Our Machine Learning Research

Dr. David Pan

Machine Intelligence and Big Data Lab
Dept. of Electrical & Computer Engineering
Univ. of Alabama in Huntsville (UAH)

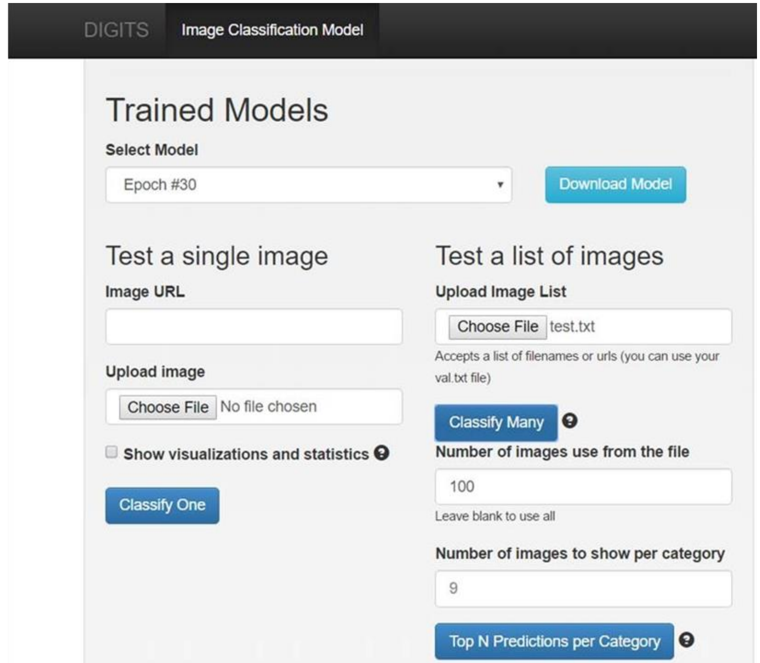
Research Group

- Dr. David Pan (Faculty Member)
  http://www.ece.uah.edu/~dwpan/

- PhD Students
  - Yuhang Dong: Biomedical Data
  - Bernard Benson: Space Science Data
  - Reetu Hooda: Image Compression
  - Myles Harthun: 3D Object Classification

- Recent PhD Alumni
  - Dr. Zhuocheng Jiang: Hyperspectral Image Compression (Postdoc, Univ. of South Carolina)
  - Dr. Hongda Shen (Machine Learning Technical Lead, Chubb, New York)
  - Dr. Amir Liaghati (Electrical Engineer, Boeing, Huntsville, AL)
  - Dr. Yi Wang (Associate Professor, Manhattan College, New York)

Dedicated Computing System for Deep Learning



The "DevBox" has four Nvidia TITAN X GPUs (12 GB of memory each) and is capable of 7 trillion floating point operations per second. (Tensorflow, Caffe, Theano, Torch, DIGITS, cuDNN Library …)

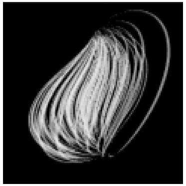
Current Research

- Modeling and Prediction of Solar Activities Using Deep Learning
- Classification of Malaria Infected Cells Using Deep Convolutional Neural Networks
- Efficient Predictive Lossless Compression on Hyperspectral Images Using Machine Learning

Modeling and Prediction of Solar Activities Using Deep Learning

In collaboration with
Center for Space Plasma and Aeronomic Research (CSPAR)
Dept. of Space Science and Center for Space
UAH

Bernard Benson

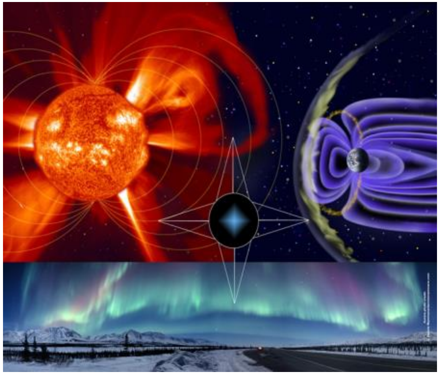


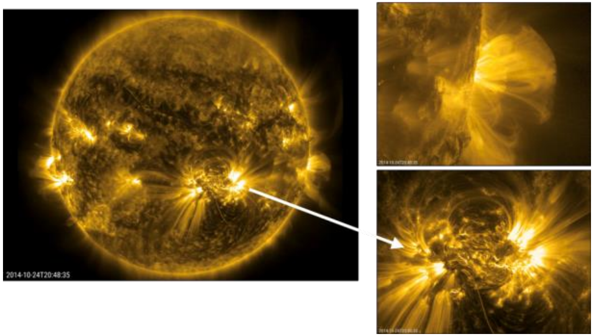
Solar Magnetic Activity Impact on Earth

- Large coronal eruptions like flares and coronal mass ejections can influence Earth's magnetic field.

- This may trigger magnetic storms.

- Coronal eruptions also can cause harmful effects:
  - Disturbances in communications
  - Damages to satellites
  - Causing power outages
  - Causing life-threatening radiation damage to astronauts in space

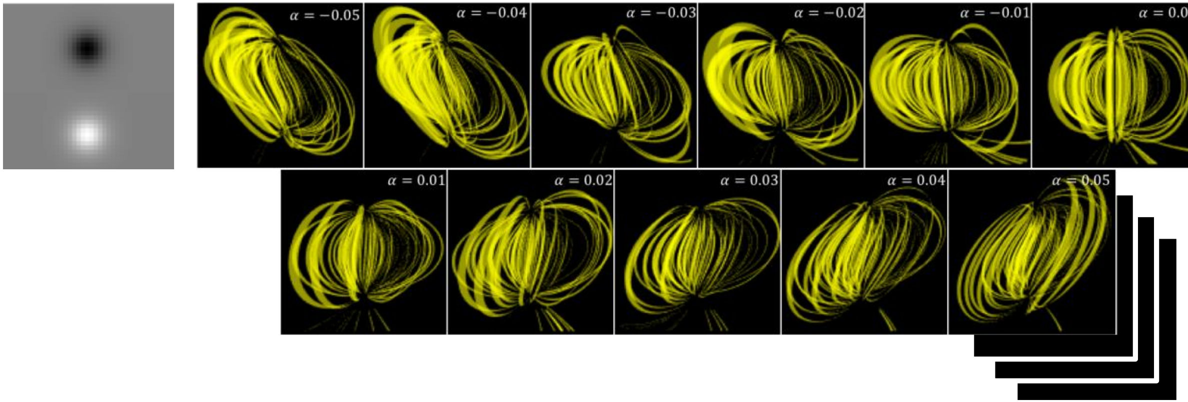




Research Objectives

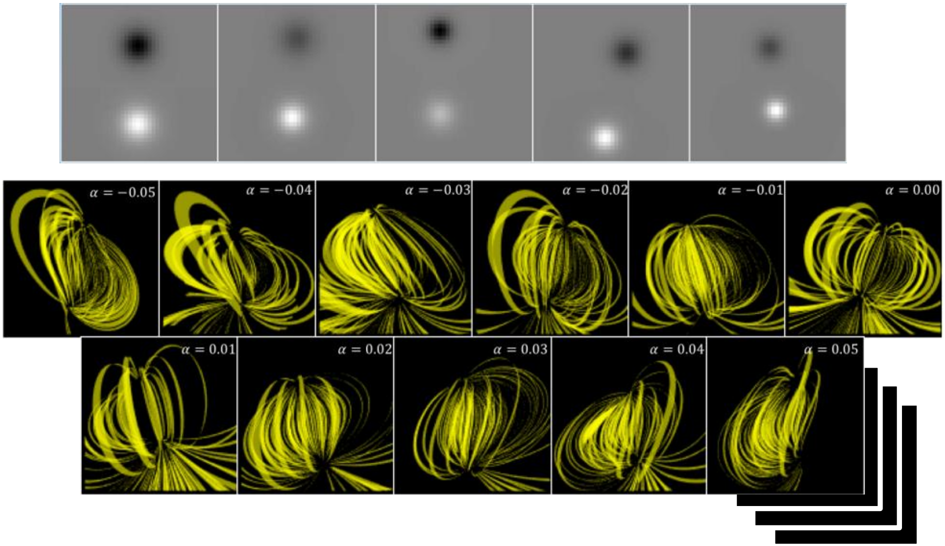
- Use a simple model to describe the nature of an active region.
- Extrapolate the magnetic field of this pseudo-active region to generate synthetic data which is representative of the observed data.
- Gauge the effectiveness of deep neural networks to determine magnetic field parameters of this model.
- Repeat the process with increasingly complex representations of synthetic data.
- Extend the process to high fidelity magnetic field models.

Image Datasets of Coronal Loops



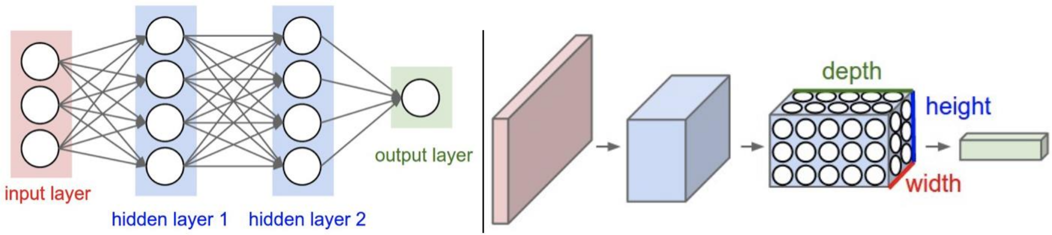
Stationary Dipole Loops with 1000 images per class of α

Datasets of Coronal Loops Images

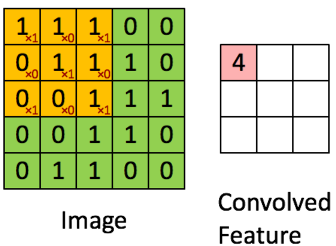


Varying Dipole Loops with 5000 images per α

Convolutional Neural Networks

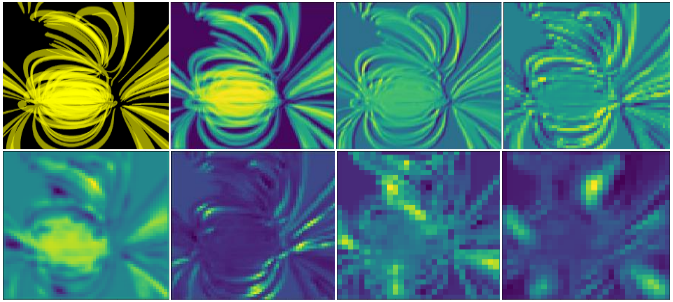


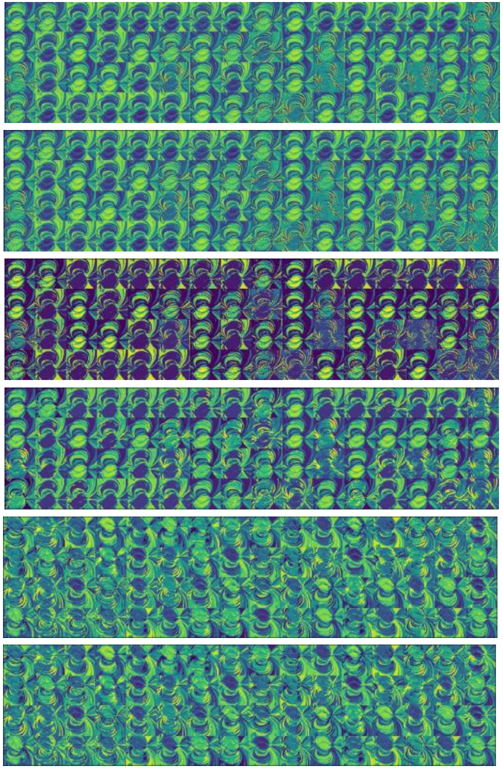
- **Convolution Layer**
- **Activation Function**
- **Pooling Layer**
- **Fully Connected Layer**



Images courtesy of http://cs231n.github.io/convolutional-networks/ and http://danielnouri.org/

Features Learned





Deep CNN Algorithms

- Classify input images into one out of 11 possible α values.

- We experimented with the well-known deep learning algorithms:
  - LeNet-5 (about 18 million parameters)
  - AlexNet (about 28 million parameters)
  - Sparse Autoencoders (to extract features automatically)

- We divided our datasets into:
  - Training (70%)
  - Validation (15%)

LeNet-5 Architecture

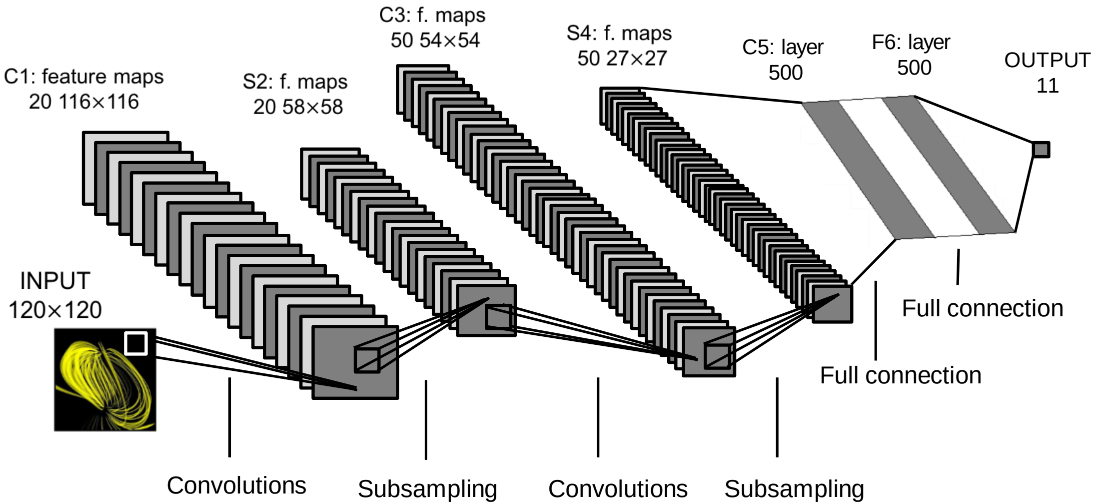
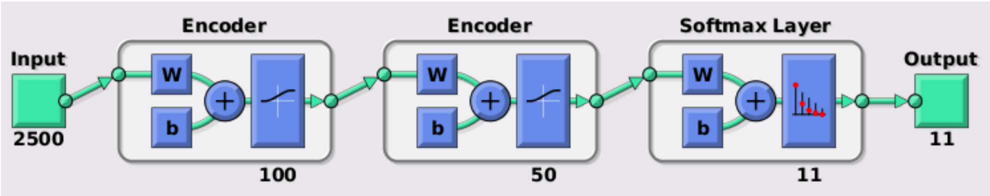


Figure from Yann LeCun et al., 1998

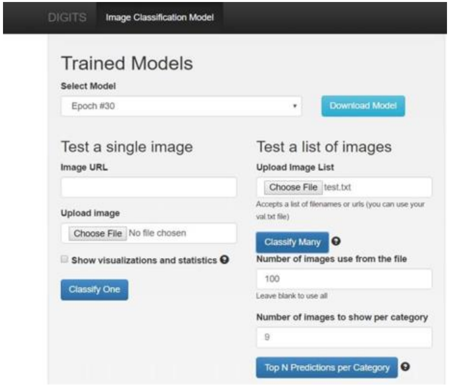
Autoencoders



- Training occurs in an unsupervised manner using training images of size 50×50.
- Two sparse autoencoders used with 100 and 50 hidden units, respectively.
- Softmax activation was used for classification.

Computational Platform and Efficiencies

- Computational Platform: Deep Learning “DevBox”
  with 4 NVIDIA Titan X GPUs, running Nvidia DIGITS software.
- Training Times
  - 2 mins for LetNet-5, 5 mins for AlexNet
- Testing Times
  - 8 seconds for LetNet-5, 10 seconds for AlexNet



Simulation Results

- Very high accuracy using AlexNet on both datasets.
  - Top-1 accuracy: 99.6% and 96.4%
  - Top-5 accuracy: 100%
- Reduced top-1 accuracy (85.49%) on varying dipole dataset due to the absence of dropout in LeNet-5, Top-5 accuracy of 99.88%.
- High sensitivity, specificity and precision show results are not skewed to uneven test cases.

CNN Model	Dataset	Top-1 Accuracy %	Top-5 Accuracy %	Sensitivity	Specificity	Precision
AlexNet	Stationary Dipole Loops	99.60%	100%	0.996	0.999	0.999
	Varying Dipole Loops	96.40%	100%	0.964	0.964	0.996
LeNet-5	Stationary Dipole Loops	99.87%	100%	0.998	0.999	0.999
	Varying Dipole Loops	85.49%	99.88%	0.855	0.986	0.856
Autoencoder	Stationary Dipole Loops	99.80%	100%	0.998	0.999	0.999

Recent Publications:

B. Benson, W. D. Pan, G. A. Gary, Q. Hu, and T. Staudinger, "Determining the Parameter for the Linear Force-Free Magnetic Field Model with Multi-Dipolar Configurations Using Deep Neural Networks," *Astronomy and Computing*, vol. 26, January 2019, pp. 50 - 60. (Link)

Classification of Malaria Infected Cells Using Deep Convolutional Neural Networks

Yuhang Dong, Zhuocheng Jiang, Hongda Shen, David Pan
Dept. of Electrical & Computer Engineering, U. of Alabama in Huntsville

THE UNIVERSITY OF ALABAMA IN HUNTSVILLE

Allen W. Bryan, Jr, Lance A. Williams, Vishnu V. B. Reddy, William H. Benjamin, Jr. ,
Dept. of Pathology, Univ. of Alabama at Birmingham

UAB SCHOOL OF MEDICINE
Knowledge that will change your world

Key Facts about Malaria Infection

- Malaria is a life-threatening disease caused by parasites that are transmitted to people through the bites of infected female Anopheles mosquitoes.
- Malaria is caused by Plasmodium parasites, which are spread to people through the bites of infected female Anopheles mosquitoes, called "malaria vectors."
- There are 5 parasite species that cause malaria in humans, and 2 of these species – *P. falciparum* and *P. vivax* – pose the greatest threat.
- *P. falciparum* is the most prevalent malaria parasite on the African continent, responsible for most malaria-related deaths globally.
- P. vivax is the dominant malaria parasite in most countries outside of sub-Saharan Africa.

Population Groups At Risk

- In 2016, nearly half of the world's population was at risk of malaria.
- Most malaria cases and deaths occur in sub-Saharan Africa. However, the WHO regions of South-East Asia, Eastern Mediterranean, Western Pacific, and the Americas are also at risk.
- In 2016, 91 countries and areas had ongoing malaria transmission.
- Population groups are at considerably higher risk of contracting malaria, and developing severe disease, than others. These include infants, children under 5 years of age, pregnant women and patients with HIV/AIDS,.

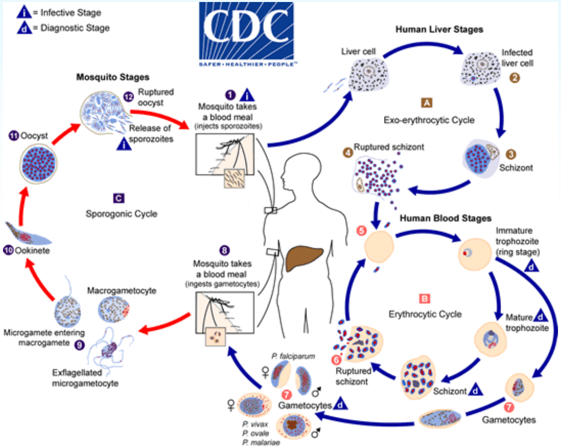
Symptoms

- In a non-immune individual, symptoms usually appear 10–15 days after the infective mosquito bite.
- The first symptoms are:
  fever, headache, and chills – may be mild and difficult to recognize as malaria.
- If not treated within 24 hours, *P. falciparum* malaria can progress to severe illness, often leading to death.

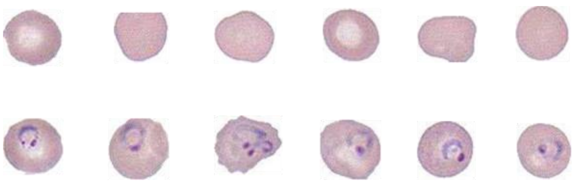
- Children with severe malaria frequently develop one or more of the following symptoms:
  - severe anemia
  - respiratory distress
  - cerebral malaria

Life Cycles of Malaria Infection

The natural ecology of malaria involves malaria parasites infecting successively two types of hosts: humans and female Anopheles mosquitoes. In humans, the parasites grow and multiply first in the liver cells and then in the **red cells of the blood**. In the blood, successive broods of parasites grow inside the red cells and destroy them, releasing daughter parasites ("merozoites") that continue the cycle by invading other red cells.



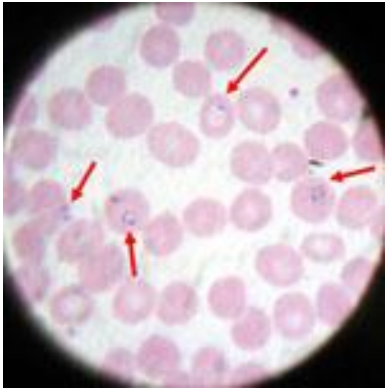
The blood stage parasites are those that cause the symptoms of malaria. When certain forms of blood stage parasites ("gametocytes") are picked up by a female Anopheles mosquito during a blood meal, they start another, different cycle of growth and multiplication in the mosquito.



Fast and Reliable Diagnosis of Malaria

- Reliable malaria diagnoses require necessary training and specialized human resources
- Unfortunately, in many malaria-predominant areas, such resources are <u>inadequate and frequently unavailable.</u>
- Fast and reliable diagnosis will be very useful.

Blood smear from a patient with malaria; microscopic examination shows *Plasmodium falciparum* parasites (arrows) infecting some of the patient's red blood cells. (CDC photo)



Whole Slide Imaging

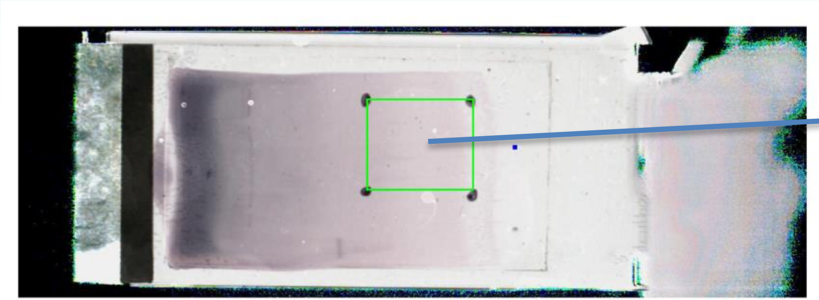
- Whole slide imaging (WSI):
  - Scans conventional glass slides
  - Produces high-resolution digital slides
  - The most recent pathology imaging modality, available worldwide
- WSI images allow for highly-accurate automated identification of malaria infected cells.
- **Challenge**: automatic identification of malaria infected red blood cells in a huge-size WSI.

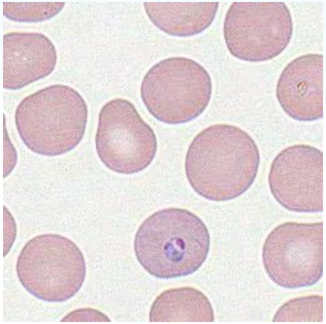
Whole Slide Image for malaria infected red blood cells from UAB

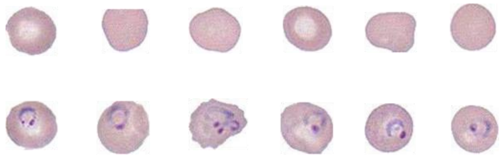
Example of WSI

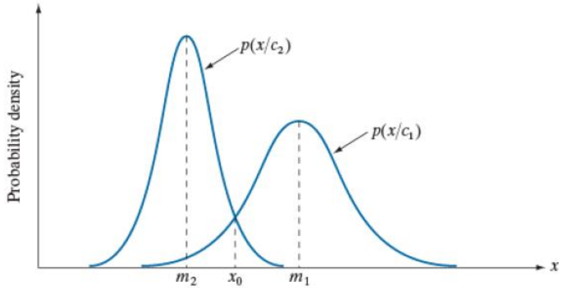
Entire slide with cropped region delineated in green

Image of 258×258 with 100X magnification









Question:

How can we automatically recognize infected cells?

A pattern classification problem

Research Goal and Topics

<u>Develop automatic and reliable algorithms for computer assisted detection of malaria infections</u>

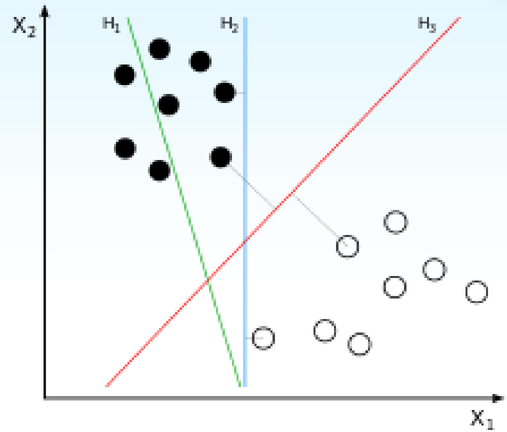
- Machine Learning for Automated Classification of Malaria Infected Cells
- Building a Dataset of Red Blood Cell Images from WSIs for Malaria Infection Detection (first such pathologist curated dataset in the world)
- Deep Convolutional Neural Networks
- Evaluation Results and Case Study
- Conclusions and Further Work

Machine Learning for Malaria Detection

- Machine learning (ML) is a subset of *artificial intelligence* in the field of computer science.
- ML techniques give computers the ability to "learn", i.e., progressively improve performance on a specific task with training data.
- **Feature-based** supervised learning algorithms have been shown to be capable for building automated diagnostic systems for malaria.
- Classification accuracy of **feature-based** supervised learning methods are relatively low:
  - 84% (SVM)
  - 83.5% (Naïve Bayes Classifier)
  - 85% (Three-layer "shallow" Neural Network)

Feature-Based Machine Learning

- In machine learning, support vector machines (SVMs) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis.
- Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other.



In the following work, we employed SVM with a set of pre-determined features (from training data), to classify malaria infected cells from non-infected cells:

- Hu's moment 2,3,5,6,7
- MinIntensity
- Shannon's Entropy

V. Muralidharan, Y. Dong, and W. D. Pan, "A comparison of feature selection methods for machine learning based automatic malarial cell recognition in wholeslide images," *IEEE BHI'16*.

Deep Learning Methods

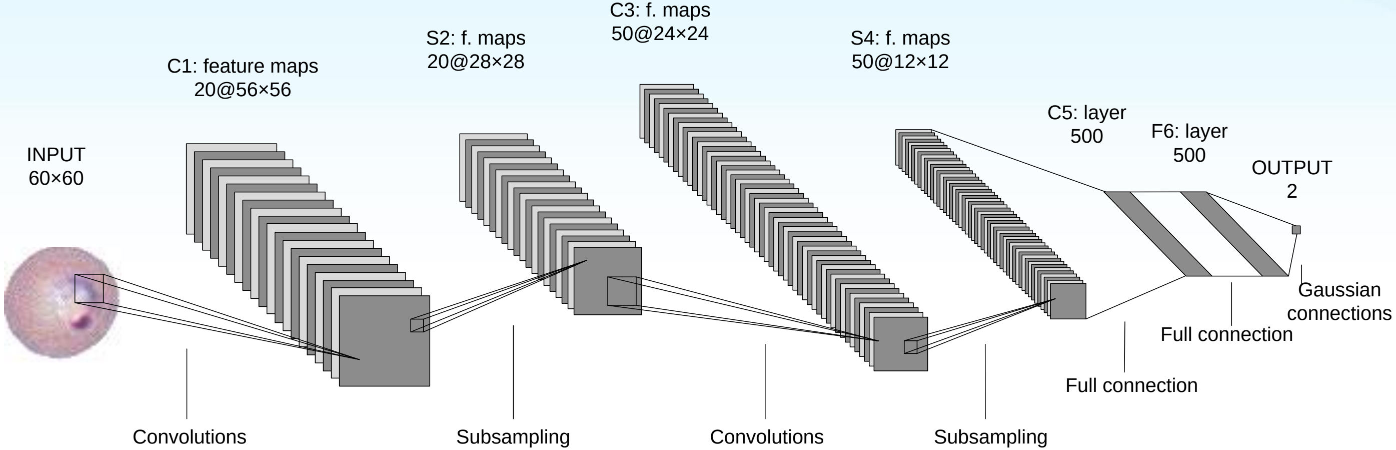
In contrast to feature-based methods, deep learning

- is useful for problems where designing features and programming explicit algorithms with good performance is difficult or infeasible
- can extract hierarchical representation of the data
- higher layers represent increasingly abstract concepts (“features”)
- higher layers become invariant to transformations and scales

We employed deep <u>convolutional</u> neural network (CNN)

- Inspired by biological processes where the connectivity pattern between **neurons** resembles the organization of the animal visual cortex.
- A CNN consists of an input and an output layer, as well as multiple hidden layers.
- The **convolution** operation emulates the response of an individual neuron to visual stimuli.

LeNet-5 CNN Architecture

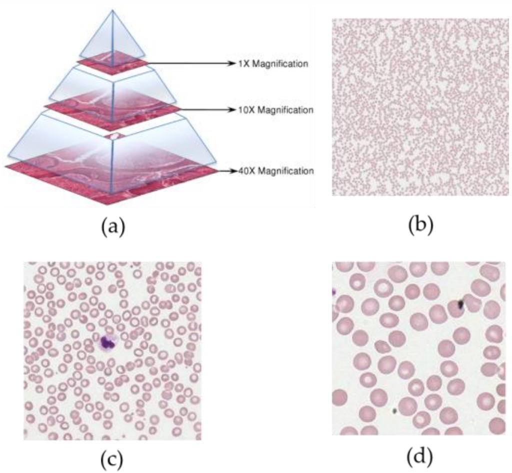


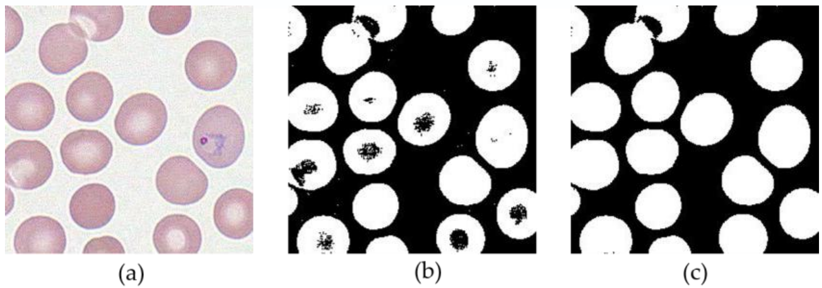
LeNet-5 consists of

- Two convolution layers (C1 and C3)
- Two subsampling (pooling) layers (S2 and S4), and
- Two fully connected layers

Training of CNNs: Data Hungry

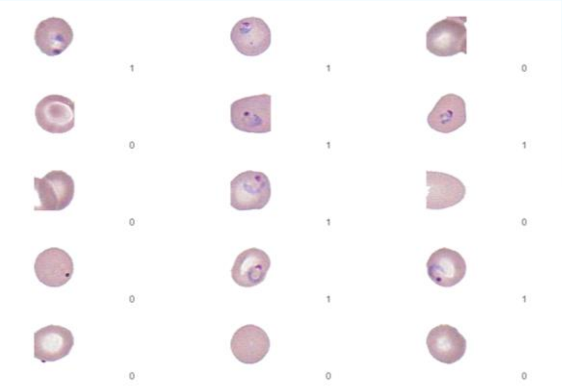
- Deep CNN Training (determining the optimal “weights” assigned for) neuron connections) requires a lot of data.
- There are no publicly available datasets of high-resolution red blood cell images for malaria infection detection.
- We collaborated with UAB pathologist to build curated datasets.





Steps of image processing and morphological operations for single-cell image extraction:
(a) An image tile of interest
(b) Otsu thresholded image
(c) Morphologically filled image

Compilation of a Pathologist Curated Dataset



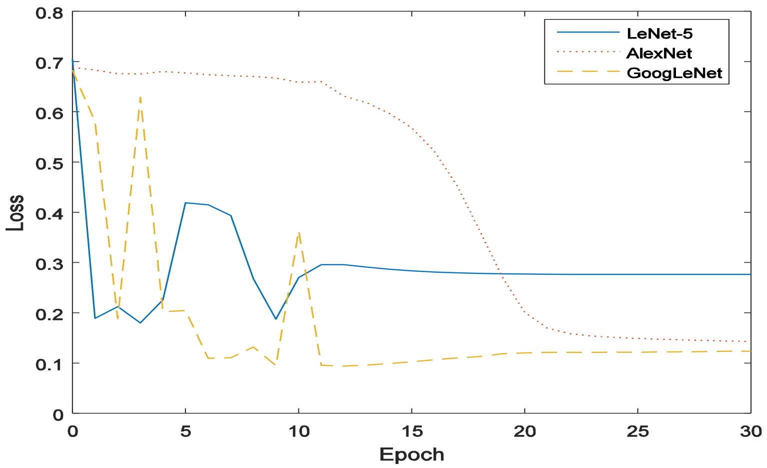
- Dataset curation:
  - Four board-certified pathologists
  - Each single-cell image scored by **at least two pathologists**
  - To include an image in “infected” set, **all reviewers must mark positively** (excluded otherwise).
  - Similarly, to be “non-infected”, all reviewers must mark negatively.
- Final dataset:
  - 1,034 infected cells
  - 1,531 non-infected cells

[Link to the dataset]

Three CNNs Used

CNN	LeNet-5	AlexNet	GoogLeNet
Year Proposed	1998	2012	2014
# of Layers	4	8	22
Top 5 Errors on ILSVRC	?	16.4%	6.7%
# of Convolutional Layers	3	5	21
Convolutional Kernel Size	5	11, 5, 3	7, 1, 3, 5
# of Fully Connected Layers	1	3	1
# of Parameters	3,628,072	20,176,258	5,975,602
Dropout	No	Yes	Yes
Data Augmentation	No	Yes	Yes
Inception	No	No	Yes
Local Response Normalization	No	Yes	Yes

Training and Verification of CNN's



- The dataset is still too small.
- Overfitting issue.
- LeNet-5 has no drop-out.

Label	Training	Testing
Infected	517	517
Normal	765	766

Note: 25% of the training set used for verification.

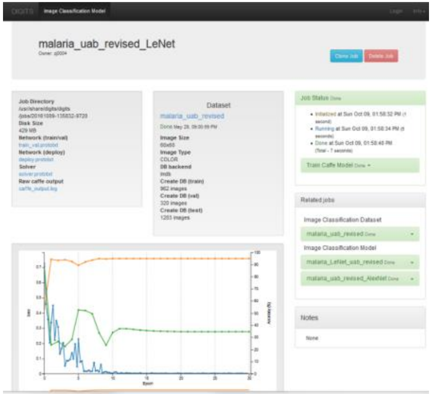
Evaluation Results

		Ground Truth		
		Positive	Negative	Accuracy
LeNet-5	Positive	493	25	**96.18%**
	Negative	24	741	
AlexNet	Positive	502	39	**95.79%**
	Negative	15	727	
GoogLeNet	Positive	503	10	**98.13%**
	Negative	14	756	
SVM	Positive	500	90	91.66%
	Negative	17	676	

Computational Efficiencies

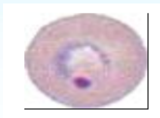
- Support Vector Machine (SVM) involves feature selection and feature extraction.
- Three CNN running times (in seconds):

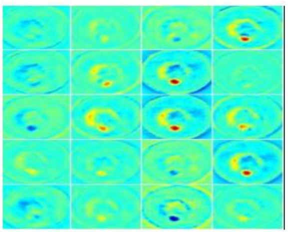
CNN	LeNet-5	AlexNet	GoogLeNet
Training-Validation	7	28	141
Testing	5	5	19

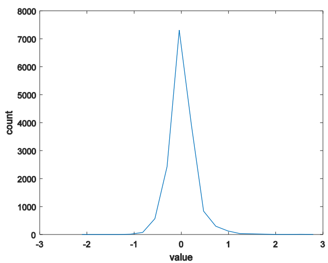


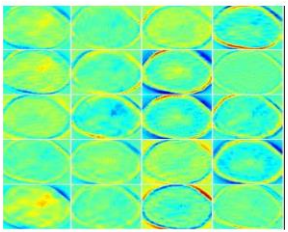
More parameters means longer training and testing time.

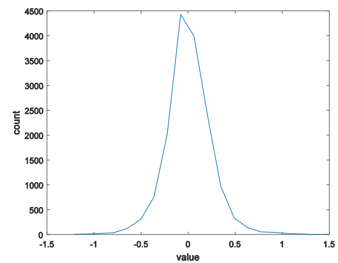
Features Learned (LeNet-5)





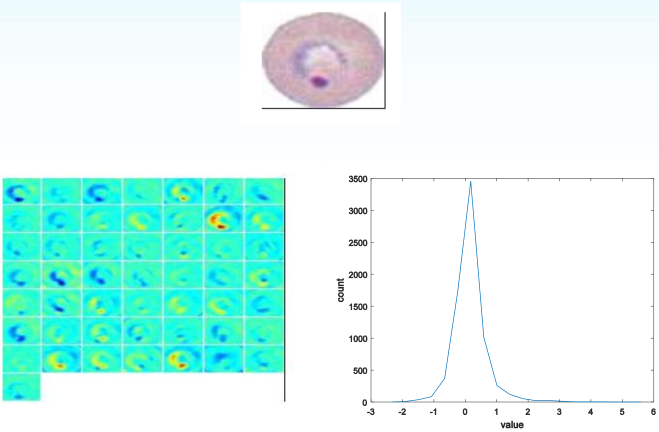


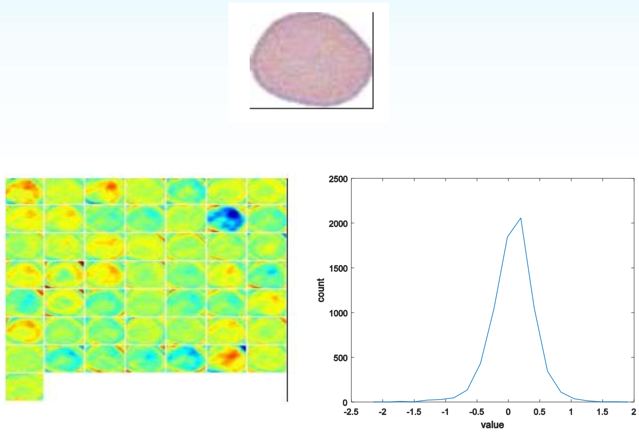




Convolutional Layer 1 and Histogram

Features Learned (LeNet-5)





Convolutional Layer 2 and Histogram

Conclusions

Advantage of using CNN:

- About 98% accuracy achieved with GoogleNet, significantly higher than SVM.
- Tradeoffs between computational complexity and classification accuracy.
- Deep learning methods allow features to be automatically extracted, which is not possible with traditional methods.

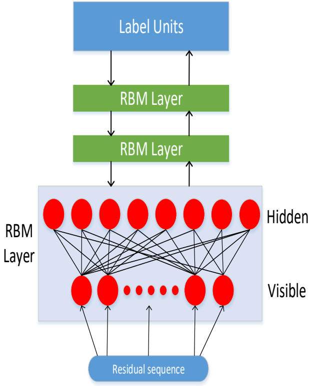
Recent Publication:

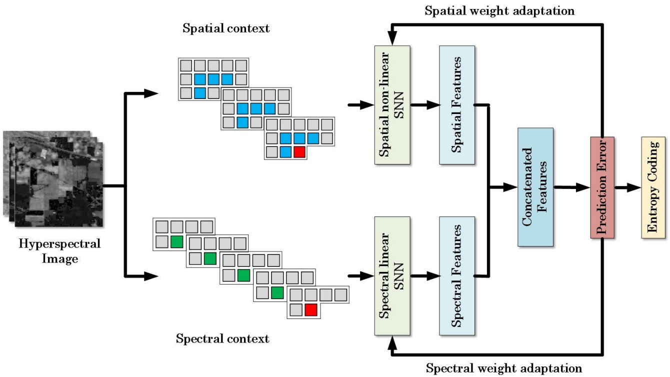
- Data augmentation

W. D. Pan, Y. Dong, and D. Wu, "Classification of Malaria Infected Cells Using Deep Convolutional Neural Networks," Book Chapter in *Machine Learning - Advanced Techniques and Emerging Applications,* IntechOpen, ISBN 978-1-78923-753-5. Sep, 2018. (Link)

Efficient Predictive Lossless Hyperspectral Image Compression Using Machine Learning

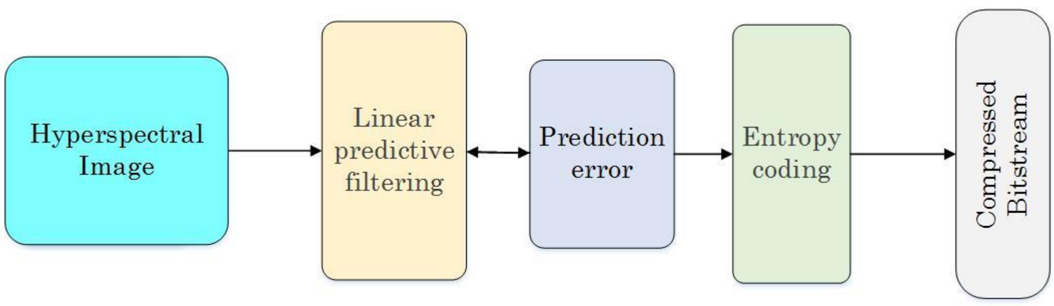
Zhuocheng (Jack) Jiang





Motivation

- Hyperspectral image compression becomes important as more and more hyperspectral data are collected over time.
- Lossless compression vs. lossy compression
  i. Lossy compression: provides lower bit rates but incurs loss on the original data
  ii. Lossless compression: guarantees perfect reconstruction on the original data



Pixel Value Prediction

- On-board lossless compression is more challenging than ground-based compression.
- Traditional machine learning based approaches rely on the availability of data from all the spectral bands during the process of training.
- Large quantity of uncompressed data are normally not available or only partially available in many real-time applications.
- Trained model can not provide an universal solution for all the hyperspectral datasets, which indicates that the model need to retrain for each new dataset

Entropy Coding

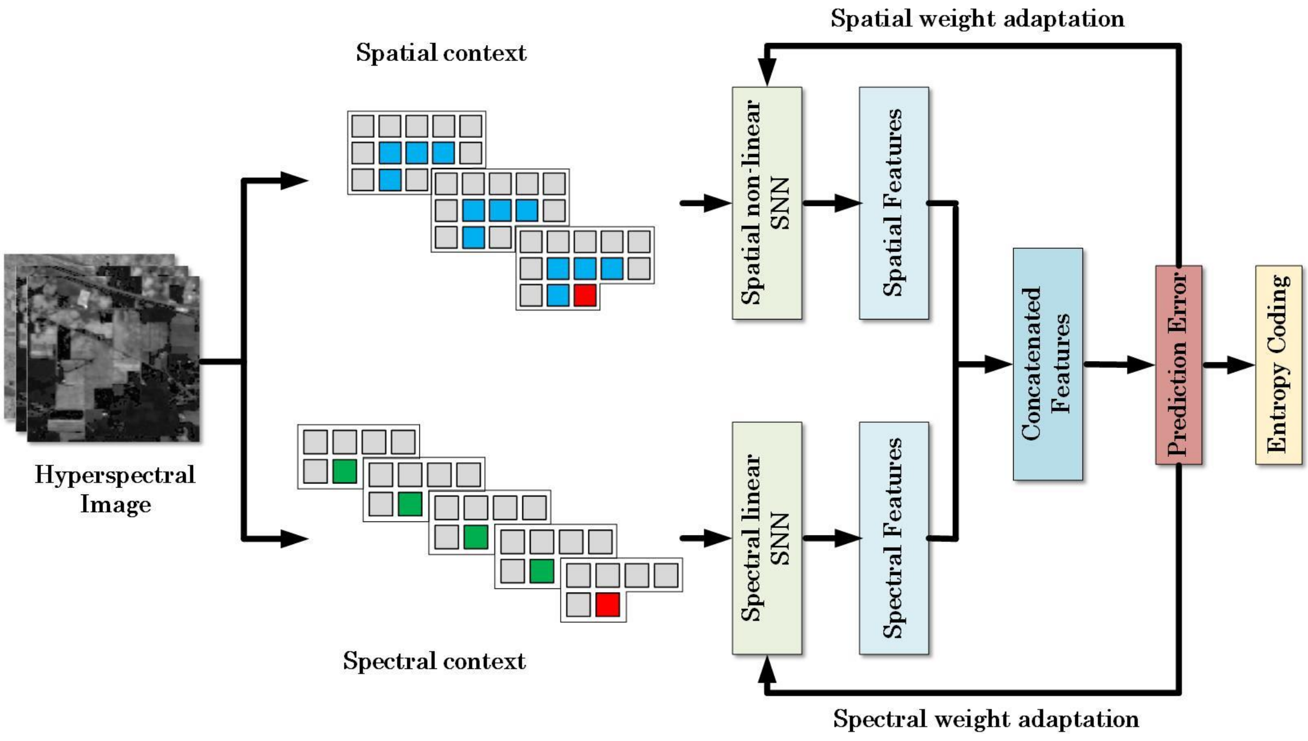
- A Golomb-Rice (GR) codes is often used to compress the residual due to its simplicity and minimal memory capacity requirement.
- Efficiency of the Golomb-Rice codes depends on the accuracy of the coding parameter estimated from the input data.
- Existing parameter estimation methods rely on estimation of the sample mean from the input data sequence (sample mean would vary with changing lengths).
- Existing parameter estimation methods assume the data to be coded follow the geometric distributions (underlying distribution might deviates from the geometric distribution).

Contributions

- We propose an adaptive prediction algorithm based on concatenated shallow neural networks (CSNN). Unlike most of neural network based methods reported in literature, the CSNN was designed as an adaptive prediction filter rather than a training-based network, thus the model need not be pre-trained before they are used for pixel value calculation.

- We formulate the problem of selecting the best coding parameter for a given input sequence as a supervised pattern classification problem. We propose a universal Golomb-Rice coding parameter estimation method using deep belief network, which does not rely on any assumption on the distribution of the input data.

Concatenated Neural Networks



Coding Parameter Estimation

- We consider the problem of parameter estimation as a supervised pattern classification problem.
- In practical implementations, there are only a finite number of parameter values to choose from for Golomb-Rice codes.
- For example, coding an image with 8 bits/pixel would require the coding parameter m to be chosen from 9 possible integers in the set of [0, 8].
- Similarly, a hyperspectral image having 16 bits/pixel would require a slight larger set of 17 integers, [0, 16].

Training and Testing

- In the training phase:

  We train a classifier, with its input being the residual data sequences, the corresponding label for this residual sequence is the value that will give the shortest Golomb-Rice codewords.

- In the testing phase:

  We feed the new residual sequences to the classifier, which will output the best value to code this residual sequence.

More details found in the paper:

Z. Jiang, W. D. Pan, and H. Shen, "Universal Golomb-Rice Coding Parameter Estimation Using Deep Belief Networks For Hyperspectral Image Compression," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (JSTARS)*, vol. 11, no. 10, pp. 3830 - 3840, Oct. 2018. (IEEE Xplore)