Video Captioning and Retrieval Models with Semantic Attention

Winner for Movie annotation & retrieval,
Movie fill in the blank

Youngjae Yu   Hyungjin Ko
Jongwook Choi   **Gunhee Kim**

SEOUL NATIONAL UNIV.
VISION & LEARNING

October 16, 2016

Outline

- Objective and Key Ideas
- Approaches
  - A Model for Fill-in-the-Blank
  - A Model for Multiple Choice Test
  - A Model for Retrieval
  - A Model for Description
- Experiments

Objective

Participate in all the tasks in three tracks of LSMDC 2016

Fortunately, we have own three of them

movie description

His vanity license plate reads 732.

movie multiple-choice

movie fill-in-the-blank

movie retrieval

Key Ideas – Semantic Attention

A separate model for each task

Take advantage of state-of-the-art techniques in our base models

Adopt ***semantic attention*** to strengthen meaning of words

- Extract concepts or attributes and selectively attend on them
- Successfully applied to image captioning [You et al. CVPR 2016]
- Input words for more semantic representation
- Output words for more accurate prediction

Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image Captioning with Semantic Attention. CVPR 2016.

Extracting Attributes (or Concepts)

Goal: Extract a set of attribute (or concept) words per video

Video

Attribute Set

{*Man*, *car*, *running*, *road* …}

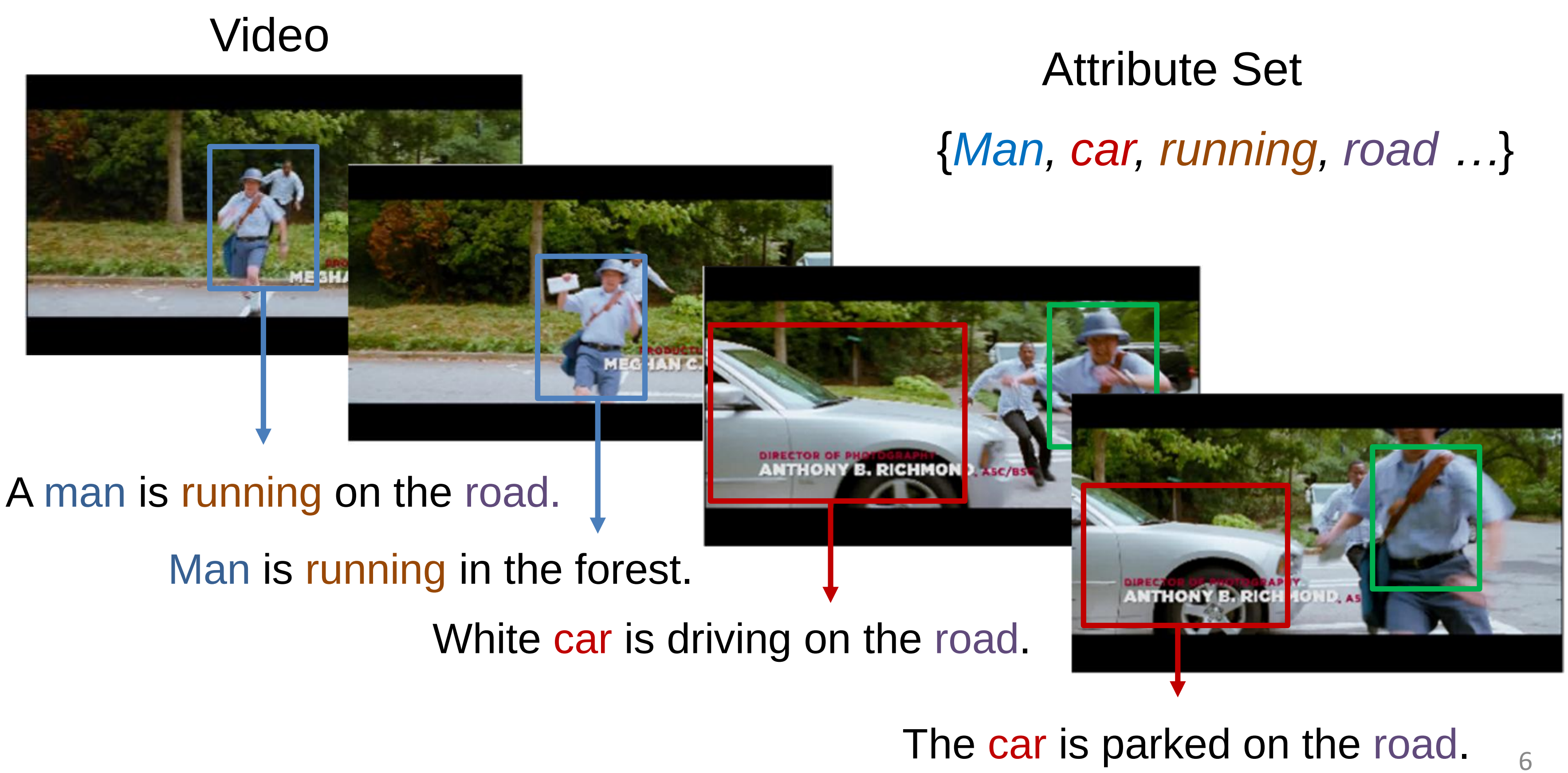
Our approach

- Obtain spatially-located sentences using the DenseCap2 model pretrained on VisualGenome dataset
- Select top-*K* words (K=20) that occur most continuously and frequently
- Much room for improvement

Justin Johnson, Andrej Karpathy, Li Fei-Fei, DenseCap: Fully Convolutional Localization Networks for Dense Captioning, CVPR 2016

Extracting Attributes (or Concepts)

Among the language descriptions, we select top K (= 20) words that occur most continuously and frequently



Extracting Attributes (or Concepts)

More detail of selecting top K (= 20) words

- If overlapped region is bigger than 20% of smaller box, it is considered as continuous detection
- We increase the word count if it appears again in the overlapped regions over consecutive 10 frames

A man is running on the road. Man is running in the forest.

White car is driving on the road. The car is parked on the road.

Representation of Video Frames and Text

Video: Use the conv5b layer of ResNet

- Sample one frame per 10 frames (max length = 40)
- Represent a video by $\{\boldsymbol{v}_i\}_{i=1}^{N}$ where each $\boldsymbol{v}_i \in \mathbb{R}^{2,048}$

Text: Use the word2vec skip-gram embedding E

- Build a dictionary *V* with size of 12,486 (the words that occur more than three times in the training set)
- Represent a word by multiplying its one-hot vector by $E \in \mathbb{R}^{d \times V}$ where $d = 300$

Outline

- Objective and Key Ideas
- Approaches
  - A Model for Fill-in-the-Blank
  - A Model for Multiple Choice Test
  - A Model for Retrieval
  - A Model for Description
- Experiments

Question for Fill-in-the-Blank

Given a video clip and a sentence with a blank in it, predict a single correct word to fill in the blank

Sentence : SOMEONE _____ over at his sullen face, then smiles.

Answer : glances

Model's input and output

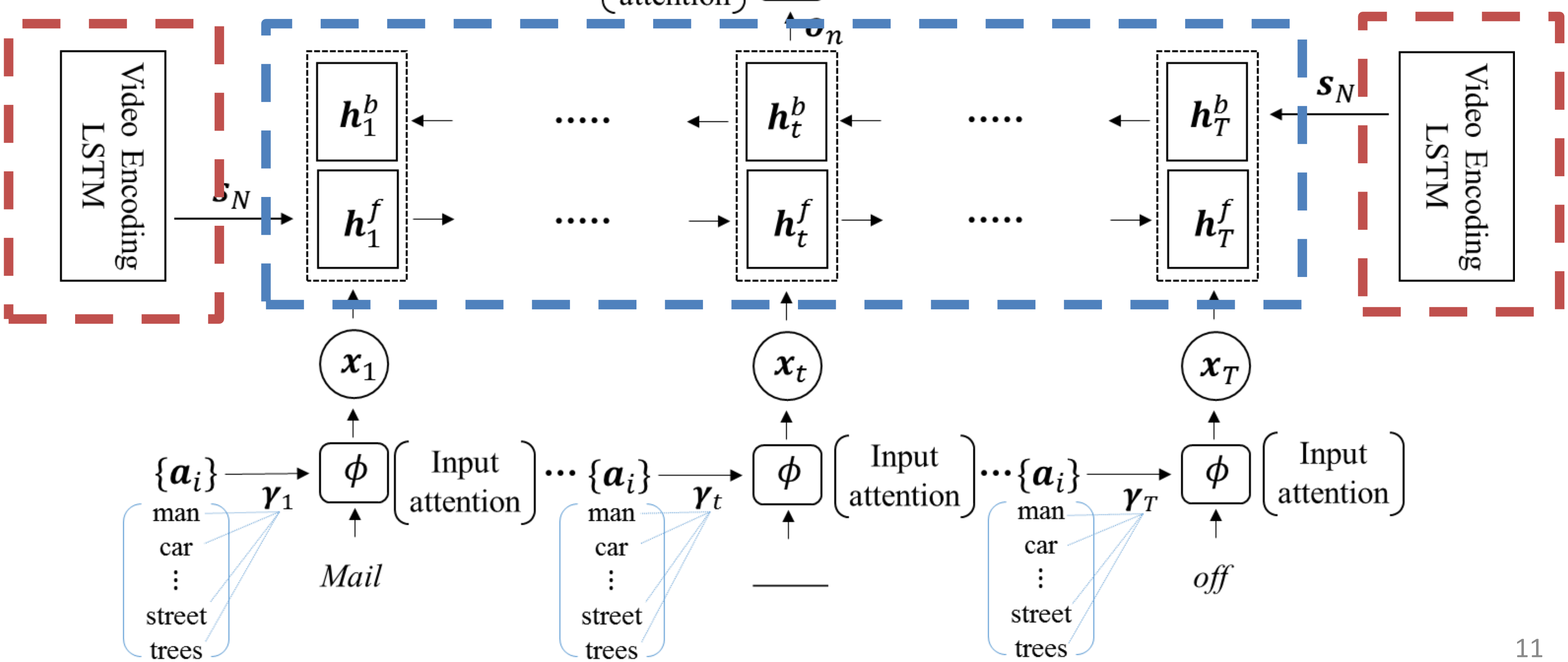
- Input: (1) a video, (2) a sequence of words with a blank
- Output: a single word

Model for Fill-in-the-Blank

The base model is Bidirectional LSTM

- Consider both forward and backward context for a blank

Input video is encoded by LSTM

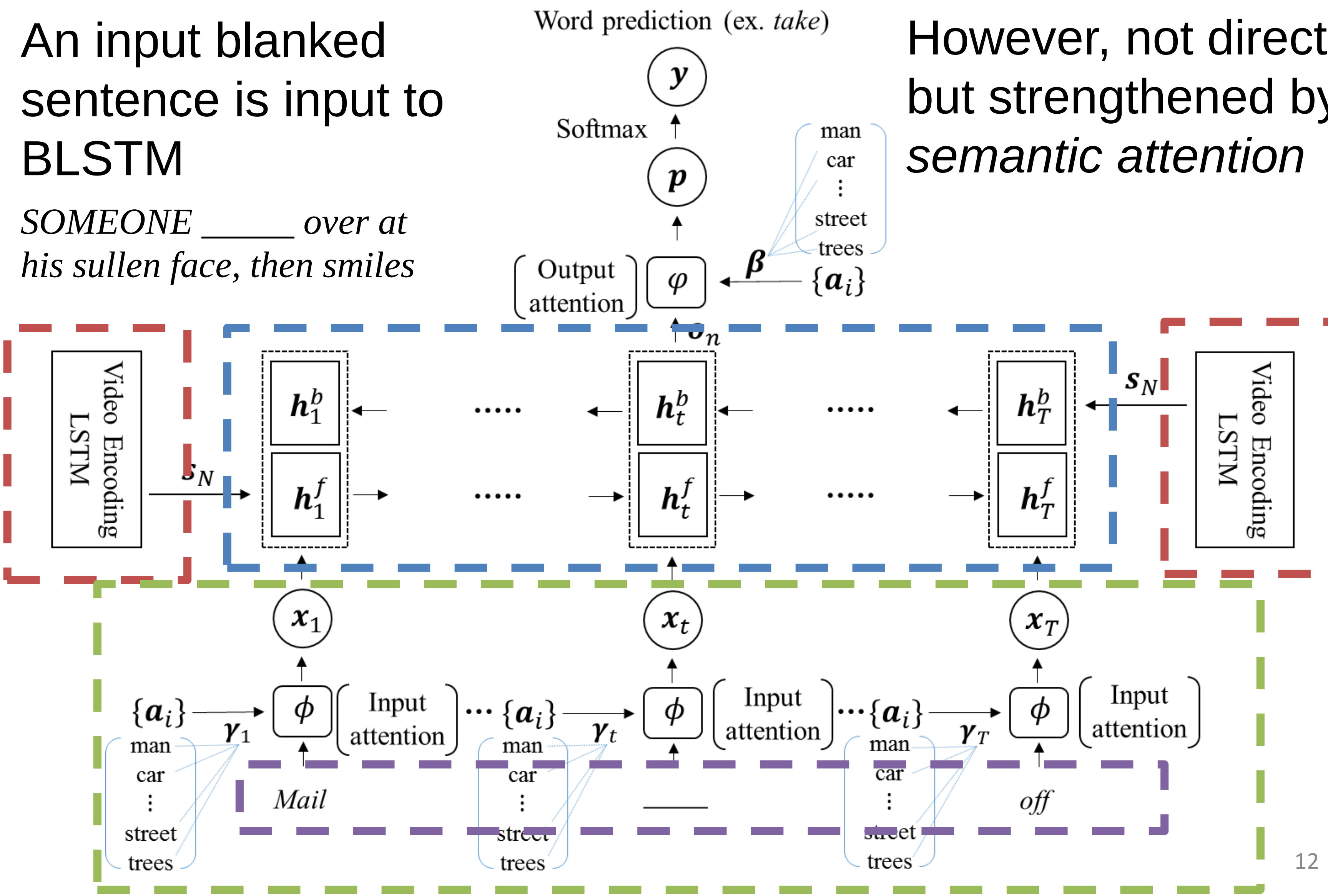


Model for Fill-in-the-Blank

An input blanked sentence is input to BLSTM

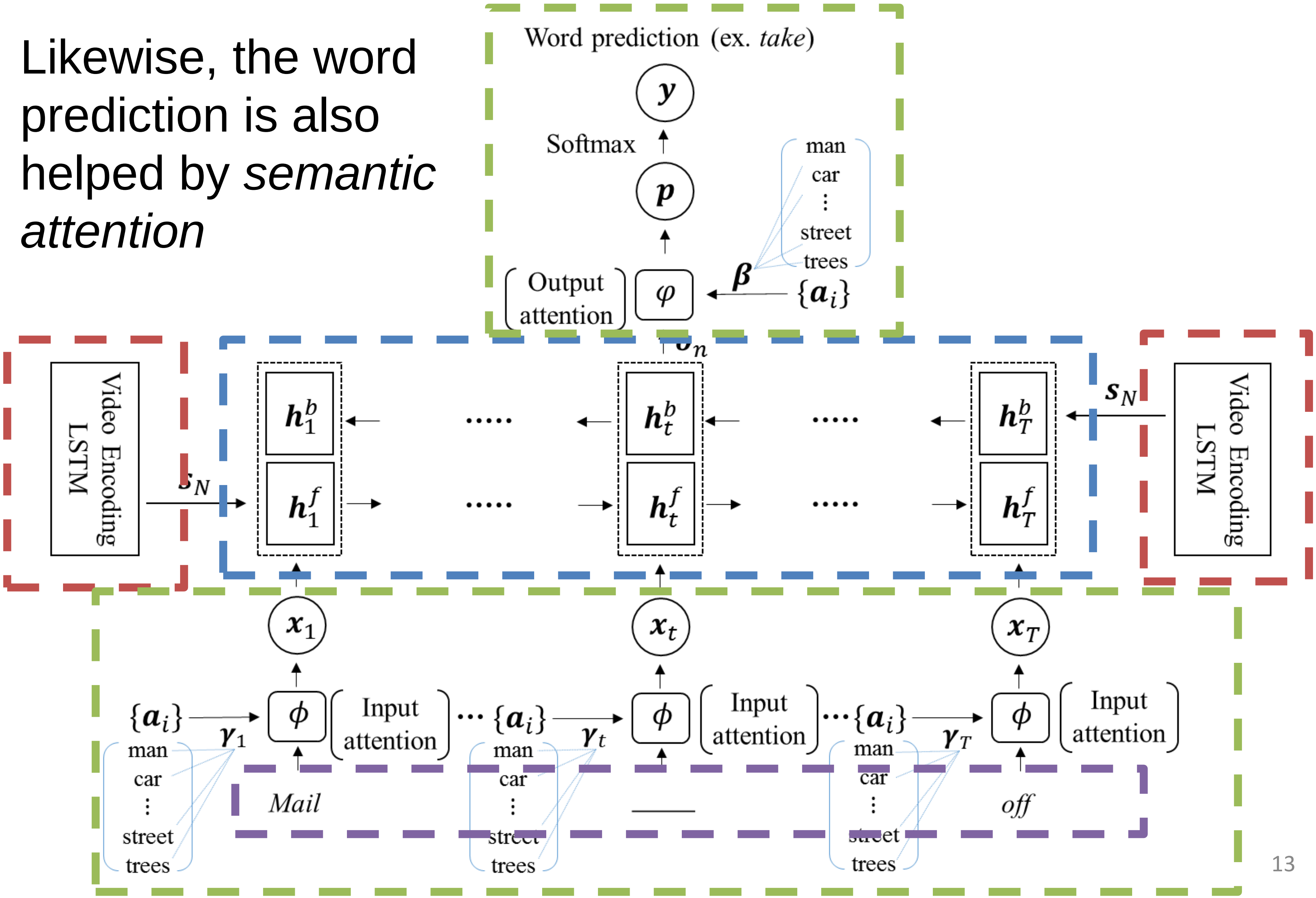
SOMEONE _____ over at his sullen face, then smiles

However, not directly but strengthened by *semantic attention*



Model for Fill-in-the-Blank

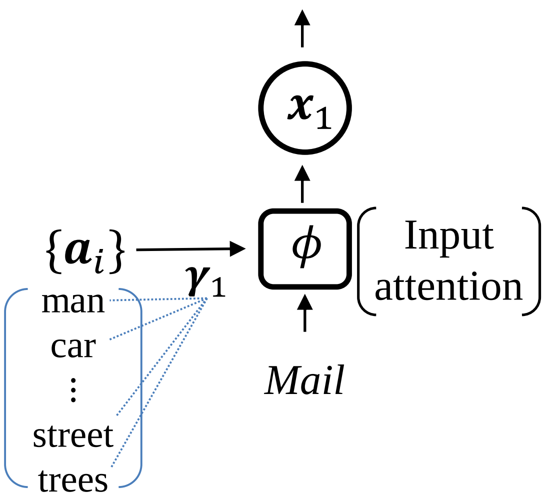
Likewise, the word prediction is also helped by *semantic attention*



Input and Output Semantic Attention

For better input representation and better output prediction

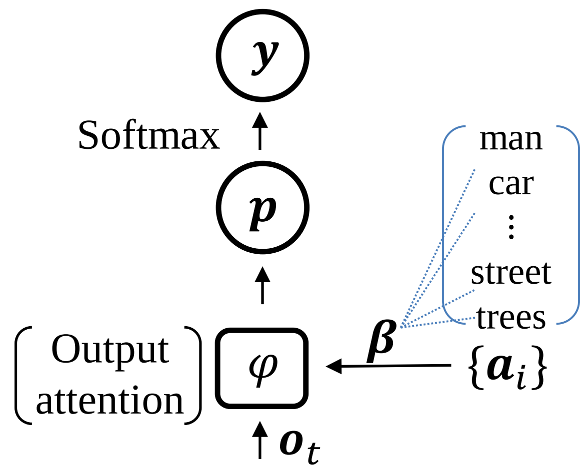
- Semantic attention ϕ/φ computes attention weights $\gamma_{t,i}/\beta_{t,i}$, which is assigned to each attribute word $\{a_i\}$



$$\gamma_{t,i} \propto \exp(\mathbf{c}_t^T \mathbf{W}_\gamma \mathbf{a}_i),$$

$$\mathbf{x}_t = \phi(\mathbf{c}_t, \{\mathbf{a}_i\}),$$

$$= \mathbf{W}_x(\mathbf{c}_t + \mathrm{diag}(\mathbf{w}_{x,a}) \sum_i \gamma_{t,i} \mathbf{a}_i)$$



$$\beta_{t,i} \propto \exp(\mathbf{o}_t^T \mathbf{W}_\beta \sigma(\mathbf{a}_i)),$$

$$\mathbf{p} = \varphi(\mathbf{o}_t, \{\mathbf{a}_i\}),$$

$$= \mathbf{o}_t + \mathrm{diag}(\mathbf{w}_{o,a}) \sum \beta_{t,i} \mathbf{W}_o \sigma(\mathbf{a}_i))$$

$$p(\mathbf{y} \mid \{\mathbf{c}_t\}_{t=1}^T) = \mathrm{softmax}(\mathbf{W}_y \mathbf{p} + \mathbf{b}_y)$$

Outline

- Objective and Key Ideas
- Approaches
  - A Model for Fill-in-the-Blank
  - A Model for Multiple-Choice Test
  - A Model for Retrieval
  - A Model for Description
- Experiments

Question for Multiple-Choice

Given a video query and five candidate captions, find the correct one out of five possible choices

Candidate Sentences

① SOMEONE sits on the corner of a desk.

② A man delivers a bouquet of red roses to SOMEONE.

③ She opens her eyes.

④ SOMEONE looks around awkwardly.

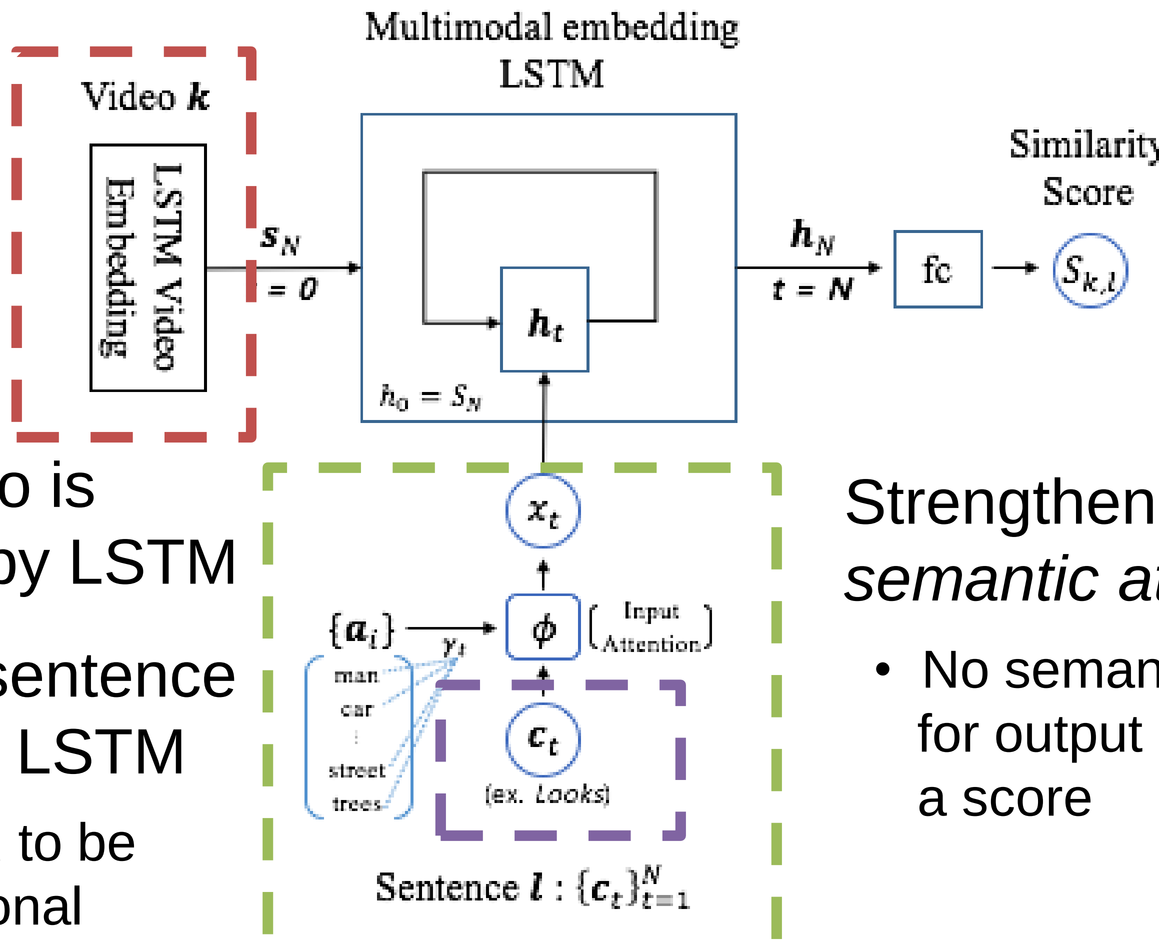
⑤ She knocks him out.

Model's input and output

- Input: (1) a video, (2) a sequence of words
- Output: a compatibility score

Model for Multiple-Choice

The base model: a multimodal LSTM that embeds a video-sentence pair into same space to calculate similarity score



Input video is encoded by LSTM

An input sentence is input to LSTM

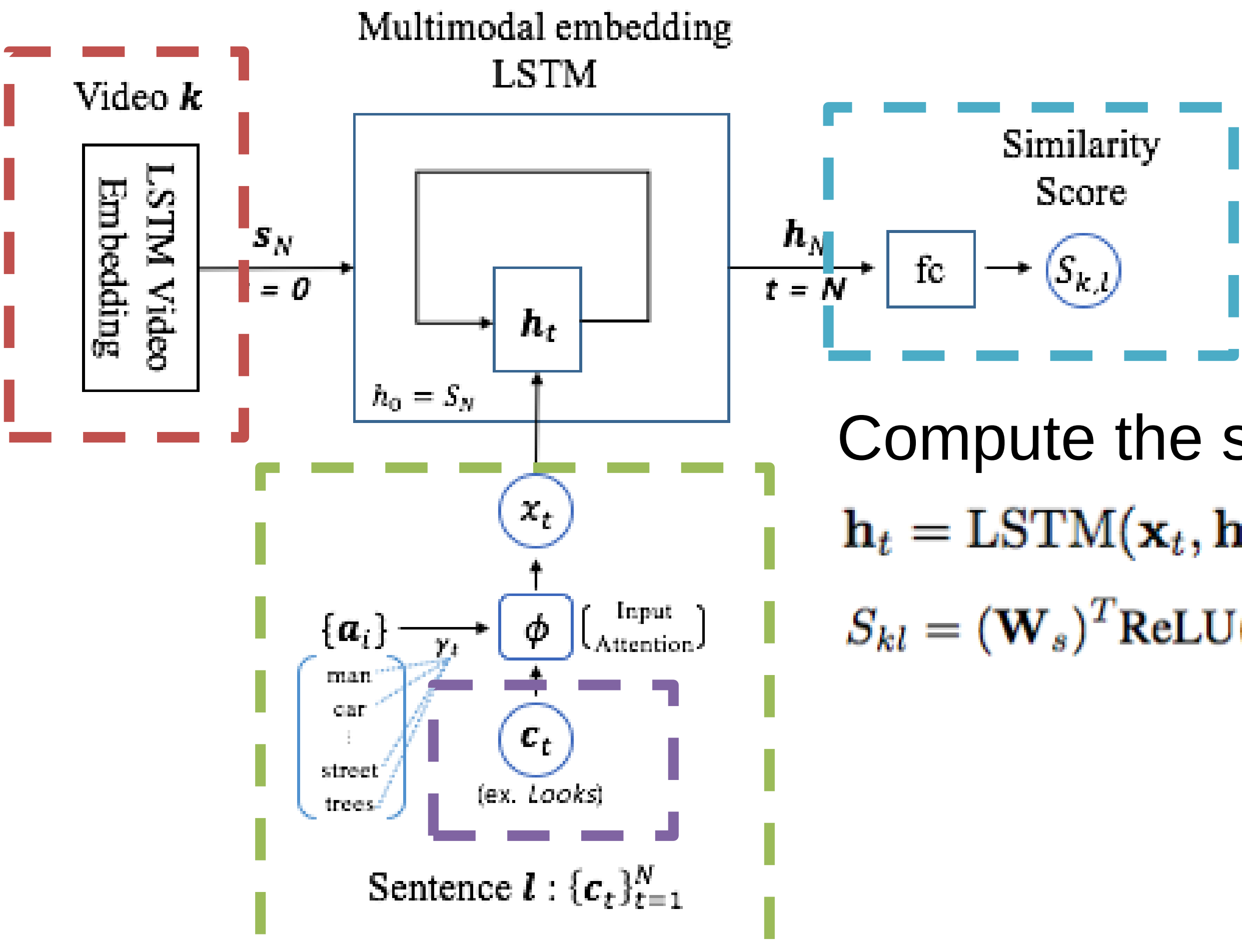
- No need to be bidirectional

Strengthened by *semantic attention*

- No semantic attention for output because it is a score

Model for Multiple-Choice

The base model: a multimodal LSTM that embeds a video-sentence pair into same space to calculate similarity score



Compute the score

$$\mathbf{h}_t = \mathrm{LSTM}(\mathbf{x}_t, \mathbf{h}_{t-1})$$

$$S_{kl} = (\mathbf{W}_s)^T \mathrm{ReLU}(\mathbf{W}_a \mathbf{h}_T + \mathbf{b}_a)$$

Outline

- Objective and Key Ideas
- Approaches
  - A Model for Fill-in-the-Blank
  - A Model for Multiple-Choice Test
  - A Model for Retrieval
  - A Model for Description
- Experiments

Question for Movie Retrieval

Given a short query text, find its corresponding video out of 1,000 candidate videos

Q : Throughout the cafeteria, students dance together and clap their hands.

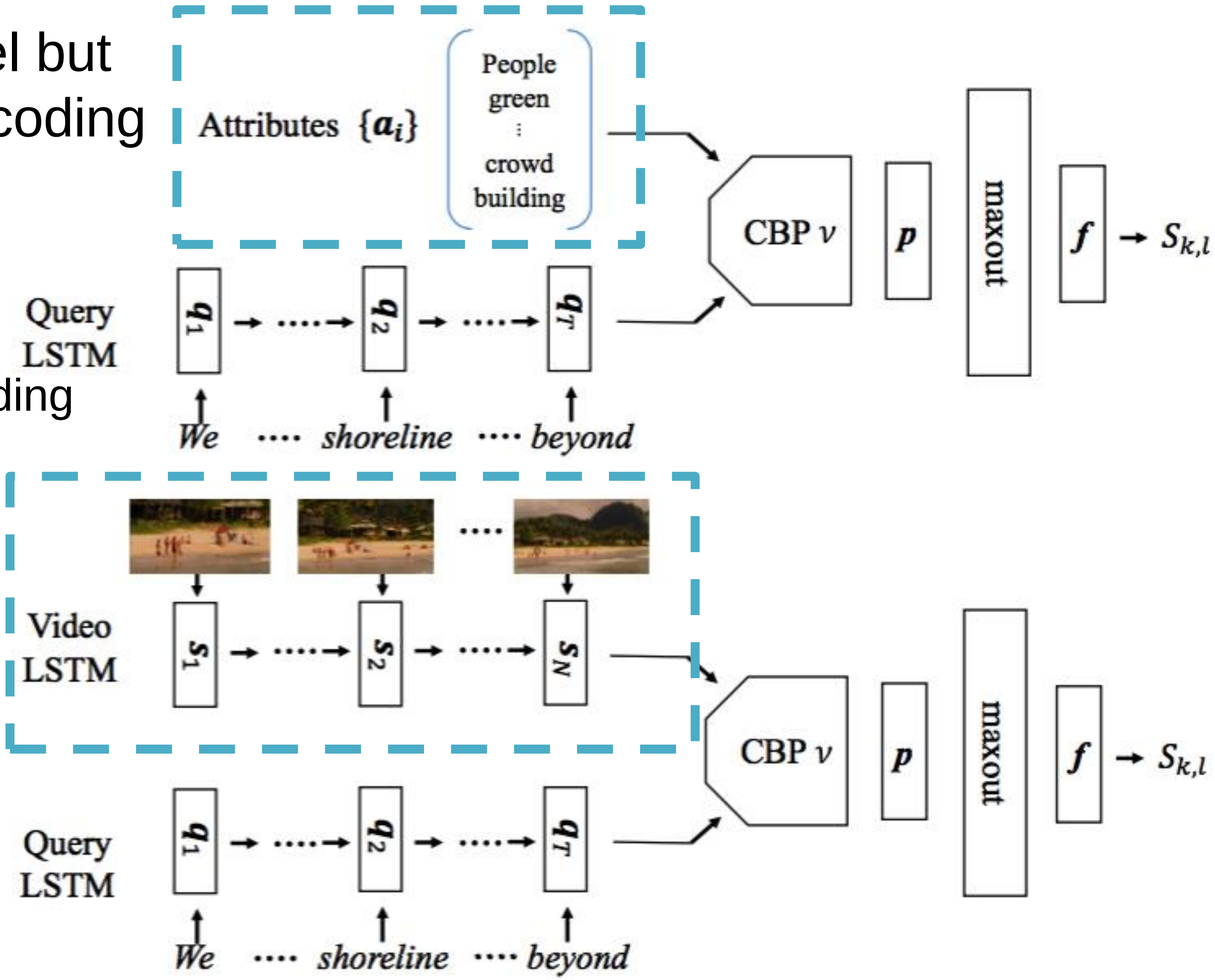
Model's input and output

- Input: (1) a video, (2) a sequence of words
- Output: a compatibility score

Models for Movie Retrieval

Same model but different encoding of a video

- Attributes of a video
- LSTM encoding



Models for Movie Retrieval

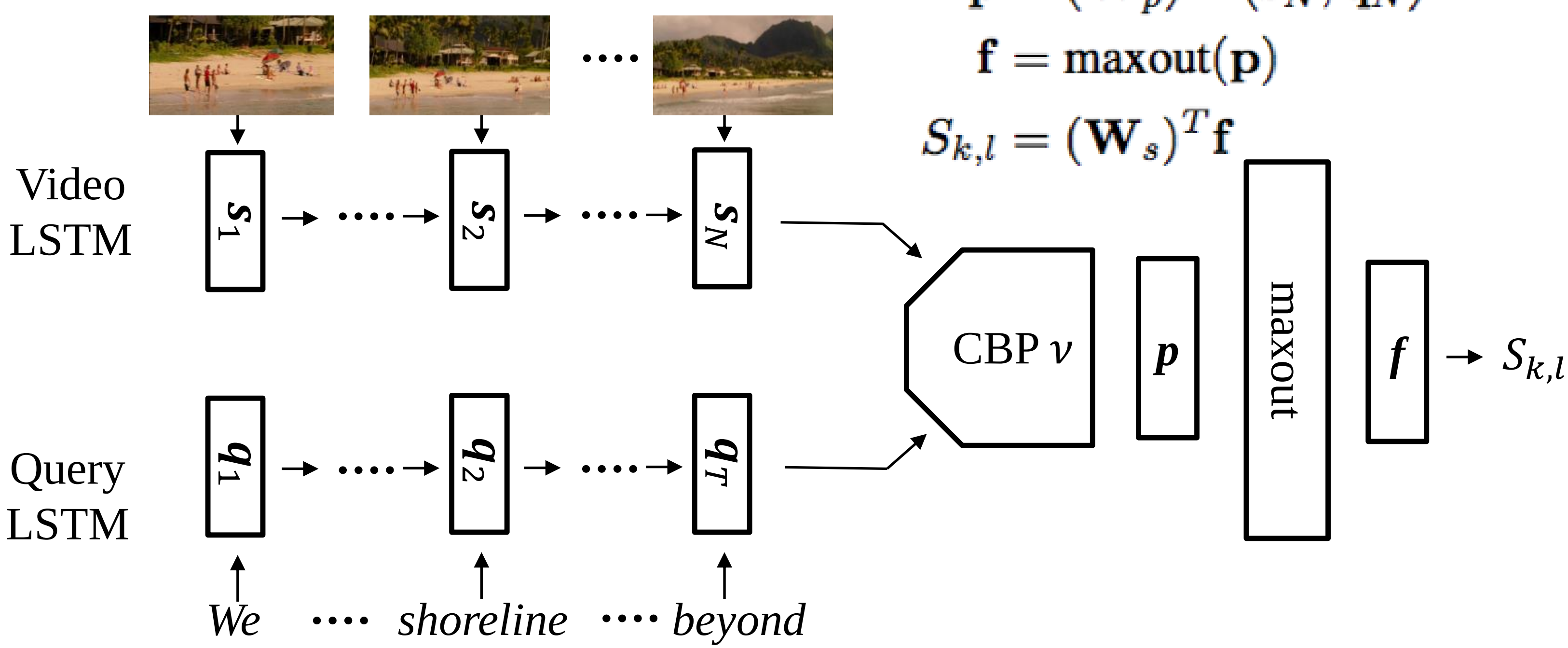
A multimodal embedding model

- Use Multimodal Compact Bilinear (MCB) model
- Use Maxout and Dropout for reducing overfitting

$$\mathbf{p} = (\mathbf{W}_p)^T \nu(\mathbf{s}_N, \mathbf{q}_N)$$

$$\mathbf{f} = \text{maxout}(\mathbf{p})$$

$$S_{k,l} = (\mathbf{W}_s)^T \mathbf{f}$$



A. Fukui et al. Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding. EMNLP, 2016.

Models for Movie Retrieval

A multimodal embedding model

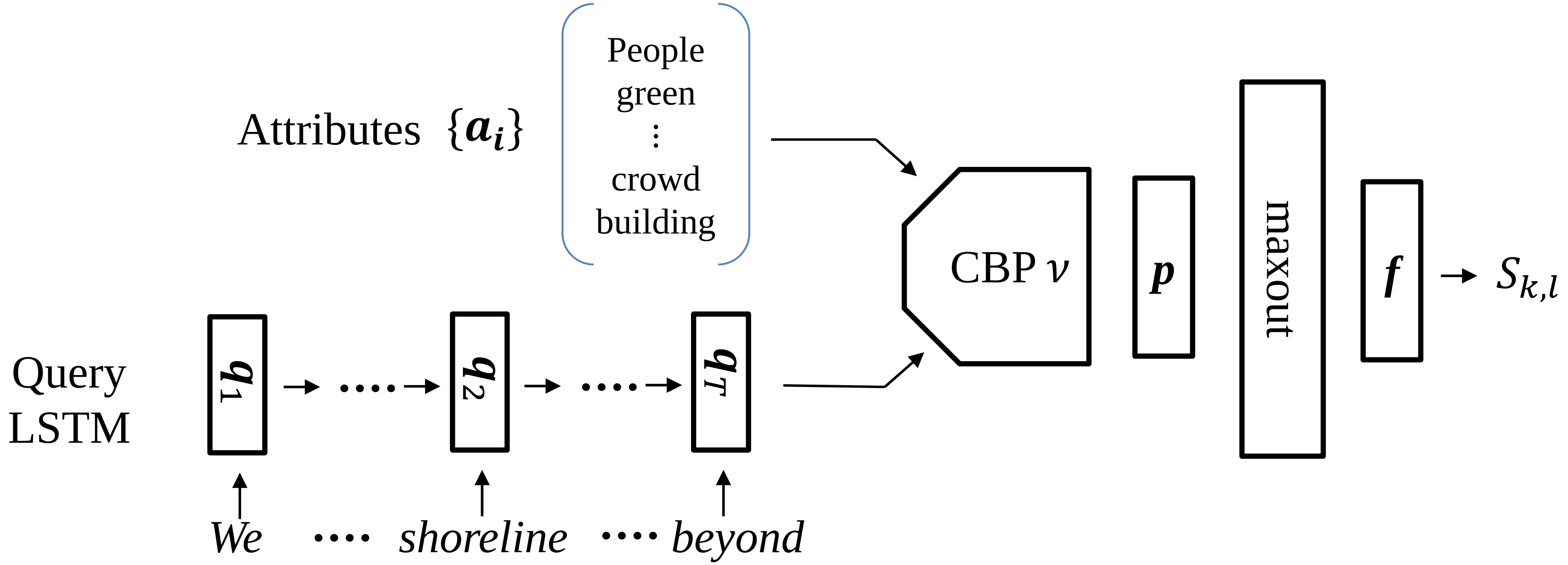
- Max-margin loss: a positive pair has higher score than a negative pair by Δ

$$\mathbf{p}_i = (\mathbf{W}_p)^T \nu(\mathbf{a}_i, \mathbf{q}_N)$$

$$\mathbf{f}_i = \text{maxout}(\mathbf{p}_i)$$

$$\mathcal{L} = \sum_k \sum_l \max(0, S_{k,l} - S_{k,l^*} + \Delta)$$

$$S_{k,l} = \frac{1}{K} \sum_{i=1}^{K} (W_s)^T \mathbf{f}_i$$



A. Fukui et al. Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding. EMNLP, 2016.

Models for Movie Retrieval

A ensemble of the following models

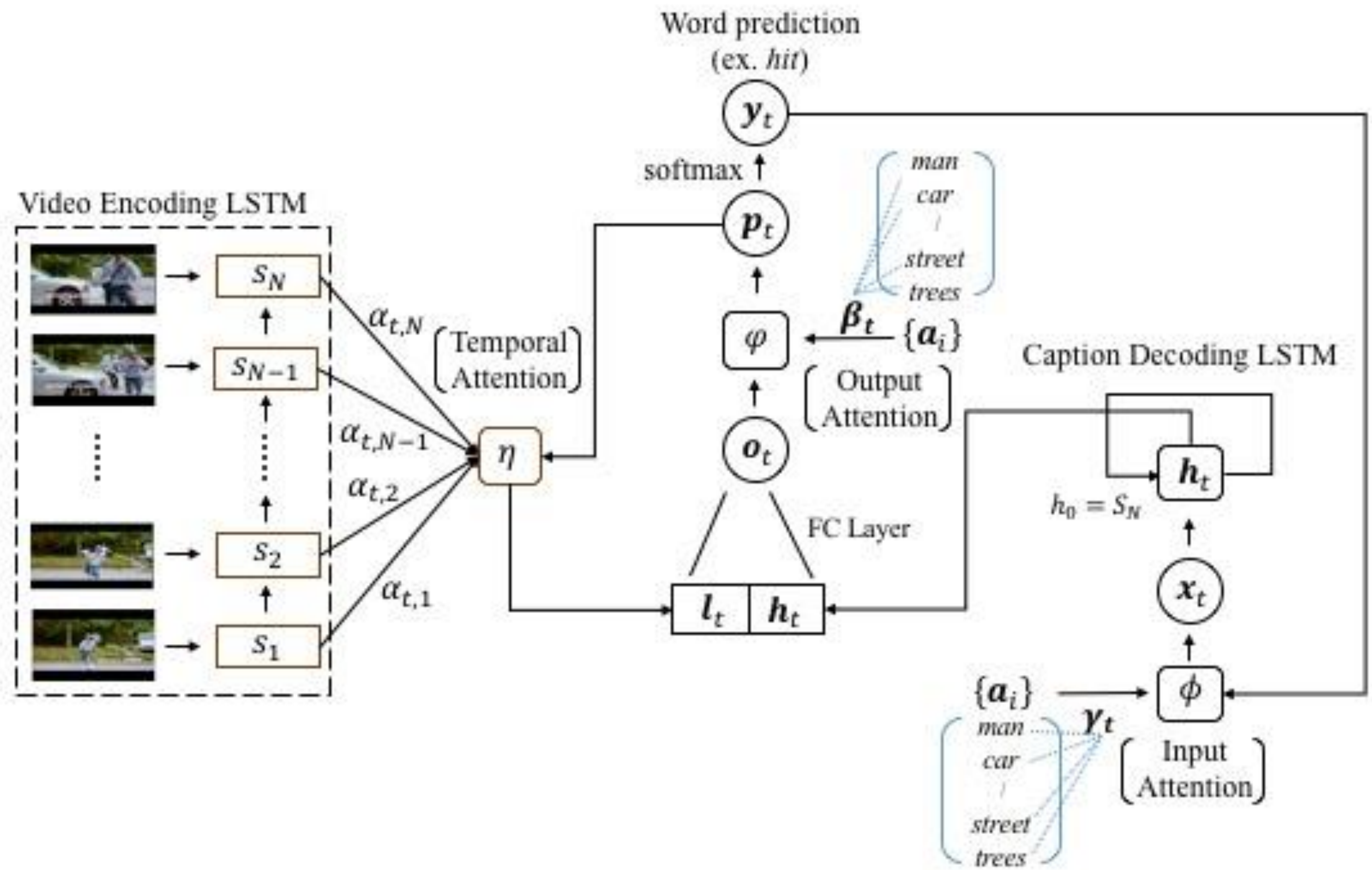
- Our two retrieval models discussed previously
- Our multiple-choice model
- The METEO score between a query sentence and a generated one by our description model

Outline

- Objective and Key Ideas
- Approaches
  - A Model for Fill-in-the-Blank
  - A Model for Multiple-Choice Test
  - A Model for Retrieval
  - A Model for Description
- Experiments

Models for Movie Retrieval

Temporal + Semantic attention models



Outline

- Objective and Key Ideas
- Approaches
  - A Model for Fill-in-the-Blank
  - A Model for Multiple-Choice Test
  - A Model for Retrieval
  - A Model for Description
- Experiments

Quantitative Results – Movie Annotation and Retrieval

Multiple-choice task

metrics	accuracy
arnavkj95	20.12
frcnnBigger	39.69
atousa	58.11
EITanque	63.71
Ours (Single Model)	63.10
Ours (Ensemble)	**65.70**

Movie retrieval

metrics	R@1	R@5	R@10	MedR
atousa	4.300	12.600	18.900	98
EITanque	**4.700**	**15.900**	23.400	64
Ours	3.600	14.700	**23.900**	**50**

Quantitative Results – Fill-in-the-Blank

Fill-in-the-blank task

metrics	accuracy
`tegan`	0.006
`arnavkj95`	0.014
`amirmazaheri`	0.342
Ours (Single Model)	**0.380**
Ours (Ensemble)	**0.407**

Quantitative Results – Movie Description

Movie description

Language metrics	B1	B2	B3	B4	M	R	Cr
sophieag	0.159 (3)	0.043 (4)	0.010 (8)	0.003 (8)	**0.080** (1)	0.150 (4)	0.048 (8)
ayush11011995	0.118 (6)	0.036 (6)	0.013 (5)	0.005 (5)	0.074 (2)	0.142 (6)	0.047 (9)
arohrbach	0.161 (2)	0.052 (2)	**0.021** (1)	**0.009** (1)	0.071 (3)	**0.164** (1)	**0.112** (1)
Ours	0.156 (4)	0.044 (3)	0.014 (3)	0.004 (6)	0.071 (4)	0.147 (5)	0.070 (6)
s2vt	**0.174** (1)	**0.053** (1)	0.018 (2)	0.007 (2)	0.070 (5)	0.161 (2)	0.091 (3)
rakshithShetty	0.110 (7)	0.034 (7)	0.013 (6)	0.006 (3)	0.061 (6)	0.156 (3)	0.090 (4)
EITanque	0.145 (5)	0.041 (5)	0.014 (4)	0.006 (4)	0.058 (7)	0.134 (8)	0.101 (2)
macmadman	0.056 (10)	0.015 (10)	0.006 (9)	0.003 (9)	0.052 (8)	0.134 (7)	0.062 (7)
fodrh1201	0.092 (8)	0.029 (8)	0.010 (7)	0.004 (7)	0.040 (9)	0.096 (9)	0.075 (5)
frcnnBigger	0.069 (9)	0.016 (9)	0.005 (10)	0.002 (10)	0.034 (10)	0.070 (10)	0.035 (10)

Qualitative Results – Fill-in-the-Blank

Good examples

Blank Sentence : Now, at night, our _____ glides over a highway, its lanes glittering from the lights of traffic below.

Answer : view **Our result** : view

Attribute : *city, scene, street, background, cars, building, sky, cloudy, tall*

Blank Sentence : The vehicle breaks the gate and _____ off.

Answer : speeds **Our result** : speeds

Attribute : *city, man, sitting, street, parked, car, glasses, building, fence, windows, train, station, day, dark*

Qualitative Results – Fill-in-the-Blank

Negative examples

Blank Sentence : SOMEONE _____ over at his sullen face, then smiles.
Answer : glances **Our result** : looks
Attribute : *car, man, window, beard, short, sitting, back, hair, mouth, background, open, men*

Blank Sentence : SOMEONE kicks him under the table, upsetting her _____.
Answer : purse **Our result** : teeth
Attribute : *woman, eating, pizza, suit, man, wearing, tie, people, restaurant, table, knife, window, food*

Qualitative Results – Multiple-Choice

Good examples

Candidate Sentences
① SOMEONE sits on the corner of a desk.
② A man delivers a bouquet of red roses to SOMEONE.
③ She opens her eyes.
④ SOMEONE looks around awkwardly.
⑤ She knocks him out.
Our result : A man delivers a bouquet of red roses to SOMEONE.
Attribute : *woman*, *sitting*, *table*, *wearing*, *people*, *long*, *hair*, *women*, *man*, *pink*, *person*

Qualitative Results – Multiple-Choice

Negative examples

Candidate Sentences

① SOMEONE puts his arms around the bikini babes.

② SOMEONEs eyes widen.

③ He gives a faint bobble of his head.

④ With people.

⑤ Later she enters her apartment.

Our result : SOMEONE puts his arms around the bikini babes.

Attribute : *man*, *cell*, *looking*, *phone*, *holding*, *standing*, *bathroom*, *wearing*, *tie*, *wall*, *metal*, *large*, *behind*, *tile*, *door*

Qualitative Results – Movie Retrieval

Good examples

Q : Throughout the cafeteria, students dance together and clap their hands.

Attributes : *wall, light, standing, looking, person, camera, background*

Qualitative Results – Movie Retrieval

Negative examples

Q : That evening, he crosses the driveway.

129th

Attributes : *looking, suit, beard, watching, person, car*

Qualitative Results – Movie Description

Good examples

GT : SOMEONE enters the classroom and closes the shutters with his wand.
Ours : someone walks through the crowd.
Attribute : taken, man, room, people, sitting, window, building, walking, woman, suit, looking, dark, chair, bench, wall

GT : His eyes flicker and close.
Ours : someones eyes are closed
Attribute : *man, wearing, head, hat, looking, person, mans, chair*

Qualitative Results – Movie Description

Failure examples

GT : The man glances around.
Ours : someone is wearing a hat.
Attribute : *man*, *wearing*, *hat*, *sitting*, *standing*, *horse*, *bench*, *hai*r

GT : They run to a storage trailer.
Ours : someones car pulls up to a main street.
Attribute : taken, car, walking, sidewalk, man, standing, people, station, ground, fence, building, wooden, bench, shadow, windows

Conclusion

Video-to-language models with semantic attention

- A separate model for each task
- Take advantage of state-of-the-art techniques as our base models
- Adopt ***semantic attention*** to strengthen meaning of words

Promising results in LSMDC 2016 Challenge

- Has won three tasks of two tacks (movie multiple-choice, movie fill-in-the-blank, and movie retrieval)

More details can be found in our arxiv paper
Video captioning and retrieval models with semantic attention
https://arxiv.org/abs/1610.02947