0	0
WELDON	k	k	1

Outline

Ongoing Issues in Deep Learning

New Tasks in Artificial Intelligence

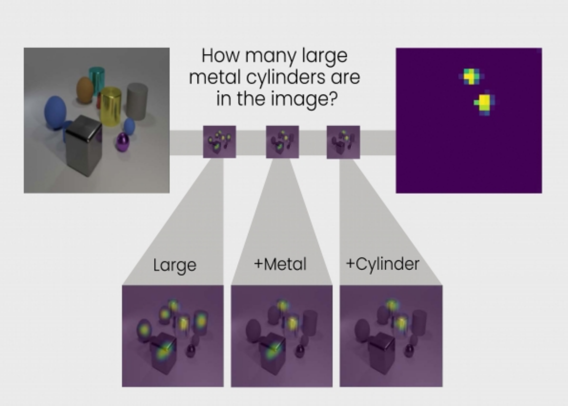
- Vision and language: leverage deep learning advances
  - Vision: use of Convolutional Neural Networks (ConvNets)
  - Language: use of Recurrent Neural Networks (RNNs)



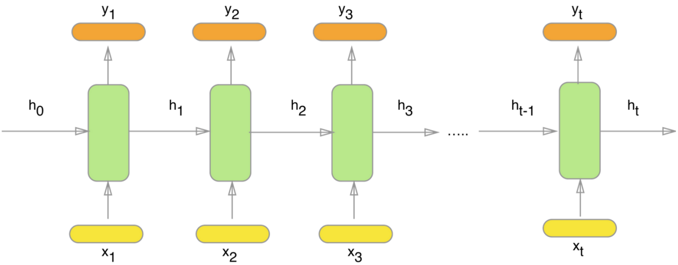
Credit: M. Malinowski [Malinowski et al., 2015]

New Tasks in Artificial Intelligence

- Vision and language: tasks requiring some form of **high level reasoning**
  - Detecting concepts/objects in images, **but also**
  - **Relationships between them**
  - **NLP descriptions/understanding of these relationships**



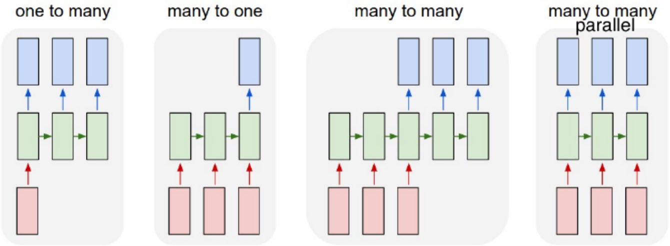
Recurrent Neural Networks (RNNs): Recap



- 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.

Recurrent Neural Networks (RNNs)

Sequence modeling with RNNs



One to Many - Image captioning

- **Input:** image
- **Output:** a sentence in natural language
- Approaches inspired by works in machine translation, *e.g.* [Sutskever et al., 2014]
  - Encoder-decoder: encode image, decode into words

Image captioning

- Show and Tell,CVPR'15 [Vinyals et al., 2015]
  - Image input represented by a deep feature, *e.g.* GoogLeNet FC
  - Text input with dense embedding from one-hot encoding
  - Different architectural variants

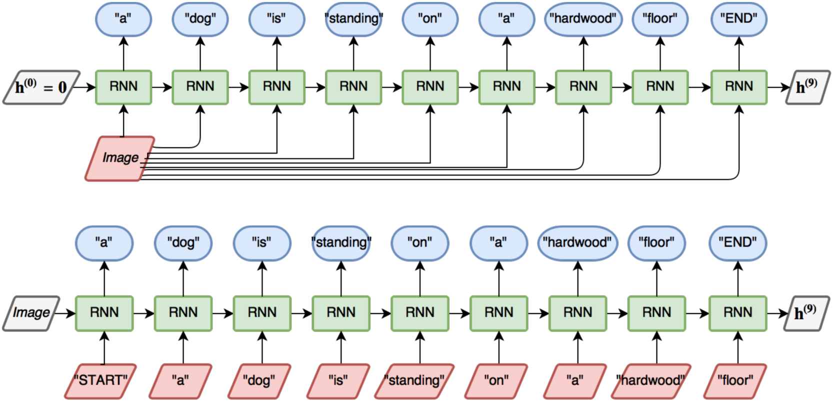
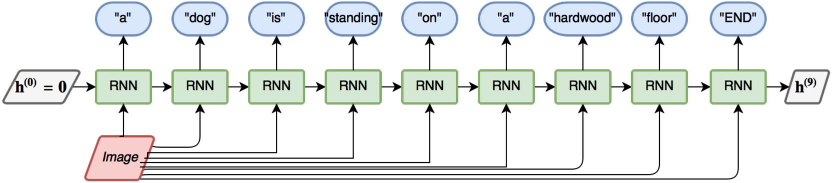


Image captioning: Practical Session

Model

- Using image & text (word) input at each time step
  - Image: VGG deep feature ⇒ 100 dim (PCA)
  - Text word: Glove embedding (100 dim + 2 for '<start>', '<end>')



- RNN layer + FC and soft-max
- No fine-tuning of image/text embeddings

Image captioning: Practical Session

Training

- Trained with cross-entropy-loss for predicting next word
- Use masking for handling sequence of different lengths

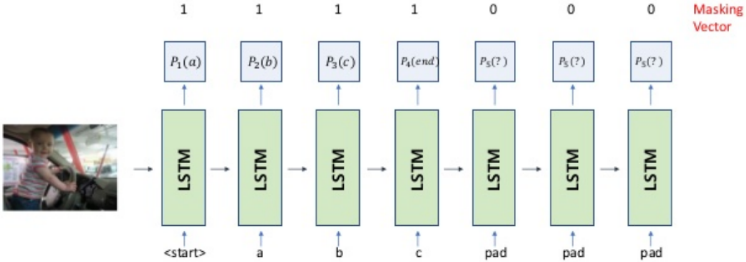


Image captioning: Practical Session

- **Evaluation** on FlickR8k : http://nlp.cs.illinois.edu/HockenmaierGroup/8k-pictures.html
  - Each image ⇔ 5 captions
  - Training: 6000 images, testing : 1000 images, validation: 1000 images
  - For speeding up limiting vocabulary size 100 ⇒ 1000
- Caption generation: soft-max (temperature) sampling (previous course)
  - Improvement: Beam search

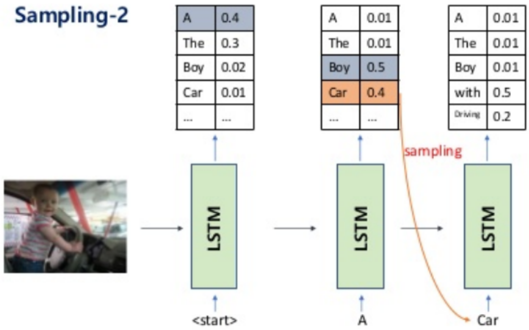


Image captioning: Practical Session

Evaluation on FlickR8k

- Some results with 1000 words, LSTM

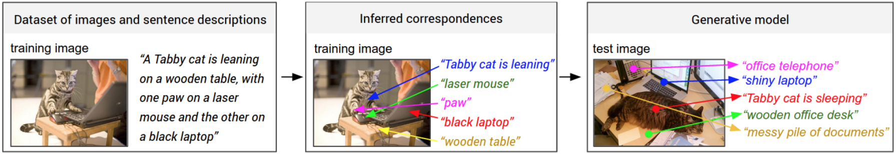
Caption n° 1: a girl in a swimsuit is swimming in the water . <end>
Caption n° 2: a girl is playing in the water . <end>
Caption n° 3: a young girl in a swimsuit is splashing in the water . <end>
Caption n° 4: a girl in a swimsuit is swimming in the water . <end>
Caption n° 5: a young girl in a swimsuit is swimming in the water . <end>

Caption n° 1: a dog is jumping over a hurdle . <end>
Caption n° 2: a dog jumps over a bar . <end>
Caption n° 3: a dog jumps over a bar . <end>
Caption n° 4: a dog jumps over a bar . <end>
Caption n° 5: a dog jumps over a hurdle . <end>

- Qualitative evaluation: compute a caption for each test image
- Compare each prediction to the five reference captions with BLUE score

Image captioning: Spatial Information

- Aligning Image regions with words [Karpathy and Li, 2015], CVPR



- Using object detector, *e.g.* R-CNN, top-19 detections (+whole) selected
- Compute max similarity between region embedding and word RNN outputs

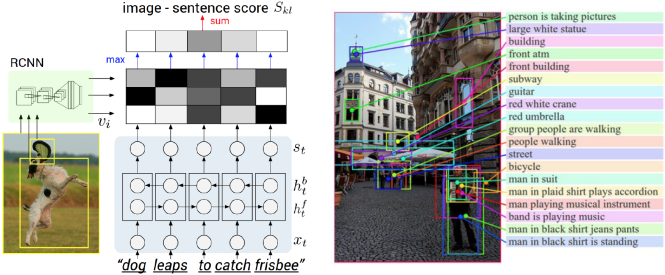
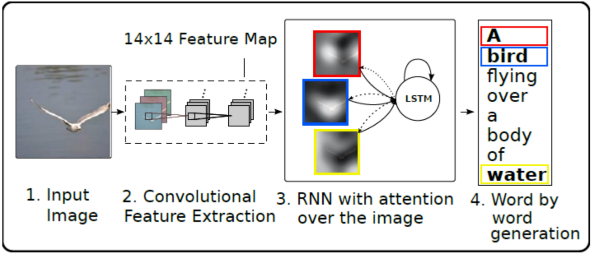


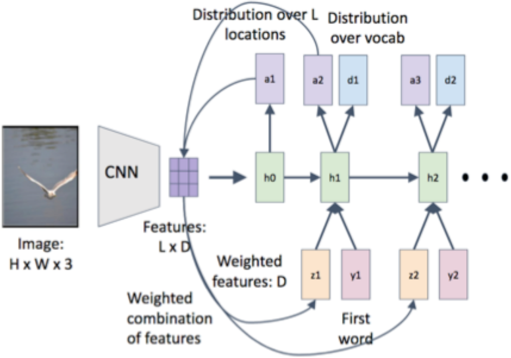
Image captioning: Spatial Information

- Attention: Show, Attend and Tell (SAT) [Xu et al., 2015], ICML
- Motivation: extends Show and Tell [Vinyals et al., 2015] by aligning image regions with word predictions
  - Use fully convolutional layer instead of full connected
- Hard attention: binary selection of region, non differentiable ⇒ reinforce
- Soft attention: weighted average of image region features



Show, Attend and Tell: Soft Attention

- a_i region feature $(L \times D)$, $(a_i, h_{t-1}) \Rightarrow$ MLP $e_{t,i} = f_{att}(a_i, h_{t-1})$
  - $+$ soft-max : $\alpha_{t,i} = softmax(e_{t,i})$
- LSTM $\hat{z}_t$ representation: context vector: $\hat{z}_t = \phi(a_i, \alpha_{t,i}) = \sum_i \alpha_{t,i} a_i$



Based on CS231n by Fei-Fei Li, Justin Johnson & Serena Yeung

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?

Visual Question Answering (VQA)

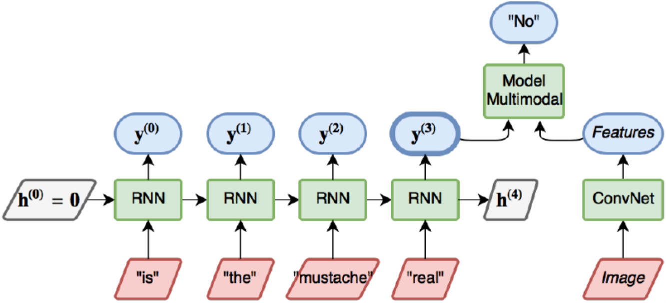
- Very complex task, that requires :
  - Precise image and text models
  - High level interaction modeling
  - Full scene understanding
  - Reasoning (*e.g.* spatial ...)

What color is the fire hydrant on the right ? yellow

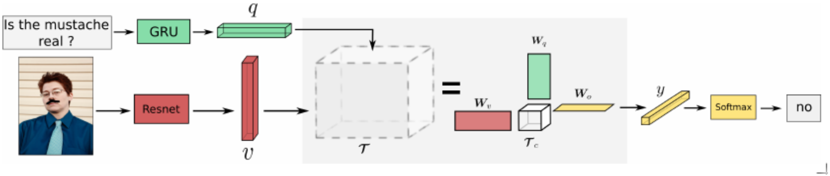
What color is the fire hydrant on the left ? green

Visual Question Answering (VQA)

- Input: question & image
- Output: answer



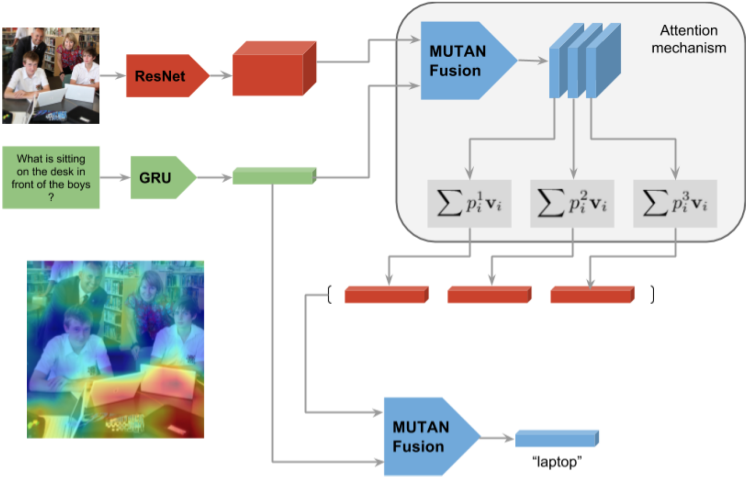
VQA : Multi-modal Fusion



- Mono-modal representations:
  - Visual representation: ResNet-152
  - Question representation: pre-trained GRU
- How to perform multi-modal fusion ⇒ Tucker decomposition [Ben-younes et al., 2017]

VQA : Attention

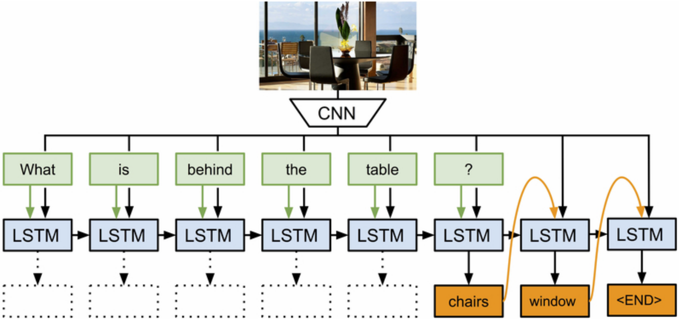
- Attention (glimpses) also used in VQA to bias spatial region analysis depending on question



Ongoing Issues in Deep Learning

New Tasks in Artificial Intelligence

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



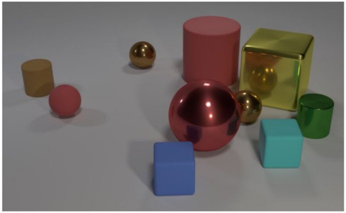
Credit: M. Malinowski [Malinowski et al., 2015]

Datasets and Biaises

- Many datasets have been used, especially for VQA
- Important biases, *e.g.* textual: What sport is? ⇒ Tennis (41%)
- VQA 1.0 ⇒ VQA 2.0: makes image needed to answer
  - VQA-CP: different prior distributions in train / test to limit biaises

Datasets and Reasoning

- Synthetic CLEVR dataset: spatial and relational reasoning
- Counterfactual reasoning: important primitive?



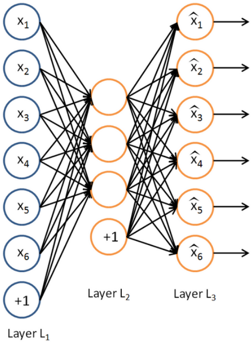
Q: Are there an equal number of large things and metal spheres? **Q:** What size is the cylinder that is left of the brown metal thing that is left of the big sphere? **Q:** There is a sphere with the same size as the metal cube; is it made of the same material as the small red sphere? **Q:** How many objects are either small cylinders or metal things?

What would happen if the flamingo folds his legs?

Outline

Unsupervised Learning

- Standard criterion for unsupervised training: reconstruction error, *e.g.* Mean Squared Error (MSE), Maximum likelihood *etc*
- Ex: Auto-encoders: $z = f(Wx)$, $\tilde{x} = g(W^t x)$
  - Auto-encoder objective function: $C = \sum_{i=1}^{N} ||x_i - \tilde{x}||^2$

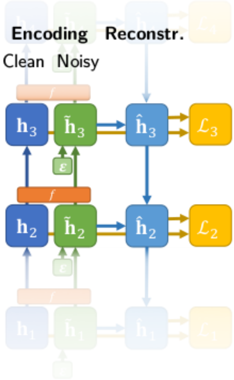


Unsupervised Learning

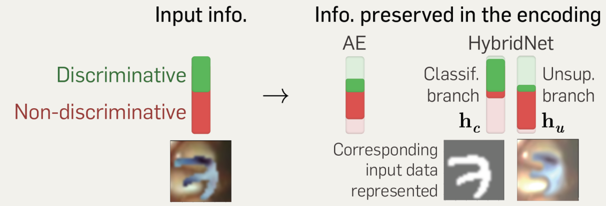
- Success of deep learning essentially for supervised tasks, *e.g.* classification
- Unsupervised deep learning no comparable breakthrough, **WHY?**
  $\Rightarrow$ Classification: clear objective (discrimination) *vs*
  $\Rightarrow$ Reconstruction: questionable
  - Fitting data well: what if ultimate goal is classification, generalization to a set of examples ?
  - Reconstruction not required, or even not a good idea
    - Deeper representation $\Leftrightarrow$ more abstract representations $\Leftrightarrow$ generalization $\Leftrightarrow$ loss of information
- Two current alternatives to unsupervised learning:
  1. Objective without reconstruction
  2. Casting unsupervised training as classification

Beyond Reconstruction: Ladder Networks [Rasmus et al., 2015]

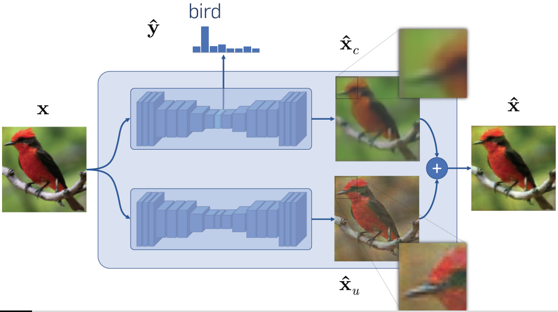
- "An autoencoder which can discard information"
- Layer above does not reconstruct layer below only with its activation
- Solution: Provide the details to learn only the abstract features
  - Decoder has a noisy version of the input to reconstruct



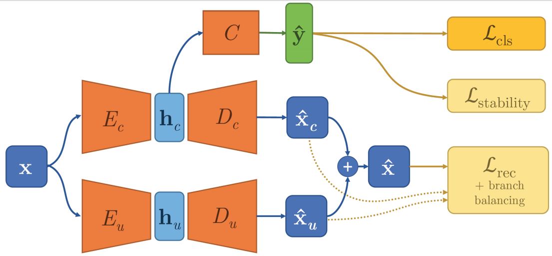
Beyond Reconstruction: HybridNet [Robert et al., 2018]



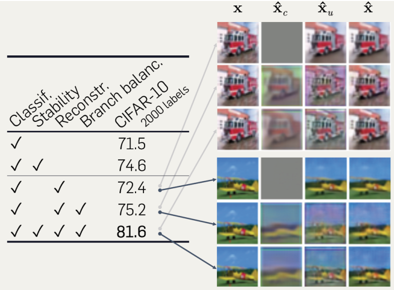
Separation of discriminative and complementary information for reconstruction into two branches



Beyond Reconstruction: HybridNet [Robert et al., 2018]



Controls the behavior of information separation



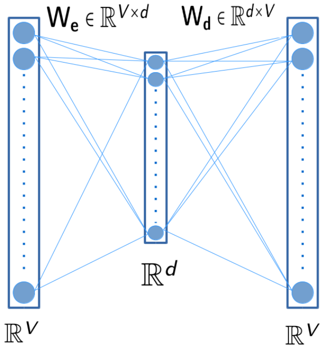
- Encourage invariant features in Ec
  - Classification + stability loss
- Additional info.
  - Reconstruction loss + branch balancing

Auto-Supervision & Predictive Learning

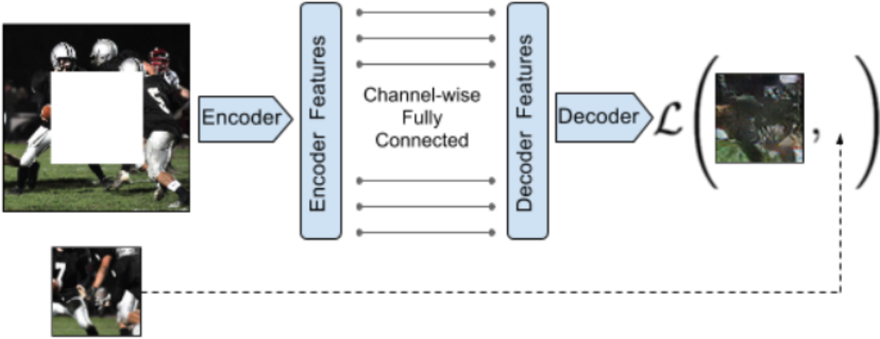
- Transformed unsupervised problem to a supervised one
- Automatically creating labels, exploiting "neighborhood", *e.g.*
  - Spatial
  - Temporal

Word2Vec [Mikolov et al., 2013]

- Embedding of words, *i.e.* assign each one-hot word $\in \mathbb{R}^V$ a vector $\in \mathbb{R}^d$
- Word2Vec principle: predict a word given its context
  - Assumption: similar words appears in similar contexts
  - Input: Bag of Words of context
  - Project to a given space, apply soft max to classify the central word



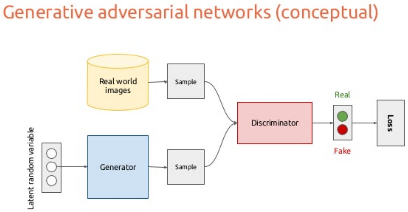
Context-Encoders [Pathak et al., 2016]: Word2Vec for Images

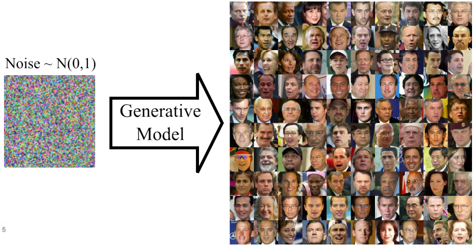


Ongoing Issues in Deep Learning

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) [Goodfellow et al., 2014]

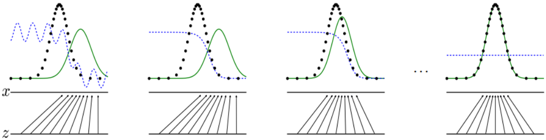




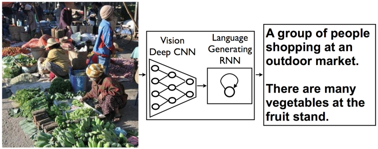
Ongoing Issues in Deep Learning

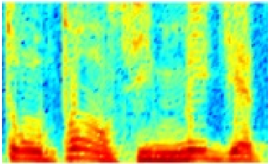
Unsupervised Training: GAN

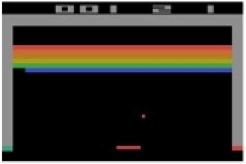
- Unsupervised problem ⇒ 2-player game theory problem
- Interesting results: optimal generator learns data distribution



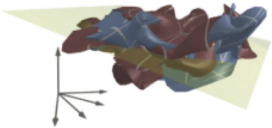
Ongoing Issues in Deep Learning: DL Theory

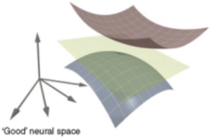






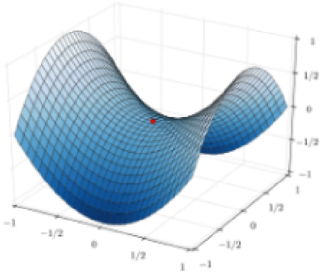
- Deep Learning: huge impact in terms of experimental results
- BUT: formal understanding still limited,
  - Optimization: non-convex problem
  - Model: ability to untangle manifold
  - Robustness to over-fitting & generalization





Non-Convex Optimization

- One of the main historical shortcoming of deep neural networks
- In pratice, not really an issue with modern neural networks, WHY?
- Some preliminary answer elements:
  - In high dimension, few local minima
    but many saddle points [Dauphin et al., 2014]
    - Empirically, gradient descent methods manage to
      escape [Goodfellow and Vinyals, 2015] saddle points



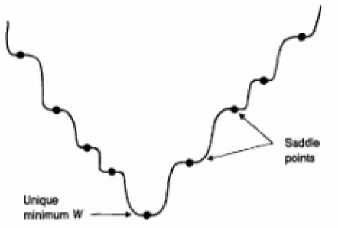
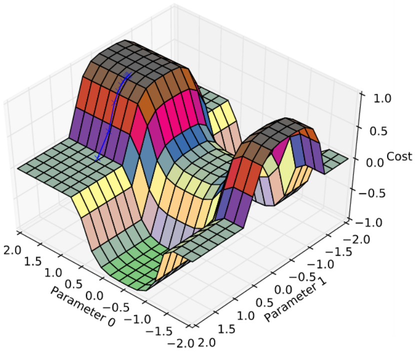


FIGURE 2. The landscape of E.

Non-Convex Optimization

- WHY non-convex optimization ist not a major practical issue for deep learning?
- Some preliminary answer elements:
  - Most of local minima have about the same objective value [Haeffele and Vidal, 2015, Choromanska et al., 2014]

(Cartoon of Dauphin et al 2014's worldview)

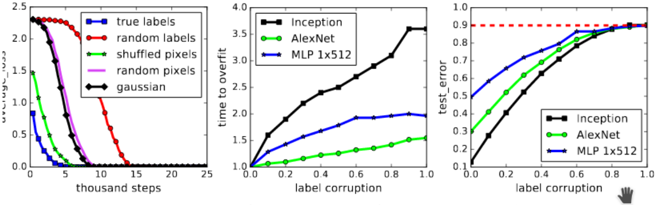


Deep Learning and generalization

- 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 [Zhang et al., 2017]

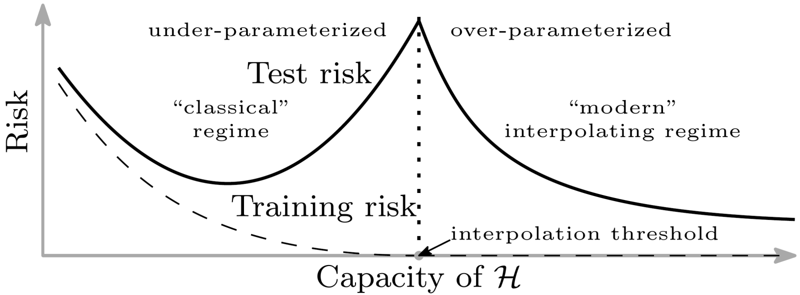


(a) learning curves (b) convergence slowdown (c) generalization error growth

  - 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**

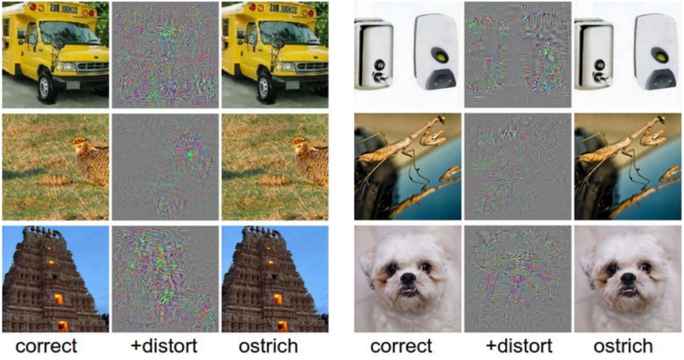
Generalization and over-parametrized models

- Double U-curve phenomena observed with deep models! [Belkin et al., 2019]



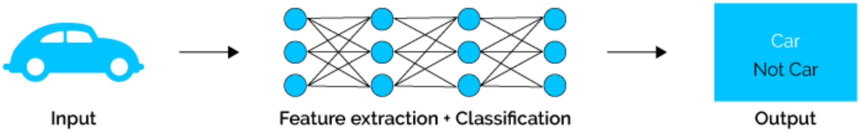
Deep Learning (DL) & Stability

- Deep Models not necessarily robust to input variations
- Ex: Adversarial Examples



Deep Learning (DL) & Uncertainty

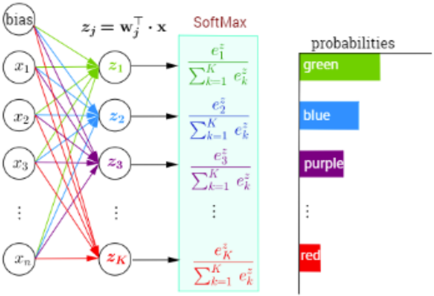
Softmax output in deep neural network $\neq$ confidence (uncertainty) measure!

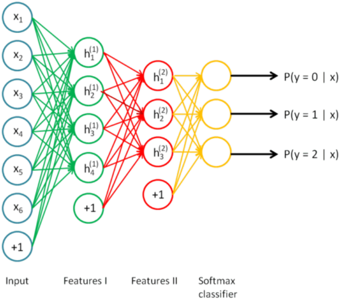


- Often wrong prediction $\leftrightarrow$ unjustified high confidence
- Uncertainty however crucial in major applicative domains:
  - Healthcare
  - Autonomous driving
  - Nuclear

Deep Neural Networks: Black Boxes

- **Lack of confidence estimate** (uncertainty): how (un)certain about decision?
- Softmax classification: probability distribution over output given input?





  - Only with single layer model, *i.e.* logistic regression
  - Bayesian Neural Nets: scalability issues

Deep Learning Theory

Formal theory explaining deep learning success: infancy

- **Optimization:** preliminary results for non-convex functions [Dauphin et al., 2014, Choromanska et al., 2014, Goodfellow and Vinyals, 2015, Haeffele and Vidal, 2015]
- **Regularization:** to be established
- **Stability:** studies under signal processing perspective [Bruna and Mallat, 2013], kernel methods [Bietti and Mairal, 2017]
- **Uncertainty:** preliminary connections between Bayesian models and dropout [Gal and Ghahramani, 2016]

TO BE CONTINUED ...

References I

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

[Belanger and McCallum, 2016] Belanger, D. and McCallum, A. (2016).
Structured prediction energy networks.
In *ICML*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 983–992. JMLR.org.

[Belkin et al., 2019] Belkin, M., Hsu, D., Ma, S., and Mandal, S. (2019).
Reconciling modern machine-learning practice and the classical bias–variance trade-off.
Proceedings of the National Academy of Sciences, 116(32):15849–15854.

[Ben-younes et al., 2017] Ben-younes, H., Cadène, R., Cord, M., and Thome, N. (2017).
MUTAN: multimodal tucker fusion for visual question answering.
CoRR, abs/1705.06676.

[Bietti and Mairal, 2017] Bietti, A. and Mairal, J. (2017).
Group invariance and stability to deformations of deep convolutional representations.
CoRR, abs/1706.03078.

[Bruna and Mallat, 2013] Bruna, J. and Mallat, S. (2013).
Invariant scattering convolution networks.
IEEE Trans. Pattern Anal. Mach. Intell., 35(8):1872–1886.

[Cao et al., 2017] Cao, Z., Simon, T., Wei, S.-E., and Sheikh, Y. (2017).
Realtime multi-person 2d pose estimation using part affinity fields.
In *CVPR*.

[Chen et al., 2015a] Chen, L., Schwing, A. G., Yuille, A. L., and Urtasun, R. (2015a).
Learning deep structured models.
In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pages 1785–1794.

References II

[Chen et al., 2015b] Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L. (2015b).
Semantic image segmentation with deep convolutional nets and fully connected crfs.
In *ICLR*.

[Choromanska et al., 2014] Choromanska, A., Henaff, M., Mathieu, M., Arous, G. B., and LeCun, Y. (2014).
The loss surface of multilayer networks.
CoRR, abs/1412.0233.

[Dauphin et al., 2014] Dauphin, Y., Pascanu, R., Gülçehre, Ç., Cho, K., Ganguli, S., and Bengio, Y. (2014).
Identifying and attacking the saddle point problem in high-dimensional non-convex optimization.
CoRR, abs/1406.2572.

[Durand et al., 2015] Durand, T., Thome, N., and Cord, M. (2015).
MANTRA: Minimum Maximum Latent Structural SVM for Image Classification and Ranking.
In *International Conference on Computer Vision (ICCV)*.

[Durand et al., 2016] Durand, T., Thome, N., and Cord, M. (2016).
WELDON: Weakly Supervised Learning of Deep Convolutional Neural Networks.
In *Computer Vision and Pattern Recognition (CVPR)*.

[Gal and Ghahramani, 2016] Gal, Y. and Ghahramani, Z. (2016).
Dropout as a bayesian approximation: Representing model uncertainty in deep learning.
In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, pages 1050–1059. JMLR.org.

[Girshick et al., 2014] Girshick, R., Donahue, J., Darrell, T., and Malik, J. (2014).
Rich feature hierarchies for accurate object detection and semantic segmentation.
In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

References III

[Goodfellow et al., 2014] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014).
Generative adversarial nets.
In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc.

[Goodfellow and Vinyals, 2015] Goodfellow, I. J. and Vinyals, O. (2015).
Qualitatively characterizing neural network optimization problems.
In *ICLR*.

[Haeffele and Vidal, 2015] Haeffele, B. D. and Vidal, R. (2015).
Global optimality in tensor factorization, deep learning, and beyond.
CoRR, abs/1506.07540.

[Karpathy and Li, 2015] Karpathy, A. and Li, F. (2015).
Deep visual-semantic alignments for generating image descriptions.
In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 3128–3137.

[Kulkarni et al., 2016] Kulkarni, P., Jurie, F., Zepeda, J., Pérez, P., and Chevallier, L. (2016).
Spleap: Soft pooling of learned parts for image classification.
In *ECCV*.

[Malinowski et al., 2015] Malinowski, M., Rohrbach, M., and Fritz, M. (2015).
Ask your neurons: A neural-based approach to answering questions about images.
In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 1–9.

[Mikolov et al., 2013] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013).
Distributed representations of words and phrases and their compositionality.
In Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

References IV

[Mordan et al., 2017] Mordan, T., Durand, T., Thome, N., and Cord, M. (2017).
WILDCAT: Weakly Supervised Learning of Deep ConvNets for Image Classification, Localization and Segmentation.
In *Computer Vision and Pattern Recognition (CVPR)*.

[Noh et al., 2017] Noh, H., Araujo, A., Sim, J., Weyand, T., and Han, B. (2017).
Large-scale image retrieval with attentive deep local features.
In *The IEEE International Conference on Computer Vision (ICCV)*.

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

[Oquab et al., 2014] Oquab, M., Bottou, L., Laptev, I., Sivic, J., et al. (2014).
Learning and transferring mid-level image representations using convolutional neural networks.
In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Pathak et al., 2016] Pathak, D., Krähenbühl, P., Donahue, J., Darrell, T., and Efros, A. (2016).
Context encoders: Feature learning by inpainting.

[Pinheiro and Collobert, 2015] Pinheiro, P. O. and Collobert, R. (2015).
From image-level to pixel-level labeling with convolutional networks.
In *CVPR*.

[Rasmus et al., 2015] Rasmus, A., Valpola, H., Honkala, M., Berglund, M., and Raiko, T. (2015).
Semi-supervised learning with ladder networks.
In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, pages 3546–3554, Cambridge, MA, USA. MIT Press.

[Robert et al., 2018] Robert, T., Thome, N., and Cord, M. (2018).
Hybridnet: Classification and reconstruction cooperation for semi-supervised learning.
In *The European Conference on Computer Vision (ECCV)*.

References V

[Sutskever et al., 2014] Sutskever, I., Vinyals, O., and Le, Q. V. (2014).
Sequence to sequence learning with neural networks.
In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc.

[Uijlings et al., 2013] Uijlings, J. R. R., van de Sande, K. E. A., Gevers, T., and Smeulders, A. W. M. (2013).
Selective search for object recognition.
International Journal of Computer Vision, 104(2):154–171.

[Vinyals et al., 2015] Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2015).
Show and tell: A neural image caption generator.
In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Wang et al., 2016] Wang, S., Fidler, S., and Urtasun, R. (2016).
Proximal deep structured models.
In *NIPS*, pages 865–873.

[Xu et al., 2015] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., and Bengio, Y. (2015).
Show, attend and tell: Neural image caption generation with visual attention.
In Blei, D. and Bach, F., editors, *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 2048–2057. JMLR Workshop and Conference Proceedings.

[Zhang et al., 2017] Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. (2017).
Understanding deep learning requires rethinking generalization.

References VI

[Zheng et al., 2015] Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., Du, D., Huang, C., and Torr, P. H. S. (2015).
Conditional random fields as recurrent neural networks.
In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, pages 1529–1537, Washington, DC, USA. IEEE Computer Society.

[Zhou et al., 2016] Zhou, B., Khosla, A., A., L., Oliva, A., and Torralba, A. (2016).
Learning Deep Features for Discriminative Localization.
CVPR.