Video-Based Inference

Objectives

Objective

Describe unique challenges in using deep networks for sequential data

Objective

Describe the difference between image-based and video-based classification tasks

Objective

Explain the value of using video action recognition to contrast the difference between image-based and video-based classification tasks

Objective

Evaluate a video-based classification example using deep learning

Going from Image to Video

Solution Naïve

- Processing each frame of a video as an independent image and then aggregating the frame-level results

Solution Better

- Extracting spatio-temporal features and an inference task will be based on such features

Video2Vec: Sample Applications

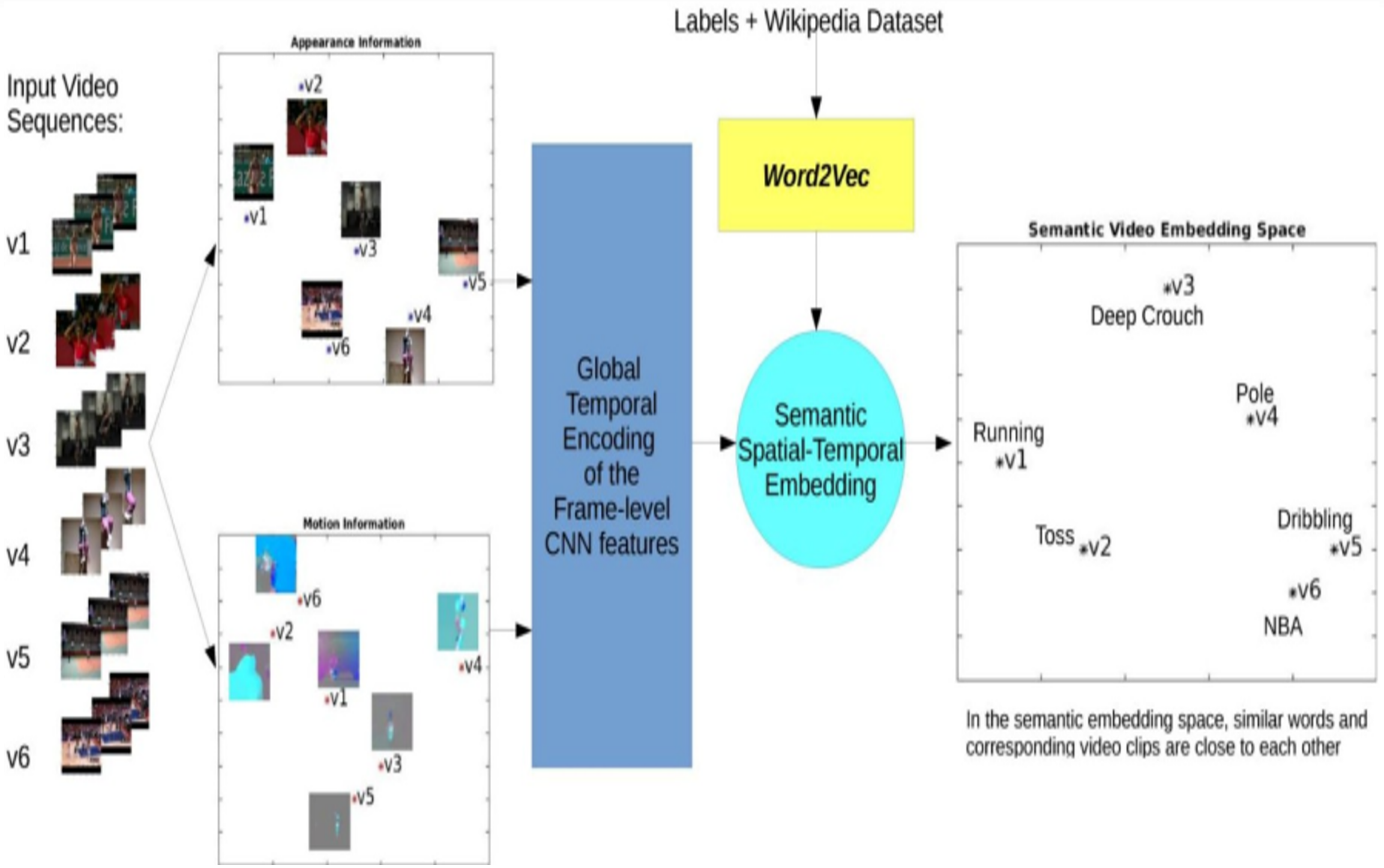
We examine a deep learning approach for finding video representations that naturally encode spatial-temporal semantics.

- Mostly based on the following papers:

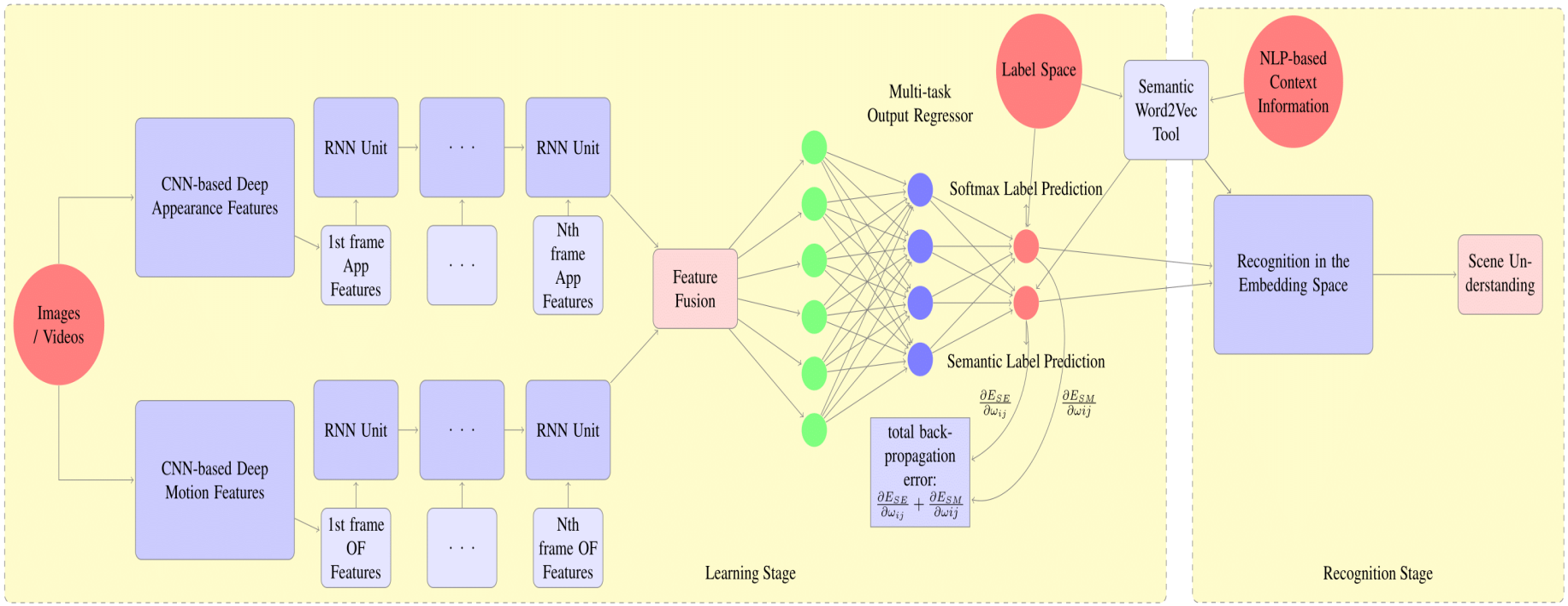
Yikang Li, Sheng-hung Hu, Baoxin Li, “Recognizing Unseen Actions in a Domain-Adapted Embedding Space”, ICIP, Sep 2016.

Yikang Li, Sheng-hung Hu, Baoxin Li, “*Video2Vec*: Learning Semantic Spatio-Temporal Embeddings for Video Representations”, ICPR, Dec 2016.

Video2Vec Deep Learning Model: Key Idea



Video2Vec Deep Learning Model: Implementation



- A two-stream CNN for extracting appearance and optical flow features
- RNNs for further global spatial-temporal encoding
- A MLP for final semantic embedding space

Applications of the Model

- Visual tasks:
  a. Video Action Recognition
  b. Zero-Shot Learning
  c. Semantic Video Retrieval

- Dataset: UCF101 dataset (13320 video clips from 101 categories; training/testing ratio is 7:3; the split list is provided by its own web)

Additional Implementation Details – 1/4

Pretraining for the component models:

- **Pre-trained Spatial CNN Model:** VGG-f trained on ImageNet
- **Pre-trained OF CNN Model:** Flow-net trained on UCF Sports
- **Pre-trained Word2Vec Model:** Wikipedia corpus contained 1 billion words

Additional Implementation Details – 2/4

Deep model parameter settings:

- **CNNs:** Pretrained model + the last layer (fc7) features (dimension: 4096x1)
- **RNNs:** Hidden layer size is 1024x1
- **MLP:** Input layer size (2048x1), hidden layer size (1200x1), output layer size (500x1)

Loss function:

- Hinge loss function for semantic embedding
- Softmax loss function for fine-tuning and classification

Additional Implementation Details – 3/4

Video processing settings:

- Dense Optical Flow and RGB frames are extracted at 10fps.
- Building Video Sequence Mask for each training batch to make each sequence the same length.

Additional Implementation Details – 4/4

Training parameter settings:

- Learning rate: initialized as 0.0001 and reduced by half each 15 epochs
- Total epoch: 60 epochs
- Batch size: 30 video clips
- Margin value for Hinge Loss function:
  - a. For zero-shot learning, 0.4
  - b. For video retrieval and action recognition, 0.55

Summaries of Key Results

Dataset: UCF101 dataset

Zero-shot learning results

- The model achieved state-of-the-art performance on ZSL even without any domain-adapted strategy.

Video action recognition

- The performance was on par with those with sophisticated fusion strategies or deeper networks.

Additional Results

- The task is to retrieve videos from training dataset by using query words that never appear in the training stage but share some information with training labels.
- The results show the top 10 retrieval video clips among video dataset.

Query Labels	Top10 Retrieve Results	Query Labels	Top10 Retrieve Results
NBA	Basketball Dunk (10)	Extreme	Rock Climbing Indoor (5), Uneven Bars (2), Soccer Juggling (2), Pole Vault (1)
Orchestra	Playing Cello (9), Playing Piano (1)	Tide	Cliff Diving (4), Surfing (2), Throw Discus (2), Sky Diving (1), Rafting (1)
Army	Military Parade (10)	India	Paying Tabla (4), Playing Sitar (2), Head Massage (1), Cricket Shot (1), Mixing (1)
Music	Playing Sitar (9), Playing Piano (1)	Celebrate	Military Parade (6), Long Jump (1), Band Marching (1), Ice Dancing (1), Blowing Candles (1)
Computer	Typing (10)	Home-run	Baseball Pitch (5), Basketball Dunk (3), Field Hockey Penalty (1), Frisbee Catch (1)
Park	Biking (9), Golf Swing (1)	Boat	Kayaking (4), Rafting (2), Rowing (2), Cliff Diving (1), Push Ups (1)
Summit	Cliff Diving (7), Skiing (2), Rope Climbing (1)	Toy	Yo-yo (4), Nun chucks (4), Pull Ups (1), Juggling Ball (1)
School	Skate Boarding (10)	Snow	Skiing (2), Ice Dancing (2), Cricket Bowling (1), Pole Vault (1), Blowing Candles (1), Blow Dry Hair (1), Rafting (1), Sky Diving (1)
Park	Biking (9), Golf Swing (1)	Acrobatics	Juggling Balls (5), Soccer Juggling (5)
Water	kayaking (10)	Ocean	Cliff Diving (4), Sky Diving (3), Kayaking (2), Rafting (1)
FIFA	Soccer Penalty (8), Soccer Juggling (2)	Hurl	Throw Discus (2), Mopping Floor (2), Baby Crawling (1), Javelin Throw (1), Cricket Shot (1), Blowing Candles (1), Pull Ups (1)
Club	Golf Swing (8), Soccer Juggling (2)	Hiking	Biking (5), Kayaking (4), Rafting (1)
Nature	Tai Chi (7), Hammering (2), Walking with Dog (1)	Swim	Diving (5), kayaking (3), Cricket Bowling (1), Sky Diving (1)
Beethoven	Playing Cello (8), Playing Voilin (2)	Jogging	Biking (5), Skate Boarding (2), Soccer Juggling (1), Skiing (1), Ice Dancing (1)
Classical	Playing Cello (7), Playing Voilin (3)	Foam	Blowing Candles (7), Pull Ups (1), Rope Climbing (1), Juggling Balls (1)
Yankees	Baseball Pitch (10)	Hip-hop	Trampoline Jumping (6), Swing (4)
Duel	Boxing Punching Bag (8), Punch(2)	Scramble	Pull Ups (6), Trampoline Jumping (2), Rope Climbing (1), Cricket Shot (1)
Lifting	Body Weight Squats (4), Rope Climbing (4), Pull Ups (2)	Mat	Rope Climbing (4), Pommel Horse (3), Trampoline Jumping (2), Javelin Throw (1)
Martial	Fencing (3), Archery (3), Boxing Punching Bag (3), Balance Beam (1)	Parachuting	Diving (6), Cricket Bowling (2), Hand Stand Walking (1), Sky Diving (1)
Tumbling	Trampoline Jumping (8), Throw Discus (1), Frisbee Catch (1)	Hunting	Horse Riding (3), Kayaking (3), Nun chucks (3), Frisbee Catch (1)