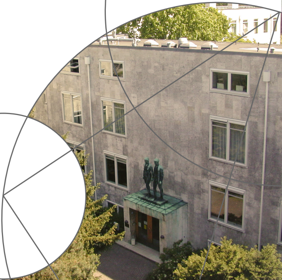
Machine Learning for Science and Society

New biomarkers for disease diagnosis and prognosis

Christian Igel
Department of Computer Science



igel@diku.dk

General mission

Machine learning gets more and more powerful –
let's apply it to solve problems of societal relevance!

Today's talk

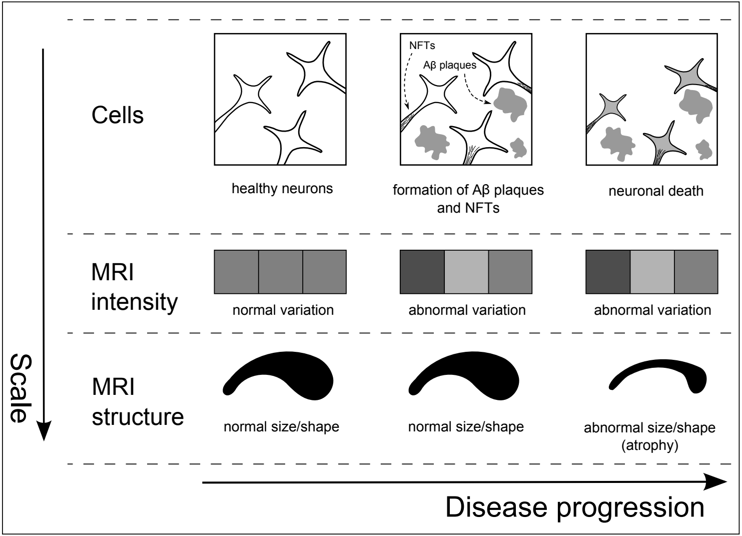
How can machine learning be used to design biomarkers (measurable indicators) for diagnosis and prognosis of diseases?

Many people think that deep learning is the future of AI, so let's discuss it in the context of medical biomarkers.

Outline

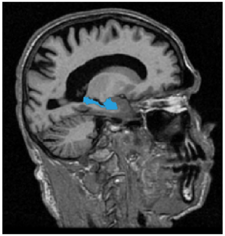
Outline

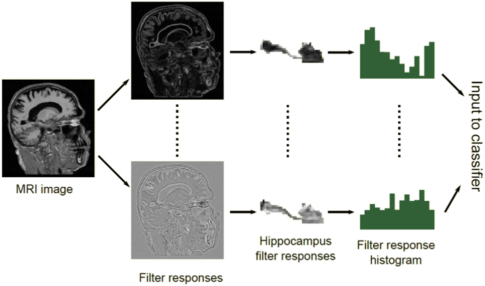
Alzheimer's Disease progression



Hippocampal texture from structural MRI

- Image analysis of brain scans
    - Segmentation of brain regions (hippocampi)
    - Extraction of image features
- Classification using SVMs





Sørensen, Igel, Hansen, Osler, Lauritzen, Rostrup, Nielsen. Early detection of Alzheimer's disease using MRI hippocampal texture. *Human Brain Mapping*, 2016

CADDementia Challenge

- *Computer-Aided Diagnosis of Dementia based on structural MRI data* (CADDementia) competition
- Goal: Diagnosis into Alzheimer's disease (AD), mild cognitive impairment (MCI), and normal controls (NC)
- No labelled training data was provided, we trained on freely available data sets.
- We combined several biomarkers (cortical thickness measurements, hippocampal shape, hippocampal texture, and volumetric measurements).

Sørensen, Pai, Anker, Balas, Lillholm, Igel, Nielsen. Dementia Diagnosis using MRI Cortical Thickness, Shape, Texture, and Volumetry. *MICCAI 2014 – Challenge on Computer-Aided Diagnosis of Dementia Based on Structural MRI Data*, 2014

Bron et al. Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: The CADDementia challenge. *NeuroImage*, 2015

Sørensen, Igel, Pai, Anker, Balas, Lillholm, Nielsen. Differential diagnosis of mild cognitive impairment and Alzheimer's disease using structural MRI cortical thickness, hippocampal shape, hippocampal texture, and volumetry. *NeuroImage: Clinical*, in revisions

CADDementia: Results

Original results table presented on 18 September 2014 at the MICCAI workshop.

<table>
<tr><th>Algorithm</th><th>Rank accuracy</th><th>Accuracy [CI] (%)</th><th>TPF_{CN} [CI] (%)</th><th>TPF_{MCI} [CI] (%)</th><th>TPF_{AD} [CI] (%)</th><th>Rank AUC</th><th>AUC_{all} [CI] (%)</th><th>AUC_{CN} [CI] (%)</th><th>AUC_{MCI} [CI] (%)</th><th>AUC_{AD} [CI] (%)</th><th>Date</th><th>Documentation</th></tr>
<tr><td>Soerensen-equal</td><td>1.0</td><td>63.0 [57.9 - 67.5]</td><td>96.9 [92.9 - 99.2]</td><td>28.7 [21.3 - 37.4]</td><td>61.2 [51.6 - 69.8]</td><td>1.5</td><td>78.8 [75.6 - 82.0]</td><td>86.3 [81.8 - 89.3]</td><td>63.1 [56.6 - 68.3]</td><td>87.5 [83.4 - 91.1]</td><td>15/06/2014</td><td>Paper | Wiki | Pres.</td></tr>
<tr><td>Soerensen-optimized</td><td>2.0</td><td>59.9 [54.8 - 64.7]</td><td>70.5 [62.8 - 77.8]</td><td>41.0 [33.3 - 50.0]</td><td>68.9 [59.6 - 77.2]</td><td>1.5</td><td>78.8 [75.5 - 82.1]</td><td>86.3 [81.9 - 89.3]</td><td>62.7 [56.8 - 68.4]</td><td>86.7 [82.3 - 90.4]</td><td>15/06/2014</td><td>Paper | Wiki | Pres.</td></tr>
<tr><td>Wachinger-enetNorm</td><td>3.0</td><td>59.0 [54.0 - 63.6]</td><td>72.1 [63.4 - 79.2]</td><td>51.6 [43.5 - 61.3]</td><td>51.5 [41.5 - 61.2]</td><td>4.0</td><td>77.0 [73.6 - 80.3]</td><td>83.3 [78.5 - 87.0]</td><td>59.4 [52.9 - 65.5]</td><td>88.2 [83.8 - 91.4]</td><td>15/06/2014</td><td>Paper | Wiki | Pres.</td></tr>
<tr><td>Ledig-ALL</td><td>4.0</td><td>57.9 [52.5 - 62.7]</td><td>89.1 [83.7 - 93.8]</td><td>41.0 [32.4 - 49.6]</td><td>38.8 [30.7 - 50.0]</td><td>5.0</td><td>76.7 [73.6 - 79.8]</td><td>86.6 [82.7 - 89.8]</td><td>59.7 [53.3 - 65.1]</td><td>84.9 [79.7 - 88.7]</td><td>15/06/2014</td><td>Paper | Wiki | Pres.</td></tr>
<tr><td>Moradi</td><td>5.0</td><td>57.6 [52.3 - 62.4]</td><td>57.4 [48.7 - 66.1]</td><td>59.8 [51.3 - 68.1]</td><td>55.3 [46.7 - 65.2]</td><td>-</td><td>-</td><td>-</td><td>-</td><td>-</td><td>15/06/2014</td><td>Paper | Wiki | Pres.</td></tr>
<tr><td>Franke</td><td>6.0</td><td>56.2</td><td>58.9</td><td>43.4</td><td>68.0</td><td></td><td></td><td></td><td></td><td></td><td>15/06/2014</td><td>Paper | Wiki | Pres.</td></tr>
<tr><td>Wachinger-step1Norm</td><td>12.5</td><td>53.7 [48.6 - 58.8]</td><td>63.6 [54.9 - 71.9]</td><td>47.5 [38.4 - 56.6]</td><td>48.5 [39.6 - 59.1]</td><td>9.5</td><td>74.3 [70.5 - 77.9]</td><td>79.3 [74.1 - 83.5]</td><td>55.5 [48.5 - 61.6]</td><td>87.7 [83.7 - 91.1]</td><td>15/06/2014</td><td>Paper | Wiki | Pres.</td></tr>
<tr><td>Ledig-MBL</td><td>15.0</td><td>53.4 [47.7 - 57.9]</td><td>82.9 [76.0 - 88.7]</td><td>43.4 [35.1 - 52.9]</td><td>28.2 [20.2 - 37.4]</td><td>7.0</td><td>75.2 [72.0 - 78.1]</td><td>82.5 [77.8 - 86.0]</td><td>57.3 [50.9 - 63.6]</td><td>86.4 [81.4 - 89.9]</td><td>15/06/2014</td><td>Paper | Wiki | Pres.</td></tr>
<tr><td>Wachinger-man</td><td>16.0</td><td>53.1 [47.7 - 57.9]</td><td>61.2 [53.5 - 69.6]</td><td>60.7 [51.7 - 70.0]</td><td>34.0 [25.7 - 44.7]</td><td>9.5</td><td>74.3 [70.9 - 77.9]</td><td>80.6 [75.7 - 84.9]</td><td>56.3 [49.7 - 63.0]</td><td>86.1 [81.7 - 90.0]</td><td>15/06/2014</td><td>Paper | Wiki | Pres.</td></tr>
<tr><td>Eskildsen-ADNI1</td><td>17.5</td><td>52.0 [46.6 - 56.8]</td><td>65.1 [56.9 - 73.2]</td><td>32.0 [24.1 - 40.9]</td><td>59.2 [49.5 - 68.3]</td><td>-</td><td>-</td><td>-</td><td>-</td><td>-</td><td>15/06/2014</td><td>Paper | Wiki | Pres.</td></tr>
<tr><td>Eskildsen-FACEADNI1</td><td>17.5</td><td>52.0 [46.9 - 57.1]</td><td>65.1 [56.6 - 73.1]</td><td>36.1 [28.1 - 45.5]</td><td>54.4 [44.6 - 63.6]</td><td>-</td><td>-</td><td>-</td><td>-</td><td>-</td><td>15/06/2014</td><td>Paper | Wiki | Pres.</td></tr>
<tr><td>Eskildsen-Combined</td><td>19.0</td><td>51.1 [45.5 - 56.2]</td><td>64.3 [56.2 - 72.3]</td><td>35.2 [27.1 - 44.3]</td><td>53.4 [43.0 - 62.9]</td><td>-</td><td>-</td><td>-</td><td>-</td><td>-</td><td>15/06/2014</td><td>Paper | Wiki | Pres.</td></tr>
<tr><td>Dolph</td><td>20.0</td><td>49.7 [44.6 - 54.8]</td><td>84.5 [77.9 - 90.4]</td><td>23.0 [16.4 - 31.2]</td><td>37.9 [28.9 - 47.3]</td><td>17.0</td><td>63.0 [59.6 - 67.2]</td><td>66.2 [61.3 - 70.3]</td><td>55.4 [50.0 - 60.0]</td><td>65.8 [60.6 - 71.3]</td><td>15/06/2014</td><td>Paper | Wiki | Pres.</td></tr>
<tr><td>[illegible]</td><td>[illegible]</td><td>49.2 [illegible]</td><td>94.6 [illegible]</td><td>11.5 [illegible]</td><td>36.9 [illegible]</td><td>-</td><td>-</td><td>-</td><td>-</td><td>-</td><td>15/06/2014</td><td>Paper | Wiki | Pres.</td></tr>
</table>

- Our system outperforms the state-of-the-art.
- Our biomarker remains significant if combined with other indicators.
- The biomarker generalizes across patient populations.

Outline

1. Invariant Features
   - Diagnosis and Prognosis of Alzheimer's Disease
   - Working in Scale-space
2. Convolutional Neural Networks
   - Deep Learning
   - Analysis of Mammograms
   - Knee Cartilage Segmentation
   - Brain Segmentation
3. Some Thoughts

Recall: Convolution

1D convolution with filter kernel $w : \mathbb{R} \to \mathbb{R}$:

$$s(t) = (x * w)(t) = \int x(a)w(t-a)\mathrm{d}a$$

Convolution of a discrete 2D image I with filter kernel K:

$$\begin{aligned} S(i,j) = (I * K)(i,j) &= \sum_m \sum_n I(m,n)K(i-m,j-n) \\ &= (K * I)(i,j) = \sum_m \sum_n I(i-m,j-n)K(m,n) \end{aligned}$$

Example: Convolution

Example of applying a horizontal derivative filter:

Goodfellow, Bengio, Courville. *Deep Learning*. MIT Press

Derivatives in scale-space

- Let $\mathbf{x} = [x, y, z]^T$ be a point in image I.
- The Gaussian scale-space representation $I_\sigma = I * G_\sigma$ at scale σ is given by convolving with:

$$G_\sigma(\mathbf{x}) = \frac{1}{((2\pi)^{1/2}\sigma)^3} \exp\left(-\frac{\|\mathbf{x}\|_2^2}{2\sigma^2}\right)$$

- Partial derivatives are now easily obtained, e.g., w.r.t. y using

$$I_{y,\sigma}(\mathbf{x}) = [I * G_{y,\sigma}](\mathbf{x})$$

with partial first-order derivative of the Gaussian, e.g.:

$$G_{y,\sigma}(\mathbf{x}) = \frac{\partial}{\partial y} G_\sigma(\mathbf{x})$$

The Hessian

- We denote the eigenvalues of the Hessian

$$H_\sigma(\mathbf{x}) = \begin{bmatrix} I_{xx,\sigma} & I_{xy,\sigma} & I_{xz,\sigma} \\ I_{xy,\sigma} & I_{yy,\sigma} & I_{yz,\sigma} \\ I_{xz,\sigma} & I_{yz,\sigma} & I_{zz,\sigma} \end{bmatrix} (\mathbf{x})$$

  by

$$\lambda_{i,\sigma}(\mathbf{x}) \quad , \quad |\lambda_{1,\sigma}(\mathbf{x})| \geq |\lambda_{2,\sigma}(\mathbf{x})| \geq |\lambda_{3,\sigma}(\mathbf{x})|$$

  for a voxel $\mathbf{x} = [x, y, z]^T$.
- Considering these eigenvalues induces *rotation invariance*.

Scale-space features

- Gradient magnitude

$$\|\nabla G_\sigma(\mathbf{x})\|_2 = \sqrt{I_{x,\sigma}(\mathbf{x})^2 + I_{y,\sigma}(\mathbf{x})^2 + I_{z,\sigma}(\mathbf{x})^2}$$

- Laplacian of the Gaussian

$$\nabla^2 G_\sigma(\mathbf{x}) = \lambda_{1,\sigma}(\mathbf{x}) + \lambda_{2,\sigma}(\mathbf{x}) + \lambda_{3,\sigma}(\mathbf{x})$$

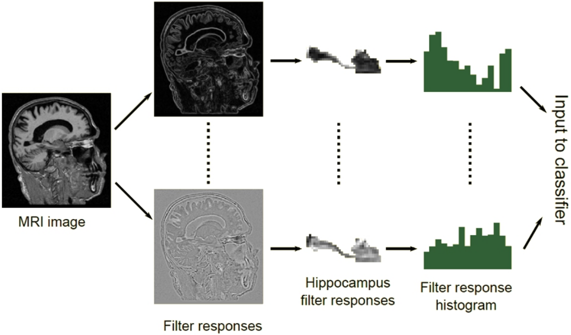
- Gaussian curvature

$$K_\sigma(\mathbf{x}) = \lambda_{1,\sigma}(\mathbf{x}) \cdot \lambda_{2,\sigma}(\mathbf{x}) \cdot \lambda_{3,\sigma}(\mathbf{x})$$

- Frobenius norm of the Hessian

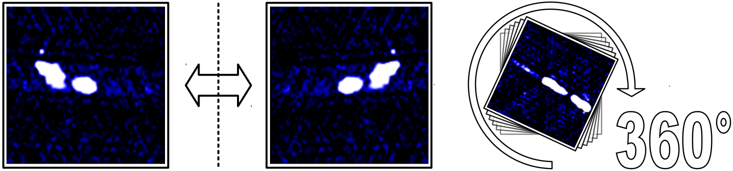
$$\|H_\sigma(\mathbf{x})\|_F = \sqrt{\lambda_{1,\sigma}(\mathbf{x})^2 + \lambda_{2,\sigma}(\mathbf{x})^2 + \lambda_{3,\sigma}(\mathbf{x})^2}$$

Histogram representation



- Features values are computed at each $\mathbf{x}$, the values are binned and the frequencies are the final representation.
- Collecting values over a region in a histogram makes the features *invariant under translation* and allows for easy handling of *variable-size input*.

Invariance



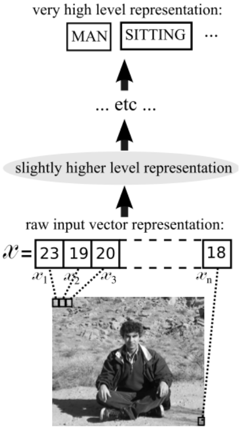
- Histograms of selected scale-space features do not change if the image is translated, flipped, or rotated.
- Feature invariance carry over to the machine learning model on top of the features.
- This fosters generalization and makes learning statistically efficient.

Outline

Deep learning

- Deep learning refers to ML architectures composed of multiple levels of non-linear transformations.
- Idea: Extracting more and more abstract features from input data, learning more abstract representations.
- Representations can be learned in a supervised and/or unsupervised manner.
- Example: Convolutional neural networks (CNNs) are popular deep learning architectures.

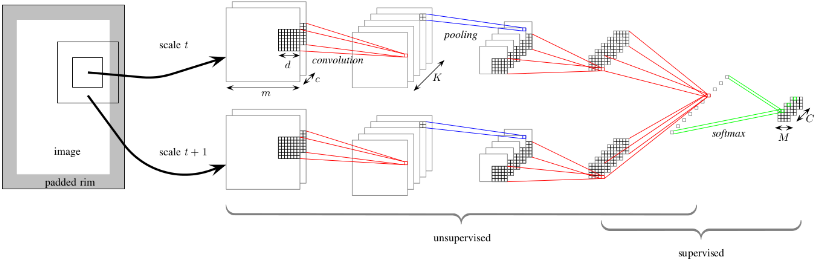
LeCun, Bottou, Bengio, Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998



Bengio. Learning Deep Architectures for AI. *Foundations and Trends in Machine Learning* 2(1): 1–127, 2009

Convolutional neural network (CNN)

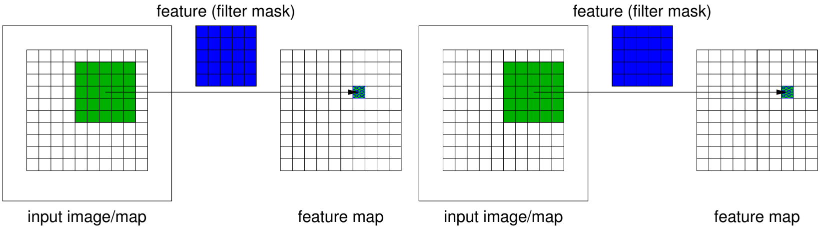
Example of a special CNN architecture we use:



A canonical CNN consists of

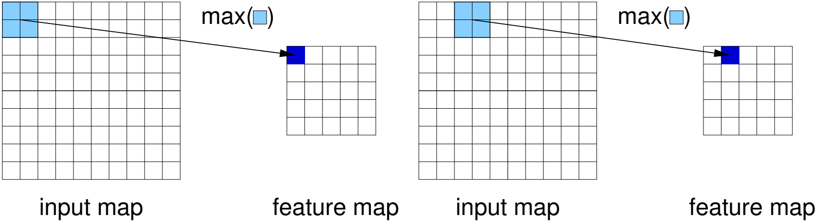
- convolutional layers, each of which producing several *feature maps*; interleaved with
- pooling layers; and
- a standard neural network on top.

Convolutional layer



- A convolutional layer computes a feature map by convolution of the input with a linear filter.
- The coefficients of each convolution kernel are learned as weights in a neural network.
- A non-linear function is applied to the feature map elements, which can be viewed as neurons with shared weights.

Pooling layer



- Pooling layers compute the maximum or average over a region of a feature map.
- Used to reduce the dimensionality, increase the scale, and to support translation invariance.

Outline

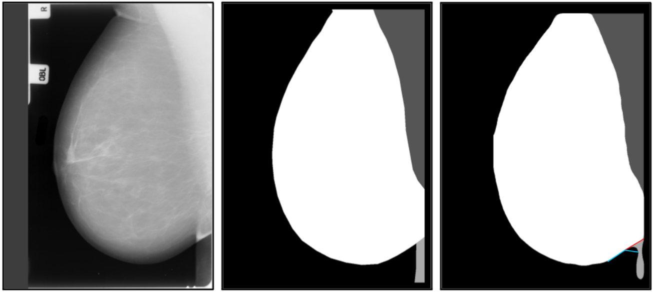
Breast cancer

- Breast cancer is a frequent cause of death among women
- Screening programs halve the risk of death[1] and reduce mortality by 28-36 %[2]
- Still: 33 % of cancers are missed,[3] 70 % of referrals are false positives,[4] and 25 % of cancers could have been detected earlier[5]
- Goal: More accurate image-based biomarkers allowing personalized breast cancer screening

[1] Otto et al. *Cancer Epidemiol Biomarkers Prev* 21, 2012
[2] Broeders et al. *J Med Screen* 19, 2012
[3] Karssemeijer et al. *Radiology* 227, 2003
[4] Yankaskas et al. *Am J Roentgenol* 177, 2001
[5] Timmers et al. *Eur J Public Health*, 2012

Breast segmentation

Breast segmentation is the first step in the analysis.



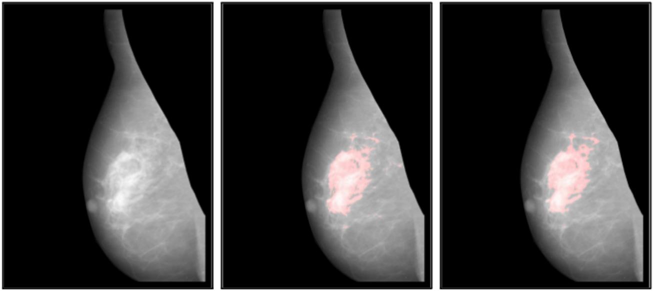
input manual CNN

Dataset	Background	Pectoral muscle	Breast tissue
Nijmegen ($N = 495$)	0.99 ± 0.01	0.95 ± 0.08	0.98 ± 0.01
Mayo ($N = 717$)	0.99 ± 0.01	0.93 ± 0.11	0.97 ± 0.02

Dice's coefficients

Breast density scoring

Breast density is related to breast cancer risk, the risk of missing breast cancer, and the risk of false positive referral.



input manual CNN

Current work: Better biomarkers using CNN density scores and CNN texture scores

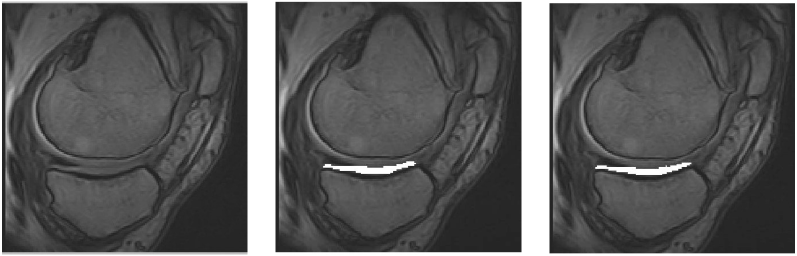
Kallenberg, Petersen, Nielsen, Ng, Diao, Igel, Vachon, Holland, Winkel, Karssemeijer, and Lillholm. Unsupervised deep learning applied to breast density segmentation and mammographic risk scoring. *IEEE Transactions on Medical Imaging*, 2016

Outline

1. Invariant Features
   - Diagnosis and Prognosis of Alzheimer's Disease
   - Working in Scale-space
2. Convolutional Neural Networks
   - Deep Learning
   - Analysis of Mammograms
   - Knee Cartilage Segmentation
   - Brain Segmentation
3. Some Thoughts

Knee cartilage segmentation

- Cartilage segmentation in knee MRI is the method of choice for quantifying cartilage deterioration.
- Cartilage deterioration implies osteoarthritis, one of the major reasons for work disability in the western world.

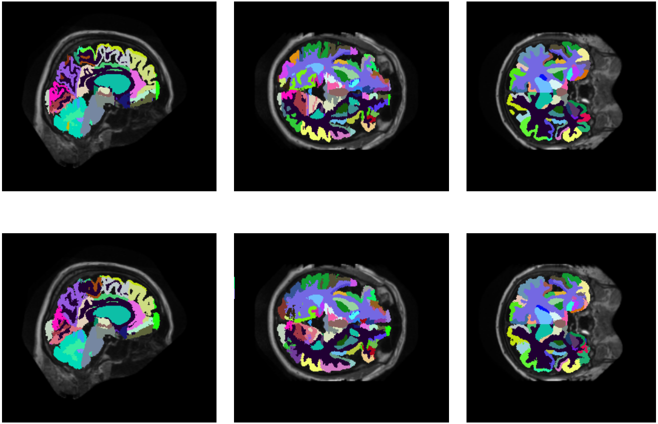


input manual CNN

Prasoon, Petersen, Igel, Lauze, Dam, Nielsen. Deep Feature Learning for Knee Cartilage Segmentation Using a Triplanar Convolutional Neural Network. *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, LNCS 8150, 2013

Outline

Brain segmentation

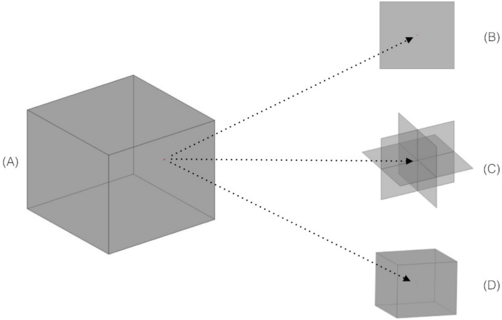


134 classes, ground truth (top) and CNN results (bottom)

Pai, Teng, Blair, Kallenberg, Dam, Sommer, Igel, Nielsen. Characterisation of errors in deep learning-based brain MRI segmentation. *Deep Learning for Medical Image Analysis*, Elsevier, in press

Tri-planar CNN

Processing 3D images (A) by 2D (B), tri-planar (C), 3D (D) convolutions:



Tri-planar CNNs are computationally and statistically efficient.

Prasoon, Petersen, Igel, Lauze, Dam, Nielsen. Deep Feature Learning for Knee Cartilage Segmentation Using a Triplanar Convolutional Neural Network. *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, LNCS 8150, 2013

Pai, Teng, Blair, Kallenberg, Dam, Sommer, Igel, Nielsen. Characterisation of errors in deep learning-based brain MRI segmentation. *Deep Learning for Medical Image Analysis*, Elsevier, in press

Some thoughts

- CNNs are well suited for image and sound processing; they have won several competitions recently.
  The best performance is usually achieved when the CNN is combined with careful image preprocessing.
- CNNs exploit "spatial" structure in the input; in particular, they incorporate translation invariance.
- Layers can be (pre-)trained sequentially.
- Unsupervised and supervised learning can nicely be combined.
- Training CNNs typically scales linearly in training set size.
- CNNs profit from massively parallel computing (e.g., GPUs)

Some more thoughts

- $\oplus/\ominus$ Deep networks are highly non-linear models.
- $\ominus$ Tuning the architecture of a deep learning system can be very difficult.
- $\ominus$ Deep learning is badly understood theoretically.

The revival of deep neural networks is partly due to the availability of fast (parallel) hardware and larger training data sets (which makes overfitting less of a problem).

Take me home . . .

Incorporating the right invariance properties in the learning system is a key principle for increasing efficiency and performance.

Machine learning can be used to design powerful biomarkers for diagnosis and prognosis of diseases.

Machine learning gets more and more powerful –
let's apply it to solve problems of societal relevance!

`http://image.diku.dk/igel`