Big Data in Network and Service Management: An Opportunity for Synergy

Stan Matwin, CRC

Institute for Big Data Analytics

Dalhousie University

Halifax, NS, Canada

stan@cs.dal.ca

Toni Morrison, Nobel Prize in Literature 1993
[1931-2019]

I think there's data,

and then there's information that comes from the data,

and then there's knowledge that comes from information.

And then, after knowledge, there's wisdom.

I'm interested how to get from data to wisdom.

Roadmap

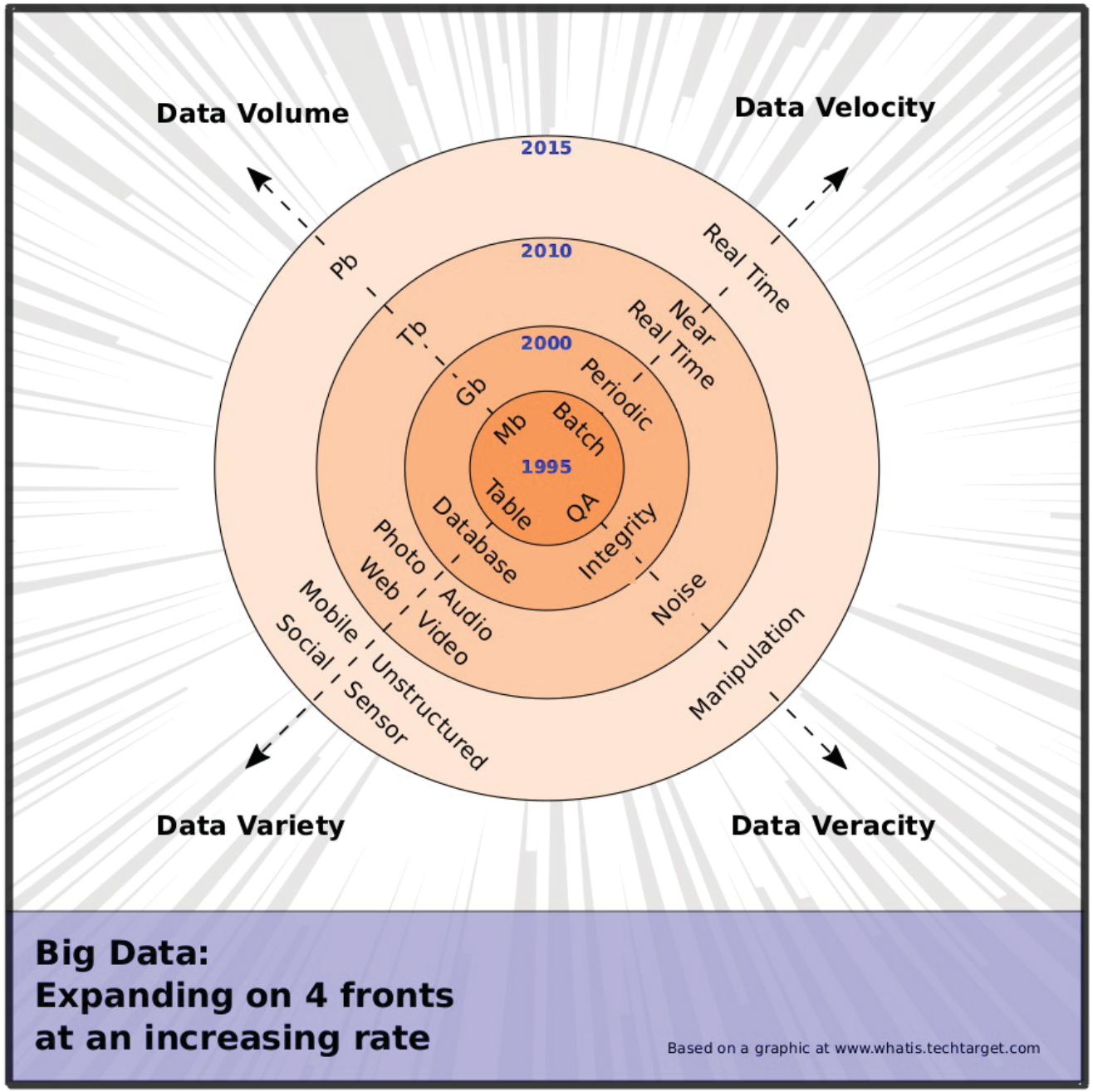
- Big Data – Birds' Eyes View
- Sample of Big Data work at Dalhousie
- Some Challenges before the Big Data field
- An outside view of the use of BD techniques in Networking
  - Traffic classification
  - QoS/QoE
  - Security
  - Data Centre mgmt
- Issues and opportunities discussion

Big Data – 5 Vs

In one minute:

- 2M Google queries
- 6M FB posts
- 100K tweets
- 1.3M video clip views
- 150 Identity theft victims
- 135 virus infections
- More than 10^{10} network-connected devices

- Volume
- Velocity
- Variety
- Veracity
- Value


Data Volume
Data Velocity
2015
2010
2000
1995
Pb
Tb
Gb
Mb
Real Time
Near Real Time
Periodic
Batch
Table
QA
Database
Integrity
Photo
Audio
Web
Video
Noise
Mobile
Unstructured
Social
Sensor
Manipulation
Data Variety
Data Veracity
Big Data:
Expanding on 4 fronts
at an increasing rate
Based on a graphic at www.whatis.techtarget.com

Deep learning (2010-…)

- “Three Musketeers”



ImageNet Classification with Deep Convolutional Neural Networks

Alex Krizhevsky University of Toronto kriz@cs.utoronto.ca
Ilya Sutskever University of Toronto ilya@cs.utoronto.ca
Geoffrey E. Hinton University of Toronto hinton@cs.utoronto.ca

Abstract

PROC. OF THE IEEE, NOVEMBER 1998

Gradient-Based Learning Applied to Document Recognition

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner

Abstract—
Multilayer Neural Networks trained with the backpropagation algorithm constitute the best example of a successful Gradient-Based Learning technique. Given an appropriate

I. Introduction

Over the last several years, machine learning techniques, particularly when applied to neural networks, have played

- promise of representation learning
  no more feature engg
- 2012 ImageNet dataset success: 72% → 85%
- contextual representations

Journal of Machine Learning Research 3 (2003) 1137–1155 Submitted 4/02; Published 2/03

A Neural Probabilistic Language Model

Yoshua Bengio BENGIOY@IRO.UMONTREAL.CA
Réjean Ducharme DUCHARME@IRO.UMONTREAL.CA
Pascal Vincent VINCENTP@IRO.UMONTREAL.CA
Christian Jauvin JAUVINC@IRO.UMONTREAL.CA
Département d'Informatique et Recherche Opérationnelle
Centre de Recherche Mathématiques
Université de Montréal, Montréal, Québec, Canada

Editors: Jaz Kandola, Thomas Hofmann, Tomaso Poggio and John Shawe-Taylor

Abstract

A goal of statistical language modeling is to learn the joint probability function of sequences of

Deep Learning toolbox

- Conv nets
- Embeddings
- Denoising autoencoders
- Transfer learning
- Generative Adversarial Networks
- tSNE
- …

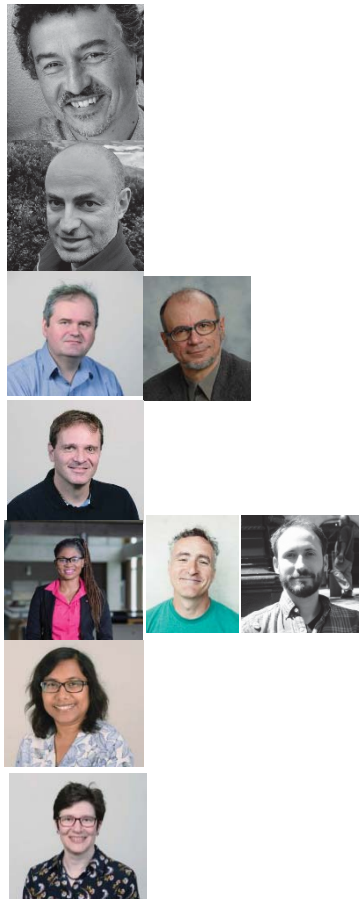
architecture engineering

Some challenges before the Big Data field

- Interpretability/transparency (data and algorithms)
- correlation/causality
- anytime algorithms
- standards
- need for [quality] data

Big Data at Institute for Big Data Analytics @ Dal

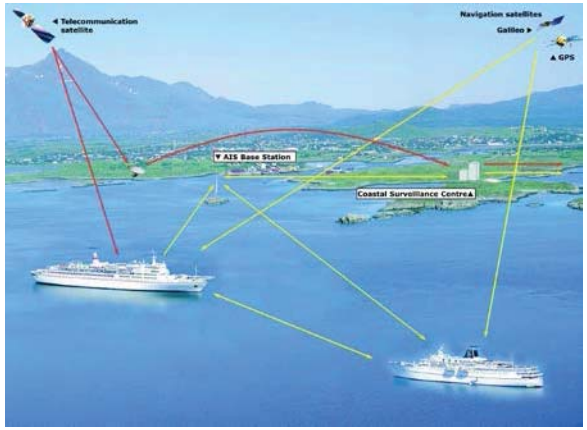
- Machine Learning [Torgo, Matwin]
- Deep Learning [Oore]
- Text/Web Analytics [Keslej, Milios, Matwin]
- Visualization [Paulovich]
- HCI [Orji, Reilly, Malloch]
- IoT [Haque]
- Applications [all of the above+ Nur ZH]

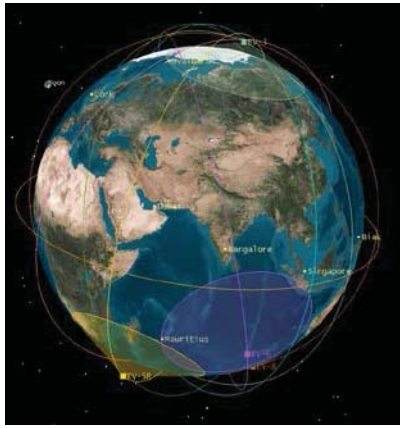


OCEAN DATA

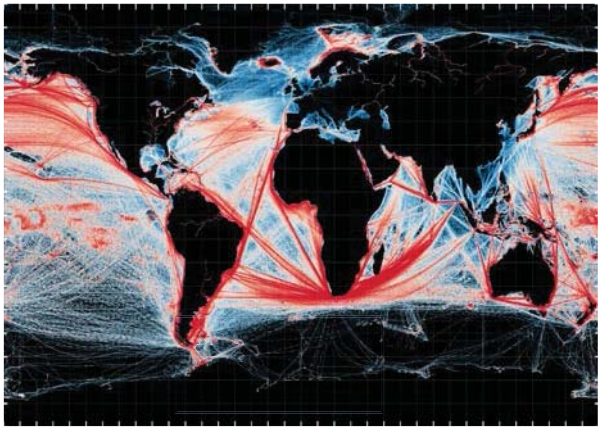
Big Data at Dal: Automatic Identification System (AIS)

IMO/ITU standard





Courtesy of ExactEarth, Inc.

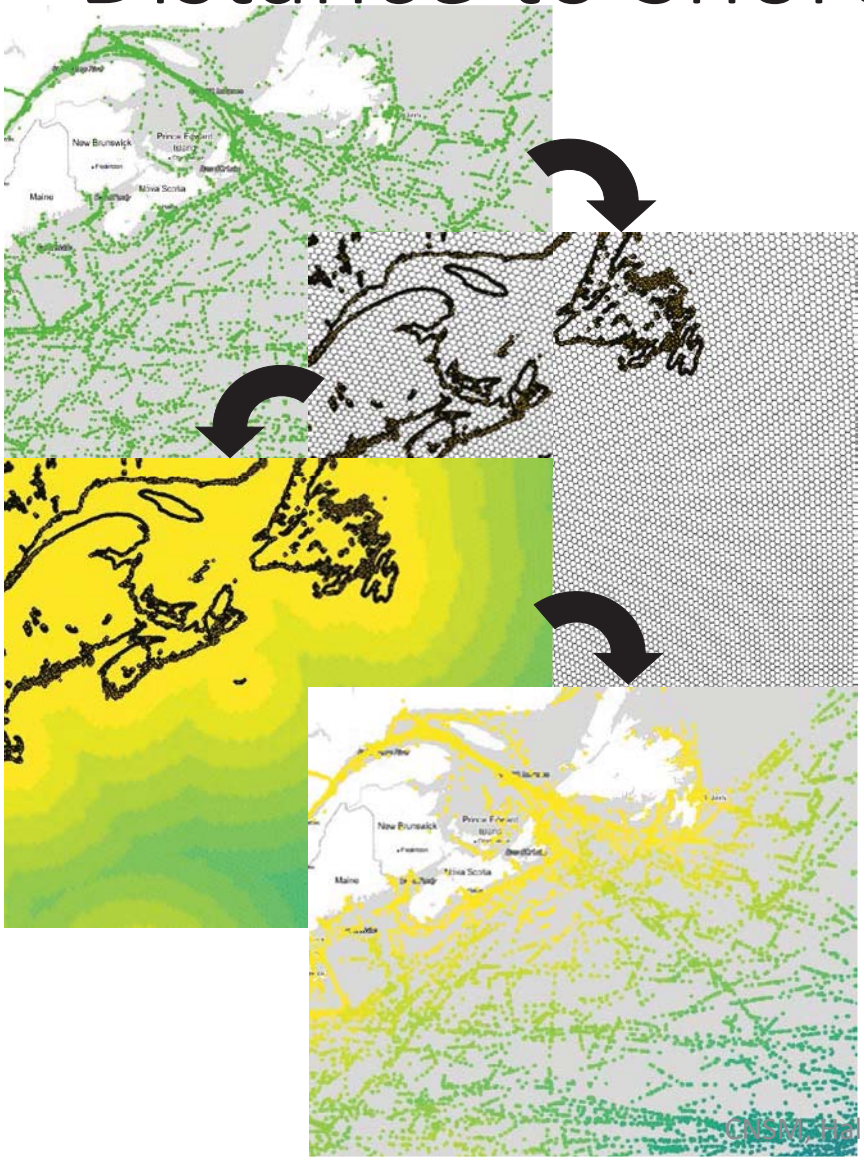


Institute for Big Data Analytics

400,000 ships
At least 100M records/day

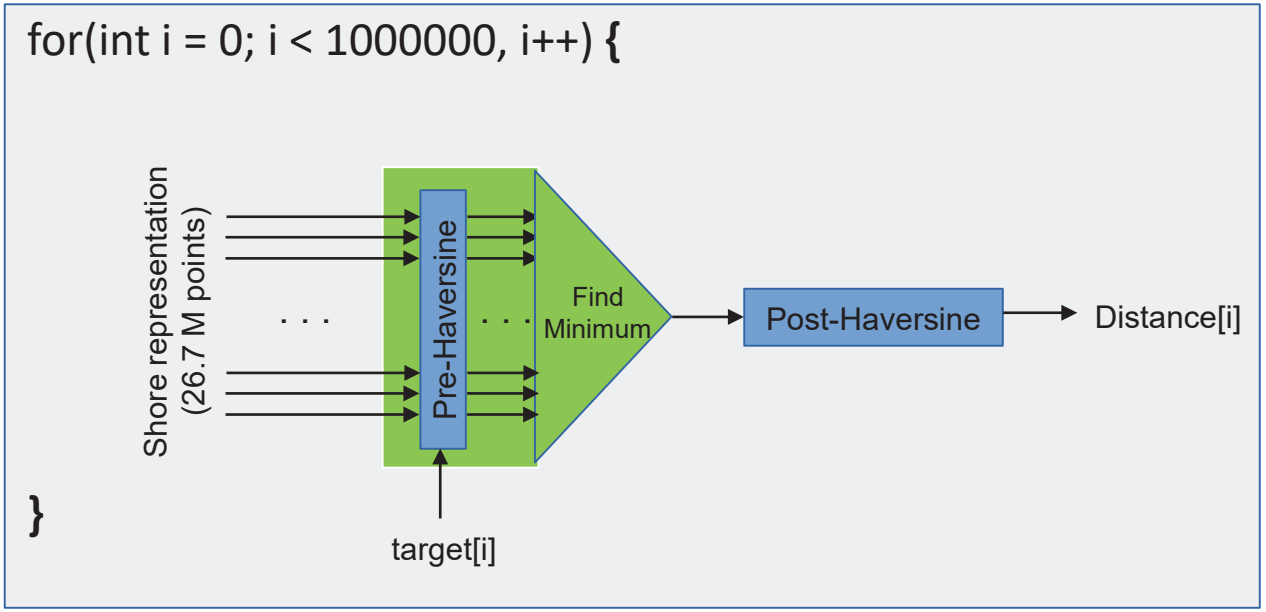
From weak to big signal

Distance to Shore Calculation



- S-AIS vessel data enrichment
- Naive (GIS) approach, on S-AIS dataset infeasible (10^9) = years of runtime
- Revised approach:
  - Calculate distance values between shore and "cells"
  - PostGIS
- Runtime: ~0.5 day for ~1M cells, but:
  - Database approach used is **not scalable further**
  - Accurate only to cell diameter (22km +/-11km)
  - Ideally distance should be calculated to individual AIS vessel positions reports directly

CUDA Implementation



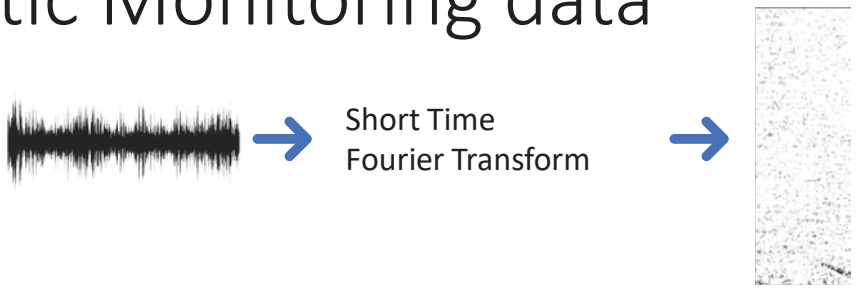
Implementation	Time for 1M targets
Numpy	**17 days**
C (OpenMP)	**2.5 days**
CUDA	**15 minutes**

Core i7-7700K
16 GB Main Memory
NVIDIA GTX 1080 Ti

- Architecture not subject to scalability issues previously encountered
- Greatly improved per-target runtime
  - Direct distance calculation on entire AIS dataset now feasible
- Distance values for 10^9 Points calculable in ~10 days
- Further gains sought through tuning of CUDA kernel size and memory streaming

Big Data at Dal: Machine Learning from Passive Acoustic Monitoring data

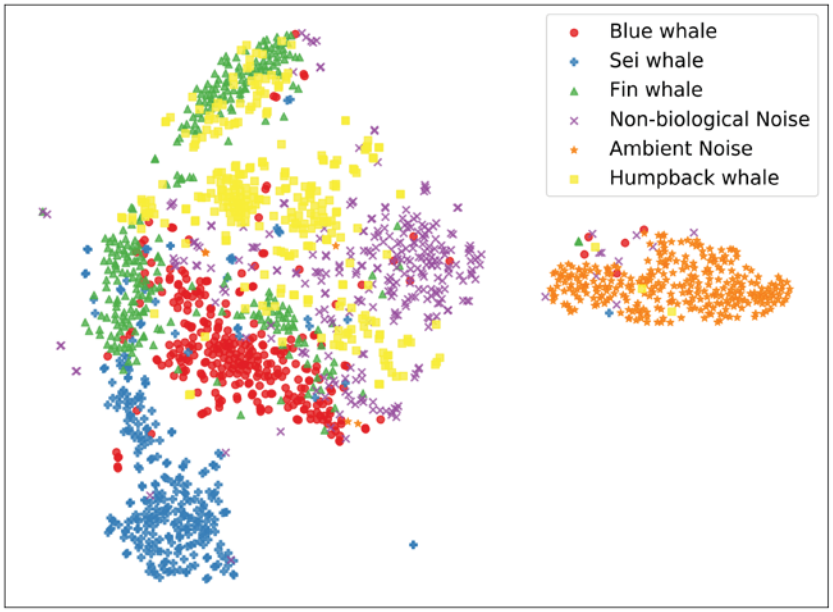
Marine mammal species detection and classification



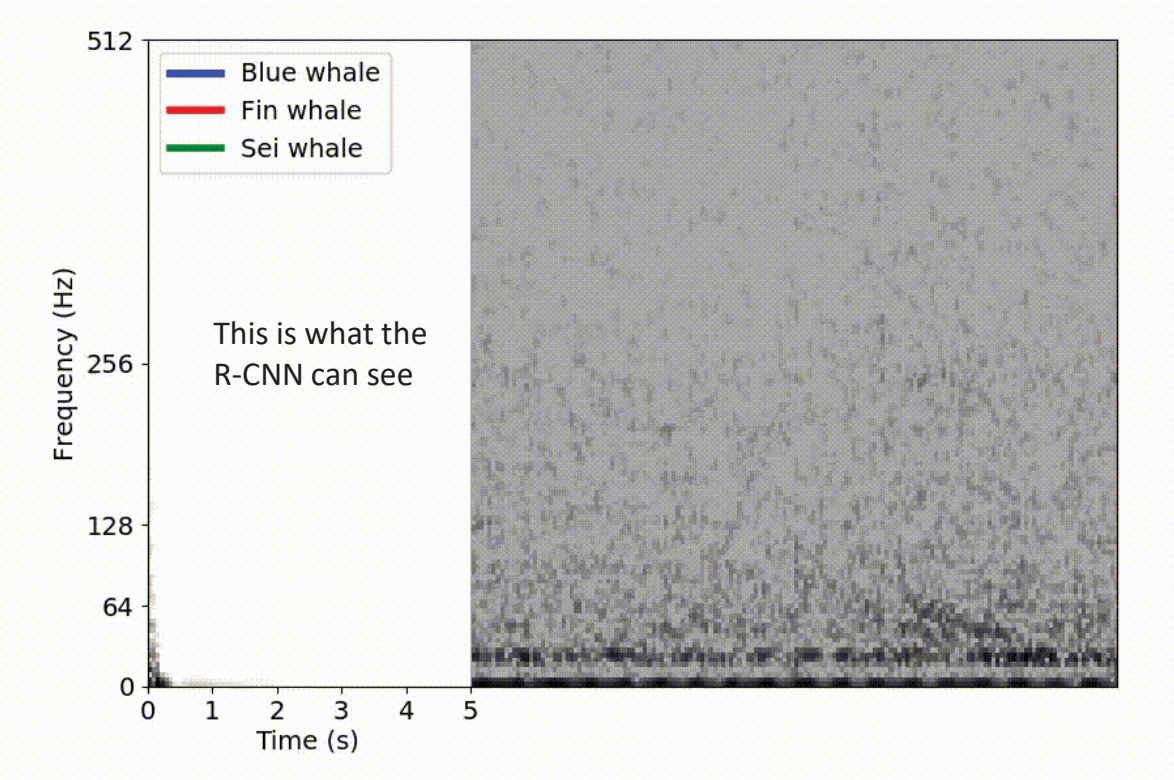
- Reframe the problem of detecting whale calls from an auditory task to a visual task: train a Convolutional Neural Network (CNN)

- Identify vocalizations from other species via transfer learning

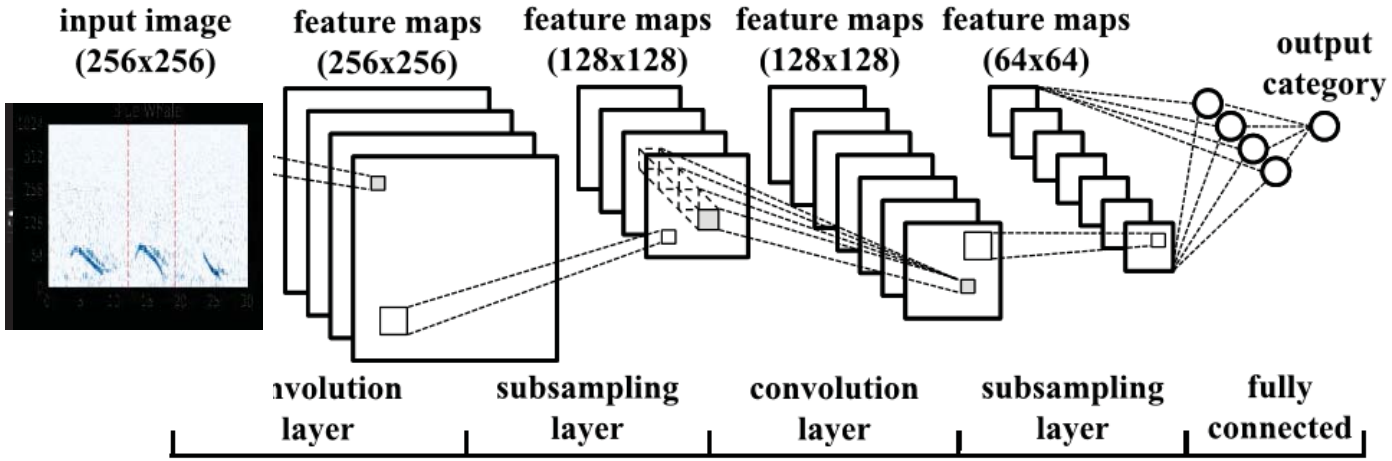
- No need to re-train the entire CNN!



Thomas, M., Martin, B., Kowarski, K., Gaudet, B., & Matwin, S. (2019). Marine Mammal Species Classificati using Convolutional Neural Networks and a Novel Acoustic Representation. *ECML-PKDD 2019*



Thomas, M., Martin, B., and Matwin., S. (2019) Detecting Endangered Baleen Whales within Acoustic Recordings using Region-based Convolutional Neural Networks. *Joint Workshop on AI for Social Good at the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*



- We did not have sufficient data to train a CNN to recognize humpback whales

  We have applied transfer learning to the CNN, obtained good results

Big Data in Networking

- Good fit, because there's MASSIVE amounts of data in all aspects of networking
- BUT [Boutaba et al. 18]:
  - networks (eg enterprise) differ a lot
  - change continuously
- Easier with Software-defined Networks
  - easier data collection
  - easier to apply resulting control actins on legacy networks

A brief look at….

- Payload-based traffic classification
- QoS/QoE
- IDS/ISP
- Data center mgmt.

Payload-based traffic classification

- many applications of different techniques on different data sets
- Lots of ingenious feature engineering
  - Bag of Flow [Zhang et al 13]
- Good results often obtained with the use of
  - K-NN
  - Random Forests and Boosting
  - SVMs
- Some methodological questions?

QoS/QoE

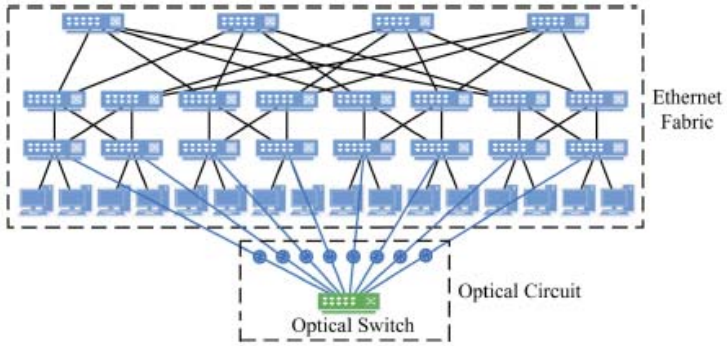
- Mapping of network flow characteristics (delay, jitter, loss ratio,…) and Mean Opinion Scores by the user
- User-labeled data: always limited
  - Use of GANs?
- Possible inspiration from internet marketing (user experience visiting a web portal)
  - Data privacy issues

Security/anomaly detection

- IDS/IPS
- Progress from using KDD 99 challenge dataset
  - Classifying network traffic into five categories of attacks
- Limitation of the classification approaches
- Clustering-based methods – unsupervised anomaly detection
- Flow-based *vs* payload-based approaches

Data Center Management with ML [Salman et al. 18]

- Data Centers - a key internet component



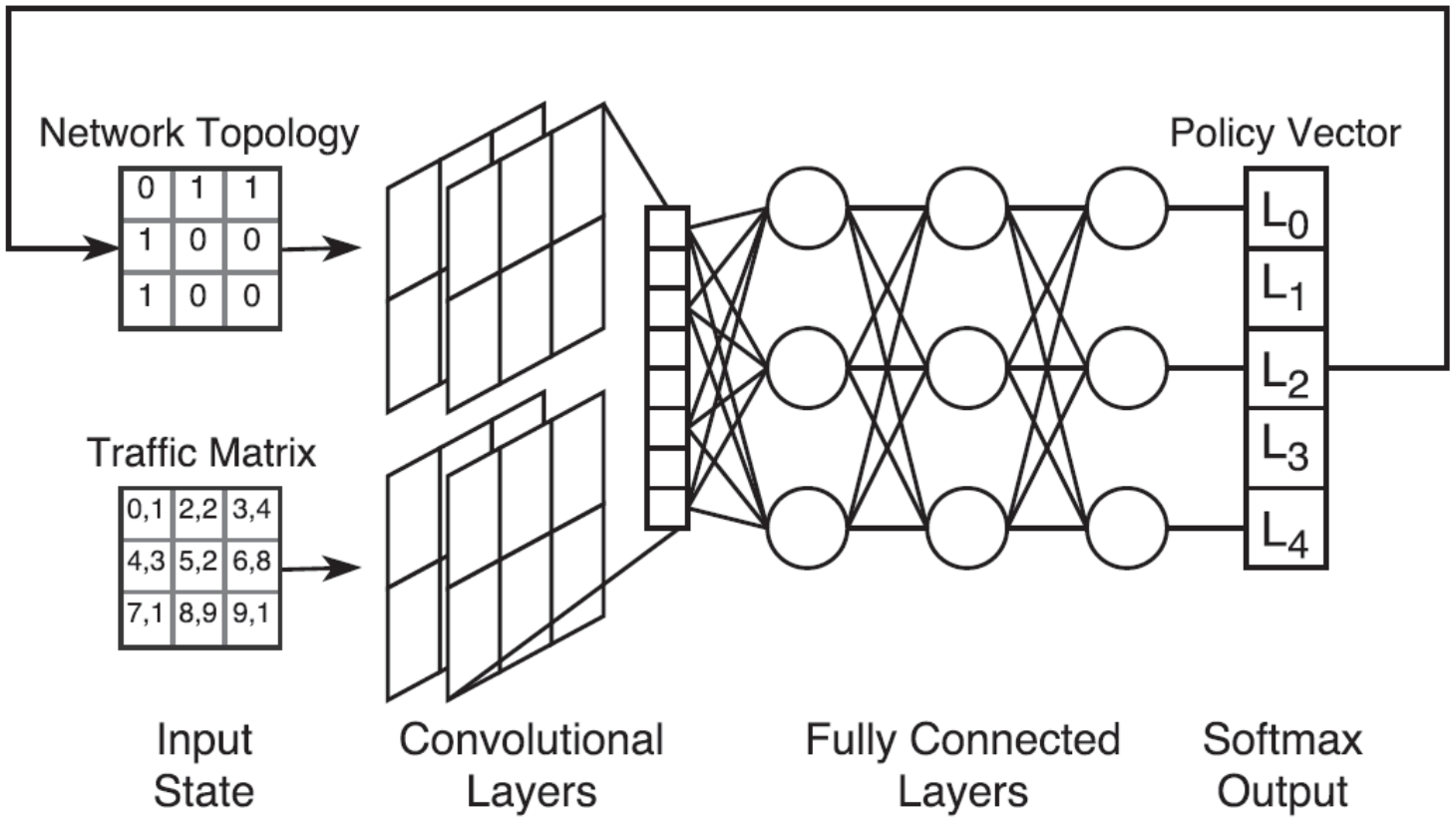
From [Salman et al 18]

- Typical optimization:
  - gather performance data
  - Run a linear programming algorithm finding a good solution
  - Take action
    - augmenting flexible links,
    - turning off links,
    - moving traffic
  - on subset of the network

reinforcement learning

- Several DL agents for different tasks:
  - Traffic engg
  - energy savings
  - …
- Each runs on top of an SDN

reward function:
maximize link utilization
minimize flow-completion time



Opportunities

- “Spectrogramming” and CNNs
- Embeddings
- Training a representation, and then
- Transfer learning?
- Semi-supervised learning and Distillation?
- Simple *vs* complex methods*?*
  - *Naïve Bayesian models?*
- Lessons from computational advertising

Some general remarks on ML in networking research

- efficiency of the learned models?
  - Are they efficient enough to be embedded in production systems?
- Combined evaluation/utility measure involving decision time?
- Lack of **standardized benchmark datasets**

Discussion …