E6893 Big Data Analytics Lecture 7:

Streaming *Big Data Analytics*

Ching-Yung Lin, Ph.D.
Adjunct Professor, Dept. of Electrical Engineering and Computer Science

October 18th, 2019
Stream Analyses on Social Media Data
Social Media Stream Monitoring

Goal 1: Detect, classify, measure and track the
(a) formation, development, and spread of ideas & concepts (memes)
(b) purposeful or deceptive messaging and misinformation

Goal 2: Recognize persuasion campaign structures and influence operations across social media sites and communities

Goal 3: Identify participants and intent, and measure effects of persuasion campaigns

Goal 4: Counter messaging of detected adversary influence operations

53+ papers published, accepted, & submitted
12+ patents filed
ACM CIKM 2012 Best Paper Award
IEEE BigData 2013 Best Paper Award
PNAS Cover Article Jan 2013
Science (1)
Nature (2)

Approach: Modeling, Tracking and Affecting Information Dissemination in Context

Thrust 1. Modeling Information Dissemination in Context:
Models of Trust and Social Capital, Information Morphing, Persuasiveness and Competition of Memes, Dynamics of Social Influence

Thrust 2. Detecting and Tracking Information Dissemination in Context:
Detecting Malicious Info Propagation, Evolution History and Authenticity of Multimedia Memes,

Thrust 3. Affecting Information Dissemination in Context:
Automated Generation of Counter Messaging, Influencing the Outcome of Competing Memes and Counter Messaging
Social Media Solution

**Live Monitoring**
Monitoring real-time tweets on keyword:  
Monitor live tweets »

**Trend Monitoring**
Analyzing trend of conversations based on hashtags  
View trends »

**Multimedia Monitoring**
Recognizing visual content and analyzing visual sentiments  
View multimedia »

**Geo Monitoring**
Monitoring the places that people are sending out tweets  
View places »

**Scope Identification**
Define user-specified sets of keywords for monitoring and analytics  
Define scopes »

**Concept Analytics**
Analyzing statistics of groups based on time, topics, etc  
Concept searches »

**Link Exploration**
Visualizing relationships, discussion sequences and graphs  
View relationships »

**Impact Prediction**
Analyzing conversations and predicting their impact to business  
View impacts »

**Story Detection**
Detecting live developing stories on social media and their evolution  
View stories »

**Person Analytics**
Analyzing a person’s personality, trustworthiness, etc.  
View person »

**Target Discovery**
Inspecting potential users for bot detection, marketing, or influencing  
Inspect targets »

**Forensic Analytics**
Analyzing retweet sequences and displaying anomalies  
View anomalies »
Social Media Stream Monitoring Applications

- **Live Monitoring**: Monitoring real-time tweets on keyword.
  - Monitor live tweets
  - View trends
  - View multimedia

- **Trend Monitoring**: Analyzing trend of conversations based on hashtags.
  - View trends

- **Multimedia Monitoring**: Recognizing visual content and analyzing visual sentiments.
  - View multimedia
  - View places

- **Geo Monitoring**: Monitoring the places that people are sending out tweets.
  - View places
  - Define scopes
  - Concept searches

- **Scope Identification**: Define user-specified sets of keywords for monitoring and analytics.
  - Define scopes
  - Concept searches

- **Concept Analytics**: Analyzing statistics of groups based on time.
  - Concept searches

- **Link Exploration**: Visualizing relationships, discussion sequences and graphs.
  - View relationships

- **Impact Prediction**: Analyzing conversations and predicting their impact to business.
  - View impacts

- **Story Detection**: Detecting live developing stories on social media and their evolution.
  - View stories
  - View person

- **Person Analytics**: Analyzing a person’s personality, trustworthiness, etc.
  - View person
  - Inspect targets

- **Target Discovery**: Inspecting potential users for bot detection, marketing, or influencing.
  - Inspect targets
  - View anomalies

- **Forensic Analytics**: Analyzing retweet sequences and displaying anomalies.
  - View anomalies

Images in tweets that belong to one story

Text of the newest tweet in this story
Bank Use Case

- **Objective:** Detect unexpected social media movements that may impact a major bank’s business

- **X-Bank:**
  - Major bank in Spain

- **Client needs:**
  - Monitor Catalan independence movement: independence may bring bankruptcy since X-Bank needs ECB support
  - Detect potential PR crisis by analyzing the formation and spreading of grassroots opinion on their employees and services

- **Challenges:**
  - Existing social media monitoring tools miss important tweets that don’t contain specified keywords and are not from specified users
  - Existing tools lack of predictive capability of tweets’ potential influence

An image tweet (without mentioning “the bank name”) sparks a lot critiques of their unfair practice
City Government Use Case

- **Objective:** Detect unexpected social media discussions that may impact a government’s campaign

- **Challenges:**
  - Detecting emerging trends as well as the powerful multimedia memes
  - Existing tools lack of predictive capability of tweets’ potential influence

- **Client needs:**
  - Foster society diversity and fusion: Need to understand the citizens’ responses to a government policy on social welfare on minority
  - Detect growing topic trends and multimedia discussions around the policy

- **X-Government:**
  - A government in Asia
Live Monitoring
Live Monitoring

- **Monitoring categories**
- **Monitoring filter**
- **Real-Time Translation, Locations, Top Retweets**
- **Growing Influential Retweet Graphs**
- **Live Tweets, Sentiment, Keywords**

Total Tweets: 231
Positive: 33 (15%)
Negative: 31 (13%)

Select CIO Category (les): EXEC, BLADE, RENTEN, TIM, Security, MATRIX, CIO

Translation: RT @Lion_King_Bhr: The traitors in Bahrain Safavid attack on public utilities and security men, 2/19/2013 "#Bahrain #Egypt #Syria #KSA #UAE #News http://t.co/1dRana 137
And One Way Or Another is also number 1 in Guatemala, Peru, Israel, Brazil, Egypt and Panama! OMG

Translation: RT @Lion_King_Bhr 44,12052 --> @SaloomButtila
19/2/2013 "#Syria..." http://t.co/1dRana 137

Translation: @Depoland 48,230 --> @Depoland 48
19/2/2013 "#Egypt..." http://t.co/1dRana 137

Translation: @Depoland 48,230 --> @Depoland 48
19/2/2013 "#Syria..." http://t.co/1dRana 137

Translation: @Depoland 48,230 --> @Depoland 48
19/2/2013 "#Egypt..." http://t.co/1dRana 137
Trend Monitoring
Trend Monitoring

Real-time Trend Analytics

The chart shows top 4 hashtags for 'isis' topic during February 27, 2015 to March 2, 2015
(285667 tweets for hashtag iraq loaded. 236532 tweets for hashtag isis loaded. 199351
tweets for hashtag eu loaded. 198716 tweets for hashtag sunnairaqqciviliansvoice loaded.)

Click on the histogram below to see raw tweets.

Showing the tweets for iraq on 2015/3/2 16:00

@giddey_up

@thetimes: #Iraq troops begin major offensive to drive #ISIS out of Tikrit http://t.co/Xyjy7x8iFm
http://t.co/4UJxjSZuU" Gogetm!
-- 2015-03-02T22:44:46Z

@rashidrashidi5

RT @paige_antoine: 'Mullahs' regime Deputy President, Foreign Minister visit #Iraq to skirt sanctions and transfer huge
http://t.co/nrwtJQ...
-- 2015-03-02T22:44:47Z

@Befree

RT @AlanTonelson: BREAKING: #Obama interviewed by @Reuters on #Netanyahu, #Iraq, #nukes: http://t.co/OZ8Ne5RhGM
-- 2015-03-02T22:44:48Z

@ChrisAnoya

RT @Matt_VanDyke: Brave soldiers of the Nineveh Plain Protection Units (#NPU) training to defend Christians against #ISIS in #Iraq.
http://...
-- 2015-03-02T22:45:29Z

@proctor_hassie
Multimedia Monitoring
To be integrated soon — Automatic Affective Comment

VAC Prediction

“wonderful,” “lovely,”
“peaceful,” “moody,”...

Candidate Comments

“Awesome and cute!”
“Lovely moody shot -
so peaceful”

PACs

C

Automatic Commenting

lovely moody shot - so peaceful!
What and Where of Image Sentiment Analysis

What’s in the image that makes a dog cute vs. scary?

What makes a beautiful landscape beautiful?
Detecting Politically Persuasive Web Content

Summary

- Extract Affective and Semantic Information
  - Audio, Visual and Textual Modalities
- Detection of Politically Persuasive Content
  - Additionally predict viewer response in advance
Scope Identification
Scope Discovery

IBM System G Social Media Solution

Select - User Initiative -- Common_Core_01

Topics and terms from **Common_Core_01**

<table>
<thead>
<tr>
<th>School</th>
<th>school x</th>
<th>high school x</th>
<th>middle school x</th>
<th>elementary school x</th>
<th>class x</th>
<th>gradeschool x</th>
<th>grade schoo. x</th>
</tr>
</thead>
<tbody>
<tr>
<td>Common Core Tes</td>
<td>educational standards x</td>
<td>state standards x</td>
<td>state testing x</td>
<td>standardized testing x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Professional Devel</td>
<td>Training x</td>
<td>curriculum materials x</td>
<td>curriculum development x</td>
<td>professional development x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teacher</td>
<td>Teacher x</td>
<td>Lecturer x</td>
<td>Tutor x</td>
<td>Instructor x</td>
<td>teach person x</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Add a topic and terms to **Common_Core_01**

<table>
<thead>
<tr>
<th>Topic Name</th>
<th>Add</th>
</tr>
</thead>
</table>

**Advance Settings for Segment Analytics**

Crystal Ball Graph Tool for Keyword Terms Exploration
Initiative: ISIS_07

Foci

- ISIS_Name_Occurrences: 21426
- ISIS_In_Iraq_or_Syria: 9734
- ISIS_Marriage_Interest: 108
- ISIS_Extreme_Military_Action: 266
- ISIS_Military_Action: 833
- Syrian_Opposition_Groups: 15162
- ISIS_Social_Media_Recruiting: 1727
- Islamic_Religion_Interest: 16895
- Western_Political_Response_Interest: 3078
- ISIS_Travel: 8969

Map

# of profiles in states

- Parent: 50%
- Not Parent: 50%
- Male: 28%
- Female: 45%
- Unknown: 27%
- Business Owner: 96%
- Not Business Owner: 98%
- Traveler: 98%
- Not Traveler: 98%

State: Ohio: 1682
Impact Prediction
# Impact Prediction

## Real-time hashtag monitoring

Predicting the business impact of tweet messages grouped by hashtags.

Please click on the "hashtags" to learn more about each conversations content.

Last updated at 2015-03-03 03:20:01 GMT

<table>
<thead>
<tr>
<th>#</th>
<th>Conversations</th>
<th>Impact</th>
<th>Impact Score</th>
<th>Prediction Detail</th>
<th>First Tweet Time</th>
<th>Last Tweet Time</th>
<th>Duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>russia, turkey</td>
<td>HIGH</td>
<td>52.93</td>
<td>URL</td>
<td>2015-02-18 17:36:59</td>
<td>2015-03-01 09:10:57</td>
<td>265 hours</td>
</tr>
<tr>
<td>2</td>
<td>isis, syria</td>
<td>HIGH</td>
<td>44.66</td>
<td>URL</td>
<td>2015-02-20 14:28:51</td>
<td>2015-03-01 08:58:20</td>
<td>210 hours</td>
</tr>
<tr>
<td>3</td>
<td>isis, syria, turkey</td>
<td>HIGH</td>
<td>42.1</td>
<td>URL</td>
<td>2015-02-20 07:12:33</td>
<td>2015-03-01 08:58:20</td>
<td>217 hours</td>
</tr>
<tr>
<td>4</td>
<td>syria, turkey</td>
<td>HIGH</td>
<td>40.77</td>
<td>URL</td>
<td>2015-02-27 14:31:05</td>
<td>2015-03-01 08:13:34</td>
<td>42 hours</td>
</tr>
<tr>
<td>5</td>
<td>erdogan</td>
<td>HIGH</td>
<td>39.92</td>
<td>URL</td>
<td>2015-02-27 23:35:06</td>
<td>2015-03-01 09:12:05</td>
<td>33 hours</td>
</tr>
<tr>
<td>6</td>
<td>syria, turkey, us</td>
<td>HIGH</td>
<td>38.28</td>
<td>URL</td>
<td>2015-02-19 02:45:53</td>
<td>2015-02-28 11:50:52</td>
<td>225 hours</td>
</tr>
<tr>
<td>7</td>
<td>erdogan, turkey</td>
<td>HIGH</td>
<td>36.89</td>
<td>URL</td>
<td>2015-02-24 02:16:10</td>
<td>2015-03-01 09:12:05</td>
<td>126 hours</td>
</tr>
</tbody>
</table>
Flow Analytics
Flow Analytics

Tracking Information Flow Propagation
Collective intelligence and predicting market movement

Data:
- instant messages and trades by employees of a large hedge fund
- 24 traders, 95 analysts, 63 portfolio managers, 8646 outside contacts
- 47K trades
- 12 million IMs (2010–2011)

Findings: We identify two behavioral patterns that correlate with changes in the market:
- Reaction to IMs containing relevant information
- Attention to people who work with the same/different stocks

Using these two signals, we can make predictions of the market movement with better accuracy than the hedge fund did.
13:11:33, I was thinking all this AAPL anti-trust might be actionable 
13:11:42', not great for AAPL 
13:11:47, When GOOG had that big issue in Europe stock underperformed right? 
13:11:52, true 
13:14:01, Also not sure if you caught, but GSCO is going to allow employees to bring own phone device for corporate email 
13:14:24, Maybe GSCO allowing that could be positive for AAPL, as security focused firm saying iPhone works 
13:14:35, But bad for RIMM 
13:14:42, Maybe all this is priced in 
13:16:50, Did you see speculation that Bing is actually quietly going to be default search on iPhone 4? 
13:17:18, heard a lot of talk of that 
13:17:23, but didn't see that specifically like that 
13:17:44, Okay let me figure out where I saw that and get back to you 
13:17:45, One sec 
Relay: How urgently do employees relay information?

\[ T_2 - T_1 < h \text{ hours} \]

- “I think AAPL is going up”
- “Pay attention to AAPL”
- “I heard AAPL is doing well”

\[ R_h(t) = \text{fraction of times users send an IM containing a stock symbol to another user within at most } h \text{ hours after having received an IM containing the stock symbol} \]
Concentration: Do employees focus on others who discuss the same stock symbols?

\[ C_{(A,B)}(t) = \text{fraction of messages between } A \text{ and } B \text{ and all messages of } A \text{ and } B \]

In example: \( 10/(20+10+5) = 0.29 \)

\[ C_k(t) = \text{average } C_{(A,B)}(t) \text{ among all pairs } (A, B) \text{ that share } k \text{ stock symbols} \]
• Motivation
Emotional states can effect how information is processed. Good information can be undermined or strengthened by emotional states.

• Approach
– Measure emotional activation in tweets using the ANEW dictionary
– Control for the number of words in the text

• Preliminary experiments
– Traders at a hedge fund are more likely to make decision errors when they are very emotionally activated or very emotionally deactivated.
– Users who retweeted the 20 detected most anomalous sequences tend to post tweets with higher level of emotion than a baseline of 20 million tweets from June 2009.
Finding optimal edges to slow down propagation

- **Task:** Minimize propagation by deleting edges of a graph

- **Problem:** Given a graph $G$ and a budget $k$, find the $k$ edges to delete in order to get the largest drop in the leading eigenvalue $\lambda_1$ of $G$

Yahoo IM!

- **Red edges** are selected for deletion
- *NetMelt* [Tong et al., CIKM 2012] tracks one eigenvalue, $\lambda_1$
- Our *MET* tracks on average 5.17 eigenvalues
Measuring Human Essential Traits in Social Media

- **Personality:** Mapping personal/organizational social media postings to scores of BIG 5 Personality (*Openness, Conscientiousness, Extraversion, Agreeableness, and Neurocism)*

- **Needs:** Mapping personal/organizational social media postings to scores of *Harmony, Curiosity, Self-expression, Ideal, Excitement, and Closeness.*

- **Values:** Mapping personal/organizational social media postings to scores of *Self-Enhance, Conservation, Open-to-Change, Hedonism, and Self-Transcend.*

- **Trustingness and Trustworthiness:** Deriving from *interaction and propagation* history between the user and his followers and the people he follows.

- **Influence:** Total *attention* received by user as leader across all discovered flows.

**Precision-Recall performance of predicting info propagation by different features (Our proposed influence index: FLOWER)**

**===>** Preliminary studies showed propagation behavior is related to these social cognitive traits
Target Discovery
Anomaly Detection
Detecting Anomalous Information Spreading

• Motivation
  – People’s dynamic reactions to information (e.g., retweeting) give clues to information credibility and quality
    • For example, trustworthy people may take time to verify uncertain information from strangers before spreading it

• Approach
  – Use one-class conditional random field to model people’s behavior in information spreading sequences and detect anomalous sequences
    • Features: content features such as emotion, network features such as tie strengths and clustering coefficients

• Preliminary experiments
  – Detect anomalies in retweeting sequences during Hurricane Sandy
    • Including hijacker, fake pictures spreaders

One-class CRF to detect temporal anomalies

Detected as top 1 anomaly in Sandy Tweets
Photo Tweet Sentiment Tracking during Hurricane Sandy
(Columbia University)

- **Goal:** Detect sentiment during Hurricane Sandy.

- **Data collection:**
  - Date: Oct 25 – Nov 02
  - Hashtags (based on popularity): #prayforusa, #frankenstorm, #nyc,#hurricane,#sandy,#hurricanesandy, #staysafe, #redcross,#myheartgoesouttoyou,…
  - 2000 Photo Tweets collected

- **Ground Truth Labeling:**
  - 1340 unanimously agreed labels from 2 individuals

- **Training Classifier:**
  - Text (SentiStrength)
  - Visual(SentiBank, Logistic Reogr.)
  - Training/Testing ratio: 4:1
  - 5-fold cross-validation
  - Accuracy(Text-Visual Combined): 72%
Forensic System

Data Ingest & prep. → Social Tie Strength Estimation → User Behavior Statistics

Social Media Data Store → User Behavioral Models → Social Network Models

Graph Analytics

Social Media Data Store

Emotion Analytics
Egonet Analytics
Graph Similarity (Rutgers)
Influence/Trust (UMN)
Human Personality & Value
Emotion Activation (UNW)

Detecting Anomalous Sequences that may be disinformation

RT sequence extraction & filtering (keyword, loc etc)

Anomalous Sequence Detection By One-Class CRF

Top K Sequences

Multimedia Sentiment Tracking

*Note: Components in blue are work in progress and applied to the top-K sequences
Users can interactively analyze the personality traits, emotion activation, value and other traits of the people involved.

This rumor sequence involves users that are highly emotionally active.
Finding Anomalies

Detect disinformation spreading (retweet) during Hurricane Sandy by people’s dynamic behavior in response to the information.

Detected rumor: Shark in NJ

The rumor sequence has low trustworthiness users involved

Human personality, value traits to show (showing Trustworthiness)

User info (e.g., personality chart)
Stream Analyses Technical Challenges
Example IP Packet Stream Instantiation

Inputs

Dataflow Graph

By IBM Dense Information Gliding Team
Semantic MM Filtering

Inputs
- ip
- tcp
- udp
- http
- rtsp
- ftp
- audio
- sess
- ntp
- video

Dataflow Graph
- Packet content analysis
- Advanced content analysis
- Interest Routing
- Interest Filtering
- keywords
- Interested MM streams

per PE rates
- 200-500MB/s
- ~100MB/s
- 10 MB/s
Resource-Accuracy Trade-Offs

Configurable Parameters of Processing Elements to maximize relevant information:
\[ Y''(X \mid q, R) > Y'(X \mid q, R), \]
with resource constraint.
Required resource-efficient algorithms for:
- Classification, routing and filtering of signal-oriented data: (audio, video and, possibly, sensor data)

- Input data \( X \) – Queries \( q \) – Resource \( R \)
  - \( Y(X \mid q) \): Relevant information
  - \( Y'(X \mid q, R) \cdot Y(X \mid q) \): Achievable subset given \( R \)
Example: Distributed Video Signal Understanding (Lin et al.)

(Distributed Smart Sensors) Block diagram of the smart sensors

- **TV broadcast, VCR, DVD discs, Video**
- **File Database, Webcam**

**Smart Cam**

- **Sensor 1**
- **Sensor 2**
- **Sensor N**
- **Sensor 3**

**Processing Elements**
- **Face**
- **Female**
- **Outdoors**
- **Male**
- **Indoors**
- **Airplane**
- **Chair**
- **Clock**
- **Control Modules**

**Concept Detection**

- **PE1: 9.2.63.66: 1220**
- **PE2: 9.2.63.67**
- **PE3: 9.2.63.68**
- **PE4: 9.2.63.66:1235**
- **PE5: 9.2.63.66: 1240**
- **PE6: 9.2.63.66**
- **PE7: 9.2.63.66**
- **PE100: 9.2.63.66**

**Resource Constraints**

- **User Interests**

**Display and Information Aggregation Modules**

- **1.5 Mbps**
- **2.8 Kbps**
- **600 bps**

**MPEG-1/2**

- **320 Kbps**
- **22.4 Kbps**
- **2.8 Kbps**

**GOP Extraction**

- **Feature Extraction**

- **Event Extraction**

**Encode**
Semantic Concept Filters

E.g.:
Complexity Reduction Introduction

- Objective: Real-time classification of instances using Support Vector Machines (SVMs)
- Computationally efficient and reasonably accurate solutions
- Techniques capable of adjusting tradeoff between accuracy and speed based on available computational resources
SVM formulation

- **Given:**
  - Training instances \( \{ \mathbf{x}_i \} \) with labels \( y_i \)

- **Objective:**
  - Find maximum margin hyperplane separating positive and negative training instances
**Decision**

- **Score of unseen instance** \( u_j : w \cdot \phi(u_j) \)
- **In terms of Lagrangian multipliers**
  \[ \sum_i \alpha_i y_i k(x_i, u_j) \]
- **Computational Cost**: \( O(n_{sv}d) \)
  - \( n_{sv} \): Number of support vectors
  - \( d \): Dimensionality of each data instance
Problems

- Number of support vectors grows quasi-linearly with size of training set [Tipping 2000]

- Inner product with each support vector of dimensionality $d$ expensive
  - Example TREC2003
    - Human : 19745 support vectors
    - Face : 18090

- High data rates (10Gbits/sec) means large number of abandoned data
Example

- Processing Power 1 Ghz
- 10000 support vectors
- 1000 / 2 features per instance
- Order of at least $10^7$ operations required per stream per sec
- Translates to less than 100 instances evaluated per sec with only one classifier
Naïve Approach I – Feature Dimension Reduction

- Experimental Results for Weather_News Detector
- Model Selection based on the Model Validation Set
- E.g., for Feature Dimension Ratio 0.22, (the best selection of features are: 3 slices, 1 color, 2 texture selections), the accuracy is decreased by 17%.
Naïve Approach II – Reduction on the Number of Support

- Proposed Novel Reduction Methods:
  - Ranked Weighting
  - P/N Cost Reduction
  - Random Selection
  - Support Vector Clustering and Centralization

- Experimental Results on Weather_News Detectors show that complexity can be at 50% for the cost of 14% decrease on accuracy
Weighted Clustering Approach

- **Basic steps**
  - Cluster support vectors
  - Use cluster center as representative for all support vectors in cluster
  - Determine scalar weight associated with each cluster center
  - Use only cluster centers to score new instances
Cluster center weight (contd.)

- Choose $\gamma_i$ minimizing square of difference in scores over all $\pm_i$ and $d$

- Sub-cases:

$$d \geq \Delta_i \quad d < \Delta_i$$
Using the weights

- For every support vector in cluster
  - Distance $\Delta_i$ known
  - Two weights computed

- Cumulative effect of all support vectors in clusters additive
  - $\Delta_i$ because of various support vectors added up at center to simulate effect of all support vectors

- $\Delta_i$ sorted, weight arrays rearranged
Experiments

• Datasets
  • TREC video datasets (2003 and 2005)
    • 576 features per instance
    • > 20000 test instances overall
  • MNist handwritten digit dataset (RBF kernel)
    • 576 features
    • 60000 training instances, 10000 test instances

• Performance metrics
  • Speedup achieved over evaluation with all support vectors
  • Average precision achieved
Results (Mnist 0-4)

![Graph showing Average Precision vs Speedup Ratio](image)
Results (Mnist 5-9)

![Graph showing average precision vs. speedup ratio for different digits (5-9). The graph indicates the performance of various models as the speedup ratio increases. The x-axis represents the speedup ratio, while the y-axis represents the average precision. Each digit (5-9) is represented by different markers on the graph.](image-url)
Results (TREC 2003)

<table>
<thead>
<tr>
<th>Concept</th>
<th>AP_fast</th>
<th>AP_original</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>0.75</td>
<td>0.675</td>
</tr>
<tr>
<td>Outdoors</td>
<td>0.5</td>
<td>0.45</td>
</tr>
<tr>
<td>Sport-Event</td>
<td>0.3</td>
<td>0.225</td>
</tr>
<tr>
<td>Crowd</td>
<td>0.1</td>
<td>0.0</td>
</tr>
<tr>
<td>People-Event</td>
<td>0.05</td>
<td>0.0</td>
</tr>
</tbody>
</table>

![Bar chart showing Average Precision for different concepts](image1)

<table>
<thead>
<tr>
<th>Concept</th>
<th>Speedup</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>1000</td>
</tr>
<tr>
<td>Studio-Setting</td>
<td>100</td>
</tr>
<tr>
<td>Crowd</td>
<td>10</td>
</tr>
</tbody>
</table>

![Bar chart showing Speedup for different concepts](image2)
Summary of Complexity Reduction

- Techniques presented demonstrate reasonable performance in terms of both speedup and average precision over multiple concepts in datasets.

- Speedups
  - MNist: All concepts at least 50 times faster with AP within 0.04 of original.
  - TREC 2003: Eight out of nine concepts speedup greater than 80 times with AP within 0.05 of original.
  - TREC 2005: APs in some cases along with speedup respectable.

- APs of most concepts close to original APs.
Acceleration of Neural Network for Streams
Summary of Acceleration of Neural Network on Mobile Devices

- Porting Deep Convolution Neural Network on iOS Device with near real-time computation
  - Reduce algorithmic complexity with ignorable performance degradation.
  - Utilize computational hardware to achieve better performance.

<table>
<thead>
<tr>
<th></th>
<th>iPhone 7</th>
<th>iPhone 7+</th>
<th>iPhone 6s</th>
<th>iPad Pro 12.9&quot;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alex Net</td>
<td>70 ms</td>
<td>70 ms</td>
<td>130 ms</td>
<td>69 ms</td>
</tr>
<tr>
<td>p-Alex Net</td>
<td>N/A</td>
<td>35 ms</td>
<td>N/A</td>
<td>28 ms</td>
</tr>
<tr>
<td>GoogLeNet</td>
<td>130 ms</td>
<td>128 ms</td>
<td>195 ms</td>
<td>110 ms</td>
</tr>
<tr>
<td>p-GoogLeNet</td>
<td>N/A</td>
<td>80 ms</td>
<td>N/A</td>
<td>70 ms</td>
</tr>
<tr>
<td>VGG16 Net</td>
<td>880 ms</td>
<td>883 ms</td>
<td>1450 ms</td>
<td>725 ms</td>
</tr>
</tbody>
</table>
Methods for Running CNNs on Mobile Devices

- Sending CNN jobs to cloud
- Acceleration CNN on Local Device

- Apple A9X SoC, 12-core GPU

- How to trade off between algorithmic complexity and performance?
- How to utilize hardware effectively
Challenges for Running CNN on Mobile Devices

<table>
<thead>
<tr>
<th>Model Size</th>
<th>Weights</th>
<th>Mult.s</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>AlexNet</strong></td>
<td>243MB</td>
<td>61M</td>
</tr>
<tr>
<td><strong>VGG-S</strong></td>
<td>393MB</td>
<td>103M</td>
</tr>
<tr>
<td><strong>VGG-16</strong></td>
<td>552MB</td>
<td>138M</td>
</tr>
<tr>
<td><strong>GoogLeNet</strong></td>
<td>51MB</td>
<td>6.9M</td>
</tr>
</tbody>
</table>

Reference:
Compression of Deep Convolutional Neural Networks for Fast and Low Power Mobile Applications
## Computational Resource on iPhone and iPad

<table>
<thead>
<tr>
<th></th>
<th>iPhone 6S (Plus)</th>
<th>iPad Air 2</th>
<th>iPad Pro (12.9/9.7)</th>
<th>iPhone 7 (Plus)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SoC</strong></td>
<td>A9</td>
<td>A8X</td>
<td>A9X</td>
<td>A10 Fusion</td>
</tr>
<tr>
<td><strong>CPU</strong></td>
<td>2x Twister @ 1.85GHz</td>
<td>3x Typhone @ 1.5GHz</td>
<td>2x Twister @ 2.26GHz</td>
<td>4-core</td>
</tr>
<tr>
<td><strong>GPU</strong></td>
<td>PVR GT7600 (6 cluster)</td>
<td>PVR GXA6850 (8 cluster)</td>
<td>PVR 12 Cluster Series 7</td>
<td>6 cluster GPU?</td>
</tr>
<tr>
<td><strong>RAM (shared memory)</strong></td>
<td>2GB LDDR4</td>
<td>2GB LDDR3</td>
<td>4GB LDDR4</td>
<td>3GB on Plus?</td>
</tr>
<tr>
<td><strong>Memory bus width</strong></td>
<td>64-bit</td>
<td>128-bit</td>
<td>128-bit</td>
<td>?</td>
</tr>
<tr>
<td><strong>Max # of threads per group</strong></td>
<td>512</td>
<td>512</td>
<td>512</td>
<td>?</td>
</tr>
</tbody>
</table>
Neuron Importance Score Propagation (NISP, Yu et al 2018)
Methods for Running CNNs on Mobile Devices

- Sending CNN jobs to cloud
- Compression (pruning) of CNN
- Speeding up CNN
Thinking Differently

- All existing methods can be viewed as approximations of an overly-redundant CNN. But do we really need such a CNN as the starting point?

![CNN Sliming!](image)
Slim CNN

• Slim CNN leads to:
  • less storage space
  • less memory usage
  • less computation
  • less power consumption
Feature Selection on CNN

- CNNs can be viewed as a set of "overly-redundant" feature extractors.

"make everything as simple as possible, but no simpler" — Albert Einstein
A method for Pruning Redundant Neurons and Kernels of

Apply thermal

A pre-trained CNN → Extract CNN Responses → Measure the Importance of Feature Extractors → Prune Model → Fine-tuning
A method for Pruning Redundant Neurons and Kernels of Deep Convolutional Neural Networks (NISP)

- Intractable → tractable
- Inconsistent → consistent

A pre-trained CNN → Extract Responses of a High-level Layer → Measure the Importance of Feature Extractors → Back-propagate the Importance & Prune Model → Fine-tuning

Forward Propagation

Input layers → Response → ... → Response → ... → Response → ... → Important Score Back Propagation and Pruning
Fine-tuning the Pruned Model

- Our method outperforms the baselines in three aspects
  - Very small accuracy loss at the beginning ==> retains the most important neurons
  - Converges much faster than baselines
  - For LeNet on MIST, our method only decreases 0.02% top-1 accuracy with a running ratio of 50% as compared to the pre-pruned network.
Fine-tuning the Pruned Model

- The pruned model consists of important feature extractors, but will suffer loss of accuracy due to loss of redundant features
- Good starting point on the learning curve due to feature selection
- Fine-tuning the pruned model with a lower learning rate to recover the performance
Spark ML Classification and Regression

MLlib: Main Guide

- Basic statistics
- Pipelines
- Extracting, transforming and selecting features
- Classification and Regression
- Clustering
- Collaborative filtering
- Frequent Pattern Mining
- Model selection and tuning
- Advanced topics

- Classification
  - Logistic regression
    - Binomial logistic regression
    - Multinomial logistic regression
  - Decision tree classifier
  - Random forest classifier
  - Gradient-boosted tree classifier
  - Multilayer perceptron classifier
  - Linear Support Vector Machine
  - One-vs-Rest classifier (a.k.a. One-vs-All)
  - Naïve Bayes

- Regression
  - Linear regression
  - Generalized linear regression
    - Available families
  - Decision tree regression
  - Random forest regression
  - Gradient-boosted tree regression
  - Survival regression
  - Isotonic regression
Spark Streaming

- Kafka
- Flume
- HDFS/S3
- Kinesis
- Twitter

- HDFS
- Databases
- Dashboards
Spark Streaming

- **Basic Concepts**
  - Linking
  - Initializing StreamingContext
  - Discretized Streams (DStreams)
  - Input DStreams and Receivers
  - Transformations on DStreams
  - Output Operations on DStreams
  - DataFrame and SQL Operations
  - MLlib Operations
  - Caching / Persistence
  - Checkpointing
  - Accumulators, Broadcast Variables, and Checkpoints
  - Deploying Applications
  - Monitoring Applications
- **Performance Tuning**
  - Reducing the Batch Processing Times
  - Setting the Right Batch Interval
  - Memory Tuning
- **Fault-tolerance Semantics**
Spark Streaming Example

First, we import `StreamingContext`, which is the main entry point for all streaming functionality. We create a local `StreamingContext` with two execution threads, and batch interval of 1 second.

```python
from pyspark import SparkContext
from pyspark.streaming import StreamingContext

# Create a local StreamingContext with two working thread and batch interval of 1 second
sc = SparkContext("local[2]", "NetworkWordCount")
ssc = StreamingContext(sc, 1)
```

Using this context, we can create a DStream that represents streaming data from a TCP source, specified as hostname (e.g. `localhost`) and port (e.g. 9999).

```python
# Create a DStream that will connect to hostname:port, like localhost:9999
lines = ssc.socketTextStream("localhost", 9999)
```

This `lines` DStream represents the stream of data that will be received from the data server. Each record in this DStream is a line of text. Next, we want to split the lines by space into words.

```python
# Split each line into words
words = lines.flatMap(lambda line: line.split(" "))
```

```python
# Count each word in each batch
pairs = words.map(lambda word: (word, 1))
wordCounts = pairs.reduceByKey(lambda x, y: x + y)
```

```python
# Print the first ten elements of each RDD generated in this DStream to the console
wordCounts.pprint()
```
Spark Streaming Example

$.bin/spark-submit examples/src/main/python/streaming/network_wordcount.py localhost 9999

Then, any lines typed in the terminal running the netcat server will be counted and printed on screen every second. It will look something like the following.

```
# TERMINAL 1:
# Running Netcat
$ nc -lk 9999
hello world
...
```

```
# TERMINAL 2: RUNNING network_wordcount.py
$ .bin/spark-submit examples/src/main/python/streaming/network_wordcount.py localhost 9999
...


(‘hello’,1)
(‘world’,1)
...
```
Discretized Streams

DStream → RDD @ time 1: data from time 0 to 1 → RDD @ time 2: data from time 1 to 2 → RDD @ time 3: data from time 2 to 3 → RDD @ time 4: data from time 3 to 4
Discretized Streams

```
lines DStream → lines from time 0 to 1 → lines from time 1 to 2 → lines from time 2 to 3 → lines from time 3 to 4

flatMap operation

words DStream → words from time 0 to 1 → words from time 1 to 2 → words from time 2 to 3 → words from time 3 to 4
```
Discretized Streams

https://www.edureka.co/blog/spark-streaming/
DStream Transforms

https://www.edureka.co/blog/spark-streaming/
Output DStreams

Transformed DStream → Output Operations → Output DStream

External Systems
- Database
- File System

https://www.edureka.co/blog/spark-streaming/
DStreams Caching

https://www.edureka.co/blog/spark-streaming/
DStreams Example — Twitter Sentiment Analysis

//Import the necessary packages into the Spark Program
import org.apache.spark.streaming.{Seconds, StreamingContext}
import org.apache.spark.SparkContext._

import java.io.File

object twitterSentiment {

def main(args: Array[String]) {
  if (args.length < 4) {
    System.err.println("Usage: TwitterPopularTags <consumer key> <consumer secret> " + "<access token> <access token secret> System.exit(1)
  }

  StreamingExamples.setStreamingLogLevels()
  //Passing our Twitter keys and tokens as arguments for authorization
  val Array(consumerKey, consumerSecret, accessToken, accessTokenSecret) = args.take(4)
  val filters = args.takeRight(args.length - 4)

  // Set the system properties so that Twitter4j library used by twitter stream
  // Use them to generate OAuth credentials
  System.setProperty("twitter4j.oauth.consumerKey", consumerKey)
  System.setProperty("twitter4j.oauth.consumerSecret", consumerSecret)

  val sparkConf = new SparkConf().setAppName("twitterSentiment").setMaster("local[2]")
  val ssc = new StreamingContext
  val stream = TwitterUtils.createStream(ssc, None, filters)

https://www.edureka.co/blog/spark-streaming/
DStreams Example — Twitter Sentiment Analysis

```scala
//Input DStream transformation using flatMap
val tags = stream.flatMap { status => Get Text From The Hashtags }

//RDD transformation using sortBy and then map function
tags.countByValue().
  foreachRDD { rdd =>
    val now = Get current time of each Tweet
    rdd
      .sortBy(._2)
      .map(x => (x, now))
  }

//Saving our output at ~/twitter/ directory
.saveAsTextFile(s"~/twitter/$now")
}

//DStream transformation using filter and map functions
val tweets = stream.filter { t =>
  val tags = t. Split On Spaces .filter(_.startsWith("#")). Convert To Lower Case
  tags.exists { x => true }
}

val data = tweets.map { status =>
  val sentiment = SentimentAnalysisUtils.detectSentiment(status.getText)
  val tagss = status.getTextHashtagEntities.map(_.getText, toLowerCase)
  (status.getText, sentiment.toString, tagss.toString())
}

data.print()
//Saving our output at ~/ with filenames starting like twitters
data.saveAsTextFiles("~/tweets", "20000")

ssc.start()
ssc.awaitTermination()
}
```

https://www.edureka.co/blog/spark-streaming/
DStreams Example — Twitter Sentiment Analysis

Results:

The following are the results that are displayed in the Eclipse IDE while running the Twitter Sentiment Streaming program.

![Eclipse IDE screenshot](https://www.edureka.co/blog/spark-streaming/)

All the tweets are categorized into Positive, Neutral and Negative according to the sentiment of the contents of the tweets.

https://www.edureka.co/blog/spark-streaming/
Homework #3 (Due 11/01/2019, 5pm)

See HW3 and Tutorial Slides
Questions?