

Beyond Full Supervision in Deep Learning

Nicolas Thome - Prof. at Cnam Paris
CEDRIC Lab, MSDMA Team

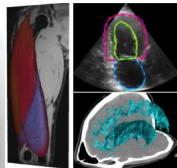
DeepImaging 2019 - PRISMES LABEX
April 18, 2019



80 years of building new worlds
through knowledge

Deep learning for medical imaging school

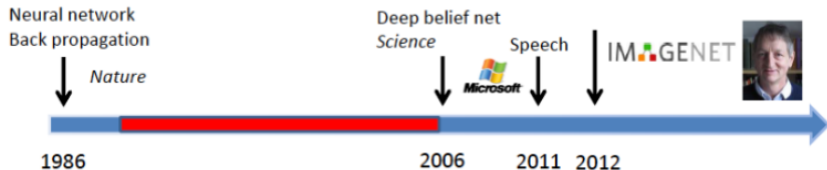
April 15—19 2019, Campus de la Doua, Lyon



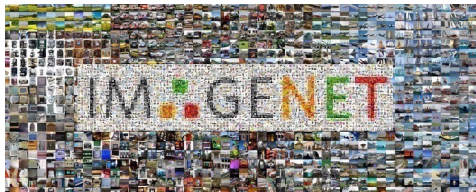
le cnam



Deep Learning Success since 2010



- ▶ ILSVRC'12: the deep revolution
⇒ outstanding success of ConvNets [Krizhevsky et al., 2012]



Rank	Name	Error rate	Description
1	U. Toronto	0.15315	Deep learning
2	U. Tokyo	0.26172	Hand-crafted features and learning models.
3	U. Oxford	0.26979	Bottleneck.
4	Xerox/INRIA	0.27058	

2012: the deep revolution

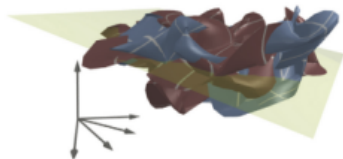
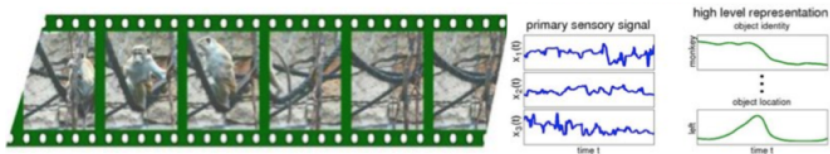
Deep ConvNet success at ILSVRC'12

Two main practical reasons:

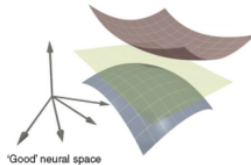
1. Huge number of labeled images (10^6 images)
 - ▶ Possible to train very large models without over-fitting
 - ▶ Larger models enables to learn rich (semantic) features hierarchies
2. GPU implementation for training
 - ▶ Relatively cheap and fast GPU
 - ▶ Training time reduced to 1-2 weeks (up to 50x speed up)



Representation Learning & Manifold Untangling



Raw data:
very tangled manifold



Deep Learning representations:
untangled manifold

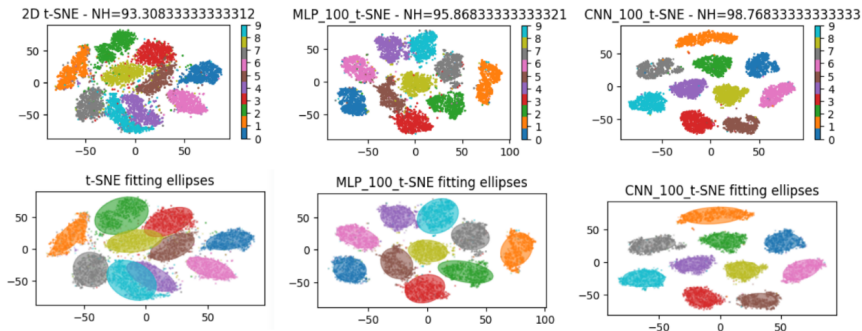
- ▶ Deep Learning models gradually disentangle data manifold
- ▶ Deformations linearized: simple classifier in disentangled space!

Manifold Disentangling and ConvNets

- ▶ Visualize data in input vs latent dimension with t-SNE [van der Maaten and Hinton, 2008]
- ▶ Ex: MNIST dataset

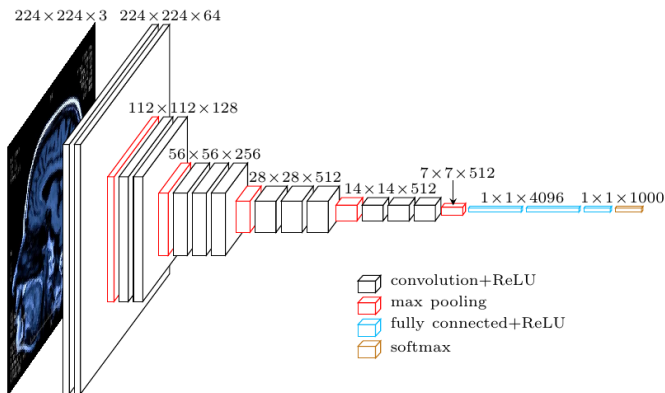


- ▶ Deep models able to disentangle data manifold!



Deep Learning (DL) for small-scale Datasets

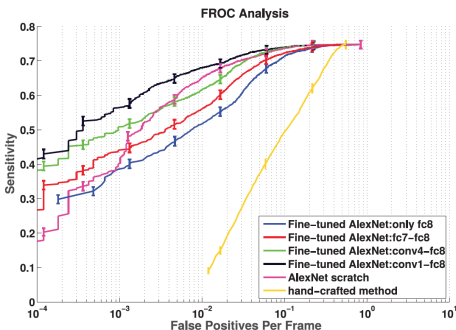
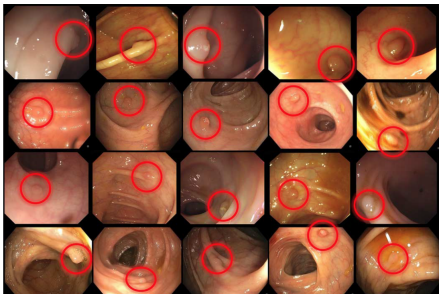
- ▶ Deep ConvNets require large-scale annotated datasets
- ▶ **Do we need to collect ImageNet scale dataset for medical image analysis?**
- ▶ **OPTION:** transferring representations learned from ImageNet:
extract layer (fixed-size vector) \Rightarrow **"Deep Features" (DF)**



- ▶ **Now state-of-the-art for any visual recognition task [Azizpour et al., 2016]**

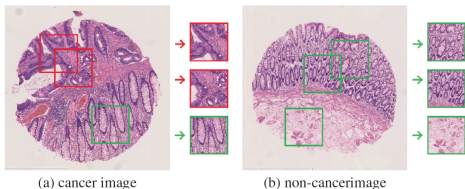
Deep Learning (DL) for Medical Image Analysis

- ▶ **Deep Features very robust to domain shifts, e.g. medical images**
- ▶ Transfer & fine-tuning (ImageNet), e.g. Polyp Detection [Tajbakhsh et al., 2016]
- ▶ ConvNets: winners of recent challenges based on deep learning: Mammography, Melanoma Detection, etc
- ▶ Using ImageNet pre-training, e.g. Liver Tumor Segmentation (LiTS'17) challenge [Li et al., 2017]

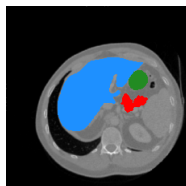


Deep Learning (DL) for Medical Image Analysis

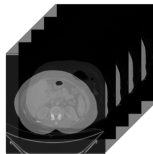
- ▶ Large-scale datasets in medical imaging: more the exception than the rule
- ▶ Data labeling expensive, especially fine-grained annotations (e.g. segmentation)
 - ▶ Exacerbated in medical context: strong expertise required for labeling
- ▶ Solutions to tackle small-scale datasets with deep learning in this context:
 - ▶ Leveraging coarse annotations to perform precise predictions
 - ▶ Using (many) unlabelled data in addition to (few) labeled data



From [Xu et al., 2014]



Few labeled data



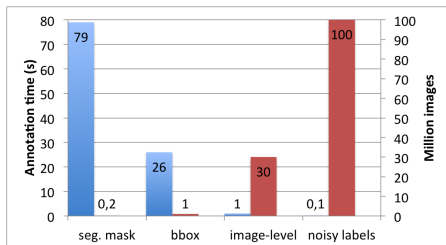
Many unlabeled

Outline

- 1 Learning with Weak Supervision
- 2 Semi-Supervised Learning

Weakly Supervised Learning

- ▶ Using full (precise) annotation, e.g. BB or segmentation masks
- ▶ **BUT:** full annotations expensive [Bearman et al., 2016]
 - ▶ Problem even more pronounced with medical images, e.g. segmentation often prohibitive
 - ▶ High resolution
 - ▶ 3D data
 - ▶ Videos
 - ▶ ⇒ **Training with weak supervision**, for performing accurate predictions
 - ▶ Ex: semantic segmentation from global labels

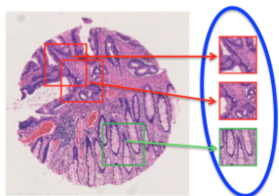


Multiple Instance Learning (MIL)

- ▶ Multiple Instance Learning (MIL) [Dietterich et al., 1997]: old model for Weakly Supervised Learning
- ▶ Model formulation: Example **b** composed of a bag of N_b instances:

$$\mathbf{b} = \{\mathbf{x}_h\}_{h \in \{1; N_b\}}$$

- ▶ **b**: image, $\{\mathbf{x}_h\}$ image regions
- ▶ **b**: text document, $\{\mathbf{x}_h\}$ paragraphs
- ▶ **b**: molecule, $\{\mathbf{x}_h\}$ molecule parts

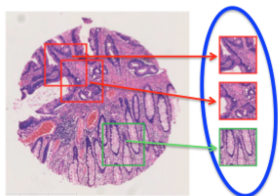


From [Xu et al., 2014]



Multiple Instance Learning (MIL)

- ▶ Example \mathbf{b} composed of a bag of N_b instances: $\mathbf{b} = \{\mathbf{x}_h\}_{h \in \{1; N_b\}}$
- ▶ Each instance \mathbf{x}_h is described by a feature vector $\phi(\mathbf{b}, h) \in \mathbb{R}^d$
- ▶ Ex: \mathbf{x}_h image region
 - ▶ $\phi(\mathbf{b}, h) \in \mathbb{R}^d$ pixels
 - ▶ $\phi(\mathbf{b}, h) \in \mathbb{R}^d$ handcrafted features (SIFT/HOG, etc)
 - ▶ $\phi(\mathbf{b}, h) \in \mathbb{R}^d$ Deep features

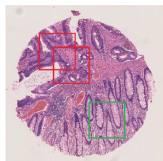


From [Xu et al., 2014]

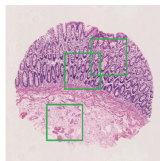
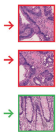


Multiple Instance Learning (MIL)

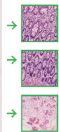
- ▶ Example \mathbf{b} composed of a bag of N_b instances: $\mathbf{b} = \{\mathbf{x}_h\}_{h \in \{1; N_b\}}$
- ▶ MIL training formulation: A set a training N pairs $(\mathbf{b}_i, \mathbf{y}_i^*)$
 - ▶ $\mathbf{b}_i = \{\mathbf{x}_{i,h}\}_{h \in \{1; N_{b_i}\}}$ i^{st} example
 - ▶ \mathbf{y}_i^* GT label, e.g. $\mathbf{y}_i^* = \pm 1$ for binary classification
 - ▶ **Weak supervision:** \mathbf{y}_i^* provided at bag level
 - ▶ **MIL goal:** performing predictions at instance level



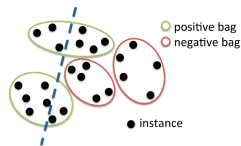
(a) cancer image



(b) non-cancer image



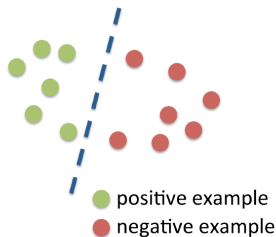
Supervised learning



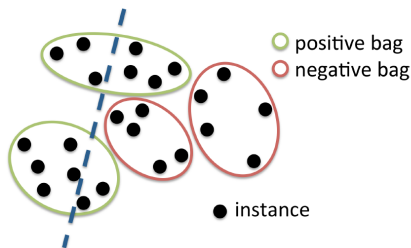
Multiple-Instance Learning (MIL)

Multiple Instance Learning (MIL)

- ▶ **MIL: Weak supervision:** \mathbf{y}_i^* provided at bag level \mathbf{b}_i , not at instance level $\mathbf{x}_{i,h}$
- ▶ **MIL hypothesis:** all instance in negative bags are negative
- ▶ **We need to pool (aggregate) over instances to train the model!**
 - ▶ Pooling over instance features: $g(\{\phi(\mathbf{b}_i, h)\}) := \phi_p(\mathbf{b}_i) \in \mathbb{R}^{d'}$, e.g. g avg or max
 - ▶ Perform bag prediction $\phi_p(\mathbf{b}_i)$ with prediction f_w : $\hat{y}_i = f_w(\phi_p(\mathbf{b}_i))$
 - ▶ Use any fully supervised learning algorithm to train f_w from \mathbf{y}_i^*
 - ▶ \ominus not straightforward to perform instance prediction for general pooling function f and learning algorithm



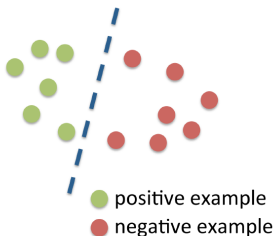
Supervised learning



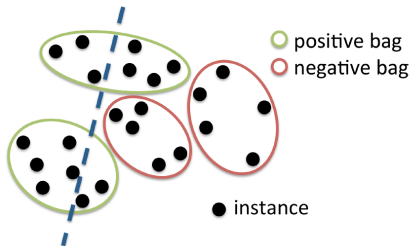
Multiple-Instance Learning (MIL)

Multiple Instance Learning (MIL)

- ▶ **MIL: Weak supervision:** y_i^* provided at bag level \mathbf{b}_i , not at instance level $\mathbf{x}_{i,h}$
- ▶ **We need to pool (aggregate) over instances to train the model!**
 - ▶ Pooling over instance prediction scores:
 - ▶ Define predictor at the instance level $f_w(\phi(\mathbf{b}_i, h)), \forall h \in \{1; N_{\mathbf{b}_i}\}$
 - ▶ Ex: binary classification: $f_w(\phi(\mathbf{b}_i, h)) \in \mathbb{R}, \text{sign}[f_w(\phi(\mathbf{b}_i, h))] \in \{-1; 1\}$
 - ▶ Pool over prediction scores to get bag prediction: $\hat{y}_i = g\{f_w(\phi(\mathbf{b}_i, h))\}$, e.g. g avg or max



Supervised learning



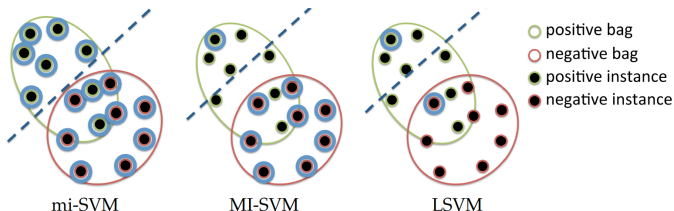
Multiple-Instance Learning (MIL)

Multiple Instance Learning

- ▶ SVM-MIL algorithms, e.g. [Andrews et al., 2003]: binary classification
 - ▶ Linear predictor on instances, i.e. $f_{\mathbf{w}}(\phi(\mathbf{b}_i, h)) = \langle \mathbf{w}; \phi(\mathbf{b}_i, h) \rangle$
 - ▶ Max pooling function g over instance scores \Rightarrow bag prediction:

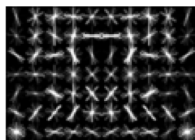
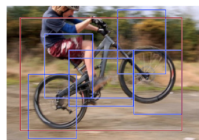
$$f_{\mathbf{w}}(\mathbf{b}_i) = \text{sign} \left[\max_{h \in N_{\mathbf{b}_i}} \langle \mathbf{w}, \phi(\mathbf{b}_i, h) \rangle \right] \quad (1)$$

- ▶ Training variants:
 - ▶ LSVM: use max prediction for \oplus and \ominus bags
 - ▶ MI-SVM: use max prediction for \oplus but all \ominus instances
 - ▶ mi-SVM: use all \ominus instances and relabel $y_{i,h}^* \in \pm 1$ all \oplus instances

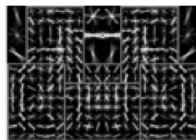


Multiple Instance Learning

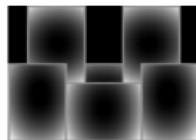
- ▶ SVM-MIL algorithms: historically applied to part-based object detection [Felzenszwalb et al., 2010] \Rightarrow **Deformable Part Model (DPM)**
- ▶ Adapted in the object detection context
 - ▶ Supervision: bounding box
 - ▶ Latent variable: position of object "parts"
 - ▶ Features for each part $\phi(\mathbf{b}_i, h)$: Handcrafted HoG



(a) Root filter



(b) Part filters in higher resolution

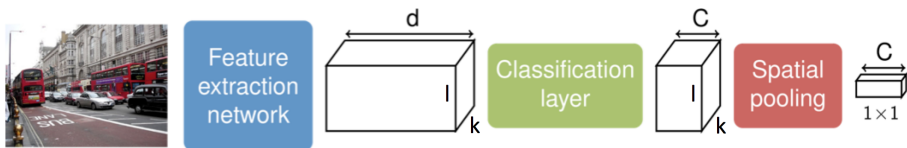


(c) A spatial model for part locations

- ▶ PASCAL VOC "Lifetime Achievement" Prize in 2010
- ▶ PAMI Longuet-Higgins Prize at CVPR'18 (Retrospective Best Paper from CVPR'08)

Multiple Instance Learning and Deep Learning

- ▶ Using MIL model in the Deep Learning era: deep architecture for WSL

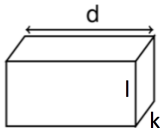


- ▶ **Feature extractor** \Rightarrow tensor of size $k \times l \times d$
- ▶ **MIL notations:** $N_b = k \times l$ instances (regions)
 - ▶ Each instance h represented by deep features $\phi(b, h) \in \mathbb{R}^d$

Multiple Instance Learning and Deep Learning



Feature
extraction
network



Classification
layer

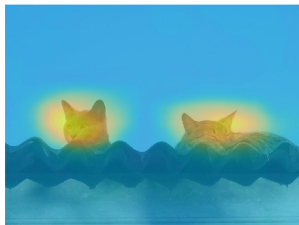


Spatial
pooling



- ▶ **Classification:** projection to get a class prediction for each instance

- ▶ $z_h^c = f_{w_c}(\phi(\mathbf{b}_i, h)), \forall h \in \{1; N_b\}, \forall c \in \{1; C\}$
- ▶ $k \times l \times C$ tensor: Class Activation Maps (CAM)



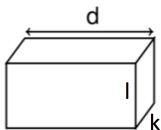
- ▶ **Pooling:** class prediction aggregation to train model from global labels

$$\hat{z}_c = g \left[\{z_h^c\}_{h \in \{1; N_b\}} \right], \forall c \in \{1; C\}$$

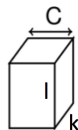
How to pool?



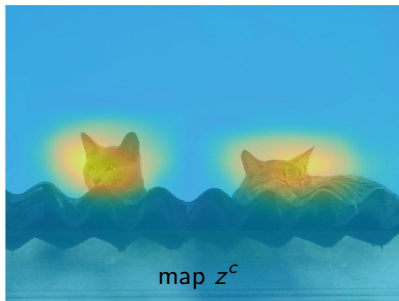
Feature
extraction
network



Classification
layer



Spatial
pooling



spatial
pooling
→ ●
score y^c

Max [Oquab et al., 2015]

$$y^c = \max_h z_h^c$$

Average (GAP) [Zhou et al., 2016]

$$y^c = \frac{1}{N} \sum_h z_h^c$$

Average pooling limitation

- ▶ Classifying with all regions
- ▶ Not efficient for small objects: lots of “noisy” regions



Max pooling

$$y^c = \max_h z_h^c \quad (2)$$

- ▶ Classifying only with the max scoring region

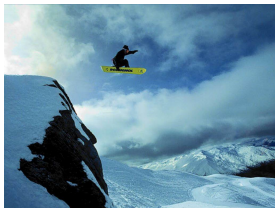


- ▶ Loss of contextual information

Max pooling

$$y^c = \max_h z_h^c \quad (2)$$

- ▶ Classifying only with the max scoring region



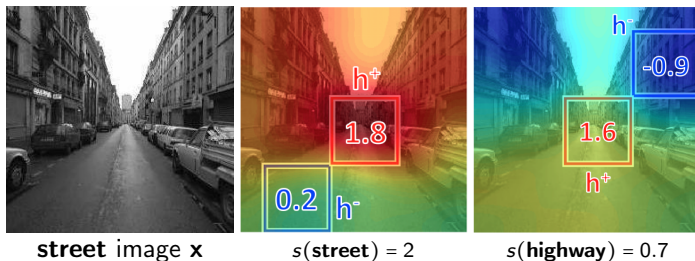
- ▶ Loss of contextual information

max+min pooling

- ▶ **MANTRA [Durand et al., 2015]: max+min pooling function**

$$y^c = \max_h z_h^c + \min_h z_h^c \quad (3)$$

- ▶ h^+ : presence of the class \rightarrow high h^+
- ▶ h^- : localized evidence of the absence of class: **negative evidence**



Generalize pooling function [Durand et al., 2019]

$$y^c = \frac{1}{2\beta_h^+} \log \left[\frac{1}{|\mathcal{H}|} \sum_{\mathbf{h} \in \mathcal{H}} e^{\beta_h^+ z_h^c} \right] + \frac{1}{2\beta_h^-} \log \left[\frac{1}{|\mathcal{H}|} \sum_{\mathbf{h} \in \mathcal{H}} e^{\beta_h^- z_h^c} \right] \quad (4)$$

- ▶ Varying β_h^+ , $\beta_h^- \Rightarrow$ recovering pooling functions used in well-known probabilistic and max-margin models
- ▶ Smoothly interpolate between these extreme cases

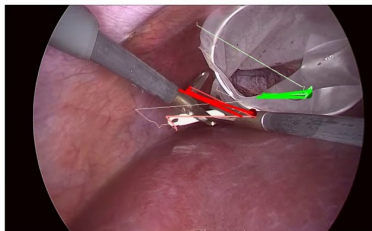
Model	Pooling Function	β_h^+	β_h^-
HCRF [Quattoni et al., 2007]	log-sum-exp	1	1
GAP [Zhou et al., 2016]	average	$\rightarrow 0$	$\rightarrow 0$
LSSVM [Yu and Joachims, 2009]	max	$\rightarrow +\infty$	$\rightarrow +\infty$
MANTRA [Durand et al., 2015]	max+min	$\rightarrow +\infty$	$\rightarrow -\infty$

Table: State-of-the-art WSL models with corresponding parameters.

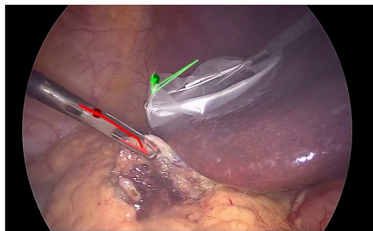
MIL for medical image analysis

- ▶ MIL directly adapted for detection of pattern from global label in medical image/videos
 - ▶ Specific lesion type in images
 - ▶ Specific surgical gesture in videos, e.g. [Nwoye et al., 2019]

Model Trained on 1-fps videos & Tested on 25-fps videos



surgery 1



surgery 2



grasper



bipolar



hook



scissors



clipper



irrigator



specimen bag

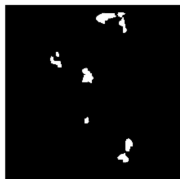
Note: the method detects only one instance per type of tool

MIL for medical image analysis

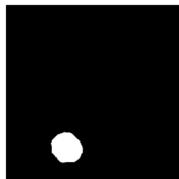
- ▶ Medical images: high resolution with small details
 - ▶ Multi-resolution adaptation MIL [Quellec et al., 2012]
 - ▶ Weighted average over scales



(a) resized image



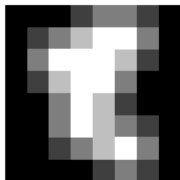
(b) CWS-segmentation



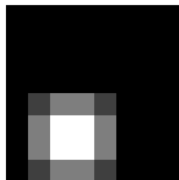
(c) IRMA-segmentation



(d) local relevance



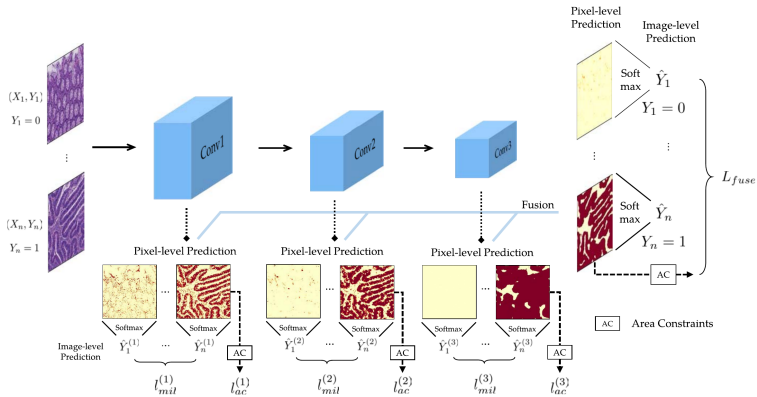
(e) CWS-label



(f) IRMA-label

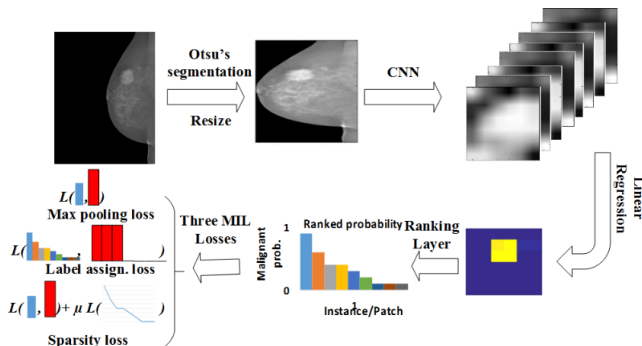
MIL for medical image analysis

- ▶ MIL with constraints [Jia et al., 2017]
 - ▶ Deep MIL (max pool) with FCN for Histopathology
 - ▶ Multi-resolution: MIL loss applied at various conv layers
 - ▶ Leveraging additional annotation, *i.e.* relative area size of the cancerous region within image



MIL for medical image analysis

- ▶ Integrating constraints from medical knowledge in deep MIL objective [Zhu et al., 2017]
 - ▶ Deep MIL (max pool) for lesion detection in mammography
 - ▶ MIL loss including sparse prior constraint on lesion classification
 - ▶ Lesion $\sim 2\%$ of image size

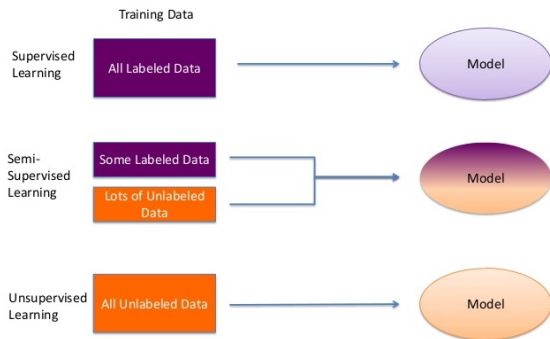


Outline

- 1 Learning with Weak Supervision
- 2 Semi-Supervised Learning**

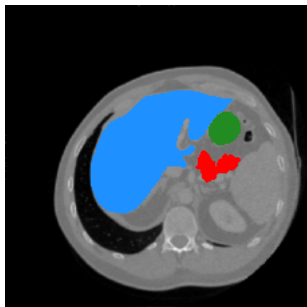
Semi Supervised Learning (SSL)

- ▶ Semi-supervised vs fully supervised vs unsupervised
- ▶ Some (few) labeled data, many unlabeled data
 - ▶ Medical context: annotations costly \Rightarrow SSL useful

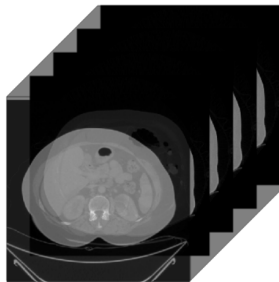


Credit: S. Jain

Semi Supervised Learning (SSL)



Few labeled data



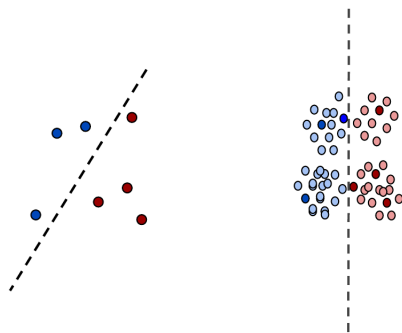
Many unlabeled

► Two main strategies :

1. Adapting supervised objective with unlabelled data
2. Use alternative objective for unlabelled data, e.g. reconstruction

SSL: Adapting supervised objective to unlabeled data

- ▶ Using unlabeled data structure, e.g. transductive SVMs [Joachims, 1999]



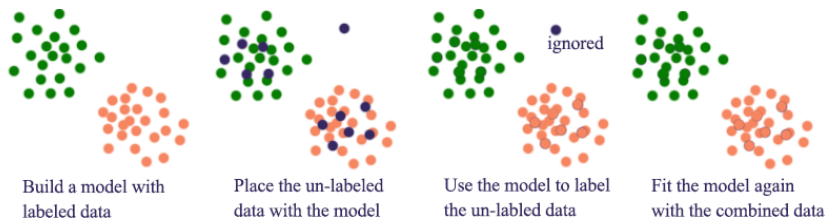
Fully supervised

SSL

- ▶ OR re-labelling each unlabelled data in training set
 - ▶ Same motivation as in mi-SVM
 - ▶ Iterative unlabelled data predictions, e.g. Curriculum learning [Bengio et al., 2009]

Curriculum learning for SSL

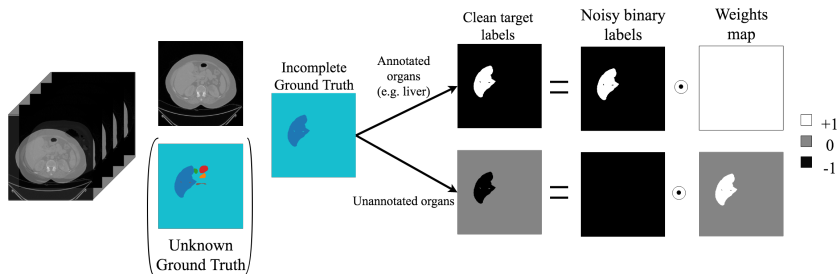
1. Train a model with labelled data \mathcal{A}
2. Until convergence:
 - Seek a sub-set of "easy" unlabelled data \mathcal{U}_e
 - Label each element in \mathcal{U}_e
 - Retrain model on $\mathcal{A} \cup \mathcal{U}_e$



Credit: J. Hui

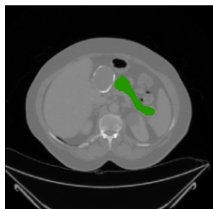
Case Study: SMILE [Petit et al., 2018]

- ▶ **Semantic segmentation of 3D abdominal CT-scans**
 - ▶ Clinical experts: focus on a subset of organs
 - ▶ Pixels with un-annotated organs \Rightarrow missing annotations
- ▶ **Semantic Segmentation with Incomplete Annotations (SMILE)**
 - ▶ **Training: only use pixels for which annotation is certain (no missing organ)**
 - ▶ K (+1 \Leftrightarrow background) classes $\Rightarrow K$ binary classifiers for each pixel
 - ▶ Organ(s) missing the whole volumes, organ present: complete annotation
 - ▶ **Missing organs in volume: only use pixels for other organs with -1 target label, ignore others**

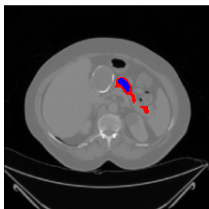


Case Study: SMILE [Petit et al., 2018]

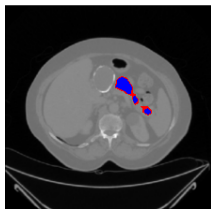
- ▶ SMILE training: labelled (certain annotations) & un-annotated pixels
- ▶ **SMILer** \Rightarrow SSL with Curriculum: take advantage of un-labelled pixels
 - ▶ Init with SMILE (easy) examples $\mathcal{A}, \mathcal{U}_e^0 \leftarrow \emptyset$
 - ▶ For $t \leftarrow 1$ to T , for each binary classifier:
 - ▶ Select \mathcal{H}^t new un-labelled positive examples
// \mathcal{H}^t : $\gamma_t = \frac{t}{T} \gamma_{max}$ top scoring pixels (blue) among predictions \hat{y}_i^+ (red)
 - ▶ $\mathcal{U}_e^t \leftarrow \mathcal{U}_e^{t-1} \cup \mathcal{H}^t$
 - ▶ Re-train model with augmented training set $\mathcal{A} \cup \mathcal{U}_e^t$



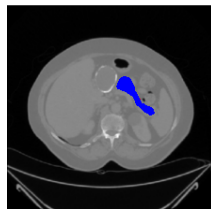
Unknown GT



$t = 1, \gamma_1 = 0.33$



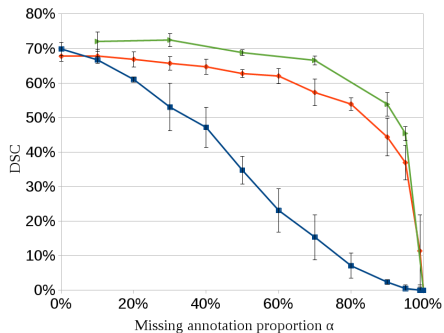
$t = 2, \gamma_2 = 0.66$



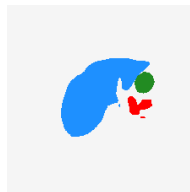
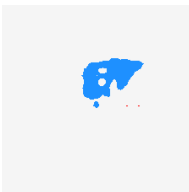
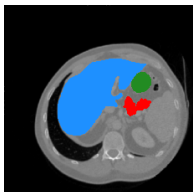
$t = 3, \gamma_3 = 1$

SMILE Results

- ▶ Experiments on 72 3D CT-scans for 3 organs: liver, pancreas and stomach
- ▶ Partial annotations generated: randomly removing $\alpha\%$ of organs

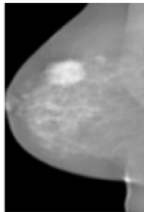
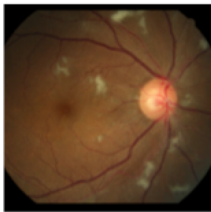
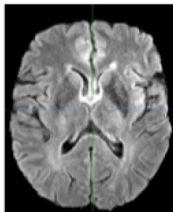


- ▶ Baseline: blue
- ▶ SMILE: orange ; SMILER: green
- ▶ **SMILER $\alpha = 70\% \sim$ baseline $\alpha = 0\%$**



Semi Supervised Learning (SSL) with Unsupervised Objective

- ▶ SSL: labelled and unlabelled data
- ▶ Simple option: combine supervised cost, e.g. classification, with unsupervised objective
- ▶ Unsupervised objective: extract (deep) representations without labels



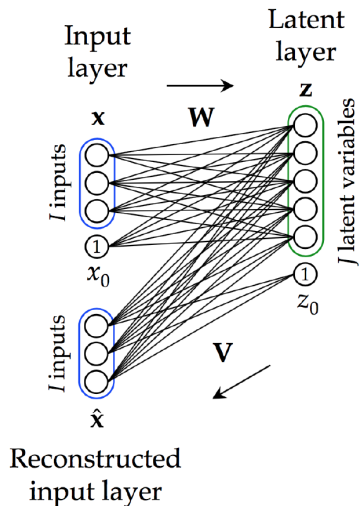
Auto-Encoders

- ▶ $\mathbf{z} = f(\mathbf{W}\mathbf{x})$
- ▶ $\hat{\mathbf{x}} = g(\mathbf{V}\mathbf{z})$
 - ▶ Often, $\mathbf{V} = \mathbf{W}^t$
- ▶ **Auto-encoder objective function: reconstruction**

$$C = \sum_{i=1}^N \|\mathbf{x}_i - \hat{\mathbf{x}}\|^2$$

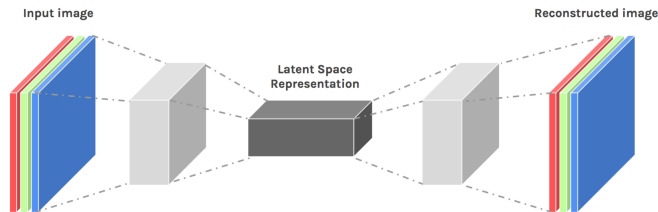
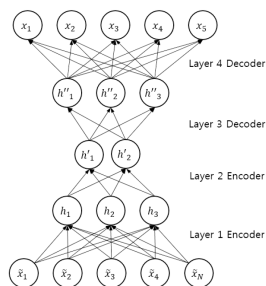
- ▶ If $f = g = Id$ (linear auto-encoder): ~ PCA:

$$C = \sum_{i=1}^N \|\mathbf{x}_i - \mathbf{W}^t \mathbf{W} \mathbf{x}_i\|^2$$



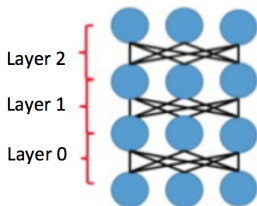
Deep Auto-Encoders

- ▶ AE: limited to linear feature extraction
- ▶ Add fully connected layers \Rightarrow more complex representations
- ▶ Add convolutional / deconvolutional layers: adapted to local feature extraction (images)

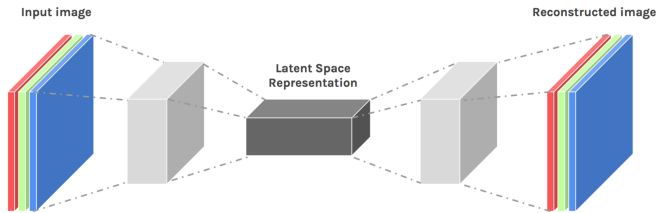


Training deep Auto-Encoders

- ▶ How to train deep unsupervised objective?
 - ▶ Fully connected deep AEs: layer-by layer tuning [Hinton et al., 2006]



- ▶ Deep conv AE: training whole architecture, *i.e.* all layers, jointly

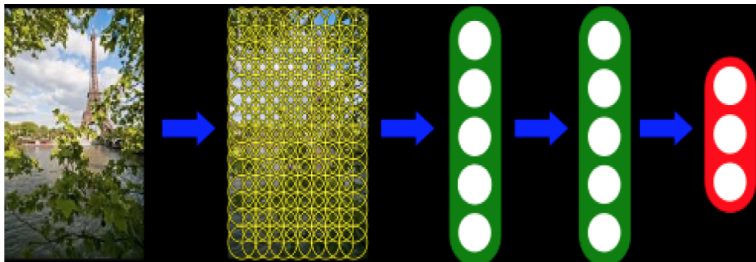


Training deep Auto-Encoders

- ▶ How to combine supervised and unsupervised objectives in SSL?
 - ▶ Used unsupervised as pre-training, supervised as fine-tuning
 - ▶ Used an hybrid objective function:

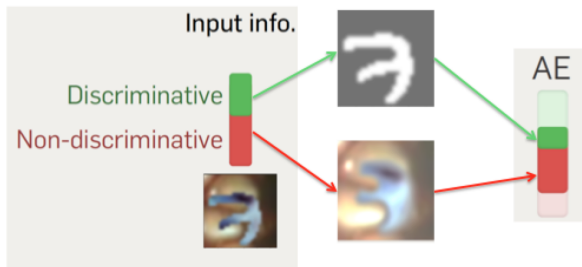
$$\mathcal{L} = \lambda_c \mathcal{L}_c + \lambda_r \mathcal{L}_r$$

- ▶ \mathcal{L}_c supervised cost, e.g. classification
- ▶ \mathcal{L}_r unsupervised cost, e.g. reconstruction
- ▶ Joint training of both tasks



Unsupervised Learning: Beyond Reconstruction

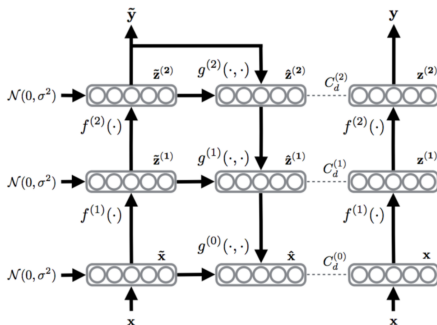
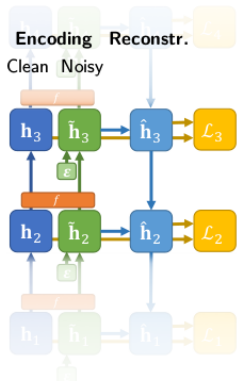
- ▶ **Unsupervised objective: why reconstruction?**
- ▶ **Reconstruction: what if ultimate goal requires generalization to a set of examples, e.g. classification?**
 - ▶ Deeper representation \Leftrightarrow more abstract \Leftrightarrow generalization \Leftrightarrow loss of information
 - ▶ **Classification & reconstruction: contradictory roles**
 - ▶ $\mathcal{L} = \lambda_c \mathcal{L}_c + \lambda_r \mathcal{L}_r$ with standard deep AE sub-optimal to disentangle discriminative from non-discriminative information



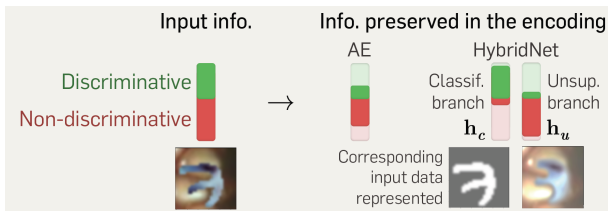
- ▶ Two current alternatives to unsupervised learning:
 1. Objective without reconstruction
 2. Casting unsupervised training as classification

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

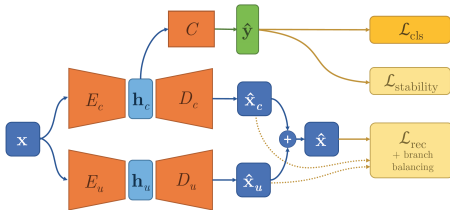
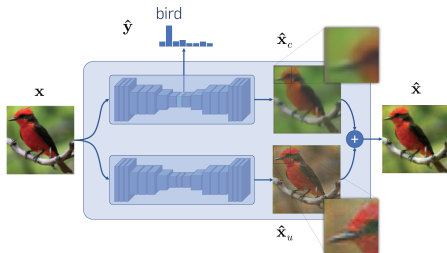
- ▶ "An autoencoder which can discard information"
- ▶ Layer above does not reconstruct layer below only with its activation
- ▶ Solution: Provide the details to learn only the abstract features
 - ▶ Decoder has a noisy version of the input to reconstruct



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

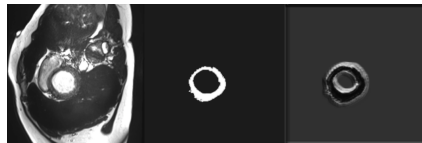


- ▶ **HybridNet: disentangling discriminative and complementary information for reconstruction**
- ▶ **Two-branch architecture**

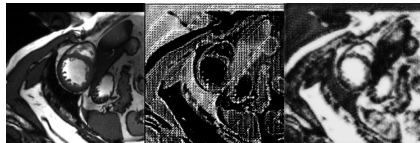


Hybrid Architectures for Medical Images

- ▶ SDNet (Spatial Decomposition) [Chartsias et al., 2018]
- ▶ SSL: Combining segmentation (cardiac MR) and reconstruction loss
 - ▶ **Motivation:** Combining losses with a single model challenging

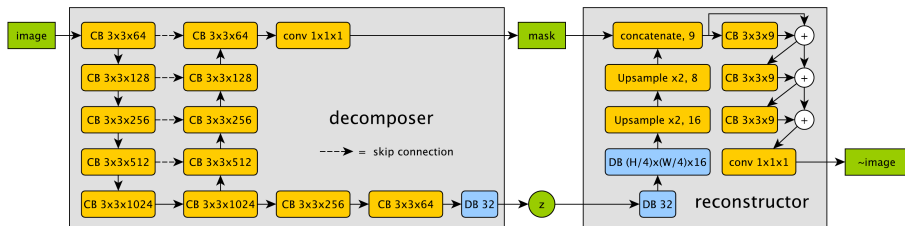


Large segmentation loss: poor reconstruction



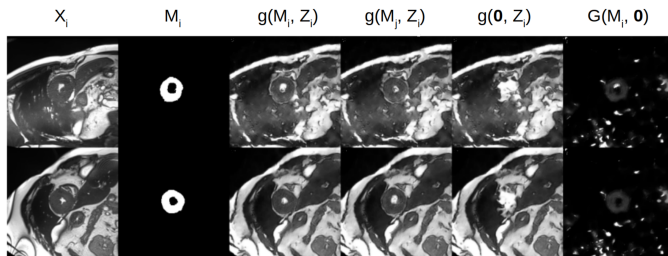
Large reconstruction loss: poor segmentation

- ▶ SDNet: 2-branch, segmentation (spatial) & global appearance layout



SDNet [Chartsias et al., 2018]

- ▶ 2-brach architecture \Rightarrow help disentangling
 - ▶ Nice latent space arithmetic properties

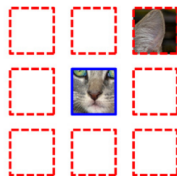
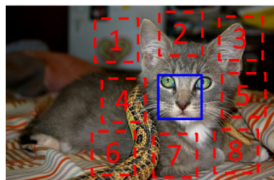


- ▶ Improvement for SSL compared e.g. U-Net [Ronneberger et al., 2015]

Labelled images	ACDC					QMRI			
	284	142	68	34	11	157	78	39	19
U-Net	0.782	0.657	0.581	0.356	0.026	0.686	0.681	0.441	0.368
GAN	0.787	0.727	0.648	0.365	0.080	0.795	0.756	0.580	0.061
SDNet	0.771	0.767	0.731	0.678	0.415	0.794	0.772	0.686	0.424

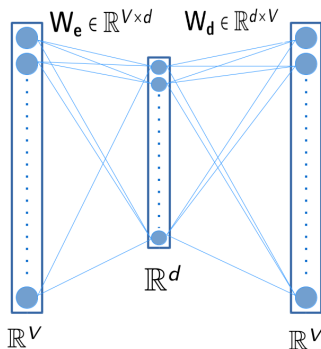
Beyond Reconstruction: Self-Supervised Training

- ▶ **Self-supervised training: unsupervised problem \Rightarrow supervised one**
- ▶ Performing prediction on data, e.g.
 - ▶ Relative position of regions
 - ▶ Temporal prediction (next frames)
- ▶ **"Auxiliary", "pretext" task**
 - ▶ Good auxiliary task requires solving high-level recognition \Rightarrow useful features for the ultimate task
 - ▶ Automatic labeling for auxiliary task \Rightarrow no manual supervision



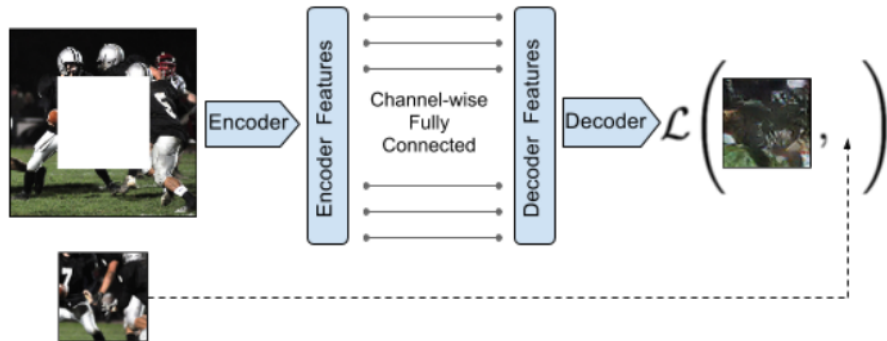
Word2Vec [Mikolov et al., 2013]

- ▶ Embedding of words: project a word in \mathbb{R}^d space
- ▶ **Word2Vec auxiliary**: predict a word given its context
 - ▶ Assumption: similar words appears in similar contexts, *i.e.* distributional hypothesis in NLP
 - ▶ Input: Bag of Words of context $\mathbf{x} \in \mathbb{R}^V$, V vocabulary size
 - ▶ $\mathbf{h} = \mathbf{W}_e \mathbf{x}$, $\hat{\mathbf{x}} = \mathbf{W}_d \mathbf{h} + \text{soft max}$: classify central word



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

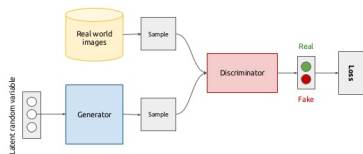
- ▶ Auxiliary task: Inpainting



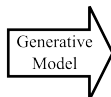
Generative Adversarial Networks (GAN) [Goodfellow et al., 2014]

- ▶ Unsupervised problem \Rightarrow 2-player game theory problem
- ▶ Interesting results: optimal generator learns data distribution

Generative adversarial networks (conceptual)



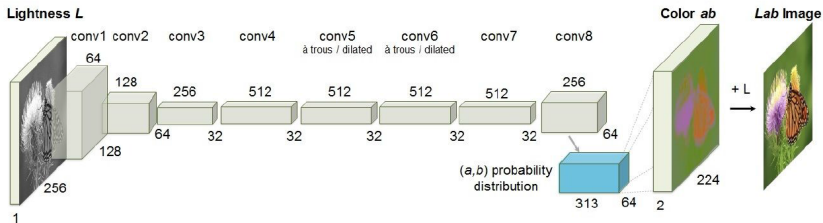
Noise $\sim N(0,1)$



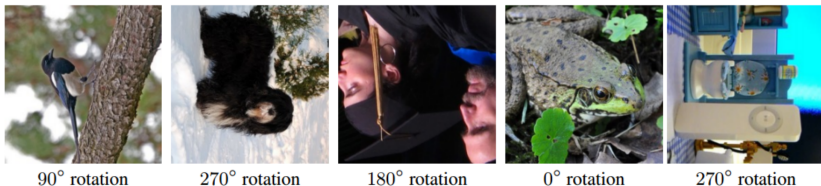
- ▶ Adversarial cost used beyond generation for distribution matching
- ▶ Next course!

Self-Supervised Training: other auxiliary tasks

- ▶ Image colorization [Zhang et al., 2016]



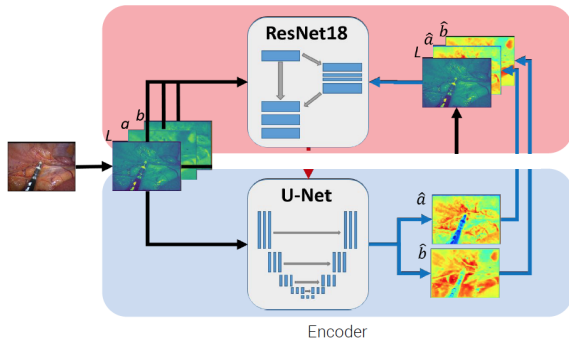
- ▶ Predicting image orientation [Gidaris et al., 2018]



Self-Supervised Training in Medical Imaging

- ▶ **Auxiliary task:** endoscopic video colorization [Ross et al., 2018] in (L,a,b) space
 - ▶ cGAN approach: predict color (a,b) from luminance L
 - ▶ Generator (U-Net): $L \rightarrow (\hat{a}, \hat{b})$
 - ▶ Discriminator (ResNet): $L, a, b \rightarrow \text{real}, L(\hat{a}, \hat{b}) \rightarrow \text{fake}$

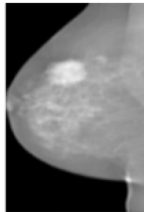
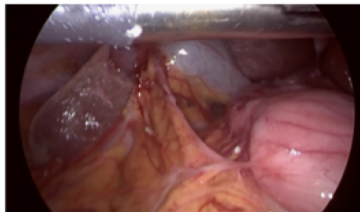
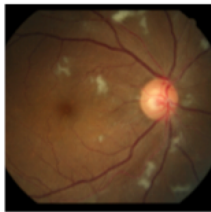
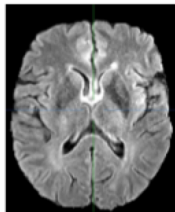
Adversarial Discriminator



- ▶ **Target task:** instrument segmentation

Conclusion

- ▶ Deep models: huge volume of annotated data
 - ▶ Annotation cost exacerbated in healthcare
- ▶ Learning from weak supervision (WSL)
 - ▶ Very relevant for localized tasks (e.g. segmentation) in medical images: high-resolution, 3D, videos, etc
 - ▶ Pooling function (local prediction → global label) crucial
 - ▶ Constraining models which medical *prior* knowledge useful
- ▶ Learning from (few) labeled data and (many) unlabeled supervision (SSL)
 - ▶ Re-labeling unlabeled data, e.g. Curriculum-based approaches
 - ▶ Beyond reconstruction with:
 - ▶ Architectures for disentangling supervised from unsupervised signals
 - ▶ Self-supervision



References I

- [Andrews et al., 2003] Andrews, S., Tsochantaridis, I., and Hofmann, T. (2003).
Support vector machines for multiple-instance learning.
In Advances in Neural Information Processing Systems (NIPS).
- [Azizpour et al., 2016] Azizpour, H., Razavian, A. S., Sullivan, J., Maki, A., and Carlsson, S. (2016).
Factors of transferability for a generic convnet representation.
IEEE Trans. Pattern Anal. Mach. Intell., 38(9):1790–1802.
- [Bearman et al., 2016] Bearman, Russakovsky, Ferrari, and Fei-Fei (2016).
What's the Point: Semantic Segmentation with Point Supervision.
ECCV.
- [Bengio et al., 2009] Bengio, Y., Louradour, J., Collobert, R., and Weston, J. (2009).
Curriculum learning.
In Proceedings of the 26th Annual International Conference on Machine Learning, ICML '09, pages 41–48.
- [Chartsias et al., 2018] Chartsias, A., Joyce, T., Papanastasiou, G., Semple, S., Williams, M., Newby, D. E., Dharmakumar, R., and Tsaftaris, S. A. (2018).
Factorised spatial representation learning: Application in semi-supervised myocardial segmentation.
In MICCAI (2), volume 11071 of *Lecture Notes in Computer Science*, pages 490–498. Springer.
- [Dietterich et al., 1997] Dietterich, T. G., Lathrop, R. H., and Lozano-Pérez, T. (1997).
Solving the multiple instance problem with axis-parallel rectangles.
Artif. Intell.
- [Durand et al., 2015] Durand, T., Thome, N., and Cord, M. (2015).
MANTRA: Minimum Maximum Latent Structural SVM for Image Classification and Ranking.
In IEEE International Conference on Computer Vision (ICCV).
- [Durand et al., 2019] Durand, T., Thome, N., and Cord, M. (2019).
Exploiting negative evidence for deep latent structured models.
IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(2):337–351.

References II

- [Felzenszwalb et al., 2010] Felzenszwalb, P. F., Girshick, R. B., McAllester, D., and Ramanan, D. (2010). Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*.
- [Gidaris et al., 2018] Gidaris, S., Singh, P., and Komodakis, N. (2018). Unsupervised representation learning by predicting image rotations. In *ICLR*, volume abs/1803.07728.
- [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.
- [Hinton et al., 2006] Hinton, G. E., Osindero, S., and Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural Comput.*, 18(7):1527–1554.
- [Jia et al., 2017] Jia, Z., Huang, X., Chang, E. I., and Xu, Y. (2017). Constrained deep weak supervision for histopathology image segmentation. *IEEE TRANSACTIONS ON MEDICAL IMAGING*, 36(11).
- [Joachims, 1999] Joachims, T. (1999). Transductive inference for text classification using support vector machines. In *Proceedings of the Sixteenth International Conference on Machine Learning, ICML '99*, pages 200–209, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.
- [Krizhevsky et al., 2012] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.
- [Li et al., 2017] Li, X., Chen, H., Qi, X., Dou, Q., Fu, C., and Heng, P. (2017). H-denseunet: Hybrid densely connected unet for liver and liver tumor segmentation from CT volumes. *CoRR*, abs/1709.07330.

References III

- [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.
- [Nwoye et al., 2019] Nwoye, C., Mutter, D., Marescaux, J., and Padoy, N. (2019). Weakly supervised convolutional lstm approach for tool tracking in laparoscopic videos. In *International Conference on Information Processing in Computer-Assisted Interventions (IPCAI)*.
- [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 *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.
- [Petit et al., 2018] Petit, O., Thome, N., Charnoz, A., Hostettler, A., and Soler, L. (2018). Handling missing annotations for semantic segmentation with deep convnets. In *Deep Learning in Medical Image Analysis - and - Multimodal Learning for Clinical Decision Support - 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings*, pages 20–28.
- [Quattoni et al., 2007] Quattoni, A., Wang, S. B., Morency, L.-P., Collins, M., and Darrell, T. (2007). Hidden conditional random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*.
- [Quellec et al., 2012] Quellec, G., Lamard, M., Abràmoff, M. D., Decencière, E., Lay, B., Erginay, A., Cochener, B., and Cazuguel, G. (2012). A multiple-instance learning framework for diabetic retinopathy screening. *Medical image analysis*, 16 6:1228–40.
- [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.

References IV

- [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)*.
- [Ronneberger et al., 2015] Ronneberger, O., P.Fischer, and Brox, T. (2015).
U-net: Convolutional networks for biomedical image segmentation.
In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 9351 of *LNCS*, pages 234–241. Springer.
(available on arXiv:1505.04597 [cs.CV]).
- [Ross et al., 2018] Ross, T., Zimmerer, D., Vemuri, A. S., Isensee, F., Wiesenfarth, M., Bodenstedt, S., Both, F., Kessler, P., Wagner, M., Müller, B., Kenngott, H., Speidel, S., Kopp-Schneider, A., Maier-Hein, K. H., and Maier-Hein, L. (2018).
Exploiting the potential of unlabeled endoscopic video data with self-supervised learning.
Int. J. Computer Assisted Radiology and Surgery, 13(6):925–933.
- [Tajbakhsh et al., 2016] Tajbakhsh, N., Shin, J. Y., Gurudu, S. R., Hurst, R. T., Kendall, C. B., Gotway, M. B., and Liang, J. (2016).
Convolutional neural networks for medical image analysis: Fine tuning or full training?
IEEE Transactions on Medical Imaging, PP(99):1–1.
- [van der Maaten and Hinton, 2008] van der Maaten, L. and Hinton, G. E. (2008).
Visualizing high-dimensional data using t-sne.
Journal of Machine Learning Research, 9:2579–2605.
- [Xu et al., 2014] Xu, Y., Zhu, J.-Y., Chang, E. I., Lai, M., and Tu, Z. (2014).
Weakly supervised histopathology cancer image segmentation and classification.
Medical Image Analysis, 18(3):591–604.
- [Yu and Joachims, 2009] Yu, C.-N. and Joachims, T. (2009).
Learning structural svms with latent variables.
In *ICML*.

References V

- [Zhang et al., 2016] Zhang, R., Isola, P., and Efros, A. A. (2016).
Colorful image colorization.
In *ECCV (3)*, volume 9907 of *Lecture Notes in Computer Science*, pages 649–666. Springer.
- [Zhou et al., 2016] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. (2016).
Learning Deep Features for Discriminative Localization.
In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [Zhu et al., 2017] Zhu, W., Lou, Q., Vang, Y. S., and Xie, X. (2017).
Deep multi-instance networks with sparse label assignment for whole mammogram classification.
In Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D. L., and Duchesne, S., editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2017*, pages 603–611, Cham. Springer International Publishing.