Web-scale Multimedia Search for Internet Video Content

Lu Jiang
Language Technologies Institute, Carnegie Mellon University
October 15th 2015.

Thesis Committee:
Dr. Alex Hauptmann (Co-chair), Carnegie Mellon University
Dr. Teruko Mitamura (Co-chair), Carnegie Mellon University
Dr. Louis-Philippe Morency, Carnegie Mellon University
Dr. Tat-Seng Chua, National University of Singapore
Outline

- Introduction

- Proposed Approaches:
  - Indexing Semantic Features
  - Semantic Search
  - Video Reranking
  - Building Semantic Concepts

- Conclusions
  - Proposed Work: hybrid search
Outline

- **Introduction**
- **Proposed Approaches:**
  - Indexing Semantic Features
  - Semantic Search
  - Video Reranking
  - Building Semantic Concepts
- **Conclusions**
  - Proposed Work: hybrid search
Introduction

• We are living in an era of big multimedia data:
  – 300 hours of video are uploaded to YouTube every minute;
  – social media users are posting 12 million videos on Twitter every day;
  – video will account for 80% of all the world's internet traffic by 2019.

• Video search is becoming a valuable source for acquiring information and knowledge.

• Existing large-scale methods are still based on text-to-text matching (user text query to video metadata), which may fail in many scenarios.
  – 66% videos on the social media site Twitter are not associated with hashtag or mention [Vandersmissen et al. 2014]
Introduction

- 66% videos on the social media site Twitter are not associated with hashtag or mention [Vandersmissen et al. 2014]
Introduction

- Much more video captured by mobile phones, surveillance cameras and wearable devices does not have any metadata at all.
Introduction

• We are living in an era of big multimedia data:
  – 300 hours of video are uploaded to YouTube every minute;
  – social media users are posting 12 millions videos on Twitter every day;
  – video will account for 80% of all the world's internet traffic by 2019.

• Video search is becoming a valuable source for acquiring information and knowledge.

• Existing large-scale methods are still based on text-to-text matching (user query to video metadata), which may fail in many scenarios.
  – 66% videos on a social media site of Twitter are not associated with meaningful metadata (hashtag or a mention)[Vandersmissen et al. 2014]
  – Much video captured by mobile phones, surveillance cameras and wearable devices does not have any metadata at all.
Introduction

• This thesis addresses a fundamental research question: how to satisfy information needs about video content at a very large scale?

• We embody this question into a concrete content-based video retrieval problem which aims at searching videos solely based on content, without using any user-generated metadata (e.g. titles or descriptions).

• We focus on two types of queries: semantic query and hybrid query.
Information need:
Find videos about birthday party.

1) Semantic concept (acoustic and visual)
2) ASR (Automatic Speech Recognition)
3) OCR (Optical Character Recognition)

text-to-video search
Hybrid Query:

- semantic query
- video examples provided by users

*text&video-to-video search*
Example Queries

• In response to a, our system should be able to:
  – find simple objects, actions, speech words;
  – search complex activities;

Information need:
people running away after an explosion in urban areas.

Query:  Boolean logical operator
urban_scene
AND (walking OR running)
OR fire OR smoke
OR audio:explosion
TBefore(audio:explosion, running)
Example Queries

• Using the query, our system should be able to
  – find simple objects, actions, speech words;
  – search complex activities;
  – answer questions by/in videos.

Information need:
What are they doing?

Query:

person AND action
AND

How to learn Tai Chi Chuan

Become a Tai Chi Master in 5 Steps:
#2 - Tai Chi Solo Form Practice
Challenges

• The problem was initiated by a TRECVID task Multimedia Event Detection (MED) in 2012 (common evaluation benchmark).
  – State-of-the-art accuracy is very low.
  – Large-scale system can only handle 200k videos (5 min to search).

• For this understudied problem, this thesis confronts the following research challenges:
  – Algorithms to boost state-of-the-art accuracy.
  – Efficient methods to search billions of videos.
Preliminary Results

• We proposed a novel and practical solution that can
  – substantially boost state-of-the-art accuracy across a number of datasets.
  – Scale up the search to hundreds of millions of Internet videos.
    • 0.2 second to process a semantic query on 100 million videos
    • 1 second to process a hybrid query on 1 million videos.

• Within a system called E-Lamp Lite, we implemented the first of its kind large-scale multimedia search engine for Internet videos:
  – Achieved the best accuracy in TRECVID MED zero-example search 2013, 2014 and 2015, the most representative task on this task. 3x better than the runner-up in 2014.
  – To the best of our knowledge, it is the first content-based retrieval system that can search a collection of 100 million videos.
Let the above videos represent the upper-bound of the current largest dataset for this problem (200k videos).
Framework

Offline Stage

Video Stream

Trajectory/MFCC/Deep learning features

Low-level features

Adjusted Semantic features

Indexes

cake
detector

0.9

kids
detector

0.8

cheering
detector

0.2

ASR

OCR

Visual
concepts

Audio
concepts

User Query

Semantic Query

An indoor party celebrating a child’s birthday

Online Stage

System Query

indoor AND party AND kid OR birthday cake

Indexes

ASR

OCR

Visual
concepts

Audio
concepts

(1) Query Generation

(2) Multimodal Search

(3) Fusion/Pseudo-Relevance Feedback (PRF)

Feedback

Reranking model on pseudo labels

Ranked list

Semantic features

High-level features

Birthday song
Key Contributions:
First-of-its-kind Framework

- The first-of-its-kind framework for web-scale content-based search over hundreds of millions of Internet videos [ICMR’15]. The proposed framework supports text-to-video, video-to-video, and text&video-to-video search [MM’12]. *(Chapter 1 and 5)*


Key Contributions: Self-paced curriculums learning theory

- The first-of-its-kind framework for web-scale content-based search over hundreds of millions of Internet videos [ICMR’15]. The proposed framework supports text-to-video, video-to-video, and text&video-to-video search [MM’12].
- A novel theory about self-paced curriculums learning and its application on robust concept detector training [NIPS’14, AAAI’15]. (Chapter7)

Key Contributions: Self-paced curriculums learning theory

- The first-of-its-kind framework for web-scale content-based search over hundreds of millions of Internet videos [ICMR’15]. The proposed framework supports text-to-video, video-to-video, and text&video-to-video search [MM’12].

- A novel theory about self-paced curriculums learning and its application on robust concept detector training [NIPS’14, AAAI’15].


Key Contributions:
Reranking

• The first-of-its-kind framework for web-scale content-based search over hundreds of millions of Internet videos [ICMR’15]. The proposed framework supports text-to-video, video-to-video, and text&video-to-video search [MM’12].

• A novel theory about self-paced curriculums learning and its application on robust concept detector training [NIPS’14, AAAI’15].

• Novel reranking algorithms for improving performance. They have concise mathematical objectives to optimize and useful properties that can be theoretically verified [MM’14, ICMR’14]. (Chapter6)


Key Contributions:

- The first-of-its-kind framework for web-scale content-based search over hundreds of millions of Internet videos [ICMR’15]. The proposed framework supports text-to-video, video-to-video, and text&video-to-video search [MM’12].

- A novel theory about self-paced curriculum learning and its application on robust concept detector training [NIPS’14, AAAI’15].

- Novel reranking algorithms for improving performance. They have concise mathematical objectives to optimize and useful properties that can be theoretically verified [MM’14, ICMR’14]. (Chapter 6)


Key Contributions: Scalable Indexing Method

• The first-of-its-kind framework for web-scale content-based search over hundreds of millions of Internet videos [ICMR’15]. The proposed framework supports text-to-video, video-to-video, and text&video-to-video search [MM’12].

• A novel theory about self-paced curriculums learning and its application on robust concept detector training [NIPS’14, AAAI’15].

• Novel reranking algorithms for improving performance. They have concise mathematical objectives to optimize and useful properties that can be theoretically verified [MM’14, ICMR’14].

• A concept adjustment method representing a video by a few salient and consistent concepts that can be efficiently indexed by the modified inverted index [MM’15] (Chapter3)

Key Contributions: Scalable Indexing Method

• The first-of-its-kind framework for web-scale content-based search over hundreds of millions of Internet videos [ICMR'15]. The proposed framework supports text-to-video, video-to-video, and text&video search [MM'12].

• A novel theory about self-paced curriculums learning and its application on robust concept detector training [NIPS'14, AAAI’15].

• Novel reranking algorithms for improving performance. They have concise mathematical objectives to optimize and useful properties that can be theoretically verified [MM’14, ICMR’14].

• A concept adjustment method representing a video by a few salient and consistent concepts that can be efficiently indexed by the modified inverted index [MM’15] (Chapter3)

Thesis Statement

- In this thesis, we approach a fundamental problem of searching information in video content at a very large scale. We address the problem by proposing an accurate, efficient, and scalable method that can search the content of billions of videos by semantic visual/acoustic concepts, speech, visible texts, video examples, or any combination of these elements.
Outline

- Introduction

- Proposed Approaches:
  - Indexing Semantic Features [95%]
  - Semantic Search [95%]
  - Video Reranking [95%]
  - Building Semantic Concepts [80%]

- Conclusions
  - Proposed Work: hybrid search [10%]
Outline

- Introduction
- Approaches:
  - Indexing Semantic Features [95%]
  - Semantic Search [95%]
  - Video Reranking [95%]
  - Building Semantic Concepts [80%]
- Conclusions and Proposed Work
  - Proposed Work: Hybrid Search [10%]
Introduction to Indexing Semantic Features

• Semantic features include ASR (speech), OCR (visible text), visual concepts and audio concepts.
• Indexing textual features like ASR and OCR is well studied.

• Indexing semantic concepts is well studied.
• Existing methods index the raw detection score of semantic concepts by dense matrices [Mazloom et al. 2014][Wu et al. 2014][Lee et al. 2014]
• We propose a scalable semantic concept indexing method. The key is a novel method called concept adjustment.

Method Overview

- Represent raw video (or video clip) by low-level features.
- Semantic concept detectors are of limited accuracy. The raw detections are meaningful but very noisy.
The raw score representation has two problems:

- **Distributional inconsistency**: every video has every concept in the vocabulary (with a small but nonzero score);
- **Logical inconsistency**: a video may contain a “terrier” but not a “dog”.

To address the problems, we introduce a novel step called concept adjustment which represents a video by **a few salient and logically consistent visual/audio concepts**.
Concept Adjustment Model

- The proposed adjustment model is:

\[
\arg \min_{\mathbf{v} \in [0,1]^m} \frac{1}{2} \| \mathbf{v} - f_p(\mathbf{D}) \|_2^2 + g(\mathbf{v}; \alpha, \beta)
\]

subject to \( \mathbf{A} \mathbf{v} \leq \mathbf{c} \)

where \( \mathbf{v} \in \mathbb{R}^{m \times 1} \) is the adjusted concept score. \( f_p(\mathbf{D}) \) is a pooling on the raw detection score matrix \( \mathbf{D} \): each row corresponds to a shot and each column corresponds to a concept.

- Our goal is to generate video representations that tends to be similar to the underlying concept representation in terms of the distributional and logical consistency.
Concept Adjustment Model: Distributional Consistency

- Our general implementation:

\[ g(v; \alpha, \beta) = \alpha \beta \|v\|_1 + (1 - \alpha) \sum_{l=1}^{q} \beta \sqrt{p_l} \|v^{(l)}\|_2, \]

- When \( \alpha = 1 \) \( \rightarrow \) concepts are independent.
- When \( \alpha = 0 \) \( \rightarrow \) groups of concepts frequently co-occur, e.g. sky/cloud, beach/ocean/waterfront, and table/chair. Multimodal concepts baby/baby_crying.
- When \( \alpha \in (0, 1) \) \( \rightarrow \) only few concepts in a co-occurring group are nonzero [Simon et al. 2013].

The choice of the model parameters depends on the underlying distribution of the semantic concepts in the dataset.

Definition 3.1. A HEX graph \( G = (N, E_h, E_e) \) is a graph consisting of a set of nodes \( N = \{n_1, \ldots, n_m\} \), directed edges \( E_h \subseteq N \times N \) and undirected edges \( E_e \subseteq N \times N \) such that the subgraph \( G_h = (N, E_h) \) is a directed acyclic graph and the subgraph \( G_e = (N, E_e) \) has no self-loop.

[Deng et al, 2014]

Theorem 3.3. The optimal solutions of Eq. (3.1) (before or after normalization) is logically consistent with its given HEX graph.

\[
\begin{align*}
    v_{\text{dog}} & \leq v_{\text{animal}} \\
    v_{\text{animal}} + v_{\text{blank_frame}} & \leq 1 \\
    v_{\text{animal}}, v_{\text{blank_frame}} & \in \{0, 1\}
\end{align*}
\]

Integer programming solved by mix-integer toolbox or by constraint relaxation.

Jia Deng, Nan Ding, Yangqing Jia, Andrea Frome, Kevin Murphy, Samy Bengio, Yuan Li, Hartmut Neven, and Hartwig Adam. Large-scale object classification using label relation graphs. In *ECCV, 2014*. 
Finally, the adjusted concept representation is indexed by an inverted index. The index structure needs to be modified to account for:

- Indexing real-valued concepts
- Indexing the shot-level scores
- Supporting Boolean logical and temporal operators.

*Detailed methods are in Chapter 3*
Experiments on MED

• Dataset: MED13Test and MED14Test (around 25,000 videos). Each set contains 20 events.
• Official evaluation metric: Mean Average Precision (MAP)
• Supplementary metrics:
  – Mean Reciprocal Rank = (1/rank of the first relevant sample)[Voorhees, 1999]
  – Precision@20
  – MAP@20
• Configurations:
  – NIST’s HEX graph is used for IACC;
  – We build the HEX graphs for the rest of the semantic concept features.
  – Raw prediction scores of the 3000+ concepts trained in Chapter 7.

Experiments on MED

Comparison of the raw and the adjusted representation

<table>
<thead>
<tr>
<th>Method</th>
<th>Index</th>
<th>Evaluation Metric</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>P@20</td>
</tr>
<tr>
<td>MED13 Raw</td>
<td>385M</td>
<td>0.312</td>
</tr>
<tr>
<td>MED13 Adjusted</td>
<td>11.6M</td>
<td>0.325</td>
</tr>
<tr>
<td>MED14 Raw</td>
<td>357M</td>
<td>0.233</td>
</tr>
<tr>
<td>MED14 Adjusted</td>
<td>12M</td>
<td>0.219</td>
</tr>
</tbody>
</table>

The accuracy of the proposed method is comparable to that of the baseline method.
Experiments on 100M Videos

- Baseline method (raw score representation) fails when the data reaches 5 million videos.
- Our method can scale to 100M videos.
  - take 0.2s on a single core (on-line search time);
  - create an on-disk inverted index of 20G;
  - Use 512MB memory.

The scalability and efficiency test on 100 million videos.

The proposed method is scalable and efficient.
Summary of Indexing Semantic Features

• We proposed a scalable semantic concept indexing methods that extends the current scale of video search by a few orders of magnitude while maintaining state-of-the-art retrieval performance.

• The key is a novel step called concept adjustment that can represent a video by a few salient and consistent concepts which can be efficiently indexed by a modified inverted index.
Outline

- Introduction
- Approaches:
  - Indexing Semantic Features [95%]
  - **Semantic Search [95%]**: the search process for semantic queries.
  - Video Reranking [95%]
  - Building Semantic Concepts [80%]
- Conclusions and Proposed Work
  - Proposed Work: Hybrid Search [10%]
Semantic Search:
Semantic Query Generation

• (1) **Semantic query generation**: how to map out-of-vocabulary query words to the concepts in our vocabulary?

  - The key is to measure the similarity between a query word and a concept in the vocabulary:
    - **Exact word matching**
    - **WordNet Similarity**: structural depths in WordNet taxonomy.
    - **Wikipedia Point-wise Mutual Information (PMI)**: calculate the mutual information of two words in Wikipedia.
    - **Word embedding mapping**: word distance in a learned embedding space in Wikipedia by word2vec.

  ![Diagram](image.png)
  - User query: Making a sandwich
  - Generated query: food, bread, cheese, kitchen, cooking, room, lunch, dinner;
Semantic Query Generation

• We empirically study the following methods.

MAP comparison on MED13Test and MED14Test datasets

<table>
<thead>
<tr>
<th>Mapping Method</th>
<th>MAP</th>
<th>Time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>13Test</td>
<td>14Test</td>
</tr>
<tr>
<td>Exact Word Matching</td>
<td>9.66</td>
<td>7.22</td>
</tr>
<tr>
<td>WordNet</td>
<td>7.86</td>
<td>6.68</td>
</tr>
<tr>
<td>PMI</td>
<td>9.84</td>
<td>6.95</td>
</tr>
<tr>
<td>Word Embedding</td>
<td>8.79</td>
<td>6.21</td>
</tr>
<tr>
<td>Mapping Fusion</td>
<td>10.22</td>
<td>9.38</td>
</tr>
</tbody>
</table>

Individual methods are comparable.

Fusion improves the mapping results.
Semantic Search: Multimodal Search

(2) Retrieval methods: what retrieval model to use for which modality?


– We studied classical four retrieval models over three modalities: ASR, OCR, and semantic concepts
  • Vector Space Model (VSM): tf and tf-idf representations.
  • BM25
  • Language Model-JM Smoothing (LM-JM)
  • Language Model-Dirichlet Smoothing (LM-DL)

– We found retrieval models have substantial impacts to the search result.
  • For ASR, **LM-JM** works the best. More than 1.5x better than the second best model.
  • For semantic concepts and OCR, **BM25** seems to be a robust and accurate retrieval model.

Summary of Semantic Search

• We empirically studied the semantic query generation and retrieval methods. We found that:
  – The fusion of mapping methods perform better than any individual methods.
  – Language Model-JM Smoothing works the best for ASR and BM25 works reasonably well for other types of features.
Outline

- Introduction
- Approaches:
  - Indexing Semantic Features [95%]
  - Semantic Search [95%]
  - Video Reranking [95%]
- Building Semantic Concepts
- Conclusions and Proposed Work
- Proposed Work: Hybrid Search [10%]
Generic Reranking Algorithm

1: \( t = 0; \) //Iteration zero
2: Choose the initial pseudo labels and weights;
3: \textbf{while} \( t \leq \text{max iteration} \) \textbf{do}
4: \hspace{1em} Train a reranking model on the fixed labels and weights;
5: \hspace{1em} Update the pseudo labels and weights;
6: \hspace{1em} \textbf{if} \( t \) \text{ is small} \hspace{1em} \textbf{then} \hspace{1em} add more pseudo positives;
7: \hspace{1em} \textbf{end while}
8: \textbf{return} The list of samples after reranking;

pseudo labels
reranking model
Iteration: 1
Generic Reranking Algorithm

1: \( t = 0; \) //Iteration zero
2: Choose the initial pseudo labels and weights;
3: \textbf{while} \( t \leq \text{max iteration} \) \textbf{do}
4: \quad Train a reranking model on the fixed labels and weights;
5: \quad Update the pseudo labels and weights;
6: \quad \textbf{if} \( t \) is small \textbf{then} add more pseudo positives;
7: \quad \textbf{end while}
8: \textbf{return} The list of samples after reranking;

\begin{align*}
\text{Iteration: 1} \quad &1 \quad 2 \quad 3 \quad 4 \\
&\vdots \quad \vdots \quad \vdots \quad \vdots \\
&n-5 \quad n-4 \quad n-3 \quad n-2 \\
&n-1 \quad n \\
\text{Iteration: 2} \quad &1 \quad 2 \quad 3 \quad 4 \\
&\vdots \quad \vdots \quad \vdots \quad \vdots \\
&n-5 \quad n-4 \quad n-3 \quad n-2 \\
&n-1 \quad n \end{align*}
Intuition

- Existing methods assign equal weights to pseudo samples.
- Intuition: samples ranked at the top are generally more relevant than those ranked lower.
- Our approach: learn the weight together with the reranking model.
Self-paced Learning

• Self-paced learning (Kumar et al 2010) is a learning paradigm that is inspired by the learning process of humans and animals.

• The samples are not learned randomly but organized in a meaningful order which illustrates from easy to gradually more complex ones.

Self-paced Learning

• In the context of reranking: easy samples are the top-ranked videos that have smaller loss.
Self-paced Reranking (SPaR)

- The propose model:

\[
\min_{\Theta_1, \ldots, \Theta_m, y, v} \mathbb{E}((\Theta_1, \ldots, \Theta_m, v, y; C, k))
\]

\[
= \min_{y, v, \Theta_1, \ldots, \Theta_m} C \sum_{i=1}^{n} v_i \sum_{i=1}^{m} \ell_{ij} + \sum_{i=1}^{m} \frac{1}{2} \|w_j\|_2^2 + mf(v; k)
\]

\text{s.t. } \forall i, \forall j, y_i (\sum_{j}^{m} \phi(x_{ij}) + b_j) \geq 1 - \ell_{ij}, \ell_{ij} \geq 0

\quad y \in \{-1, +1\}^n,

\quad v \in [0, 1]^n,

\Theta_1, \ldots, \Theta_m \quad \text{Reranking models for each modality.}

y \in \{-1, 1\}^n \quad \text{The pseudo label.}

v \in [0, 1]^n \quad \text{The weight for each sample.}

Hinge loss function

Function determines the weighting scheme

The self-paced is implemented by a regularizer.

The loss in the reranking model is discounted by a weight.
Proposed Weighting Schemes

**Existing**

Binary weighting [Kumar et al 2010]

\[ f(v; k) = -\frac{1}{k} \|v\|_1 = -\frac{1}{k} \sum_{i=1}^{n} v_i. \]

**Proposed**

Linear weighting

\[ f(v; k) = \frac{1}{k} \left( \frac{1}{2} \|v\|_2^2 - \sum_{i=1}^{n} v_i \right). \]

Logarithmic weighting

\[ f(v; k) = \sum_{i=1}^{n} \left( \zeta v_i - \frac{\zeta v_i}{\log \zeta} \right). \]

Mixture weighting

\[ f(v; k, k') = -\zeta \sum_{i=1}^{n} \log(v_i + \zeta k), \]
Reranking in Optimization and Conventional Perspective

Optimization perspective

1. \( t = 0; \) //Iteration zero
2. Choose starting values for \( y, v; \)
3. \( \text{while} \ t \leq \text{max iteration} \) do
4. \( \Theta_1^{(t+1)}, ..., \Theta_m^{(t+1)} = \arg \max E_{y,v}(\Theta_1^{(t)}, ..., \Theta_m^{(t)}; C); \)
5. \( y^{(t+1)}, v^{(t+1)} = \arg \max E_{\theta}(y^{(t)}, v^{(t)}; k); \)
6. \( \text{if} \ t \text{ is small then increase } 1/k; \)
7. \( \text{end while} \)
8. \( \text{return} \ [v_1y_1, \cdots, v_ny_n]^T; \)

Conventional perspective

1. \( t = 0; \) //Iteration zero
2. Choose the initial pseudo labels and weights;
3. \( \text{while} \ t \leq \text{max iteration} \) do
4. Train a reranking model on the fixed labels and weights;
5. Update the pseudo labels and weights;
6. \( \text{if} \ t \text{ is small then add more pseudo positives;} \)
7. \( \text{end while} \)
8. \( \text{return} \ \text{The list of samples after reranking;} \)

SPaR solution

Optimization perspective

- Optimization perspective \( \rightarrow \) theoretical justifications
- Conventional perspective \( \rightarrow \) practical lessons

Q: Does the process converge? If so, to where?
A: For the proposed weighting, yes, to the local optimum.

Theorem 6.2. The algorithm in Fig. 6.2 converges to a stationary solution for any fixed \( C \) and \( k. \)

See the proof in Appendix D
Experiments on MED13Test

MAP (x100) comparison with baseline methods

<table>
<thead>
<tr>
<th>Method</th>
<th>NIST’s split</th>
<th>10 splits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Without Reranking</td>
<td>3.9</td>
<td>4.9 ± 1.6</td>
</tr>
<tr>
<td>Rocchio</td>
<td>5.7</td>
<td>7.4 ± 2.2</td>
</tr>
<tr>
<td>Relevance Model</td>
<td>2.6</td>
<td>3.4 ± 1.0</td>
</tr>
<tr>
<td>CPRF</td>
<td>6.4</td>
<td>8.3 ± 1.8</td>
</tr>
<tr>
<td>Learning to Rank</td>
<td>3.4</td>
<td>4.2 ± 1.4</td>
</tr>
<tr>
<td>MMPRF</td>
<td>10.1</td>
<td>13.6 ± 2.4</td>
</tr>
<tr>
<td>SPaR</td>
<td>12.9</td>
<td>15.3 ± 2.6</td>
</tr>
</tbody>
</table>

Mixture weighting is used.

AP comparison with baseline methods on each event

Significant improvement!
Outperforms baseline methods on 15/20 events.
Experiments on Web Query

- Web image (353 queries over 71,478 images)
- Densely sampled SIFT are extracted.
- Parameters are tuned on a validation set.
- Mixture self-paced function is used.

**MAP and MAP@100 comparison with baseline methods**

<table>
<thead>
<tr>
<th>Method</th>
<th>MAP</th>
<th>MAP@100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Without Reranking [17]</td>
<td>0.569</td>
<td>0.431</td>
</tr>
<tr>
<td>CPRF [38]</td>
<td>0.658</td>
<td>-</td>
</tr>
<tr>
<td>Random Walk [10]</td>
<td>0.616</td>
<td>-</td>
</tr>
<tr>
<td>Bayesian Reranking [33, 32]</td>
<td>0.658</td>
<td>0.529</td>
</tr>
<tr>
<td>Preference Learning Model [32]</td>
<td>-</td>
<td>0.534</td>
</tr>
<tr>
<td>BVLS [26]</td>
<td>0.670</td>
<td>-</td>
</tr>
<tr>
<td>Query-Relative(visual) [17]</td>
<td>0.649</td>
<td>-</td>
</tr>
<tr>
<td>Supervised Reranking [39]</td>
<td>0.665</td>
<td>-</td>
</tr>
<tr>
<td>SPaR</td>
<td>0.672</td>
<td>0.557</td>
</tr>
</tbody>
</table>

SPaR also works for image reranking (single modality)
Discussions on Video Reranking

• We proposed SPaR, a novel and general framework for multimodal reranking.
• It has theoretical justification, e.g. convergence properties.
• We found two scenarios where SPaR may fail:
  – Initial top-ranked samples are completely off-topic (bad starting values).
  – Features used in reranking are not discriminative to the queries.
Outline

- Introduction
- Approaches:
  - Indexing Semantic Features [95%]
  - Semantic Search [95%]
  - Video Reranking [95%]
  - Building Semantic Concepts [80%]
- Conclusions and Proposed Work
  - Proposed Work: Hybrid Search [10%]
Introduction: Building Semantic Concepts

• Training concept detectors need lots of labeled training data. Annotated video data are hard to collect.
• Our solution is to train detectors from weakly labeled video data (metadata) downloaded from the Internet.
  – Pros: no manual annotations
  – Cons: weakly labeled data are very noisy
• We are interested in approaching this problem in a more principled and theoretically sound way.
  – Derive a theory from paradigms of curriculum learning and self-paced learning.
  – Use proposed theory to train concept detectors on noisy data.
Curriculum Learning and Self-paced Learning

Learning philosophy [Bengio et al. 2009, Kumar et al. 2010]:

- Learning is an iterative process.
- Samples should be organized in a meaningful order (called curriculum).
- Model complexity increases in each iteration.

*The above of real examples in the TRECVID SIN dataset (http://trecvid.nist.gov/).
Curriculum Learning and Self-paced Learning

• **Curriculum Learning (CL):** assign learning priorities to training samples, according to prior knowledge or heuristics about specific problems [Bengio et al. 2009].
  
  – parsing from shorter sentences to longer sentence [Spitkovsky et al. 2009].

• **Self-paced Learning (SPL):** the curriculum is determined by the learned models. Solving a joint optimization problem of the learning objective with the latent curriculum [Kumar, Packer, and Koller 2010].

  – Broadly used in many learning problems such as tracking [Supancic et al. 2013], domain adaptation [Tang et al. 2012], segmentation [Kumar et al. 2011], etc.


Kevin Tang, Vignesh Ramanathan, Li Fei-Fei, and Daphne Koller. Shifting weights: Adapting object detectors from image to video. In NIPS, 2012


Curriculum Learning versus Self-paced Learning

Curriculum Learning (CL)

• Pros
  – Flexible to incorporate prior knowledge/heuristics.

• Cons
  – Curriculum is determined beforehand which may not be consistent with dynamically learned models.

Self-paced Learning (SPL)

• Pros
  – Learn consistent models.
  – Concise optimization problem.

• Cons
  – Cannot use prior knowledge.
  – Random starting values (can significantly affect performance).

Unified in a single framework:
Self-paced Curriculum Learning
Self-paced Learning

- Formulated as an optimization problem (based on SPL).

\[
\arg\min_{\mathbf{w}, \mathbf{v} \in [0,1]^n} \sum_{i=1}^{n} v_i L(y_i, g(x_i, \mathbf{w})) - \lambda \sum_{i=1}^{n} v_i
\]

\(\mathbf{w} \Rightarrow\) parameters in the off-the-shelf model

\(L(y_i, g(x_i, w)) \Rightarrow\) loss for the ith sample

\(\mathbf{v} = [v_1, \ldots, v_n] \Rightarrow\) latent weight vector for all samples

- While fixing \(\mathbf{w}\), the solution is:

\[
v_i^* = \begin{cases} 
1, & L(y_i, g(x_i, \mathbf{w})) < \lambda, \\
0, & \text{otherwise}.
\end{cases}
\]

\(\lambda \Rightarrow\) model age
Self-paced Curriculum Learning

• Proposed learning objectives:

\[
\arg \min_{w, v \in [0,1]^n} \sum_{i=1}^{n} v_i L(y_i, g(x_i, w)) + f(v, \lambda)
\]

subject to \( v \in \Psi \)

\( f(v, \lambda) \) \( \Rightarrow \) regularizer determines the learning scheme

Generalize a single learning scheme to multiple learning schemes. For different problems, we can use different learning schemes.
“Rock Climbing”
Learning Easy and Diverse Samples

\[ f(v, \lambda) = -\lambda \sum_{i=1}^{n} u_i \]

Favor diverse examples

(a) \(\lambda = 0.15\)  
\(\gamma = 0.0\)
Curriculum: a, b, c, d

(b) \(\lambda = 0.03\)  
\(\gamma = 0.2\)
Curriculum: a, j, g, b
Self-paced Curriculum Learning

• Proposed learning objectives:

\[
\arg \min_{w,v} \sum_{i=1}^{n} v_i L(y_i, g(x_i, w)) + f(v, \lambda) \\
\text{subject to } v \in \Psi
\]

• The shape of the feasible region weakly implies a prior learning sequence of samples.

Prior knowledge in curriculum learning
Self-paced Curriculum Learning

• Proposed learning objectives:

\[ \arg \min_{w,v \in [0,1]^n} \sum_{i=1}^{n} v_i L(y_i, g(x_i, w)) + f(v, \lambda) \]

subject to \( v \in \Psi \)

• A new learning theory:
  • Flexible learning schemes to fit various problems;
  • Easy to incorporate prior knowledge;
  • Support any loss function.
Self-paced Curriculum Learning

Curriculum Learning (CL)

Self-paced Learning (SPL)

Self-paced Curriculum Learning (SPCL)

Unified in a single framework: SPCL
Preliminary Experiments

Comparison of SPL and SPCL with diversity learning scheme on MED

<table>
<thead>
<tr>
<th>Run Name</th>
<th>RandomForest</th>
<th>AdaBoost</th>
<th>BatchTrain</th>
<th>SPL</th>
<th>SPLD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Best Run</td>
<td>3.0</td>
<td>2.8</td>
<td>8.3</td>
<td>9.6</td>
<td>12.1</td>
</tr>
<tr>
<td>10 Runs Average</td>
<td>3.0</td>
<td>2.8</td>
<td>8.3</td>
<td>8.6±0.42</td>
<td>9.8±0.45</td>
</tr>
</tbody>
</table>

Comparison of SPL and SPCL with diversity learning scheme on Hollywood2 and Olympic Sports

<table>
<thead>
<tr>
<th>Run Name</th>
<th>RandomForest</th>
<th>AdaBoost</th>
<th>BatchTrain</th>
<th>SPL</th>
<th>SPLD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hollywood2</td>
<td>28.20</td>
<td>41.14</td>
<td>58.16</td>
<td>63.72</td>
<td>66.65</td>
</tr>
<tr>
<td>Olympic Sports</td>
<td>63.32</td>
<td>69.25</td>
<td>90.61</td>
<td>90.83</td>
<td>93.11</td>
</tr>
</tbody>
</table>

*Proposed method*

See more experiments in Section 7.4
Preliminary Experiments

• Using the proposed theory, we build detectors using the YFCC videos (videos sampled from Flickr) with no labels.
• We derive the curriculum from metadata (using language models) and train SPCL with diversity learning scheme.
• Train 609 detectors over 400K weakly labeled videos.
• We manually evaluate their P@10 on a third dataset (MED).

\[
\begin{array}{|c|c|c|}
\hline
\text{YFCC 609} & \text{ImageNet 1000} & \text{UCF 101} \\
0.37608 & 0.2063 & <0.1 \\
\hline
\end{array}
\]

Detectors built on a large weakly labeled data set are more accurate than those built on a small labeled dataset.
Outline

- Introduction
- Related Work
- Proposed Approaches:
  - Indexing Semantic Features
  - Semantic Search
  - Video Reranking
  - Building Semantic Concepts
- Conclusions

Conclusions

- Proposed Work: hybrid search
Proposed Work

• Processing hybrid queries:
  – Preliminary studies showed hybrid query with 10 examples can be done efficiently on compressed semantic features.
    • Shoou-I Yu, Lu Jiang, Zhongwen Xu, Yi Yang, Alexander Hauptmann. Content-Based Video Search over 1 Million Videos with 1 Core in 1 Second. In ACM International Conference on Multimedia Retrieval (ICMR), 2015.
  – The method, however, is not scalable as it needs preloading lots of data into the memory.
  – We plan to integrate semantic search methods into hybrid search
    • Use the compressed semantic features.
    • Apply concept adjustment.
    • Apply semantic search to filter out irrelevant samples.
  – We will test the proposed methods on MED and YFCC datasets.

• Training concept detectors on the whole YFCC dataset (about 0.8 million videos.)
Schedule

• October – Jan, 2015. Study the efficient search model for hybrid search.
• February – March, 2016. Test the model and finish the experiments.
• April – September, 2016. Thesis writing and defense.
Published papers on the thesis topic


Key Contributions:

• The first-of-its-kind framework for web-scale content-based search over hundreds of millions of Internet videos [ICMR’15]. The proposed framework supports text-to-video, video-to-video, and text&video-to-video search [MM’12].

• A novel theory about self-paced curriculums learning and its application on robust concept detector training [NIPS’14, AAAI’15].

• Novel reranking algorithms for improving performance [MM’14, ICMR’14].

• A concept adjustment method representing a video by a few salient and consistent concepts that can be efficiently indexed by the modified inverted index [MM’15].
References


• Masoud Mazloom, Xirong Li, and Cees GM Snoek. Few-example video event retrieval using tag propagation. In ICMR, 2014.


• Jia Deng, Nan Ding, Yangqing Jia, Andrea Frome, Kevin Murphy, Samy Bengio, Yuan Li, Hartmut Neven, and Hartwig Adam. Large-scale object classification using label relation graphs. In ECCV, 2014.


References

• Kevin Tang, Vignesh Ramanathan, Li Fei-Fei, and Daphne Koller. Shifting weights: Adapting object detectors from image to video. In NIPS, 2012.
Applications

- It can benefit a variety of related tasks such as video summarization [7], video recommendation, video hyperlinking [8], social media video stream analysis [9], in-video advertising [10], etc.
Distributional Consistency: A Toy Example
Experiments on MED

Comparison of the full adjustment model with its special case Top-$k$ Thresholding on using IACC features.

<table>
<thead>
<tr>
<th>Method</th>
<th>$k$</th>
<th>Evaluation Metric</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>P@20</td>
</tr>
<tr>
<td>Our Model</td>
<td>50</td>
<td>0.0392</td>
</tr>
<tr>
<td>Top-$k$</td>
<td>50</td>
<td>0.0342</td>
</tr>
<tr>
<td>Our Model</td>
<td>60</td>
<td>0.0388</td>
</tr>
<tr>
<td>Top-$k$</td>
<td>60</td>
<td>0.0310</td>
</tr>
</tbody>
</table>
Example Queries

- Using the query, our system should be able to
  - Find simple objects, actions, speech words.
  - Search complex activities.
  - Answer questions by/in videos.

Information need:
What did we talk about in the last year’s forest camp?

Query (search videos in last year):
forest
AND (walking OR hiking)
OR tree
AND faces
AND asr:speech != empty
Concept Adjustment Model: Distributional Consistency

• A naive implementation $\rightarrow$ infeasible to solve.

\[ g(v; \alpha, \beta) = \frac{1}{2} \beta^2 \| v \|_0 \]

• Our general implementation:

\[ g(v; \alpha, \beta) = \alpha \beta \| v \|_1 + (1 - \alpha) \sum_{l=1}^{q} \beta \sqrt{p_l} \| v^{(l)} \|_2, \]

  – When $\alpha = 1$ $\rightarrow$ lasso (approximate $l_0$ norm).
  – When $\alpha = 0$ $\rightarrow$ group lasso (nonzero entries in a sparse set of groups)
  – When $\alpha \in (0, 1)$ $\rightarrow$ sparse group lasso (group-wise sparse solution, but only few coefficients in the group will be nonzero)
Experiments on YFCC

- We manually created queries for 30 products.
- Put commercials about the product to related video (in-video ads.)
- Evaluate the relevance of the top 20 returned results.

Queries and more results are available at:
https://sites.google.com/site/videosearch100m/
Experiments on YFCC
Self-paced Reranking (SPaR)

- The propose model:

\[
\min_{\Theta_1, \ldots, \Theta_m, y, v} \mathbb{E}(\Theta_1, \ldots, \Theta_m, v, y; C', k)
\]

\[
= \min_{y, v, w_1, \ldots, w_m, b_1, \ldots, b_m, \{\ell_{ij}\}} C \sum_{i=1}^{n} v_i \sum_{j=1}^{m} \ell_{ij} + \sum_{j=1}^{m} \frac{1}{2} \|w_j\|^2 + \text{regularizer}
\]

\[
\text{s.t. } \forall i, \forall j, y_i(w_j^T \phi(x_{ij}) + b_j) \geq 1 - \ell_{ij}, \ell_{ij} \geq 0
\]

\[
y \in \{-1, 1\}^n,
\]

\[
v \in [0, 1]^n,
\]

For example the Loss in the SVM model.

\[
\ell_{ij} = \max\{0, 1 - y_i \cdot (w_j^T \phi(x_{ij}) + b_j)\}
\]

\[
\Theta_1, \ldots, \Theta_m \quad \text{Reranking models for each modality.}
\]

\[
y \in \{-1, 1\}^n \quad \text{The pseudo label.}
\]

\[
v \in [0, 1]^n \quad \text{The weight for each sample.}
\]
Self-paced Reranking (SPaR)

- The propose model:

\[
\min_{\Theta_1, \ldots, \Theta_m, y, v} \mathbb{E}(\Theta_1, \ldots, \Theta_m, v, y; C, k) = \min_{y, v, \Theta_1, \ldots, \Theta_m} \quad \text{loss-function} \quad + mf(v; k)
\]

\[\quad \text{s.t.} \quad y \in \{-1, +1\}^n, \quad v \in [0, 1]^n, \quad \Theta_1, \ldots, \Theta_m\]

Reranking models for each modality.

The pseudo label.

The weight for each sample.

The self-paced is implemented by a regularizer.
Physically corresponds to learning schemes that human use to learn different tasks.

\(m\) is the total number of modality.

\(f\) is the self-paced function in self-paced learning."
Reranking in Optimization and Conventional Perspective

Q1: Why the reranking algorithm performs iteratively?
A: Self-paced learning mimicking human and animal learning process (from easy to complex examples).

Q2: Does the process converge? If so, to where?
A: Yes, to the local optimum. See the theorem in our paper.

Q3: Does the arbitrarily predefined weighting scheme converge?
A: No, but the weights by self-paced function guarantees the convergence.

Optimization perspective

1: \( t = 0; //\text{Iteration zero} \)
2: Choose starting values for \( y, v \);
3: while \( t \leq \text{max iteration} \) do
4: \( \Theta_1^{(t+1)}, ..., \Theta_m^{(t+1)} = \arg \max \mathbb{E}_{y,v}(\Theta_1^{(t)}, ..., \Theta_m^{(t)}; C) \);
5: \( y(t+1), v(t+1) = \arg \max \mathbb{E}_\Theta(y(t), v(t); k) \);
6: if \( t \) is small then increase \( 1/k \);
7: end while
8: return \([v_1 y_1, \cdots, v_n y_n]^T\);

Conventional perspective

1: \( t = 0; //\text{Iteration zero} \)
2: Choose the initial pseudo labels and weights;
3: while \( t \leq \text{max iteration} \) do
4: Train a reranking model on the fixed labels and weights;
5: Update the pseudo labels and weights;
6: if \( t \) is small then add more pseudo positives;
7: end while
8: return The list of samples after reranking;
Reranking in Optimization and Conventional Perspective

1: $t = 0$; //Iteration zero
2: Choose starting values for $y, v$;
3: while $t \leq \text{max iteration}$ do
4: $\Theta_1^{(t+1)}, ..., \Theta_m^{(t+1)} = \arg \max \mathbb{E}_{y,v}(\Theta_1^{(t)}, ..., \Theta_m^{(t)}; C)$;
5: $y^{(t+1)}, v^{(t+1)} = \arg \max \mathbb{E}_{\Theta}(y^{(t)}, v^{(t)}; k)$;
6: if $t$ is small then increase $1/k$;
7: end while
8: return $[v_1 y_1, \cdots, v_n y_n]^T$;

1: $t = 0$; //Iteration zero
2: Choose the initial pseudo labels and weights;
3: while $t \leq \text{max iteration}$ do
4: Train a reranking model on the fixed labels and weights;
5: Update the pseudo labels and weights;
6: if $t$ is small then add more pseudo positives;
7: end while
8: return The list of samples after reranking;

**Optimization perspective**

Q1: Why the reranking algorithm performs iteratively?
A: Self-paced learning mimicking human and animal learning process (from easy to complex examples).

Q2: Does the process converge? If so, to where?
A: Yes, to the local optimum.

Q3: Does the arbitrarily predefined weighting scheme converge?
A: No, but the weights by self-paced function guarantees the convergence.
Reranking in Optimization and Conventional Perspective

Q1: Why the reranking algorithm performs iteratively?
A: Self-paced learning mimicking human and animal learning process (from easy to complex examples).

Q2: Does the process converge? If so, to where?
A: Yes, to the local optimum. See the theorem in our paper.

Q3: Does the arbitrarily predefined weighting scheme converge?
A: Not guaranteed, but the discussed weights guarantees the convergence.
Finally, the adjusted concept representation is indexed by the inverted index. Indexing the real-valued score. Our index supports:

- **Modality search**: visual:dog, ocr:dog
- **Score range search**: score(dog, >=, 0.7)
- **Basic temporal search**: tbefore(dog, cat), twindow(3s,dog, cat)
- **Boolean logical search**: dog AND NOT score(cat, >=, 0.5)
Related Work

• Categorization of reranking methods:
  – Classification-based
    • [Yan et al. 2003] [Hauptmann et al. 2008][Jiang et al. 2014]
  – Clustering-based
    • [Hsu et al. 2007]
  – LETOR(LEarning TO Rank)-based
    • [Liu et al. 2008][Tian et al. 2008][Tian et al. 2011]
  – Graph-based
    • [Hsu et al. 2007][Nie et al. 2012]

Impact of the model parameters

![Graph showing the impact of model parameters on P@20, MRR, MAP@20, and MAP.]
The official results released by NIST TRECVID 2014 on MED14Eval (200,000 videos).
Limitations

• The learning philosophy may not apply to
Related Work

• Related problems:
  – Content-based Image Retrieval
  – Copy Detection
  – Semantic Concept Indexing / Action Detection
  – Multimedia Event Detection

(Disclaimer: brief overview of related problems)
Content-based Image Retrieval

- **Goal:** find visually similar images [Sivic et al 2006]
- **Query:** a single image (query-by-example)
- **Single Modality. Minimum semantic understanding.**
- **Instance search:** search the key frames about a specific instance [Zhu et al 2012]
Copy Detection/
Near Duplicate Detection

- **Goal**: find video copies derived from the input video, usually by means of transformations such as addition, deletion, formatting modification, etc [Over et al 2008].
- **Query**: a segment of video.
- **Multimodal. Minimum semantic understanding**.
Semantics Concept Detection/Action Detection

- **Goal:** find segments of video that contains the concept.
- **Query:** a concept name or ID.
  - **Simple Query.**
  - The key is to build accurate individual detectors.
  - **Need a lot of training data.**
Multimedia Event Detection (MED)

**Goal:** find video about certain complex events [Over 2014]. Initiated by NIST TRECVID in 2012.

**Query:** text or example videos about an event.

**Complex query.**

**Solving the problems need semantic understanding about video content (especially for semantic queries).**
Generalized MED Problem

- The proposed problem is a generalized Multimedia Event Detection (MED) problem.
- It is similar to MED but with the following differences:
  - The query can be about everything, not necessarily just an event.
  - Expand the boundary from large-scale to web-scale.