Image Caption

- 2016
  - Guiding Long-Short Term Memory for Image Caption Generation
  - Show, Attend and Tell: Neural Image Caption Generation with Visual Attention
  - Watch What You Just Said: Image Captioning with Text-Conditional Attention
- 2017
  - Knowing When to Looking: Adaptive Attention via. A Visual Sentinel for Image caption
  - SCA-CNN: Spatial and Channel Attention in Convolution Networks for Image Captioning
  - Skeleton key: Image Captioning by Skeleton-Attribute Decomposition

Video Caption

- 2015
  - Sequence to Sequence - video to text (S2VT)
  - Video Description Generation Incorporating Spatio-Temporal Features and a Soft-Attention Mechanism (SA)
- 2016
  - Jointly Modeling Embedding and Translation to Bridge Video and Language
  - Frame and Segment-Level Features and Candidate Pool Evaluation for Video Caption Generation
- 2017
  - Weakly Supervised Dense Video Captioning (SOTA)
- 2018
  - TVT: Two-View Transformer Network for Video Captioning
  - End-to-End Video Captioning with Masked Transformer
  - Reconstruction Network for Video Captioning
  - *LiveBot: Generating Live Video Comments Based on Visual and Textual contexts*

Guiding Long-Short Term Memory for Image Caption Generation

Xu Jia
KU Leuven ESAT-PSI, iMinds
Xu.Jia@esat.kuleuven.be

Efstratios Gavves*
QUVA, University of Amsterdam
E.Gavves@uva.nl

Basura Fernando
ACRV, The Australian National University
basura.fernando@anu.edu.au

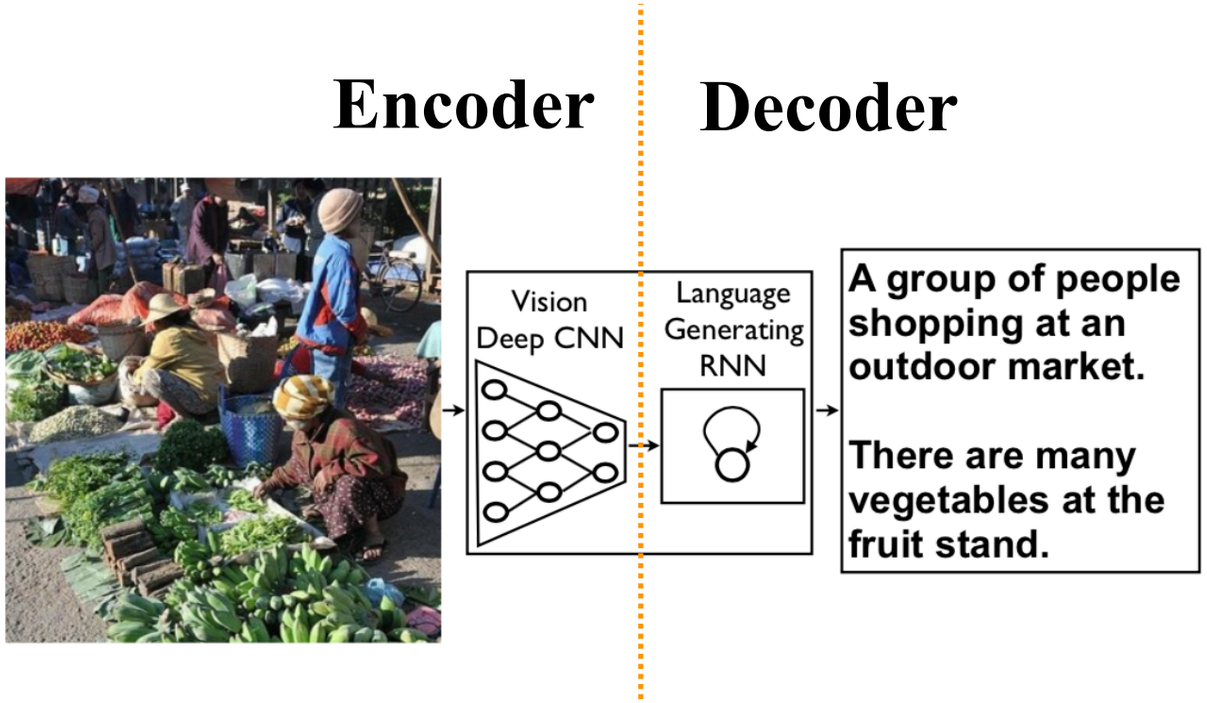
Tinne Tuytelaars
KU Leuven ESAT-PSI, iMinds
Tinne.Tuytelaars@esat.kuleuven.be

Presenter: WENWEI KANG

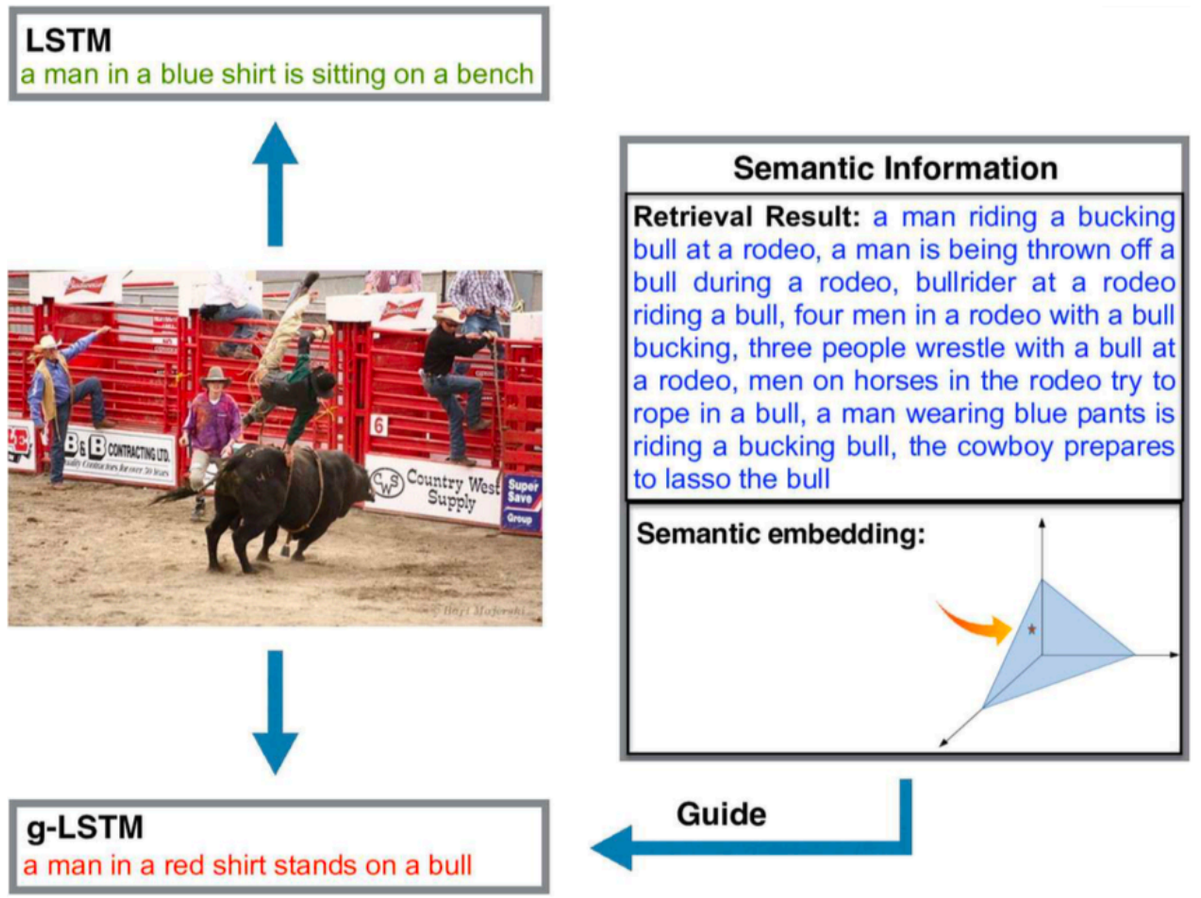
- Introduction
- Related work
- Guiding LSTM
- Canonical Correlation Analysis (CCA)
- Guiding strategy
- Beam search with Length Normalization
- Evaluation

Introduction

Show and tell: NIC
(Vinyals *et al.*)



guiding LSTM
(Jia *et al.*)

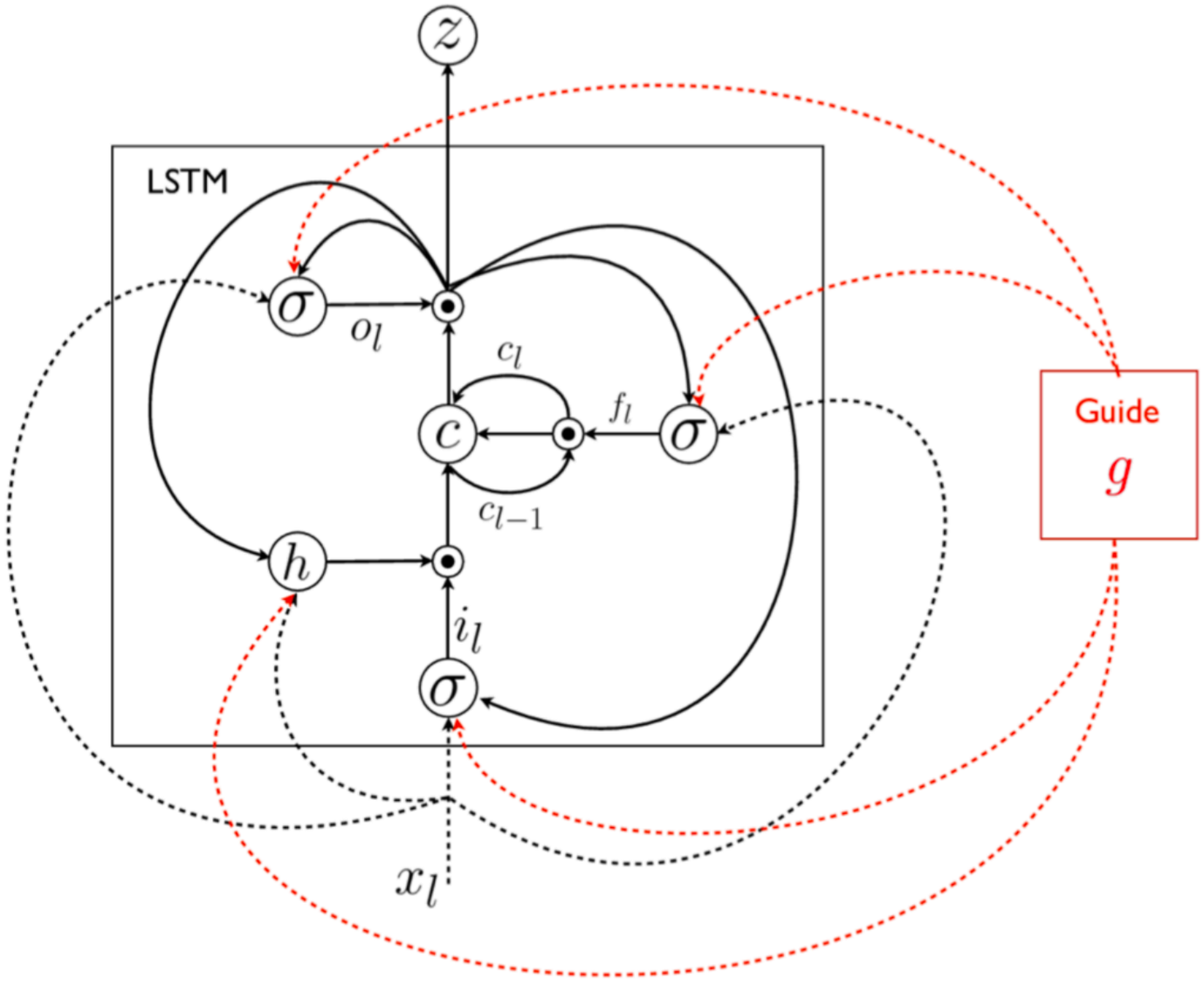


Related work

Caption generation

- **Template-based methods:** First detect objects, actions, scenes and attributes, then fill them in a fixed sentence template. (require explicit annotations)
- **Transfer-based captions:** First retrieve visually similar images, then transfer captions of those images to the query image. (directly rely on retrieval results among image)
- **Neural Network based:** Similar encoding-decoding framework (jointly learns visual attention and caption generations).

Guiding LSTM



$$
\begin{aligned}
i'_l &= \sigma(W_{ix}x_l + W_{im}m_{l-1} + W_{iq}g) \\
f'_l &= \sigma(W_{fx}x_l + W_{fm}m_{l-1} + W_{fq}g) \\
o'_l &= \sigma(W_{ox}x_l + W_{om}m_{l-1} + W_{oq}g) \\
c'_l &= f'_l \odot c'_{l-1} + i'_l \odot h(W_{cx}x_l + \\
&\quad + W_{cm}m_{l-1} + W_{cq}g) \\
m_l &= o'_l \odot c'_l
\end{aligned}
$$

g: semantic information (no timestep)

Canonical Correlation Analysis (1/3)

Pearson correlation coefficient

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

$$\underset{U_1, U_2}{\arg\max} \frac{U_1 \Sigma_{X_1 X_2} U_2}{\sqrt{U_1 \Sigma_{X_1 X_1} U_1}\sqrt{U_2 \Sigma_{X_2 X_2} U_2}}$$

$$X_1 = U_{11}^T A_1 + U_{12}^T A_2,\ X_2 = U_{21}^T B_1 + U_{22}^T B_2$$

X	Y
x1	y1
x2	y2
x3	y3
x4	y4
x5	y5
...	...
xn	yn

Correlation

A1	A2	B1	B2
A11	A23	B11	B21
A12	A22	B12	B22
A13	A23	B13	B23
A14	A24	B14	B24
A15	A25	B15	B25
...	...	...	...
A1n	A2n	B16	B2n

Correlation

Canonical Correlation Analysis (2/3)

$$X_1 = U_{11}^T A_1 + U_{12}^T A_2,\ X_2 = U_{21}^T B_1 + U_{22}^T B_2$$

$$X_1 = U_1^T A\ (A = \begin{bmatrix} A_1 \\ A_2 \end{bmatrix}),\ X_2 = U_1^T B\ (B = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix}) \longrightarrow \rho_{X_1,X_2} = \frac{Cov(X_1, X_2)}{std(X_1) \times std(X_2)}$$

1. $Var(X_1) = Var(U_1^T A) = \frac{1}{N}\sum_{i=1}^{N}(U_1^T A_i - U_1^T \bar{A})^2 = U_1^T \frac{1}{N}\sum_{i=1}^{N}(A_i - \bar{A})^2 U_1 = U_1^T \sum_{X_1,X1} U_1$
2. $Var(X_2) = U_2^T \sum_{X_2,X_2} U_2$
3. $Cov(X_1, X_2) = U_1^T \sum_{X1,X_2} U_2$

$$\rho_{X_1,X_2} = \frac{Cov(X_1, X_2)}{Var(X_1) \times Var(X_2)} = \frac{U_1^T \sum_{X1,X2} U_2}{\sqrt{U_1^T \sum_{X1,X1} U_1}\sqrt{U_2^T \sum_{X2,X2} U_2}}$$

$$\arg\max_{U_1,U_2} \frac{U_1 \Sigma_{X_1X_2} U_2}{\sqrt{U_1 \Sigma_{X_1X_1} U_1}\sqrt{U_2 \Sigma_{X_2X_2} U_2}}$$

Canonical Correlation Analysis (3/3)

$$\rho_{X_1,X_2} = \frac{Cov(X_1, X_2)}{Var(X_1) \times Var(X_2)} = \frac{U_1^T \sum_{X1,X2} U_2}{\sqrt{U_1^T \sum_{X1,X1} U_1} \sqrt{U_2^T \sum_{X2,X2} U_2}}$$

Lagrange multiplier :

$$Maximum\ U_1^T \sum_{X1,X2} U_2,\ subject\ to\ U_1^T \sum_{X1,X1} U_1 = 1,\ U_2^T \sum_{X2,X2} U_2 = 1$$

$$L = U_1^T \sum_{X1,X2} U_2 - \frac{\lambda}{2}(U_1^T \sum_{X1,X1} U_1 - 1) - \frac{\theta}{2}(U_2^T \sum_{X2,X2} U_2 - 1)$$

$$\frac{\partial L}{\partial U_1} = \sum_{X_1,X_2} U_2 - \lambda \sum_{X_1,X_1} U_1 = 0,\ \frac{\partial L}{\partial U_2} = \sum_{X_2,X_1} U_1 - \theta \sum_{X_2,X_2} U_2 = 0$$

$$\vdots$$

eigen value : λ^2, eigen vector : U_1, U_2

$$X_1 = U_1^T A,\ X_2 = U_1^T B$$

$$g_1 = \frac{X_1 U_1 D^p}{||X_1 U_1 D^p||},\ g_2 = \frac{X_2 U_2 D^p}{||X_2 U_2 D^p||}$$

where D is a diagonal matrix whose elements are set to the eigenvalues

Guiding strategy(1/2)

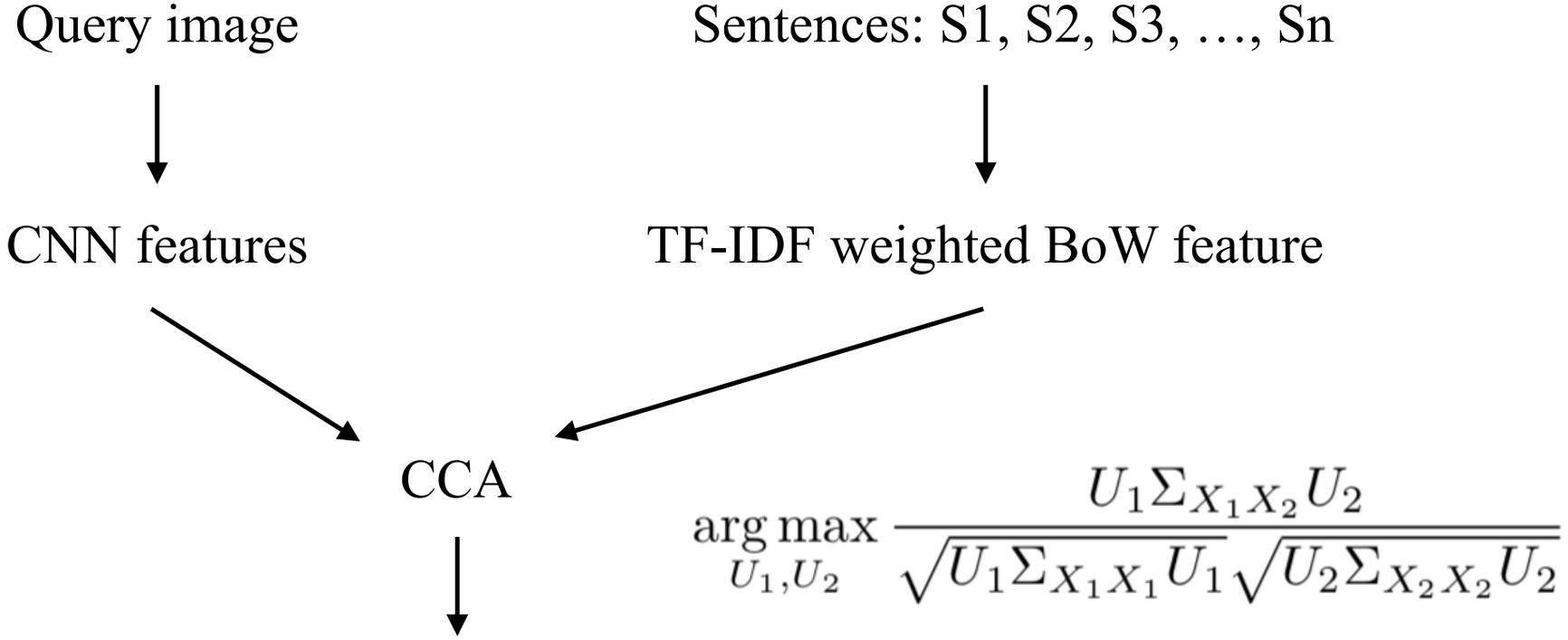
- **Retrieval-based guidance(ret-gLSTM):**

  Given an image, we first do the cross- modal retrieval whose aim is to find relevant texts to the query image.

  cross-model: normalized CCA method.

  relevant texts: TF-IDF weighted BoW features.

  query image: CNN features.



Paper: **TOP 15** sentences BoW features.

Guiding strategy(2/2)

- **Semantic embedding guidance(emb-gLSTM):**

  An image is mapped into the common semantic space by the learned projection matrix and the computed semantic embedding is fed to gLSTM model as the guide.

  $$g = X_1 U_1$$

- **Image as guidance(img-gLSTM):**

  We experimentally verify this by simply feeding the image feature itself to the gLSTM model.

  $g = x(\text{image feature})$

Beam search with Length Normalization

$$P = \frac{1}{\Omega(L)} \sum_{l=1}^{L} \log P(s_l \mid x, s_{1:l}, \theta)$$

Length normalized

- Polynomial normalization:

  $\Omega(L) = |L|^m$, L : *sentence length*

- Min-hinge normalization:

  $\Omega(L) = \min\{L, \mu\}$, μ : average length of the training sentences

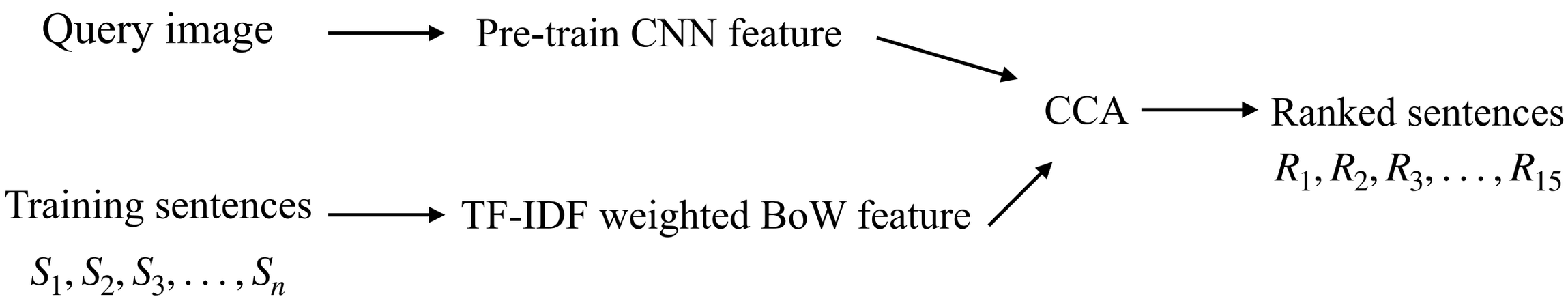
- Max-hinge normalization:

  $\Omega(L) = \max\{L, \mu\}$

- Gaussian normalization:

  $\Omega(L) \sim N(\mu, \sigma)$, σ : standard deviation of the sentence lengths

1. Retrieval result



2. Guiding LSTM

$$
\begin{aligned}
i'_l &= \sigma(W_{ix}x_l + W_{im}m_{l-1} + W_{iq}g) \\
f'_l &= \sigma(W_{fx}x_l + W_{fm}m_{l-1} + W_{fq}g) \\
o'_l &= \sigma(W_{ox}x_l + W_{om}m_{l-1} + W_{oq}g) \\
c'_l &= f'_l \odot c'_{l-1} + i'_l \odot h(W_{cx}x_l + \\
&\quad + W_{cm}m_{l-1} + W_{cq}g) \\
m_l &= o'_l \odot c'_l
\end{aligned}
\qquad
g : \begin{cases} \text{ret} - \text{gLSTM} \\ \text{emb} - \text{gLSTM} \\ \text{img} - \text{gLSTM} \end{cases}
$$

2. Test

Beam search with Length normalization

- Polynomial norm.
- Min-hinge norm.
- Max-hinge norm.
- Gaussian norm.

Evaluation(1/4)

Dataset (each image is accompanied with 5 captions):

- Flickr8k: training: 6,000, validation: 1,000, testing: 1,000
- Flickr30k: training: 29,000, validation: 1,000, testing: 1,000
- MSCOCO: training: 82,783, validation: 5,000, testing: 5,000

Metric:

- BLEU@1~4
- METEOR

Evaluation(2/4)

LSTM

Normalization	B@1	B@2	B@3	B@4	METEOR
Baseline	59.6	40.4	26.1	17.0	17.45
Polynomial	57.8	39.2	26.0	17.6	**18.86**
Min-hinge	60.4	41.4	27.6	**18.6**	18.53
Max-hinge	57.6	38.8	25.2	16.7	17.65
Gaussian	**60.7**	**41.7**	**27.8**	**18.6**	18.35

Table 1: The performance of different length normalization strategies on Flickr8k.

GT Refs	*Baseline*	*Polynom.*	*Min-hinge*	*Max-hinge*	*Gaussian*
10.87(3.74)	8.75(2.44)	11.07(2.62)	9.64(1.92)	9.55(1.69)	9.57(3.30)

Table 2: The average and the standard deviation of the sentence length for the ground truth references, and different normalization strategies on Flickr8k.

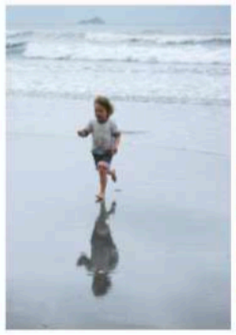
	B@1	B@2	B@3	B@4	METEOR
Baseline, Original	59.6	40.4	26.1	17.0	17.45
Baseline, Polynomial	57.8	39.2	26.0	17.6	18.86
Baseline, Min-hinge	60.4	41.4	27.6	18.6	18.53
Baseline, Gaussian	60.7	41.7	27.8	18.6	18.35
Baseline 512, Original	61.0	42.4	28.6	18.9	18.21
Baseline 512, Polynomial	58.2	40.2	27.1	18.1	19.83
Baseline 512, Min-hinge	61.3	42.9	29.2	19.6	19.13
Baseline 512, Gaussian	61.3	42.8	29.1	19.5	19.07
ret-gLSTM, Original	63.4	43.7	29.2	19.3	18.54
ret-gLSTM, Polynomial	58.8	40.4	27.5	18.6	19.86
ret-gLSTM, Min-hinge	63.0	43.8	29.9	20.2	19.46
ret-gLSTM, Gaussian	63.5	44.2	30.2	20.6	19.38
emb-gLSTM, Original	63.7	44.7	30.2	20.2	19.10
emb-gLSTM, Polynomial	61.0	43.0	29.6	20.1	**20.60**
emb-gLSTM, Min-hinge	64.3	45.7	31.6	21.5	20.28
emb-gLSTM, Gaussian	**64.7**	**45.9**	**31.8**	**21.6**	20.19
img-gLSTM, Original	61.5	42.5	27.2	16.7	17.10
img-gLSTM, Polynomial	55.7	38.1	24.9	15.8	17.69
img-gLSTM, Min-hinge	60.4	41.9	27.6	17.7	17.76
img-gLSTM, Gaussian	60.1	41.4	27.2	17.3	17.69

Table 3: Comparison between gLSTM with different semantic information on Flickr8k. We denote the gLSTM model with retrieval-based guidance as ret-gLSTM, the one with semantic embedding guidance as emb-gLSTM, and the one with image as guidance as img-gLSTM.

Evaluation(3/4)

	Flickr8k					*Flickr30k*				
	B@1	B@2	B@3	B@4	METEOR	B@1	B@2	B@3	B@4	METEOR
LogBilinear [19]	65.6	42.4	27.7	17.7	17.31	60.0	38.-	25.4	17.1	16.88
multimodal RNN [17]	57.9	38.3	24.5	16.0	16.7	57.3	36.9	24.0	15.7	15.3
Google NIC [36]	63.-	41.-	27.-	—	—	66.3	42.3	27.7	18.3	—
LRCN-CaffeNet [7]	—	—	—	—	—	58.7	39.1	25.1	16.5	—
m-RNN-AlexNet [26]	—	—	—	—	—	54.-	36.-	23.-	15.-	—
m-RNN [26]	—	—	—	—	—	60.-	41.-	28.-	19.-	—
Soft-Attention [37]	**67.-**	44.8	29.9	19.5	18.93	66.7	43.4	28.8	19.1	18.49
Hard-Attention [37]	**67.-**	45.7	31.4	21.3	20.3	**66.9**	43.9	29.6	19.9	18.46
emb-gLSTM, Polynomial	61.0	43.0	29.6	20.1	**20.60**	59.8	41.3	29.3	19.2	**18.58**
emb-gLSTM, Min-hinge	64.3	45.7	31.6	21.5	20.28	63.8	44.1	30.2	20.5	18.13
emb-gLSTM, Gaussian	64.7	**45.9**	**31.8**	**21.6**	20.19	64.6	**44.6**	**30.5**	**20.6**	17.91

Evaluation(4/4)



a young boy is running on the beach, a man in a blue shirt is riding a dirt bike, a little boy runs away from the approaching waves of the ocean, a little girl runs across the wet beach, a little girl runs on the wet sand near the ocean, a young girl runs across a wet beach with the ocean in the background, child running on the beach, two children are running towards the ocean on a beach, a dog is running in the ocean beside the beach, a dog playing in the ocean on the beach , a boy running through surf on a beach, boy running through the water at the beach, a girl runs down a beach, a boy standing on a beach, a man riding his bike on the beach by the ocean, a young girl running on the beach, a dog is running on the beach, a young child running along the shore at a beach, boy and girl running along the beach, a dog running on the beach, a dog running on the beach, a dog running on the beach

a group of dogs are running on a track, a group of people racing on a track, a a dog with a muzzle is leading several other dogs in a race, a greyhound leaps in a race, a muzzled dog in a race with four dogs following, five dogs are racing, five dogs are racing on a dirt track, two greyhounds with muzzles race along the inside curb of a railed dirt track, the greyhound racing dogs are running around a bend in the track, three muzzled greyhounds race around a turn in a track, several muzzled greyhound dogs racing around a track, two muzzled greyhounds dogs racing around a track, two greyhounds race around a track, greyhounds racing chasing a mechanical rabbit around the track, three greyhounds are racing on a track at night, three greyhound dogs race around a dark track, muzzled greyhounds are racing along a dog track at night, three greyhounds racing around the corner of a track, greyhounds racing on a track, greyhounds race on a track, greyhounds race on a track, three greyhounds are in a dog race at the track

a woman in a black shirt and sunglasses smiles, a man and a woman pose for a picture, a brunette girl wearing sunglasses and a yellow shirt, a girl in sunglasses smiles, a girl wearing a yellow shirt and sunglasses smiles, a girl wearing sunglasses smiles for the camera, a woman with a yellow shirt wears sunglasses and smiles, a woman wearing sunglasses smiles, young man with upturned hair posing with young man with sunglasses and woman with glasses, a blonde woman wearing sunglasses and dice earrings smiles, a woman wearing black sunglasses looks to the right and smiles, a smiling woman is wearing sunglasses on a day with sparse clouds, a smiling woman with long dark hair wearing sunglasses on top of her head, a man and woman wearing sunglasses and white t-shirts smile for the camera, a man in sunglasses smiles, a blonde lady with sunglasses smiles, women in hat and sunglasses smiles, a woman wearing sunglasses, man and woman wearing sunglasses posing for picture, woman with green sweater and sunglasses smiling, a woman in a sunhat is wearing sunglasses and laughing, a woman wearing sunglasses on her head looking down

Figure 3: Results of the LSTM and gLSTM model. We mark the generated sentence by LSTM and gLSTM respectively in green and red, the ground truth references in black and the most relevant retrieval results in blue. We observe that the retrieval results are helpful to caption generation. Notice that for the third example, the result of our model is not that accurately but still much better than the one of the LSTM model.