SAM Joint Imaging-Therapy Scientific Symposium

AAPM Annual Meeting 2018

Case Studies in Deep Learning

Arvind Rao, Ph.D. (joint with S. Barua, Rice University)
Associate Professor,
Department of Computational Medicine and Bioinformatics,
Department of Radiation Oncology, Michigan Medicine
& MIDAS
The University of Michigan, Ann Arbor
Aug 1, 2018

Disclosures

- Member, Voxel Analytics LLC
- Consulting for Deoxylitics LLC

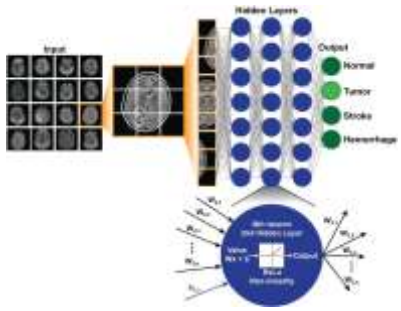
Rule of thumb

https://hbr.org/2016/11/what-artificial-intelligence-can-and-cant-do-right-now

"If a typical person can do a mental task with less than one second of thought, we can probably automate it using AI either now or in the near future.

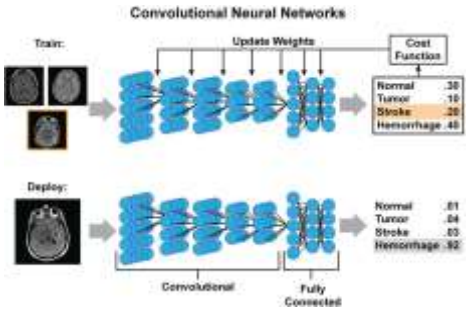
A lot of valuable work currently done by humans — examining security video to detect suspicious behaviors, deciding if a car is about to hit a pedestrian, finding and eliminating abusive online posts — can be done in less than one second. These tasks are ripe for automation. "

Example of a simple deep network architecture.

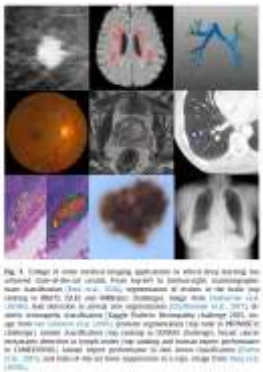


G. Zaharchuk et al. AJNR Am J Neuroradiol
doi:10.3174/ajnr.A5543

Example of training and deployment of deep convolutional neural networks.



G. Zaharchuk et al. AJNR Am J Neuroradiol
doi:10.3174/ajnr.A5543



Overview

Site: brain, lung, breast, abdomen, cardiac, eye

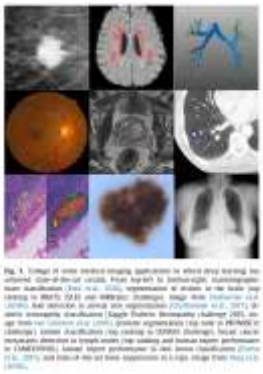
✖

Modality: MR, CT, mammography, PET, ultrasound

✖

Task: Detection, segmentation, recognition, pixel classification, molecular status prediction

Litjens, 2017



Litjens, 2017

Overview

Site: brain, lung, breast, abdomen, cardiac, eye

Modality: MR, CT, mammography, PET, ultrasound

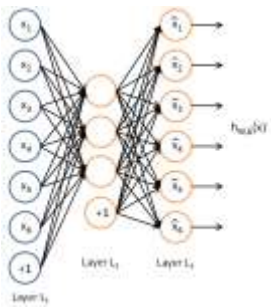
Task: Detection, segmentation, recognition, pixel classification, molecular status prediction

Picked neuroimaging for ease of illustration:
1. some aspects like preprocessing (registration, etc.) may not be relevant here
2. principles about application domains are transferable

Architectures

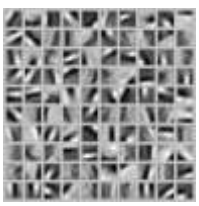
- Autoencoders
- CNN
- RNN/LSTM
- Resnet
- GANs

Architectures: Autoencoders



http://ufldl.stanford.edu/tutorial/unsupervised/Autoencoders/

AEs used for:
1. Unsupervised learning
2. Learning low-dimensional (latent/compressed) representations



CNNs

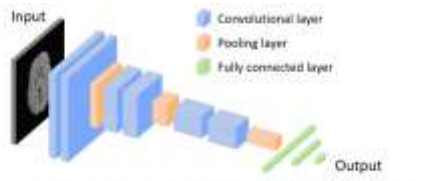
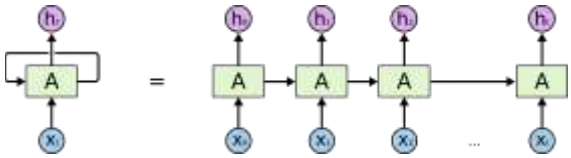


Figure 2: A diagram illustrating a typical architecture of a convolutional neural network.

Architectures: learning temporal dependencies (short/long horizons)

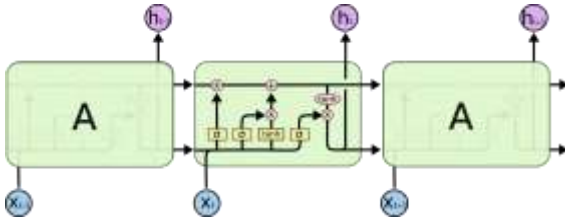
- RNN



http://colah.github.io/

Architectures: learning temporal dependencies (short/long horizons)

- LSTM

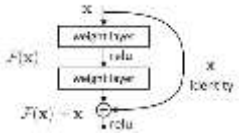


Model long range temporal dependencies

http://colah.github.io/

Architectures

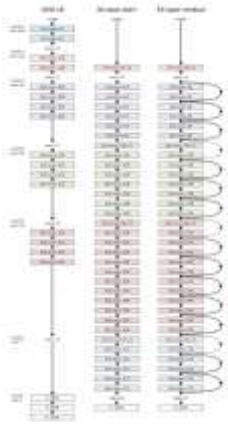
- Resnet



Residual block

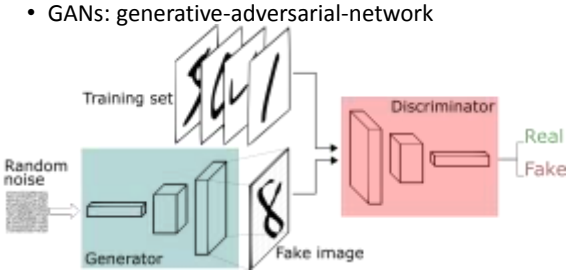
So that's the difference between ResNet and traditional neural nets: Where traditional neural nets will learn H(x) directly, ResNet instead models the layers to learn the residual of input and output of subnetworks. This will give the network an option to just skip subnetworks by making F(x)=0, so that H(x)=x. In other words, the output of a particular subnetwork is just the output of the last subnetwork.

https://wiseodd.github.io/techblog/2016/10/13/residual-net/



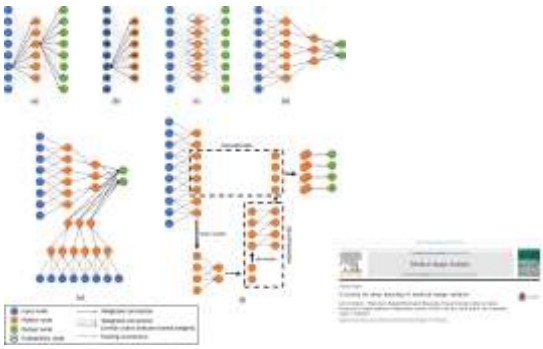
Architectures

- GANs: generative-adversarial-network



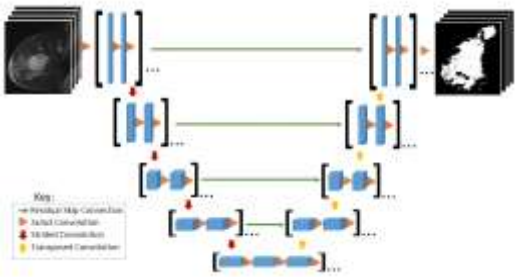
Learn how to synthesize structured data from noise (learning probability distributions)

https://deeplearning4j.org



Node graphs of 1D representations of architectures commonly used in medical imaging. (a) Auto-encoder, (b) restricted Boltzmann machine, (c) recurrent neural network, (d) convolutional neural network, (e) multi-stream convolutional neural network, (f) U-net (with a single downsampling stage). Litjens et.al, 2017

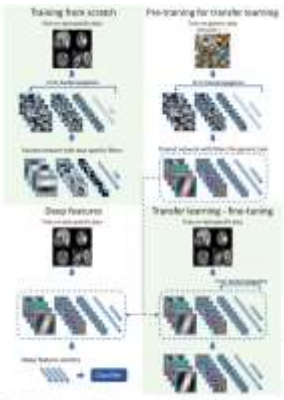
U-net



https://lmb.informatik.uni-freiburg.de/resources/opensource/unet.en.html

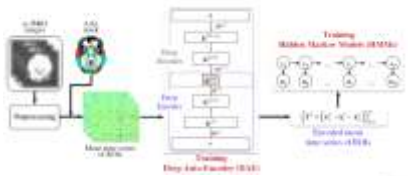
Training Strategies with Data

- Training from scratch
- Transfer learning
- CNN as feature extractor



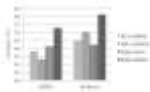
CASE STUDY 1: DISORDER CLASSIFICATION

(ALZHEIMER'S, MILD COGNITIVE IMPAIRMENT, SCHIZOPHRENIA)



DL tool:
SAE

- **Problem:** Mild Cognitive Impairment vs Healthy Control classification in rs-fMRI
- **Solution:** SAE for feature extraction, hidden Markov models as a generative model on top
- **Architecture:** 4-layer deep auto-encoder
- **Method of training:** from scratch; 8060 ROIs from ADNI2 dataset*
- **Performance:** Outperforms classic Principal Components-based approaches by >10 percentage points

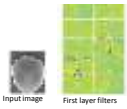


Classification performance

Sak, H.-i., et al. 2016. State-space model with deep learning for functional dynamics estimation in resting-state fMRI. NeuroImage 129, 292–307. doi: 10.1016/j.neuroimage.2016.01.005 .
* Available at 'http://www.loni.ucla.edu/ADNI'

DL tool:
CNN

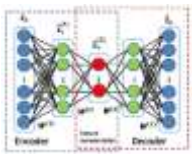
- **Problem:** Alzheimer brain vs healthy brain from fMRI images in ADNI dataset
- **Solution:** Standard CNN
- **Architecture:** Le-Net5
- **Method of training:** Transfer learning from Le-Net5 learnt on MNIST; 270900 samples from ADNI dataset*
- **Performance:** Obtained 96% classification accuracy



Input image First layer filters

Sarraf, S., et al. 2016. Classification of Alzheimer's disease using fmri data and deep learning convolutional neural networks. arxiv: 1603.08631
* from http://adni.loni.usc.edu/.

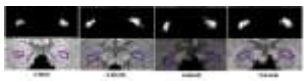
CASE STUDY 2: SEGMENTATION
(TISSUE, LESION, ANATOMY, TUMOR)



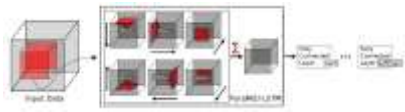
A 2-layer SAE

DL tool:
SAE

- **Problem:** Hippocampus segmentation
- **Solution:** SAE for representation learning used for target/atlas patch similarity measurement
- **Architecture:** 4-layer SAE
- **Method of training:** from scratch; 144 sagittal and 64 axis slices from 10 infants
- **Performance:** Mean Dice ratio(%) of 70.2, outperformed intensity features and handcrafted features based segmentation

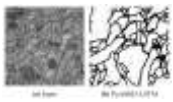

Red contours: manual, blue contours: proposed method

Guo, Y., et al. 2014. Segmenting hippocampus from infant brains by sparse patch matching with deep-learned features. In: Proceedings of the Medical Image Computing and Computer-Assisted Intervention. doi: 10.1007/978- 3- 319- 10470- 6 _ 39 .

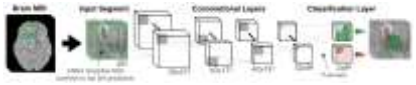


DL tool:
LSTM

- **Problem:** Tissue segmentation in MR and EM images
- **Solution:** 2D LSTM
- **Architecture:** Novel Parallel multi-dimensional LSTM
- **Method of training:** from scratch; number of training images were 30 slices for EM dataset; 240 slices for MRBrainS13 dataset
- **Evaluation:** Rand index, Dice coefficient
- **Performance:** best brain segmentation results on MRBrainS13 (and competitive results on EM-ISBI12) datasets

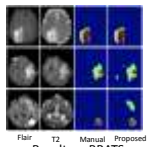


Stollenga, M.F., et al. 2015. Parallel multi-dimensional LSTM, with application to fast biomedical volumetric image segmentation. In: Proceedings of the Advances in Neural Information Processing Systems, pp. 2998–3006 .



DL tool:
CNN

- **Problem:** Lesion segmentation (TBI, Tumor, ischemic stroke)
- **Solution:** 3D multi-scale fully convolutional network with Conditional Random Field for label consistency
- **Architecture:** Novel 11-layer, multi-scale, 3D CNN with Conditional Random Fields for prediction
- **Method of training:** from scratch; 46 multi-channel MRIs for training with data augmentation
- **Performance:** Obtained state of the art DSC, beating semi-automated and 2D CNNs

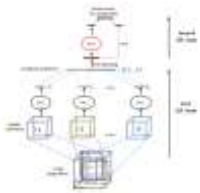


Flair T2 Manual Proposed

Result on BRATS

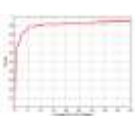
Kamnitsas, K, et al. 2017. Efficient multi-scale 3D CNN with fully con- nected CRF for accurate brain lesion segmentation. Med. Image Anal. 36, 61–78. doi: 10.1016/j.media.2016.10.004 .

CASE STUDY 3: DETECTION AND CLASSIFICATION
(LESION, TUMOR)

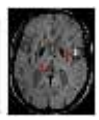


DL tool:
CNN (ISA)

- **Problem:** Microbleed detection in SWI MR images
- **Solution:** 3D stacked Independent Subspace Analysis (ISA) for candidate feature extraction + SVM classification
- **Architecture:** 2-layer stacked ISA
- **Method of training:** from scratch; number of training images equal to 454 slices from 25 patients
- **Performance:** Halved the false positives for a given detection rate compared to best existing method



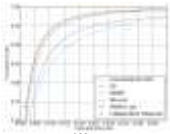
FROC curve



Red: true detections
White: Removed false candidates

Dou, Q, et al. 2015. Automatic cerebral microbleeds detection from MR images via independent subspace analysis based hierarchical features. doi: 10.1109/EMBC.2015.7320232 .

DL tool:
CNN



ROC curves

- **Problem:** Lacune detection in brain MR images
- **Solution:** Fully connected network for candidate segmentation then a multi-scale 3D CNN with anatomical features as false positive reduction
- **Architecture:** Novel 3D CNN + spatial features
- **Method of training:** from scratch; patches from MRI images from 378 cases dataset used for training*
- **Performance:** Multi-scale CNN integrated with spatial location information outperforms other CNN and a conventional method using handcrafted features

Ghafoorian, M. et al. 2016a. Location-sensitive deep convolutional neural networks for segmentation of white matter hyperintensities. arxiv: 1610.04834 .
* Van Norden, A.G., et al. Causes and consequences of cerebral small vessel disease. The RUN DMC study: a prospective cohort study. Study rationale and protocol. BMC Neurol 11 (2011) 29.

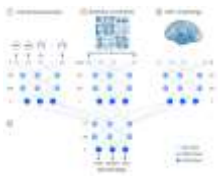
CASE STUDY 4: CLINICAL OUTCOME PREDICTION
(SURVIVAL, DISEASE ACTIVITY, DISEASE DEVELOPMENT)

DL tool:
CNN

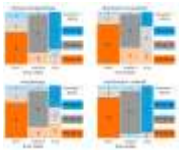
CNN discovers connections predictive of outcome

- **Problem:** Neurodevelopment prediction
- **Solution:** CNN with specially-designed edge-to-edge, edge-to-node and node-to-graph conv. layers for brain nets
- **Architecture:** 6-layer CNN with convolutional filters modified as 'edge to edge' filters
- **Method of training:** from scratch; 112 DTI scans from 115 infants, with data augmentation, is used for training
- **Performance:** Outperforms competing approaches based on PCA, Fully connected networks and clinical predictors in predicting Bayley-III motor and cognitive scores

Kawahara, J., et al. 2016b. Brainnetcnn: convolutional neural networks for brain networks; towards predicting neurodevelopment. Neuroimage doi: 10.1016/j. neuroimage.2016.09.046 .



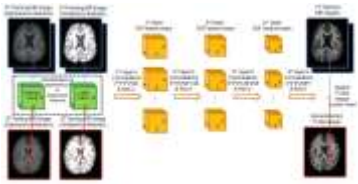
DL tool:
ANN



- **Problem:** Survival prediction of 135 ALS patients using MR images
- **Solution:** ANN + clinical characteristics and structural connectivity data
- **Architecture:** 4-layer ANNs for clinical, structural connectivity, and brain morphology, followed by a 3-layer survival class ANN
- **Method of training:** from scratch; 83 MR scans used for training
- **Performance:** Achieved 84.4% accuracy compared to 69% using a deep network learned on clinical characteristics alone

van der Burgh, H.K.et al. 2017. Deep learning predictions of survival based on MRI in amyotrophic lateral sclerosis. Neuroimage Clin. 13, 361–369. doi: 10.1016/j.nicl.2016.10.008 .

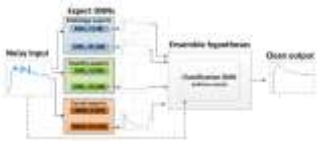
CASE STUDY 5: IMAGE CONSTRUCTION AND ENHANCEMENT
(MODALITY TRANSFORMATION, DENOISING, SUPER-RESOLUTION)



DL tool:
CNN

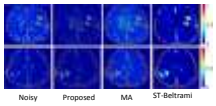
- **Problem:** Constructing higher-resolution MR images
- **Solution:** 3D CNN for constructing 7T-like images from 3T MRI
- **Architecture:** 4-layer CNN (MatConvNet)
- **Method of training:** from scratch; 14 pairs of 3T/7T images used for training
- **Performance:** Obtains PSNR improvement of 1.1dB over state of the art Canonical Correlation Analysis (CCA) method

Bahrami, K, et al. 2016. Convolutional neural network for re- construction of 7T-like images from 3T MRI using appearance and anatomical features. In: Proceedings of the Deep Learning in Medical Image Analysis (DLMIA). doi: 10.1007/ 978- 3- 319- 46976- 8 _ 5 .

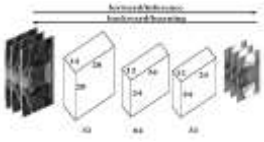


DL tool:
SAE+RBM

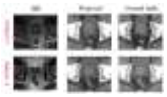
- **Problem:** Denoising DCE-MRI
- **Solution:** using an ensemble of denoising SAE (pretrained with RBMs)
- **Architecture:** Ensemble of expert deep autoencoders (11 layers) and a classification deep autoencoder (4 layers)
- **Method of training:** Pre-training using RBMs, training on 33 DCE-MRI sequences
- **Performance:** MRIs denoised using proposed approach outperform Moving Average (MA) and state of the art ST-Beltrami methods



Benou, A, et al. 2016. De-noising of contrast-enhanced MRI sequences by an ensemble of expert deep neural networks. In: Proceedings of the Deep Learning in Medical Image Analysis (DLMIA).



DL tool:
FCN

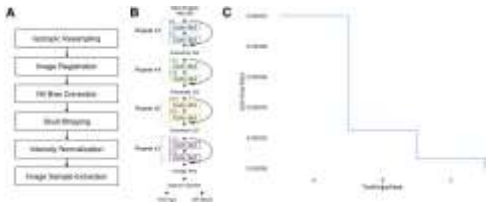


- **Problem:** Constructing CT from MR images (pelvis)
- **Solution:** 3D fully convolutional network for patch-wise CT reconstruction
- **Architecture:** 3D FCN, i.e. 3D CNN minus pooling layers
- **Method of training:** from scratch; using CT/MR pairs from 22 subjects
- **Performance:** Obtained PSNR improvement of 1.3dB than atlas-based and existing state of the art structured random forest methods

Nie, D, et al. 2016a. Estimating CT image from MRI data using 3D fully convolutional networks. In: Proceedings of the Deep Learning in Medical Image Analysis (DLMIA). doi: 10.1007/978-3-319-46976-8_18.

Classifying molecular status

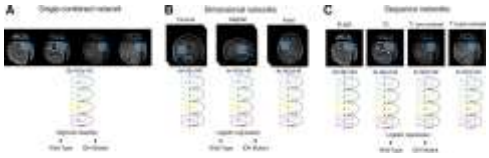
A, Image preprocessing steps in our proposed approach.



Ken Chang et al. Clin Cancer Res 2018;24:1073-1081

Clinical Cancer Research

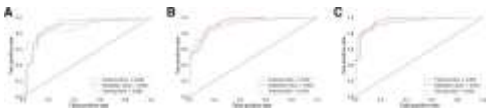
The training heuristics tested include a single combined network (A), dimensional networks (B), and sequence networks (C).



Ken Chang et al. Clin Cancer Res 2018;24:1073-1081

Clinical Cancer Research

ROC curves for training, validation, and testing sets from training on three patient cohorts for age only (A), combining sequence networks (B), and combining sequence networks + age (C).



Ken Chang et al. Clin Cancer Res 2018;24:1073-1081

Clinical Cancer Research

Other ideas

- Exam Classification: image credentialling
- Automated protocolling
- Content based retrieval
- Image generation/synthesis/image transformation
- Classifying molecular status
- Captioning
- Pose-estimation (axial/saggital?)
- Quality assessment
- Object tracking

variations

Weight initialization

Considerations:
- If the weights in a network start too small, then the signal shrinks as it passes through each layer until it's too tiny to be useful.
- If the weights in a network start too large, then the signal grows as it passes through each layer until it's too massive to be useful.

Common initialization strategies:

- Random initialization: $W = 0.01 * np.random.randn(layer\_size[l], layer\_size[l-1])$
  - Works fine for small networks, but can lead to non-homogeneous distributions of activations across the layers of a network.
- Xavier initialization: $W = np.random.randn(layer\_size[l], layer\_size[l-1]) * np.sqrt(2/layer\_size[l-1])$
  - the weights are initialized keeping in mind the **size of the previous layer** which helps in attaining a global minimum of the cost function faster and more efficiently.
- Other approaches:
  - Random walk initialization for training very deep feedforward networks [Sussillo and Abbott, 2014]
  - All you need is a good initialization [Mishkin and Matas, 2015]

Activation function

Sigmoid

1. Saturated neurons kill gradients
2. Sigmoid output not zero-centered
3. Exp() a bit compute intensive

Tanh

1. Saturated neurons still kills gradients
2. Zero-centered

ReLU

1. Does not saturate
2. Computationally efficient
3. Converges much faster than sigmoid/tanh in practice
4. Not zero-centered

https://cpb-us-e1.wpmucdn.com/blogs.rice.edu/dist/1/7022/files/2017/11/ELEC-677-Lec-4-15wt03m.pdf

Optimization and Learning rate

Gradient Descent (GD) is one of the most popular algorithms to perform optimization and by far the most common way to optimize neural networks
The learning rate determines *how fast* we approach the minima in the direction suggested by GD

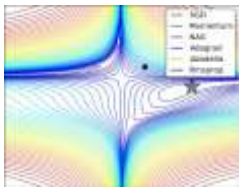
- Batch GD:
  - Guaranteed to converge to a global minima or a local minima
  - Compute gradient for whole training set→ intractable for big datasets
- Stochastic GD:
  - Computes gradient for every training example, making it much faster computationally
  - Can possibly overshoot the minima, but that can be addressed by carefully choosing learning rate
- Mini-batch GD:
  - Combines the best of both worlds, by a) reducing the parameter update fluctuation of SGD, and b) utilizes state of the art matrix computation algorithms to speed up gradient computations with respect to batch GD

The success of mini-batch GD lies in choosing a **proper learning rate**, which can be difficult.
- If too small, painfully slow convergence
- If too large, can hinder convergence and cause the loss function to fluctuate around the minimum or even to diverge

Optimization and Learning rate

Various techniques have been proposed to address how we choose learning rates:
- Momentum [Qian, 1999]
  - helps **accelerate SGD** in the **relevant direction** and dampens oscillations
- Nesterov Accelerated Gradient (nesterov, 1983]
  - Gradient is computed with respect to approximate **future value of the parameters**
  - Pre-emptively **prevents overshooting** and makes convergence faster
- Adagrad [Duchi, 2011]
  - **Adapts** learning rate to the parameters
  - performing smaller updates (i.e. low learning rates) for parameters associated with frequently occurring features, and larger updates (i.e. high learning rates) for parameters associated with infrequent features
  - Well-suited for **sparse data**
- Adadelta [Zeiler, 2012]
  - The way Adagrad is formulated, causes the learning rate to become very small, slowing down training
  - Adadelta prevents learning rate to become small. Further it **eliminates need to manually set** a learning rate
- RMSprop [Hinton, 2012]
  - Functionally similar to Adadelta, with an additional exponential decaying term in the mathematical formulation
- Adam, AdaMax, and Nadam [Kingma, 15]
  - Adaptive learning rate optimization strategies that find more stable minima.

Optimization and Learning rate

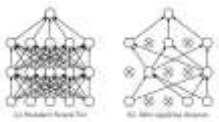


Visual comparison of various SGD optimization techniques

Visualization courtesy http://ruder.io/optimizing-gradient-descent/index.html#stochasticgradientdescent

Regularization

- Batch normalization [Ioffe, 2015]
  - We update parameters to different extents, we lose parameter normalization => slows down training and amplifies changes as the network becomes deeper
  - Batch normalization reestablishes these normalizations for every mini-batch and changes are back-propagated through the operation => higher learning rates , pay less attention to the initialization parameters, and regularization
- Dropout [Srivastava, 2014]
  - Randomly set some neurons to zero in forward pass
  - Reduces overfitting
  - Increases time to convergence, however training time for each pass is lower



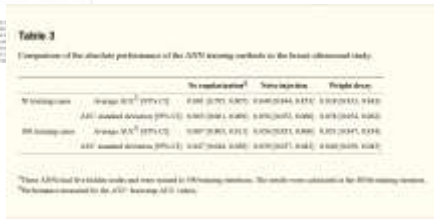
Assessing Robustness

- Noise injection & regularization
- Grad-CAM
- Deepfool: adversarial perturbations

robustness

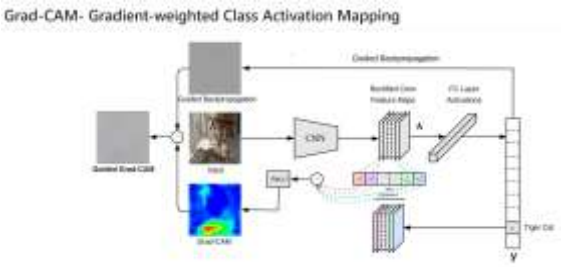
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2771718/

robustness



https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2771718/

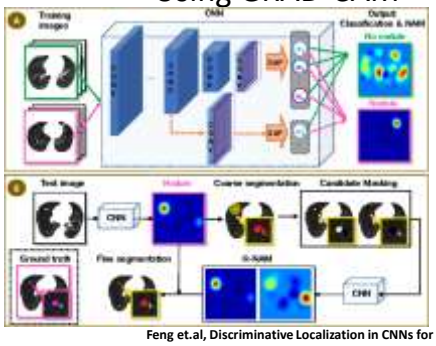
Using GRAD-CAM



https://ramprs.github.io/2017/01/21/Grad-CAM-Making-Off-the-Shelf-Deep-Models-Transparent-through-Visual-Explanations.html#guided-grad-cam

https://ramprs.github.io/2017/01/21/Grad-CAM-Making-Off-the-Shelf-Deep-Models-Transparent-through-Visual-Explanations.html#guided-grad-cam

Using GRAD-CAM



Feng et.al, Discriminative Localization in CNNs for Weakly-Supervised Segmentation of Pulmonary Nodules, MICCAI 2017

Adversarial Perturbations in GANs

A Domain Guided CNN Architecture for Predicting Age from Structural Brain Images

Pascal Sturmfels
University of Michigan
Ann Arbor, MI

Sage Rutherford
University of Michigan
Ann Arbor, MI

Claudia Sripada
University of Other Kinds of Awesome
Sure Fire City, State, Country

Mike Angstadt
University of Michigan
Ann Arbor, MI

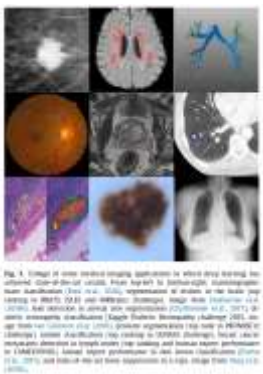
Mark Peterson
University of Other Kinds of Awesome
Sure Fire City, State, Country

Jenna Weiss
University of Michigan
Ann Arbor, MI

Multiple Tools available

- CNTK: Microsoft Cognitive Toolkit
- Keras
- TensorFlow (tf)
- Theano
- PyTorch
- DLTK: Deep Learning Toolkit
- NiftyNet: tf based
- MXNet

(some deployable on Cloud resources like Azure, AWS)



Litjens, 2017

Overview

Site: brain, lung, breast, abdomen, cardiac, eye

Modality: MR, CT, mammography, PET, ultrasound

Task: Detection, segmentation, recognition, pixel classification, molecular status prediction

(Selected) References

- Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, van der Laak JAWM, van Ginneken B, Sánchez CI. A survey on deep learning in medical image analysis. Med Image Anal. 2017 Dec;42:60-88.
- Mazurowski MA, Buda M, Saha A, Bashir MR, Deep learning in radiology: an overview of the concepts and a survey of the state of the art, arxiv:1802.08717
- Zaharchuk G, Gong E, Wintermark M, Rubin D, Langlotz CP. Deep Learning in Neuroradiology. AJNR Am J Neuroradiol. 2018 Feb 1.
- Erickson BJ, Korfiatis P, Kline TL, Akkus Z, Philbrick K, Weston AD. Deep Learning in Radiology: Does One Size Fit All? J Am Coll Radiol. 2018 Mar;15(3 Pt B):521-526. doi: 10.1016/j.jacr.2017.12.027.
- Dreyer K, Geis JR, When Machines Think: Radiology's Next Frontier, Radiology, Vol. 285, No. 3, 2017.
- Other references are on slides