

Introduction

Arthur J. Redfern
arthur.redfern@utdallas.edu
Jan 14, 2019

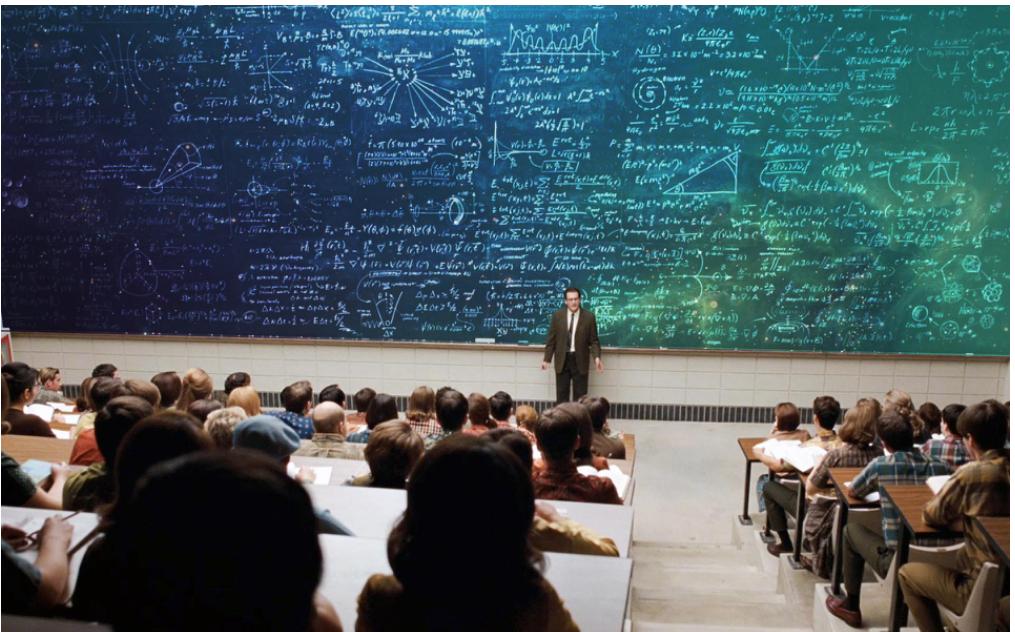
Outline

- Welcome to the course
- Motivation
- Course preview
- Logistics
- Expectations
- Opportunities
- Questions

Welcome To The Course

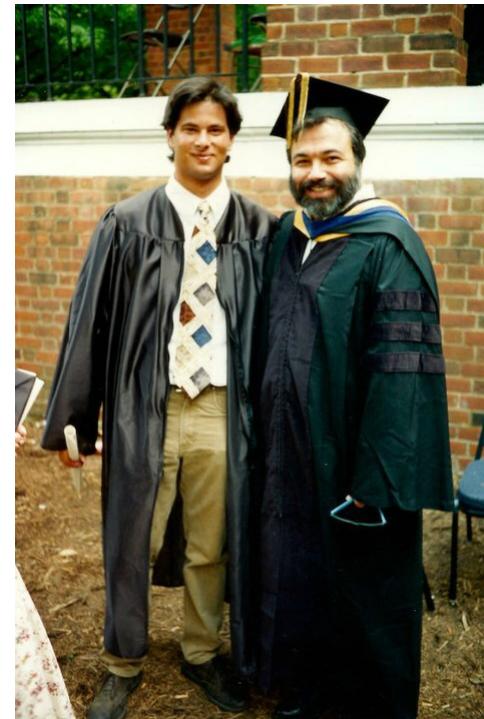
Thank You

- You're taking an important step in your academic and professional careers
- I appreciate you taking it with me
- I'm looking forward to a fun semester

Figure from <http://radscreens.com/i/9709/>

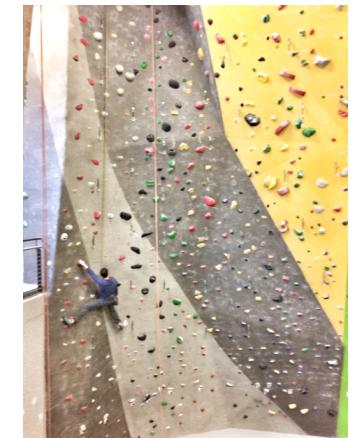
About Me

- Grew up a little south of Richmond, Virginia
- BS in EE from University of Virginia in 1995
 - Started in chemical engineering
 - Moved to electrical engineering
 - Specialized in signal processing
- PhD in ECE from Georgia Tech in 1999
 - My thesis was on Volterra systems (a nonlinear generalization of convolution)
- I'm ok with you calling me any of {Arthur, Dr. Redfern, Professor Redfern}



About Me

- Moved to Dallas to work at Texas Instruments in 2000
 - Physical layer communication system design
 - Signal processing for analog systems
 - Machine learning
- Currently I manage a machine learning lab in the TI Embedded Processors organization
 - Algorithms, software and hardware for different applications
 - This class will cover much of the same (that's not an accident)
- Live in Plano, Texas with my wife and son
- I like to add a new hobby every few years
 - Cars, guitar, poker, golf, running, biking, yoga, climbing, ...



Motivation

Information Extraction

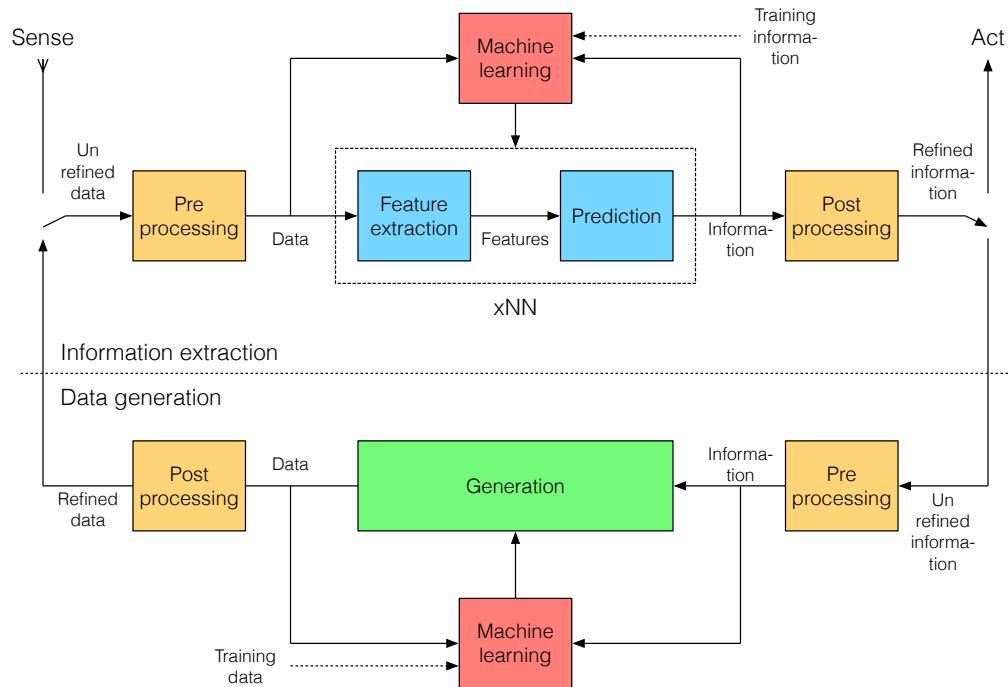
Data to information

- Sense → compute → act
- Transformation from un refined data space to data space to feature space to information space to refined information space

Definitions (not Webster quality)

- Intelligence is the ability to acquire and apply knowledge
 - Artificial intelligence is intelligence exhibited by algorithms
- Learning is the acquisition of knowledge from experience
 - Machine learning is learning from data (experience) applied to an algorithm such that it exhibits artificial intelligence
 - Deep learning is machine learning applied to a deep structure

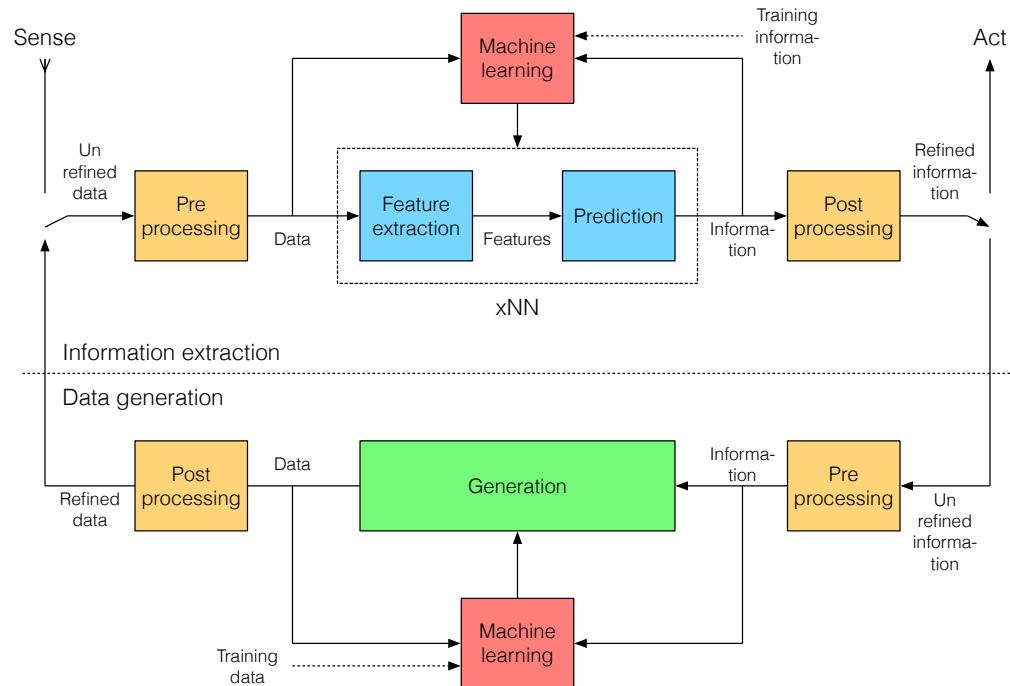
xNNs are deep structures trained using deep learning to exhibit artificial intelligence



Information Extraction

Data to information

- Pre processing
 - Make feature extraction easier
 - Data cleaning, dimensionality reduction, ... frequently via application specific side information
- Feature extraction
 - Make prediction easier
 - Hand engineered or learned
- Prediction
 - Classification (discrete)
 - Regression (continuous)
- Post processing
 - Clean up predictions frequently via application specific side information



Trend: instrumentation and automation of everything

Why Use xNNs For Information Extraction?

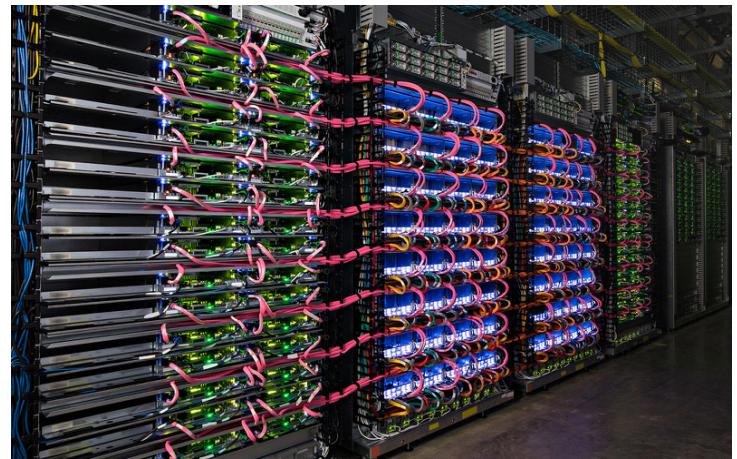
- Many tasks can be cast as a classification problem
 - Probability of input x belonging to class y
 - Given an input x predict a pmf $P(y|x)$
- Neural networks are universal approximations
 - Use them to approximate the mapping from x to $P(y|x)$
- It's possible to design xNNs structures to exploit structure in the data
 - Spatial with CNNs
 - Sequential with RNNs
 - Localized with attention



Figure from <https://www.mapillary.com/dataset/vistas> 10

Why Use xNNs For Information Extraction?

- It's possible to train xNNs from input output data
 - End to end supervised learning allows feature learning vs feature engr (though implicit in network design and training hyper params)
 - Automatic differentiation with reverse mode accumulation
 - Gradient descent variants
 - Keys are convergence and regularization for generalization
- Software and hardware exists for efficient implementations
 - High level graph specification and transformations
 - Low level graph software runtime
 - Primitives for compute and data movement
 - Lowering of tensor ops to matrix matrix mult where possible
- They provide state of the art results in many applications
 - Vision, speech and language are the biggest successes
 - Games, art, control, genomics, ...
 - A general tool with logical extensions to new apps going forward

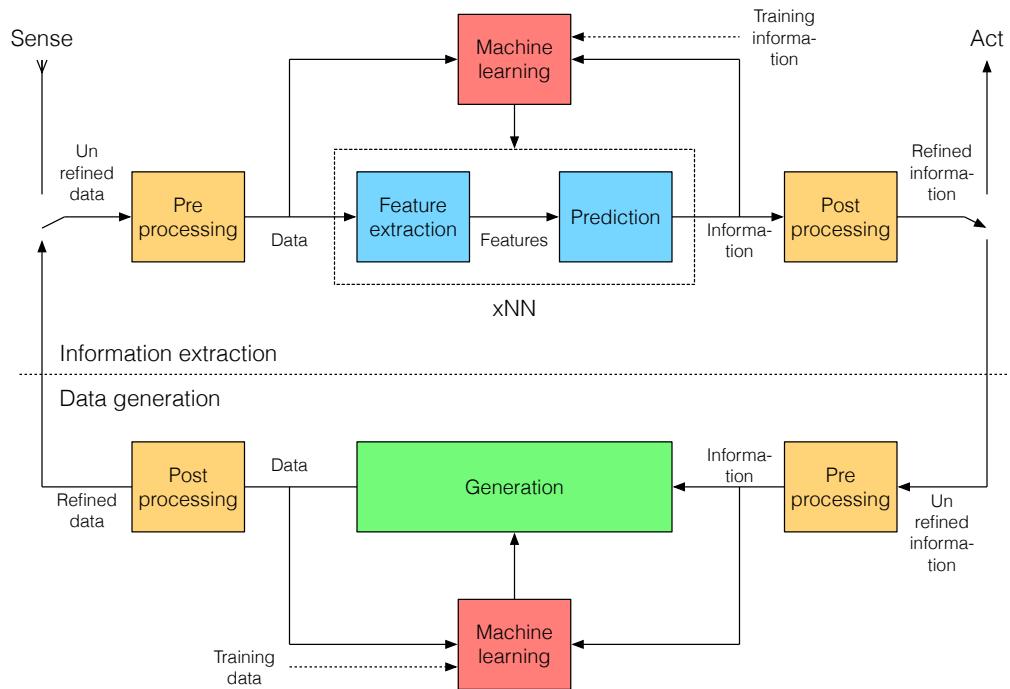
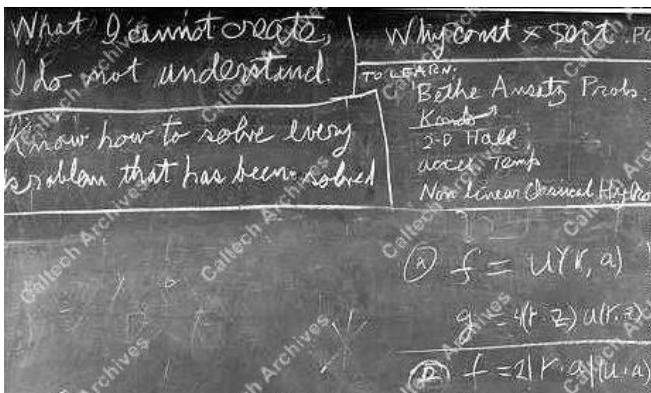


You currently can't discuss top performing vision, speech or language algorithms without xNNs

Data Generation

Information to data

- Richard Feynman: What I cannot create, I do not understand.
- Generative models are learning to create better and better data from information
 - The complement to info extraction
 - Currently less mature



Picture from <http://archives.caltech.edu/pictures/1.10-29.jpg>

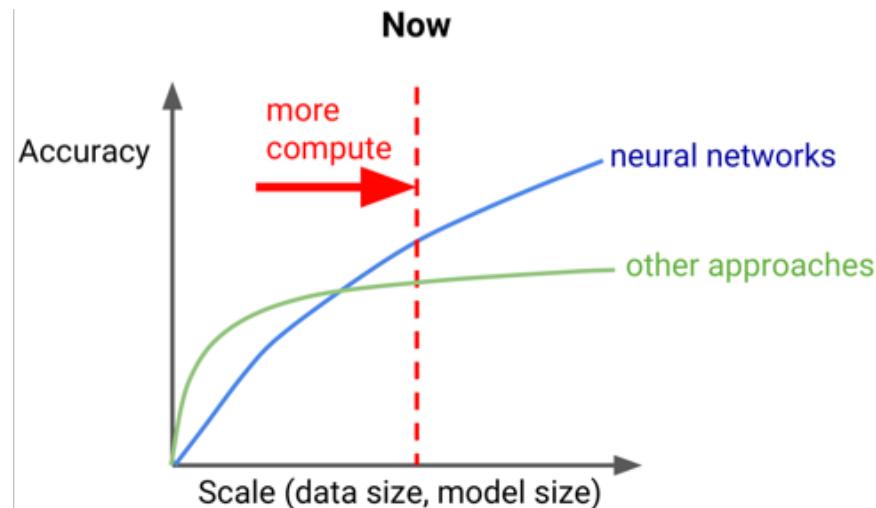
Data Generation

While less mature than information extraction, data generation via GANs and other methods is getting better all the time



Why xNNs Now?

- A confluence of developments
 - More data
 - More compute
 - Better network designs
 - Better training algorithms
 - More efficient information distribution
 - More people, companies and funding
- Successes leading to positive feedback



Course Preview

3 Parts To The Course

Math

- Linear algebra
- Calculus
- Probability
- Algorithms

Networks

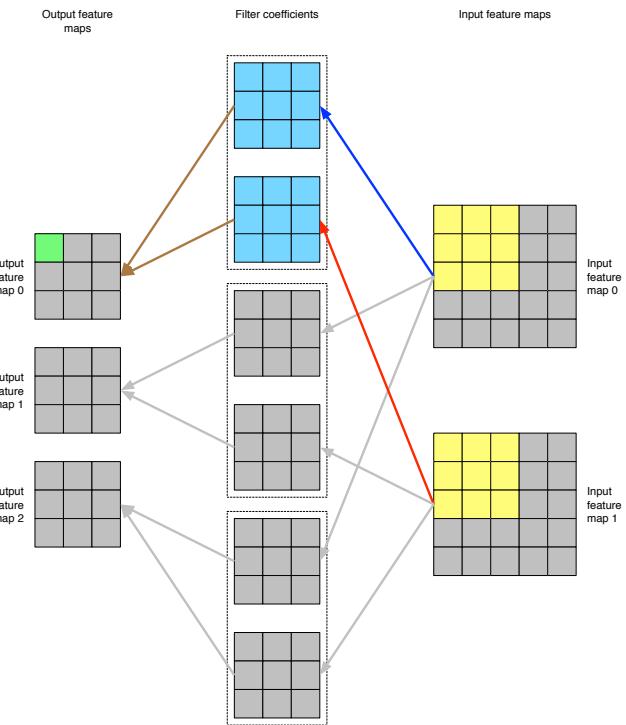
- Design
- Training
- Implementation

Applications

- Vision
- Speech
- Language
- Games
- Art

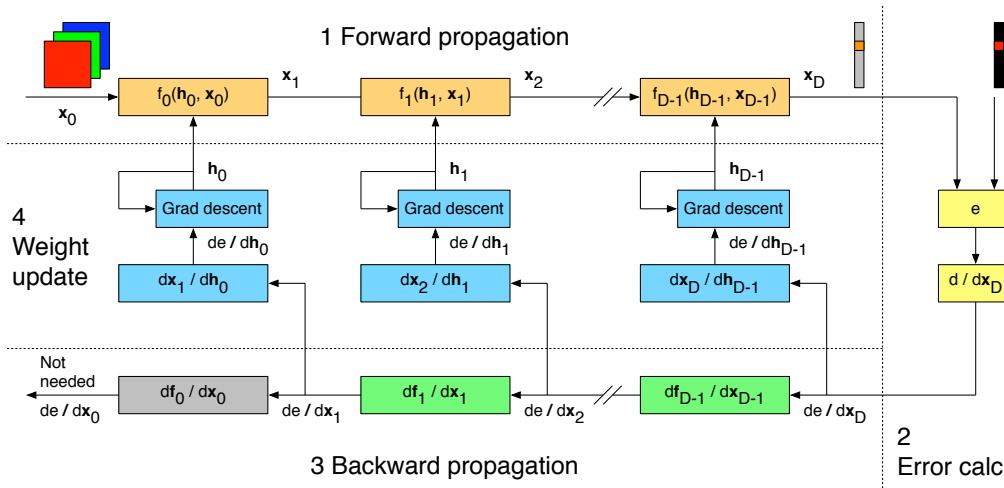
Math: Linear Algebra

- Dimensionality reduction
 - Frequently part of pre processing
 - DFTs and PCA for dimensionality reduction
 - Matrix embeddings
- Transformations from data to features to classes
 - xNNs are compositions of nonlinear functions (layers)
 - Linear transformations are key components of the nonlinear functions with learnable parameters that control the network mapping
 - Examples: densely connected, CNN style 2D convolution, RNN and attention based layers
- A laundry list of topics
 - Sets, fields, vectors, matrices, tensors, functions, vector spaces, normed vector spaces, inner product spaces, matrix vector multiplication, matrix matrix multiplication, CNN style 2D convolution, RNNs, attention, DFTs and PCA



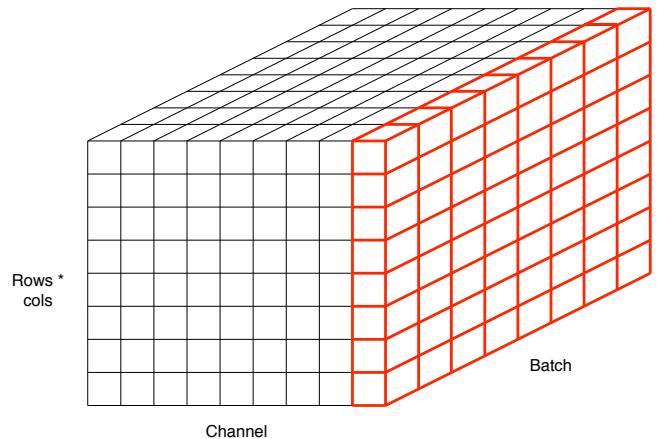
Math: Calculus

- Universal approximation
 - Proof
- Training
 - Automatic differentiation with reverse mode accumulation to back propagate error gradients
 - Gradient descent variants to update learnable parameters
- A laundry list of topics
 - Derivatives, sub derivatives, partial derivatives, gradients, Jacobians, chain rule, critical points, gradient descent, automatic differentiation with reverse mode accumulation and universal approximation



Math: Probability

- Training
 - Training vs testing data
 - Parameter initialization
 - Information flow
 - Normalization of feature maps between layers
 - Stochastic width / depth based regularization to improve generalization
 - Loss functions based on distances between pmfs
- Modeling
 - Next element prediction
 - Conditioned on inputs
- A laundry list of topics
 - Probability spaces, events, random variables, expected value, normalization, law of large numbers, central limit theorem, random processes, stationarity, time averages, ergodicity, entropy, mutual information, Kullback Leibler divergence, data processing inequality, compression, Huffman and arithmetic coding



Math: Algorithms

- Pooling
 - Common down sampling strategies include max operations
 - Max, spatial pyramid and RoI pooling
- Non maximal suppression
 - A common post processing for multiple object detection and object based image segmentation
 - Dependent on sorting
- A laundry list of topics
 - Comparison sorts, sequential merge sort, parallel merge sort, pooling, median and rank order filtering

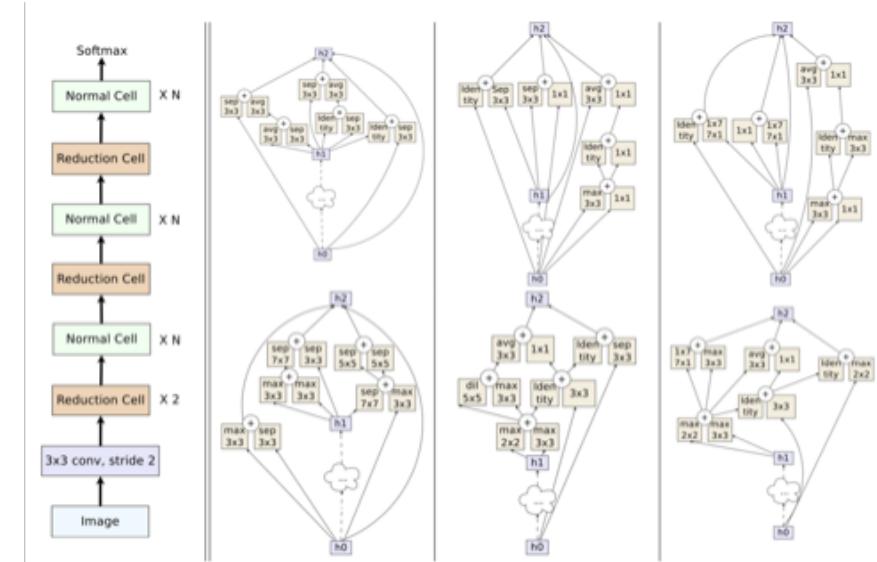
31	21	33	34	5	2	15
10	29	32	6	27	16	13
7	4	28	20	24	30	26
25	18	14	35	22	1	3
17	23	12	8	19	9	11

Max pool ↓ $3 \times 3 / 2$

33	34	30
28	35	30

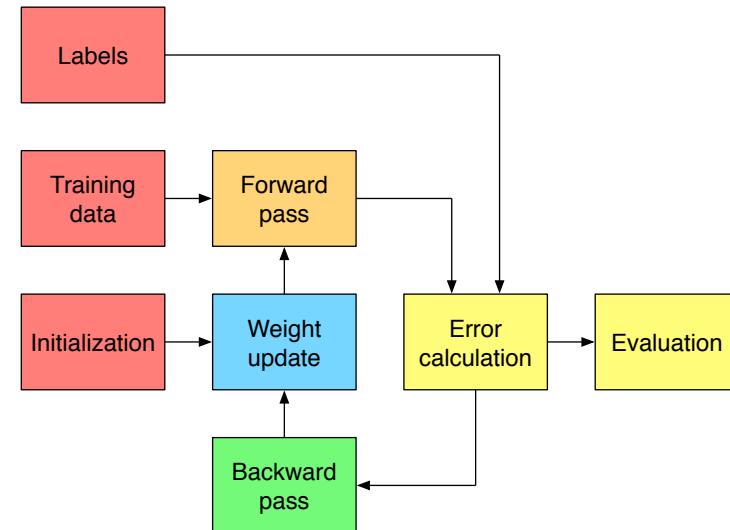
Networks: Design

- The start of the 1st spiral presentation of material
 - The math lectures are a 1st look at everything needed for network design and training
 - The network design, training and implementation lectures look at everything again in more detail
- How to design a network to achieve a goal
 - How to select and combine the layers described in the linear algebra and algorithm lectures
 - Encoder decoder style architectures
 - Lots of examples of different backbone networks
- A laundry list of topics
 - Goals, size considerations of the network, feature maps and filter coefficients, problem complexity, graph specification, layer types, tail body head decomposition, tail designs, head designs, body designs including chain, parallel, dense and residual structures, optimized architecture search and visualization

Figure from <https://arxiv.org/abs/1802.01548> 21

Networks: Training

- Training
 - Convergence speed and accuracy of result
 - Regularization to improve generalization
 - Builds on the material in the calculus and probability lectures
- A laundry list of topics
 - Supervised learning, differences with function optimization, convergence, overfitting, regularization and generalization, the curse of dimensionality, training validation testing data splits, natural data, labeling, cleaning, synthetic data, hand engineered generation, learned generation, data augmentation, random initialization, transfer learning, curriculum learning, batch normalization, group normalization, stochastic width and depth, classification and regression losses, loss surface shapes, unequal class weighting, aux network heads, multiple network heads, check pointing, reversible architectures, batch size, weight update methods including SGD, momentum, AdaGrad, RMSProp and Adam, learning a solver, weight decay, gradient noise, gradient clipping, synchronous stochastic gradient descent and hyper parameter selection



Networks: Implementation

- Bigger networks enable better performance but require more data movement and compute
 - Network modifications to improve implementations, hardware design and software to connect algorithms to hardware
- A laundry list of topics
 - Networks: theoretical complexity, precision, hardware size vs model size, training vs testing, data formats, quantization, network sizing and network simplification
 - Hardware: Moore's law, Dennard scaling, dark silicon and dark memory, power, roofline models, SoC architectures, domain specific architectures, control, memory, data movement, compression, Amdahl's law, computational basis, matrix multiplication primitive, inner, outer, Strassen style, input power of 2, sparse and analog matrix multiplication, Winograd style convolution, sort primitive, tree configurations, torus configurations and hardware design examples
 - Software: high level graph specification, high level graph transformations, static vs dynamic graphs, sessions, graph compilers, low level graphs, runtime initialization, runtime execution, software design examples, predicting performance and benchmarking

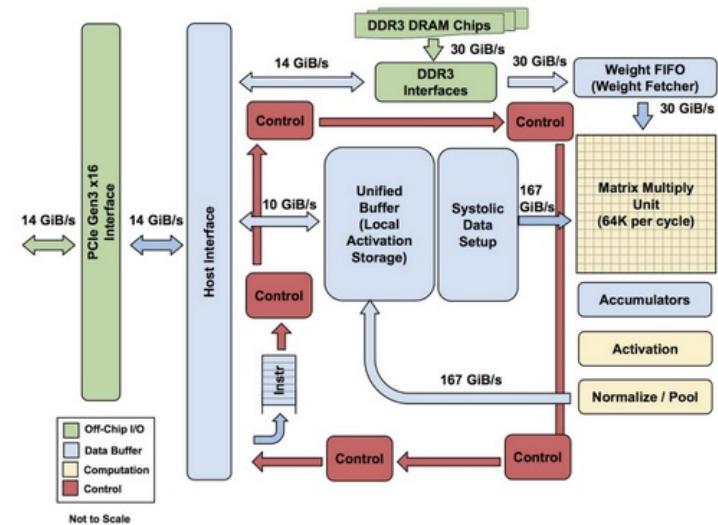
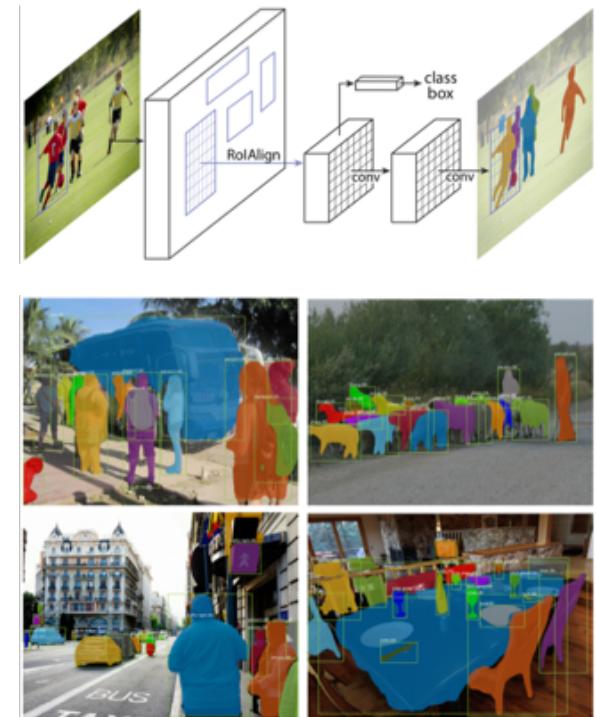


Figure from <https://www.nextplatform.com/2017/04/05/first-depth-look-googles-tpu-architecture/>

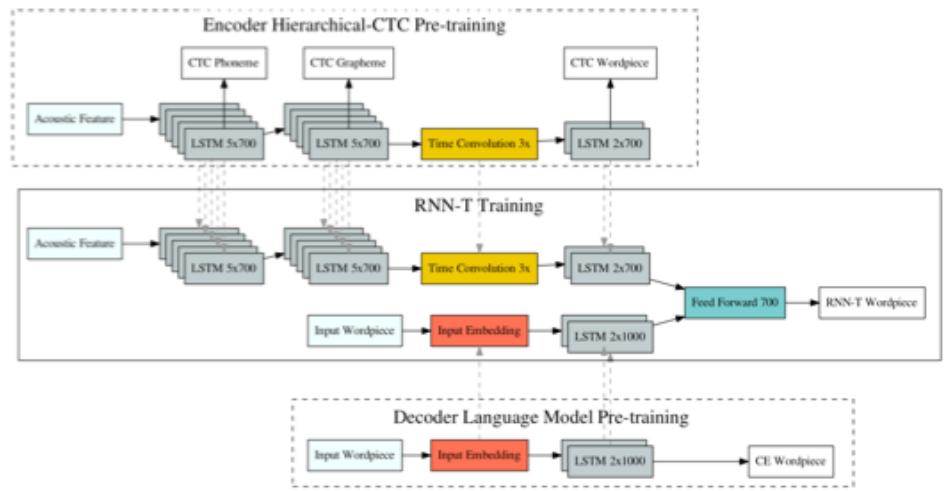
Applications: Vision

- The start of the 2nd spiral presentation of material
- The vision lectures focus on CNNs
 - Exploit spatial structure
 - Transform all sorts of vision algorithms into classification problems
 - Encoder decoder architectures with special attention paid to the spatial localization of information localization
- A laundry list of topics
 - Image capture and processing, hardware design examples, classification, pixel segmentation, up sampling, encoder decoder with skip connections, Atrous convolution, spatial pyramid pooling, 1 and 2 stage approaches to multiple object detection, feature pyramids, anchor boxes, spatial pyramid pooling, RoI pooling, region proposal networks, iterative methods, non maximal suppression, confidence threshold, intersection over union, precision, recall, precision recall curve, 2 and 3 stage approaches to object based segmentation, RoI align, depth estimation, stereo fundamentals, motion estimation and motion fundamentals



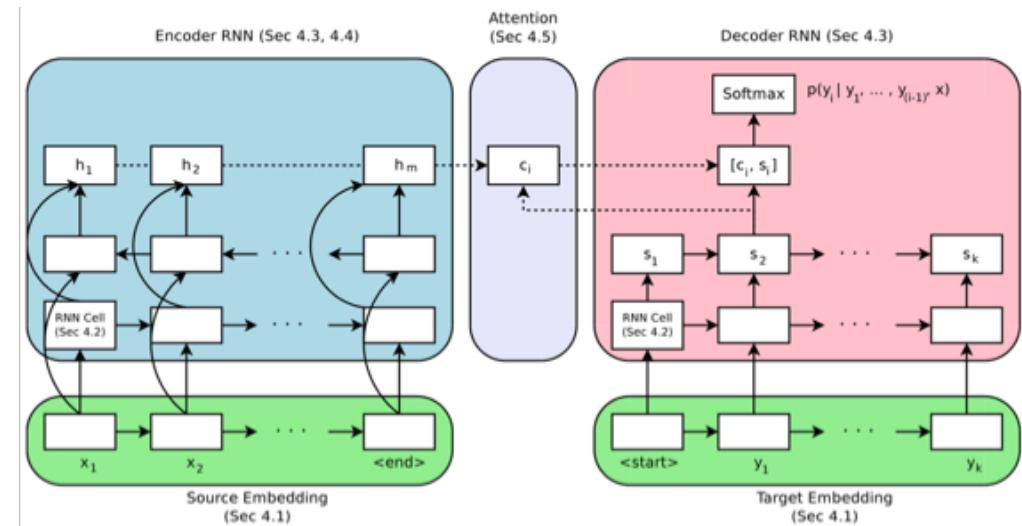
Applications: Speech

- The speech lectures focus on RNNs
 - Exploit sequential structure
 - But also include CNNs and attention
 - Encoder decoder architectures with special attention paid to the alignment of sequential information
- A laundry list of topics
 - Speech and audio signal chain, sampling, pre emphasis, windowing and spectrograms, MFCC, RNNs, GRUs, LSTMs, bi directional, pyramidal, sources of variability, speaker verification, speaker recognition, wake up, limited vocabulary speech recognition, confusion matrix, speech to text, sequence to sequence models, CTC, beam search, language model, auto segmentation, RNN transducer, attention, transition possibilities, alignments, text to speech, intermediate representation conversion and audio signal conversion



Applications: Language

- The language lectures focus on attention
 - Exploit localized information
 - But also include RNNs
 - Encoder decoder architectures with special attention paid to the localization of sequential information
- A laundry list of topics
 - Word embeddings, the distributional hypothesis, SVD based, continuous bag of words, skip grams, visualization, word similarity and analogies, task specific optimization, language modeling, N grams, neural language models, perplexity, character based, translation, sequence to sequence, greedy and beam search decoding, structured prediction, attention, architecture exploration, self attention, transformer and BLEU



Applications: Games And Art

- This is new this semester
 - So the below topics have a higher level of variance with respect to what will actually be discussed
- Games
 - A reasonable place to introduce reinforcement learning
 - Perfect information games
 - Atari and other video games
 - Go, chess and other board games
- Art
 - A reasonable place to introduce GANs

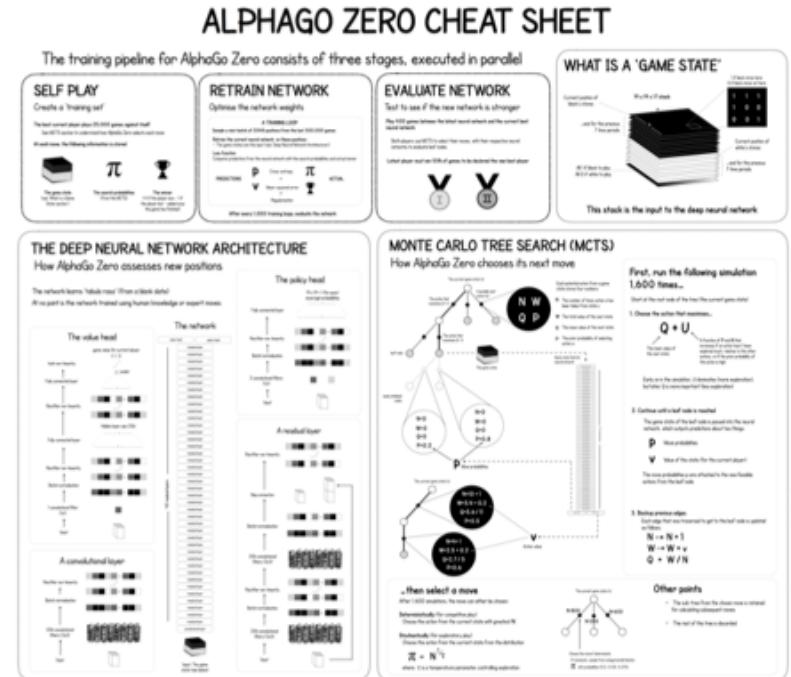


Figure from https://applied-data.science/static/main/res/alpha_go_zero_cheat_sheet.png 27

So What's Missing?

- The amount of stuff that's not included >> the amount of stuff that's included
 - Missing sub topics within the topics that are covered
 - Missing topics altogether
- 1 role of a teacher is to be a guide
 - You see what I think is most important
 - But your only hope is to learn how to learn
 - The last paper is not written

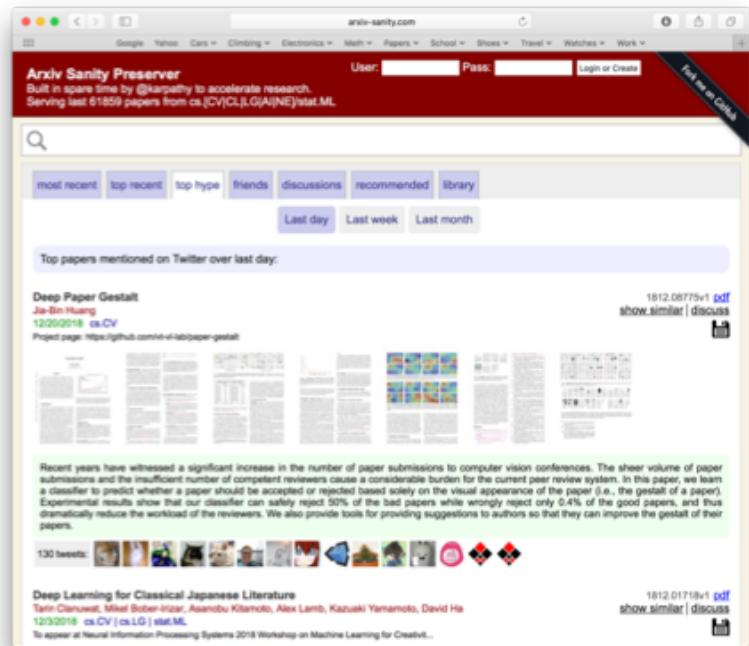


Figure from <http://www.arxiv-sanity.com/toptwtr> 28

Logistics

Unofficial Course Objections

- Understand why xNNs work
- Understand how to design, train and implement xNNs
- Understand how xNNs are applied to vision, speech, language and other applications
- Understand how to apply xNNs to new applications
- Learn how to learn

Grades

- 3 in class closed everything pencil and paper only tests at the end of each section
 - 25% math test
 - 25% networks test
 - 25% applications test
- Homework throughout the semester
 - 25% total for all assignments
- No final
- eLearning for grade distribution



This is a
change
from last
semester

Highlighting The Role Of Homework

- Homework will serve multiple purposes
 - Extending in class knowledge and learning to learn via paper reading
 - Evaluation and reinforcing of the concepts discussed in class
 - An avenue to gain familiarity with implementations
- Highlighting the implementation component of homework
 - The goal is to get everyone up and running as soon as possible with high level software libraries
 - You build skills via the relatively close reproduction of results of others
 - Which will lead to confidence in your ability to design, train, evaluate and revise networks that contain new ideas of your own
- This semester everyone will use Python / TensorFlow for the software implementation portion of their homework assignments
 - This isn't a negative reflection on other libraries (many are excellent)
 - This is a reflection on my bandwidth and need for simplification



Communication

- UT Dallas email
 - Individual emails
- eLearning
 - Full class emails
 - Discussion board
- GitHub
 - Web site <https://github.com/arthurredfern/UT-Dallas-CS-6301-CNNs>
 - Syllabus Course syllabus with approximate schedule
 - Lectures Will continually update
 - References Links to many
 - Code Will add as appropriate

Office Hours

- Typically the 1/2 hour before class each day
- Will complement with an additional 1 hour on some Fridays at the student union Starbucks (morning coffee) or food court (lunch)
 - Will confirm in class on Wed
- Will complement with a discussion board on eLearning to facilitate student to student discussions

Classroom Citizenship

- Attendance is expected
- No computers in class
- Very limited use of mobile phones
 - If it becomes annoying to me I'll restrict it
 - Don't let it become annoying to me

Expectations

Of Me

- My best every class
- My opinions
 - It's a special topics class
- That I'll speak to adults like adults
 - It's a grad class
 - I want to be precise
 - But I don't want to make things unnecessarily complicated

Of Me

- A logically laid out plan for both the whole course and individual lectures
- A willingness to modify the plan as needed
- While this is my 2nd time teaching the course and have a better idea on how things will go than the 1st time teaching ...
 - I'm also making changes to improve things which re introduces some ambiguity

Of You

- Honesty
 - In your work
 - In your interactions with other students
 - In your interactions with me
- Hard work
 - Nothing meaningful in life is easy
 - This won't be an exception
- Preparation
 - Review the lecture on your own before class
 - Listen to the lecture in class
 - Re review the lecture after class until it makes sense

Of You

- Contribute to a friendly environment
 - It's great to shine as an individual through individual accomplishments
 - It's great to shine by helping others shine
 - This is a characteristic of a leader
- Be engaged
 - Ask questions freely
 - Speak up if something is unclear
 - Correct me if I'm wrong
- Help me learn your name
 - Say it when you ask a question

Opportunities

A Suggestion

- Look for opportunities beyond this course
 - Your own research
 - Your own company
 - UT Dallas HackAI or similar events
 - ...

Questions?