Video and Language
Bridging Video and Language with Deep Learning

Tao Mei
Senior Researcher, Microsoft Research Asia
http://research.microsoft.com/en-us/people/tmei/
ECCV-MM Invited Tutorial, 3:30-4:30PM, Oct 15, 2016
Computer Vision
Since the beginning of Artificial Intelligence

“Connect a television camera to a computer and get the machine to describe what it sees.”

—Marvin Minsky (1966)
Computer vision: 50 years of progress

1973: Fischler & Elschlager (Structural Representation)
1999: SIFT (Scaled Invariant Feature Transform), 35K citations
2001: Boosting + Cascade = Speed, 13K citations
2003: Constellation Model
2005: HOG (Histograms of Oriented Gradients), 9K citations
2007: SURF (Speed Up Robust Features), 7K citations
2008: PASCAL VOC, LabelMe
2009: ImageNet, Caltech Pedestrian
2012: DPM (Deformable Parts Model)
2010: SUN
2012: DNNs for ImageNet 1K, 7K citations
Present: DNN, Everywhere, MS COCO, MSR VTT, Visual Genome, ActivityNet, Sports 1M, YFCC100M, Youtube-8M, Open Images, ...

3
Image and video understanding: core problems

Segmentation (pixel)

Detection (region)

Classification (image/frame)

Captioning/VQA (image/clip)

Storytelling (collection)

Vision and Language

“person riding a horse in a field”

“we have a good time in the party...”
Deep learning to

“describe what a 3-year-old child sees”
  • Image/video recognition: classification, detection, segmentation

“describe what a 5-year-old child sees”
  • Vision to language
    • Image captioning
    • Video captioning & commenting
  • Visual question-answering
Image Captioning

“I think it's a boat is docked in front of a building.” [Microsoft CaptionBot]

“Sasha Obama, Malia Obama, Michelle Obama, Peng Liyuan et al. posing for a picture with Forbidden City in the background.” [Xiaodong He, 2016]
Video Captioning

“a group of people are dancing” [Pan and Mei, CVPR’16]

Video Commenting

“I love baseball”
“That’s how to play baseball”
“That’s an amazing play” [Li, Yao, Mei, MM’16]

“Not just beautiful”
“You are so beautiful”
“Goddess doesn’t need plastic surgery” [Li, Yao, Mei, MM’16]
Vision to Language

- Robotic vision
- Assist for blinded
- Incident report for surveillance
- Multimedia search
- Movie description for blinded
- Seeing chat bot
This tutorial will talk about

**vision**
- image

**language**
- caption
- comment
- alignment
- sentiment

**Additional topics:**
- datasets
- evaluations
- open issues
- materials
Outline

• Image and video captioning
  • caption = object localization/recognition + object relationship + language
    • nouns (objects, people, scenes)
    • adjectives (attributes)
    • verbs (actions)
    • prepositions (relationships)

• Video commenting
• Video and language alignment
• Datasets and evaluations
• Open issues
• Learning materials
Image captioning: basic idea

- Transforming an image to a vector in visual space
  - CRF, CNN, Semantic Vector, CNN+Attention
- Transforming description to a vector in semantic space
  - Collection of words (BoW), sequence of words (RNN)
- Creating an embedding space
  - Language template (FGM, ME), RNNs (Encoder-Decoder), LSTM

- Methodologies
  - Search-based
  - Language template-based
  - Sequence learning-based
    - Generation: learning-decoder
    - Translation: encoder-decoder
Image captioning: basic idea

Convolutional Neural Networks

AlexNet [A. Krizhevsky, 2012]

Inception [C. Szegedy, 2014]

VGG [K. Simonyan, 2015]

ResNet [K. He, 2015]

Inception-ResNet [C. Szegedy, 2016]
Image captioning: basic idea

Recurrent Neural Networks

- classification
- captioning
- translation (seq-2-seq)
- generative model

"a dog leaps a Frisbee on the grass"
Image captioning

- Search-based approach [Farhadi, ECCV10; Ordonez, NIPS11; Frome, NIPS13; Socher, NIPS14; Karpahy, CVPR15; Devlin, ACL15]
Image captioning

• Search-based approach [Farhadi, ECCV10; Ordonez, NIPS11; Frome, NIPS13; Socher, NIPS14; Karpahty, CVPR15; Devlin, ACL15]
Image captioning

- **Language template-based approach** [Feng, ACL10; Yang, EMNLP11; Kulkarni, PAMI13; Fang, CVPR15]

  **Image word detection (s-v-o)**
  Woman, crowd, cat, camera, holding, purple.

  **Language generation (maximum entropy)**
  A purple camera with a woman.
  A woman holding a camera in a crowd.
  A woman holding a cat.

  **Semantic re-ranking (deep embedding)**
  A woman holding a camera in a crowd.
Image captioning

- **Sequence learning-based approach**
  [Google15, Stanford15, Berkeley15, Baidu/UCLA15, UdeM15, Rochester]

*Note that this figure only shows prediction process.*
Image Captioning with X

**X = visual attention**
[Xu, ICML’15]

**X = visual attributes**
[You, CVPR’16, Wu, CVPR’16, Yao, arxiv’16]

**X = entity recognition**
[Tran, CVPR’16]

**X = dense caption**
[Johnson, CVPR’16]
Image Captioning with Visual Attention

- Image captioning with attention mechanism [Xu, ICML’15; Cho, 2015]
- Learning stochastic “hard” vs. deterministic “soft” attention

A woman is throwing a frisbee in a park.

Image Captioning with **Visual Attributes**

- **Visual attributes**: a high-level representation w/ concept detector responses
  - Video search with high-level concepts [TRECVID, 2006]
  - Object bank for image classification [Li & Fei-Fei, NIPS’10]
  - High-level concepts for captioning and question-answering [Wu & Shen, CVPR’16]

Images:

**Attributes:**
- [piano: 0.930] [hand: 0.71] [music: 0.672] [keyboard: 0.624]
- **LSTM**: a man is playing a **guitar**
- **LSTM-E**: a man is playing a **piano**

**Attributes:**
- [bananas: 1] [market: 0.995] [bunch: 0.553] [table: 0.51] [flowers: 0.454] [people: 0.431] [yellow: 0.377]
- **LSTM**: a group of people standing around a market.
- **A-LSTM**: a group of people standing around a bunch of **bananas**.

- **Joint learning w/ recognizable attributes**: relevance + coherence [Pan, CVPR’16]
  - Image captioning [A-LSTM]: explicitly emphasize attributes together with visual content
  - Video captioning [LSTM-E]: implicitly emphasize video content with “relevance” regularizer
A-LSTM: image captioning w/ attribute-LSTM [Yao & Mei, arxiv16]

\[ x^{-1} = T_v I \]
\[ x^t = T_s w_t + T_a A \]
\[ h^t = f(x^t) \]
Image captioning

- **Leaderboard** of MS COCO image captioning

  - Rank 1 in both external and internal ranking lists, in terms of all performance metrics (July 21)

- COCO dataset
  - 123,287 images (82,783 for training + 40,504 for validation)
  - 5 sentences per image (AMT workers)
### Attributes
- **dog**: 0.555
- **body**: 0.527
- **floating**: 0.484

### Generated Sentences
**LSTM**: a group of people on a boat in the water.

**CaptionBot**: I think it's a man with a small boat in a body of water.

**A-LSTM**: a man and a dog on a boat in the water.

### Ground Truth
1. an image of a man in a boat with a dog
2. a person on a rowboat with a dalmatian dog on the boat
3. old woman rowing a boat with a dog

---

### Attributes
- **bananas**: 1
- **market**: 0.995
- **outdoor**: 0.617
- **bunch**: 0.553
- **table**: 0.51
- **flowers**: 0.454
- **people**: 0.431
- **yellow**: 0.377

### Generated Sentences
**LSTM**: a group of people standing around a market.

**CaptionBot**: I think it's a bunch of yellow flowers.

**A-LSTM**: a group of people standing around a bunch of bananas.

### Ground Truth
1. bunches of bananas for sale at an outdoor market
2. a person at a table filled with bananas
3. there are many bananas layer across this table at a farmers market

---

### Attributes
- **flying**: 0.877
- **plane**: 0.598
- **airplane**: 0.528
- **lake**: 0.495
- **water**: 0.462
- **sky**: 0.443
- **red**: 0.426
- **small**: 0.365

### Generated Sentences
**LSTM**: a group of people flying kites in the sky.

**CaptionBot**: I think it's a plane is flying over the water.

**A-LSTM**: a red and white plane flying over a body of water.

### Ground Truth
1. a plane with water skies for landing gear coming in for a landing at a lake
2. a plane flying through a sky above a lake
3. a red and white plane is flying over some water
Image Captioning with **Semantic Attention (Attributes)**

- Instead of using the same set of attributes at every step, select attributes at each step. [You, CVPR’16]
Rich Image Captioning in the Wild [Tran, CVPR’16]

- Entity recognition: extreme classification w/ large set of celebrities (precision 99% coverage ~60%) [Guo, 2016]
- Language model: maximum entropy [Fang, CVPR15]
- Word tagging & feature: ResNet [He, CVPR16]
- Deep Structured Semantic Model [He, CIKM13]
“Sasha Obama, Malia Obama, Michelle Obama, Peng Liyuan et al. posing for a picture with Forbidden City in the background.” [Xiaodong He, 2016]
Dense Image Captioning [Johnson & Karpathy, CVPR16]

Figure courtesy of [Johnson, Karpathy, and Fei-Fei, CVPR16]
Challenges for video captioning

- Video captioning is much more complicated

- Learning video representation
  - frame: visual objects (AlexNet, GoogLeNet, VGG)
  - segment: temporal dynamics (3D CNN, optical flow)
  - video: pooling/alignment on frame and/or segment

- Sentence generation
  - multi-layer RNN (LSTM)
  - semantic relationship between entire sentence and video content
What if simply applying image captioning to video?

**Video-to-sentence:**

LSTM-E: a man is riding a motorcycle

---

**Image-to-sentence (keyframe-based):** [http://deeplearning.cs.toronto.edu/i2t](http://deeplearning.cs.toronto.edu/i2t)

- there is a black motorcycle sitting in front of a small amount of cars
- someone is holding a hole in the background
- a close up of a pair of scissors with his hand
- a man wearing a helmet is **racing**
- a flock of birds flying over the rock of water on a cliff
Video captioning

• Search (embedding)-based approach [Xu, AAAI15; Yu, ACL13 & AAAI15]

• Deep visual model to learn video representation

• Compositional language model to capture semantic compatibility among concepts

• Joint embedding model to minimize distance of the above two models in video-text space [Xu, AAAI15]

\[ J(V, T) = \sum_{i=1}^{N} (E_{\text{embed}}(V, T) + \sum_{p \in \mathcal{N}T} E_{\text{rec}}(p|W_m, W_r)) + r \]
Video captioning

- **Language model-based approach** [Thomason, COLING14; Barbu, UAI12; Rohrbach, ICCV13; Krishnamoorthy, AAAI13]

  - Predicting visual words (CRF): subject (S) – verb (V) – object (O)
  - Generating sentence with FGM [UAI12]: “determiner (a/the) – S – V – Prep (optional) – D – O (optional)”
  - Translating semantic representation (SR) to sentence [ICCV13]

  - **S:** man
  - **V:** play
  - **O:** guitar

  - “a man is playing a guitar”

• UC Berkeley [Donahue, CVPR’15]: CRF + LSTM encoder-decoder + LSTM (A/B) (GoogleNet + 3D CNN) + Soft-Attention + LSTM (B) (VGG + Optical Flow) + LSTM Encoder-Decoder + LSTM (A)
• UdeM [Yao, ICCV’15]: AlexNet + Mean Pooling + LSTM (B) (VGG + 3D CNN) + Mean Pooling + Relevance Embedding + LSTM (A)
• UT Austin [Venugopalan, ICCV’15]:
• UT Austin [Venugopalan, NAACL-HLT’15]:
Video Captioning with Attention

Encoder-decoder LSTM Networks with Temporal Attention [Yao, CVPR’15]

Video Paragraph Captioning with Hierarchical RNNs with Spatiotemporal Attention [Yu, CVPR’16]
Video Captioning with Semantics

• Key issues in sentence generation
  • relevance: relationship between sentence (S, V, O) semantics and video content
  • coherence: sentence grammar

LSTM: a man is playing a guitar
LSTM-E: a man is playing a piano

LSTM: a man is dancing
LSTM-E: a group of people are dancing

• Joint learning (LSTM-E): relevance + coherence [Pan, CVPR’16]
  • Explicitly and holistically emphasize video content with “relevance” regularizer
LSTM-E for video captioning [Pan & Mei, CVPR’16]

Input video → Frames → 2D CNN → Spatio-temporal descriptor $V$ → LSTM → Multi-view embedding → Relevance and Coherence Loss

$E(V, S) = (1 - \lambda) \times ||TvV - TsS||_2^2 - \lambda \times \sum_{t=0}^{N_s} \log Pr(w_t | v, w_0, \ldots, w_{t-1}; \theta; Tv; Ts)$

Relevance

Coherence

Sequence Learning:

Coherence Loss: $E_s = -\sum_{t=1}^{N_s} \log Pr_t(w_t)$

Relevance Loss: $E_e = ||TvV - TsS||_2^2$

Joint Learning: Relevance + Coherence ($Es + Ee$)
Evaluations

- **Dataset** ([MSR Video Description Corpus](https://www.microsoft.com/en-us/research/project/msr-vision-language-dataset/), a.k.a. YouTube2Text)
  - 1,970 Youtube video snippets (1,200 training, 100 validation, 670 testing)
  - 10-25 sec for each clip
  - ~40 human-generated sentences for each clip (by AMT)
  - dictionary: 15,903 -> 7,000; 45 S-groups, 218 V-groups, 241 O-groups
- **Training**: 12 hrs in one single CPU; testing: ~5 sec per clip

---

1. a man is petting a dog
2. a man is petting a tied up dog
3. a man pets a dog
4. a man is showing his dog to the camera
5. a boy is trying to see something to a dog

1. a man is playing the guitar
2. a man is playing instrument
3. a man plays a guitar
4. a man is singing and playing guitar
5. the boy played his guitar

1. a kitten is playing with his toy
2. a cat is playing on the floor
3. a kitten plays with a toy
4. a cat is playing
5. a cat tries to get a ball

1. a man is singing on stage
2. a man is singing into a microphone
3. a man sings into a microphone
4. a singer sings
5. the man sang on stage into the microphone
Performance of video captioning [Sept 2016]

The accuracy of S-V-O triplet prediction.

<table>
<thead>
<tr>
<th>Model</th>
<th>Team</th>
<th>Subject%</th>
<th>Verb%</th>
<th>Object%</th>
</tr>
</thead>
<tbody>
<tr>
<td>FGM</td>
<td>UT Austin, COLING (2014/08)</td>
<td>76.42</td>
<td>21.34</td>
<td>12.39</td>
</tr>
<tr>
<td>CRF</td>
<td>SUNY-Buffalo, AAAI (2015/01)</td>
<td>77.16</td>
<td>22.54</td>
<td>9.25</td>
</tr>
<tr>
<td>CCA</td>
<td>Stanford, CVPR (2010/06)</td>
<td>77.16</td>
<td>21.04</td>
<td>10.99</td>
</tr>
<tr>
<td>JEM</td>
<td>SUNY-Buffalo, AAAI (2015/01)</td>
<td>78.25</td>
<td>24.45</td>
<td>11.95</td>
</tr>
<tr>
<td>LSTM</td>
<td>UC Berkeley, NAACL (2014/12)</td>
<td>71.19</td>
<td>19.40</td>
<td>9.70</td>
</tr>
<tr>
<td>LSTM-E</td>
<td>MSRA, arxiv (2015/05)</td>
<td>80.45</td>
<td>29.85</td>
<td>13.88</td>
</tr>
</tbody>
</table>

The performance of sentence generation.

<table>
<thead>
<tr>
<th>Model</th>
<th>Team</th>
<th>METEOR%</th>
<th>BLEU@4%</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSTM</td>
<td>UC Berkeley, NAACL (2014/12)</td>
<td>26.9</td>
<td>31.2</td>
</tr>
<tr>
<td>SA</td>
<td>UdeM, arxiv (2015/02)</td>
<td>29.6</td>
<td>42.2</td>
</tr>
<tr>
<td>S2VT</td>
<td>UC Berkeley, arxiv (2015/05)</td>
<td>29.8</td>
<td>--</td>
</tr>
<tr>
<td>LSTM-E</td>
<td>MSR Asia, CVPR 2016</td>
<td>31.0</td>
<td>45.3</td>
</tr>
<tr>
<td>H-RNN</td>
<td>Baidu, CVPR 2016</td>
<td>32.6</td>
<td>49.9</td>
</tr>
<tr>
<td>HRNE</td>
<td>UTS, CVPR 2016</td>
<td>33.1</td>
<td>43.8</td>
</tr>
<tr>
<td>GRU-RCN</td>
<td>UdeM, ICLR 2016</td>
<td>31.6</td>
<td>43.3</td>
</tr>
</tbody>
</table>
### Dataset Information

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Organizer</th>
<th>Context</th>
<th>Source</th>
<th>#Video</th>
<th>#Clip</th>
<th>#Sentence</th>
<th>#Word</th>
<th>Vocabulary</th>
<th>Duration (hr)</th>
</tr>
</thead>
<tbody>
<tr>
<td>YouCook</td>
<td>SUNY Buffalo</td>
<td>Cooking</td>
<td>Labeled</td>
<td>88</td>
<td>-</td>
<td>2,668</td>
<td>42,457</td>
<td>2,711</td>
<td>2.3</td>
</tr>
<tr>
<td>TACos</td>
<td>MP Institute</td>
<td>cooking</td>
<td>Labeled</td>
<td>123</td>
<td>7,206</td>
<td>18,227</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>TACos M-L</td>
<td>MP Institute</td>
<td>cooking</td>
<td>Labeled</td>
<td>185</td>
<td>14,105</td>
<td>52,593</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>M-VAD</td>
<td>UdeM</td>
<td>movie</td>
<td>DVS</td>
<td>92</td>
<td>48,986</td>
<td>55,905</td>
<td>519,933</td>
<td>18,269</td>
<td>84.6</td>
</tr>
<tr>
<td>MPII</td>
<td>MP Institute</td>
<td>movie</td>
<td>DVS+Script</td>
<td>94</td>
<td>68,337</td>
<td>68,375</td>
<td>653,467</td>
<td>24,549</td>
<td>73.6</td>
</tr>
<tr>
<td>MSVD</td>
<td>MSR</td>
<td>multi-category</td>
<td>AMT workers</td>
<td>-</td>
<td>1,970</td>
<td>70,028</td>
<td>607,339</td>
<td>13,010</td>
<td>5.3</td>
</tr>
<tr>
<td>MSR-VTT (10K)</td>
<td>MSRA</td>
<td>20 categories</td>
<td>AMT workers</td>
<td>5,942</td>
<td>10,000</td>
<td>200,000</td>
<td>1,535,917</td>
<td>28,528</td>
<td>38.7</td>
</tr>
<tr>
<td>MSR-VTT (20K)</td>
<td>MSRA</td>
<td>20 categories</td>
<td>AMT workers</td>
<td>14,768</td>
<td>20,000</td>
<td>400,000</td>
<td>4,284,032</td>
<td>49,436</td>
<td>87.8</td>
</tr>
</tbody>
</table>

### Baselines

- **MP-LSTM** (VGG, AlexNet)
- **MP-LSTM** (C3D + VGG)
- **SA-LSTM** (VGG, AlexNet)
- **SA-LSTM** (C3D + VGG)
- LSTM-E

---

1. A black and white horse runs around.
2. A horse galloping through an open field.
3. A horse is running around in green lush grass.
4. There is a horse running on the grassland.
5. A horse is riding in the grass.

1. A woman giving speech on news channel.
2. Hillary Clinton gives a speech.
3. Hillary Clinton is making a speech at the conference of mayors.
4. A woman is giving a speech on stage.
5. A lady speak some news on TV.

1. A child is cooking in the kitchen.
2. A girl is putting her finger into a plastic cup containing an egg.
3. Children boil water and get egg whites ready.
4. People make food in a kitchen.
5. A group of people are making food in a kitchen.

1. A player is putting the basketball into the post from distance.
2. The player makes a three-pointer.
3. People are playing basketball.
4. A 3 point shot by someone in a basketball race.
5. A basketball team is playing in front of spectators.
Microsoft Video to Language Challenge

77 teams registered challenge
22 teams submitted results
Awards will be announced at ACM MMM
Summary from Video to Language Grand Challenge 2016

- CNN-LSTM [1, 2, 4, 5, 7]
- Sequence-to-Sequence (encoder-decoder) [3, 6, 9, 10]

- Image features
  - VGG-19 [1][2][5][6][9][10]
  - GoogleNet [2][4][5]
  - ResNet [3][5][8]
  - VGG-16 [5][7][8]
  - PlaceNet [5][9]

- Motion features
  - C3D [1][2][3][4][5][9][10]
  - IDT [1][2]
  - Optical flow [8]
- Acoustic features
  - MFCCs [1][3][7]

- Text features
  - ASR [1]
  - Video category [3][4]

<table>
<thead>
<tr>
<th></th>
<th>B@4</th>
<th>MET.</th>
<th>ROU.</th>
<th>CID.</th>
</tr>
</thead>
<tbody>
<tr>
<td>VGG+C3D</td>
<td>32.3</td>
<td>25.8</td>
<td>56.7</td>
<td>29.6</td>
</tr>
<tr>
<td>VGG+C3D+Aug.</td>
<td>33.3</td>
<td>26.6</td>
<td>57.2</td>
<td>32.5</td>
</tr>
<tr>
<td>VGG+C3D+Res.</td>
<td>34.6</td>
<td>26.9</td>
<td>58.3</td>
<td>37.9</td>
</tr>
<tr>
<td>VGG+C3D+Res.+Aug.</td>
<td>35.3</td>
<td>27.4</td>
<td>58.9</td>
<td>38.3</td>
</tr>
<tr>
<td>VGG+C3D+Res.+Tra.</td>
<td>36.5</td>
<td>27.1</td>
<td>59.2</td>
<td>40.3</td>
</tr>
<tr>
<td>VGG+C3D+Res.+Aug.+Tra.</td>
<td>35.6</td>
<td>27.0</td>
<td>58.9</td>
<td>38.1</td>
</tr>
</tbody>
</table>


<table>
<thead>
<tr>
<th>Descriptors</th>
<th>BLEU@4</th>
<th>METEOR</th>
<th>CIDEr</th>
<th>ROUGE-L</th>
</tr>
</thead>
<tbody>
<tr>
<td>categories</td>
<td>0.298</td>
<td>0.228</td>
<td>0.236</td>
<td>0.548</td>
</tr>
<tr>
<td>audio</td>
<td>0.301</td>
<td>0.222</td>
<td>0.184</td>
<td>0.544</td>
</tr>
<tr>
<td>C3D</td>
<td>0.374</td>
<td>0.264</td>
<td>0.389</td>
<td>0.594</td>
</tr>
<tr>
<td>ResNet</td>
<td>0.389</td>
<td>0.269</td>
<td>0.400</td>
<td>0.605</td>
</tr>
<tr>
<td>+C3D</td>
<td>0.385</td>
<td>0.267</td>
<td>0.411</td>
<td>0.601</td>
</tr>
<tr>
<td>+categories</td>
<td>0.381</td>
<td>0.270</td>
<td>0.418</td>
<td>0.597</td>
</tr>
<tr>
<td>+audio</td>
<td>0.395</td>
<td>0.277</td>
<td>0.442</td>
<td>0.610</td>
</tr>
<tr>
<td>committee</td>
<td>0.407</td>
<td>0.286</td>
<td>0.465</td>
<td>0.610</td>
</tr>
</tbody>
</table>

• Other observations
  • Additional training data from MS-COCO [2][7][8]
  • Additional data from FCVID [4]
  • Additional data from Youtube2Text [9]
  • Captioning with tag based sentence reranking [4]
  • Data augmentation (sampling from different frames and horizontally flipped frames) [5]
  • Use PCA to reduce the dimensionality of low-level feature [8]
Outline

• Image and video captioning
• **Video commenting**
• Video sentiment analysis
• Video and language alignment
• Datasets and evaluations
• Open issues
• Learning materials
Video commenting [Li, MM’16]

Output comments:

- It is amazing!
- Haha haha lol.
- Wow sooo cool!
- hahaha this is awesome!
- This is so good.
- OMG!

- General-purpose phases often appear
  “It is amazing.” “OMG that was awesome!” “That is cool!”
- Comments in the training data are very diverse
  “I love how you ride a skateboard.” “After I saw this I wish I could skate board.”
- Difficult to establish a mapping from video to comments
Video commenting

- Video Commenting by Search and Multi-View Embedding [Li, MM’16]
  - Similar video search (VS)
  - Comment dynamic ranking (DR)
Video commenting

- Video Commenting by Search and Multi-View Embedding [Li, MM’16]
  - Similar video search (VS)
  - Dynamic ranking of comments (DR)

- Ranking loss

\[
\min \sum_{(c_k, v_k^+, v_k^-) \in T} \max(0, 1 + \| F_i - F_j \|_p^2 - \| F_{i,neg} - F_j \|_p^2)
\]

s.t.  \( i, j = 1, ..., 3 \),  \( i \neq j \),  \( i \neq 3 \).

- Prediction

\[
r(\hat{v}, \hat{c}) = \| F_1(\hat{v}) - F_3(\hat{c}) \|_p^2 + \| F_2(\hat{v}) - F_3(\hat{c}) \|_p^2.
\]
Video commenting

• Dataset
  • 102K videos from vine.com
  • 10.6M comments from 12 categories
  • 5~15 sec for each video clip

• Video representation
  • Video content: C3D, VGG, C3D + VGG
  • Comments: TF, word2vector
  • Visual sentiment: ANP (adj-noun pairs)

• Approaches
  • Random Selection (RS)
  • Two-view CCA (CCA-VT)
  • Three-view CCA (CCA-VST)
  • Deep Two-view Embedding (DE-VT)
  • Deep Three-view Embedding (DE-VST)

<table>
<thead>
<tr>
<th>Approach</th>
<th>mAP@1</th>
<th>mAP@2</th>
<th>mAP@3</th>
<th>mAP@4</th>
<th>mAP@5</th>
</tr>
</thead>
<tbody>
<tr>
<td>RS</td>
<td>0.259</td>
<td>0.244</td>
<td>0.219</td>
<td>0.203</td>
<td>0.191</td>
</tr>
<tr>
<td>CCA-VT</td>
<td>0.458</td>
<td>0.421</td>
<td>0.399</td>
<td>0.389</td>
<td>0.382</td>
</tr>
<tr>
<td>CCA-VST</td>
<td>0.501</td>
<td>0.465</td>
<td>0.439</td>
<td>0.429</td>
<td>0.419</td>
</tr>
<tr>
<td>DE-VT</td>
<td>0.504</td>
<td>0.469</td>
<td>0.447</td>
<td>0.433</td>
<td>0.422</td>
</tr>
<tr>
<td>DE-VST</td>
<td>0.549</td>
<td>0.513</td>
<td>0.486</td>
<td>0.471</td>
<td>0.459</td>
</tr>
</tbody>
</table>

"Haha so cute and funny at the same time"
"Glad she is better. So cute"
"Such outstanding piano pieces and you play them sublimely :)
"Amazing. I was listening to this while studying!"

The mAP@1 performance for all the 12 categories.
Test video:

* 不止漂亮 0.522
Not just beautiful
* 你好漂亮 0.497589
You are so beautiful
* 好美 喜欢看自拍视频的 0.4942
Gorgeous. Love to watch homemade video
* 心目中的女神是不整容的 0.4904
Goddess doesn't need plastic surgery
* 美丽！ 0.4857
Beautiful

Top-K similar videos:

* 很漂亮 so beautiful
* 正在笑着的美女 beautiful smile
* 哥哥的漂亮 pretty
* 那里出的美女 where did this beautiful lady come from
* 好美啊 so beautiful

* 不止漂亮 Not just beautiful
* 好美，喜欢看自拍视频的 Goddess doesn't need plastic surgery
* 有点韩国人的感觉 Looks a bit like Korean
* 闪眼，真美 Catches the eyes, so pretty
* 美丽的 Beautiful

* 你好漂亮 You are so beautiful
* 心目中的女神是不整容的 Goddess doesn't need plastic surgery
* 很好看，没有大浓妆 but it catches the eyes
* 女神 Goddess
* 美哒哒 Beautiful

* 五官真看得到 Beautiful facial
* 美女耶 Pretty lady
* 你好自恋哦！美女 You are such a narcissist
* 美女 Beautiful lady
* 大众美女脸 Generally beautiful face

* 美丽！ Beautiful
* 美美哒 Beautiful
* 白衬衣美哭 The white shirt is so pretty
* 太阳女神美美哒 The Goddess of Sun is beautiful
* 美翻了啦 Outrageously beautiful

* 今天吃得好淑女 0.4519
Eating like a lady with great manner
* 今天吃得特别干净了 0.4238
Getting better at learning how to eat
* 好想亲下momo的小嘴嘴 0.3901
Want to kiss momo's little lips
* 今天吃得特别干净了 0.3600
Eating very enjoyable
* 不在吃饭的时间教他说话 Don't teach him talking while eating
* 不要太吃 过度吃
It is enjoyable just to watch someone eats

* 吃的真香 Enjoying the yummy food
* 吃得吧唧吧唧 Eating very enjoyable
* 美食 与美食家 Beautiful food
* 吃的东西吃得多 Does it include rice noodles?
* 吃的真文明 Eating with such great manner
* 吃的很干净 Like momo's clean eating
* 吃的很干净 Like momo's clean eating
* 吃的真的很好吃 Enjoying the yummy food
* 吃的真文明 Eating with such great manner

* 吃的越来越干净了 Getting better at learning how to eat
* 他被momo的小嘴嘴吻了 Want to kiss momo's little lips
* 看着momo 吃的很干净 Enjoyable just to watch someone eats
* momo喜欢她的食物 Momo is enjoying her food
* 14 months 14 months
Results: auto-commenting

* The eyebrow is pretty 0.5613
  * Beautiful 0.5388
  * Still looks so pretty 0.5314
  * Candy to the eyes 0.5285
  * Very beautiful 0.5189

* Such a beautiful daughter 0.4469
  * What a cute and beautiful baby 0.4335
  * It's too pretty 0.4274
  * Such a beautiful baby 0.4237
  * Baby is the most beautiful gift of the whole world 0.4181

* What kind of dog is this? very cute 0.4884
  * Is this a dog? 0.4714
  * It looks exactly like my dog. Even the way they look at you is alike 0.4588
  * Your dog is so cute, beautiful lady 0.4573
  * Cute puppy 0.4571

* Beautiful manicure takes you into spring 0.4156
  * Bohemian manicure 0.4014
  * Will do this manicure next time 0.3654
  * Beautiful manicure 0.3626
  * How do you call those tools used for manicure? 0.3572

* Behave so much better than my Samoyed 0.6156
  * This is Samoyed, right? 0.5723
  * So cute that I miss my own Samoyed 0.5272
  * The puppy Samoyed is the cutest 0.4863
  * I want a Samoyed indeed 0.4768

* Little cute 0.4643
  * The hat is so cute 0.4201
  * The eyes are so beautiful. It's too cute and I love it so much 0.4102
  * Baby looks so handsome with the hat on. So cute 0.3950
  * Such a cute little baby 0.3927

* Mr. Guitar is enjoying it too much 0.4779
  * Sounds wonderful, hope that I can hear the whole version of each song 0.4715
  * I am moved by the guitar player 0.4507
  * Want to hear the final version 0.4373
  * Sounds fantastic when put together 0.4341

* It's pretty and I love ancient cloth too 0.4610
  * Beautiful Goddess 0.4395
  * Super beautiful 0.4253
  * it is beautiful 0.4145
  * Beautiful 0.4142

* Such a cute kitty 0.6174
  * What kind of cat is this? Too cute 0.6095
  * It looks too comfortable and makes me want to be a cat too 0.5817
  * Is it Garfield? 0.5575
  * What cat is this? So cute 0.5537
Outline

• Image and video captioning
• Video commenting
• Video and language alignment
• Datasets and evaluations
• Open issues
• Learning materials
Alignment of Video and Language

- Alignment of language instructions with video segments [Naim, AAAI’14]
  - Aligning nouns to video blobs
  - Model: HMM + IBM 1 [Brown, CL’93]
Outline

• Image and video captioning
• Video commenting
• Video and language alignment
• **Datasets and evaluations**
• Open issues
• Learning materials
Datasets for Captioning

Note: The class information is unknown for Flickr 8K/30K, SBU, and MSVD, MPII-MD, M-VAD, TGIF.
That's so cute where he's waving the flag
Poor Baby but it was so funny
he's so cute

Haha so cute and funny at the same time
Glad she is better. So cute
Soo awesome and cute

I love baseball
That's how to play baseball
That an amazing play!

Such outstanding piano pieces and you play them sublimely :)
Amazing. I was listening to this while studying!
Keep it up that's wonderful!
Evaluation metrics for captioning

- **Objective metrics**
  - Accuracy of $S\%$, $V\%$, $O\%$
  - ROUGE-L (Recall-Oriented Understudy for Gisting Evaluation) [Lin, 04]
  - BLEU@4 (BiLingual Evaluation Understudy) [Papineni, ACL’02]  
    modified n-gram precision
  - METEOR (Metric for Evaluation of Translation with Explicit ORdering) [Banerjee, ACL05]  
    similar with $f$-score combining precision and recall with a weight
  - CIDEr (Consensus-based Image Description Evaluation) [Vedantam, 2014; COCO evaluation]

- **Subjective metrics – human evaluations**
  - Coherence, Relevance, Helpful for Blind [MSR Video to Language]
Open issues for vision to language

- Rule-based vs. Model-based vs. Data-driven approaches
  - More accurate object/action detection/recognition from videos
- Leveraging more powerful language models
  - Attention model
  - Bi-directional RNN
- Diversity/Natural
  - Sentiment analysis (e.g., adjective-noun pair)
  - Attributes of object (e.g., human body parsing, age)
  - Entity recognition (e.g., celebrity naming, face recognition)
- Multimodal data analysis (e.g., script, speech, audio, comments)
- Visual relationship modeling [Lu, ECCV’16]
- Complex and long videos
  - Data collection from weakly supervised Web data
Reference

- ... ...
Learning materials

- Source codes for image captioning:
  - [https://github.com/karpathy/neuraltalk](https://github.com/karpathy/neuraltalk), [https://github.com/karpathy/neuraltalk2](https://github.com/karpathy/neuraltalk2)
  - LRCN for image caption: [https://github.com/jeffdonahue/caffe/tree/54fa90fa1b38af14a6fca32ed8aa5ead38752a09/examples/coco_caption](https://github.com/jeffdonahue/caffe/tree/54fa90fa1b38af14a6fca32ed8aa5ead38752a09/examples/coco_caption)
  - Show attend and tell: [https://github.com/kelvinxu/arctic-captions](https://github.com/kelvinxu/arctic-captions)

- Source codes for video captioning:
  - Sequence to Sequence - Video to Text: [https://github.com/vsubhashini/caffe/tree/recurrent/examples/s2vt](https://github.com/vsubhashini/caffe/tree/recurrent/examples/s2vt)
  - Soft-attention: [https://github.com/yaoli/arctic-capgen-vid](https://github.com/yaoli/arctic-capgen-vid)
Acknowledgement

• **Ting Yao, Jianlong Fu, Yong Rui, Xiaodong He**
  Microsoft Research

• **Yingwei Pan, Jun Xu, Houqiang Li**
  University of Science and Technology of China

• **Jiebo Luo**
  University of Rochester, USA
Thanks!

tmei@microsoft.com