Multimodal Neural Language Models

Ryan Kiros, Ruslan Salakhutdinov, Richard Zemel
Department of Computer Science, University of Toronto

Review by: David Carlson

April 24, 2015

Introduction

- This paper focuses on multimodal neural language models: models of natural language that can be conditional on other modalities
- Can perform image retrival from text, as well as generating text or text retrival given an image
- Can generate sentence description for images without the use of templates or syntactic trees
- Extend ideas from the log bilinear model and feature extraction from images

Figure 1. Left two columns: Sample description retrieval given images. Right two columns: description generation. Each description was initialized to 'in this picture there is' or 'this product contains a', with 50 subsequent words generated.

Word Embeddings

- Instead of dealing with the words directly, many recent approaches embed words in $\mathbb{R}^D$ space
- The idea is that words with similar semantic meanings will be close in feature space
  - I.E. "cat" and "dog" near each other in embedded space
- The distance used is cosine similarity

$$D(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x}^T \boldsymbol{y}}{||\boldsymbol{x}||_2 \ ||\boldsymbol{y}||_2} \tag{1}$$

The Log-Bilinear (LBL) Model

- Let $\mathbf{R}$ denote the $K \times D$ matrix of word representation vectors where K is the vocabulary size, and r_i represents the i^{th} column of $\mathbf{R}$
- Let $(w_1, \ldots, w_{n-1})$ be a tuple of $n-1$ words, with $n-1$ the context size
- Let $\mathbf{C}^{(i)}$ be a $D \times D$ context parameter matrix
- The LBL makes a linear prediction of the next word representation as

$$\hat{\boldsymbol{r}} = \sum_{i=1}^{n-1} \mathbf{C}^i \boldsymbol{r}_{w_i} \tag{2}$$

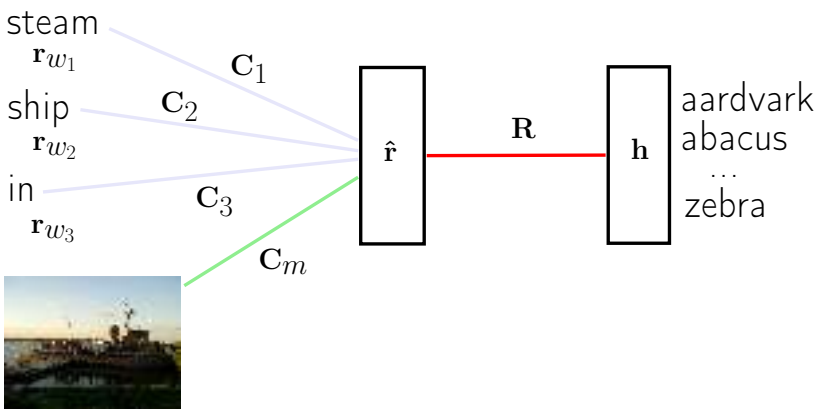
The conditional probability of the next word is given by

$$P(w_n | w_{1:n-1}) \propto \exp(\hat{\boldsymbol{r}}^T \boldsymbol{r}_i + b_i) \tag{3}$$

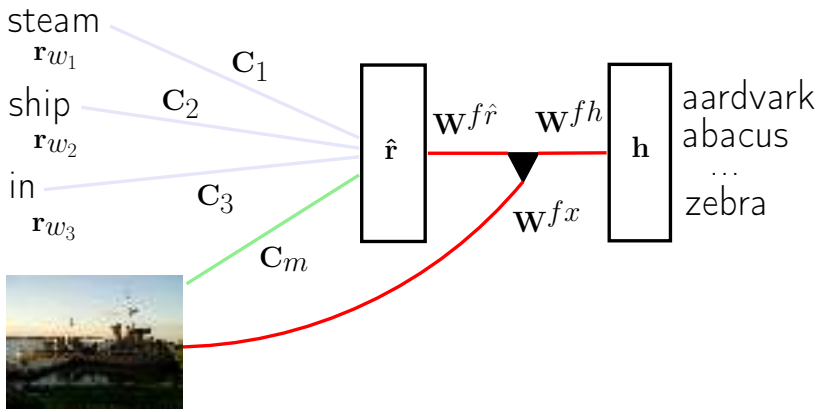
with $\boldsymbol{b} \in \mathbb{R}^K$ a word-specific bias vector.

Multimodal Log-Bilinear Models (MLBL)

- As before, we have previous tuple of words $(w_1, \ldots, w_n)$
- Add in a new associated vector $\boldsymbol{x} \in \mathbb{R}^M$
  - Vector comes from features of an associated modality
  - Assume features are precomputed
- Several MLBM models are proposed to handle this multimodal approach
  - Modality-Biased Log-Bilinear Model (MLBL-B): Adds an additional linear term on $\boldsymbol{x}$ to predict future words
  - Factored 3-way Log-Bilinear Model (MLBL-F): Changes word representation structure based on $\boldsymbol{x}$



(a) Modality-Biased Log-Bilinear Model (MLBL-B)



(b) Factored 3-way Log-Bilinear Model (MLBL-F)

Figure 2. Our proposed models. Left: The predicted next word representation $\hat{\mathbf{r}}$ is a linear prediction of word features $\mathbf{r}_{w_1}, \mathbf{r}_{w_2}, \mathbf{r}_{w_3}$ (blue connections) biased by image features $\mathbf{x}$. Right: The word representation matrix $\mathbf{R}$ is replaced by a factored tensor for which the hidden-to-output connections are gated by $\mathbf{x}$.

Modality-Biased Log-Bilinear Model (MLBL-B)

- The MLBL-B makes a linear prediction of the next word representation as

$$\hat{\boldsymbol{r}} = \left(\sum_{i=1}^{n-1} \mathbf{C}^i \boldsymbol{r}_{w_i} \right) + \mathbf{C}^{(m)} \boldsymbol{x} \tag{4}$$

- $\mathbf{C}^{(m)}$ is a $D \times M$ context matrix
- The conditional probability of the next word is the same as in the LBL model

$$P(w_n | w_{1:n-1}, \boldsymbol{x}) \propto \exp(\hat{\boldsymbol{r}}^T \boldsymbol{r}_i + b_i) \tag{5}$$

Factored 3-way Log-Bilinear Model (MLBL-F)

- Incorporates the features $\boldsymbol{x}$ by changing the word representation matrix
- Instead of a single word representation $\mathbf{R}$, have

$$\mathbf{R}^{(x)} = \sum_{m=1}^{M} x_m \mathbf{R}^{(m)} \tag{6}$$

- Instead of learning the tensor $\mathbf{R}$,factor $\mathbf{R}$ into three lower rank F matricies, $\mathbf{W}^{f\hat{f}} \in \mathbb{R}^{F \times D}$, $\mathbf{W}^{f\hat{f}} \in \mathbb{R}^{F \times D}$, and $\mathbf{W}^{f\hat{f}} \in \mathbb{R}^{F \times D}$, then

$$\mathbf{R}^{x} = (\mathbf{W}^{fh})^{T} \text{diag}(\mathbf{W}^{fx}\boldsymbol{x})\mathbf{W}^{f\hat{f}} \tag{7}$$

Factored 3-way Log-Bilinear Model (MLBL-F)

- The marginal distribution for the next word is

$$\mathbf{E} = (\mathbf{W}^{f\hat{r}})^T \mathbf{W}^{fh} \tag{8}$$

$$\hat{\boldsymbol{r}} = \left(\sum_{i=1}^{n-1} \mathbf{C}^{(i)} \mathbf{E}(:, w_i) \right) + \mathbf{C}^{(m)} \boldsymbol{x} \tag{9}$$

$$\boldsymbol{f} = (\mathbf{W}^{f\hat{r}} \hat{\boldsymbol{r}}) \odot (\mathbf{W}^{fx} \boldsymbol{x}) \tag{10}$$

$$P(w_n = i | w_{1:n-1}, \boldsymbol{x}) \propto \exp((\mathbf{W}^{fh}(:, i))^T \boldsymbol{f} + b_i) \tag{11}$$

Joint Image-Text Feature Learning

- Consider a corpus of text and images
- Want to jointly learn both image and word features
- For the images, set up a 2-layer convolutional neural network (CNN) with max pooling and Rectified Linear Unit
- Backprop gradients from the model to do training of the CNN

Generation and Retreival

- A standard approach to evaluating language models is perplexity

$$\log_2 C(w_{1:n}|x) = \frac{-1}{N} \sum_{w_{1:n}} \log_2 P(w_n = i|w_{1:n-1}, x) \tag{12}$$

- To retrieve an image given a text query $w_{1:N}$, compute $C(w_{1:N}|x)$ and return the images with lowest perplexity
- To retrieve text from an image query is "tricky"
  - "Easy" sentences can be returned for all the images
  - Instead of returning the lowest perplexity sequence, return the lowest ratio with the "average" image $\bar{\boldsymbol{x}}$

$$C(w_{1:N}|\boldsymbol{x})/C(w_{1:N}|\bar{\boldsymbol{x}}) \tag{13}$$

  - Alternatively, instead of the global average image, calculate a average over the k_r closest images in feature space, $\tilde{\boldsymbol{x}}$

Experiments

- SGD with momentum is used to learn the models
- Context size was set to 5
- 80% training, 20% testing
- Evaluated with Perplexity, Retreival, and Bleu (Papineni et al, 2002)
- Different image features are used in the experiments, Gist, DeCAF, and the proposed convolutional net

Table 1. Sample neighbors (by cosine similarity) of words learned from the SBU dataset. First row: neighbors from Collobert & Weston (2008) (C&W). Second row: neighbors from a LBL model (without images). Third row: neighbors from a MLBL-F model (with images).

gloomy	tranquil	sensuous	somber	bleak	cheerful	dreary
	dismal	slower	feeble	realistic	brighter	strong
	hazy	stormy	foggy	crisp	cloudless	dull
classroom	laptop	dorm	desk	computer	canteen	darkroom
	pub	cabin	library	bedroom	office	cottage
	library	desk	restroom	office	cabinet	kitchen
flower	bamboo	silk	gold	bark	flesh	crab
	bird	tiger	monster	cow	fish	leaf
	plant	flowers	fruit	green	plants	rose
lighthouse	breakwater	icefield	lagoon	nunnery	waterway	walkway
	monument	lagoon	kingdom	mosque	skyline	truck
	pier	ship	dock	castle	marina	pool
cup	championship	trophy	bowl	league	tournament	cups
	cider	bottle	needle	box	fashion	shoe
	bag	bottle	container	oil	net	jam
terrain	shorelines	topography	vegetation	convection	canyons	slopes
	seas	paces	descent	yards	rays	floors
	headland	chasm	creekbed	ranges	crest	pamagirri

Table 2. Results on IAPR TC-12. PPL refers to perplexity while B-n indicates Bleu scored with n-grams. Back-off GTn refers to n-grams with Katz backoff and Good-Turing discounting. Models which use a convolutional network are indicated by -conv, while -conv-R indicates using random images for conditioning. skmeans refers to the features of Kiros & Szepesvári (2012).

MODEL TYPE	PPL.	B-1	B-2	B-3
BACK-OFF GT2	54.5	0.323	0.145	0.059
BACK-OFF GT3	55.6	0.312	0.131	0.059
LBL	20.1	0.327	0.144	0.068
MLBL-B-CONV-R	28.7	0.325	0.143	0.069
MLBL-B-SKMEANS	**18.0**	0.349	0.161	0.079
MLBL-F-SKMEANS	20.3	0.348	0.165	0.085
MLBL-B-GIST	20.8	0.348	0.164	0.083
MLBL-F-GIST	28.8	0.341	0.151	0.074
MLBL-B-CONV	20.6	0.349	0.165	0.085
MLBL-F-CONV	21.7	0.341	0.156	0.073
MLBL-B-DECAF	24.7	**0.373**	**0.187**	**0.098**
MLBL-F-DECAF	21.8	0.361	0.176	0.092
GUPTA ET AL.	-	0.15	0.06	0.01
GUPTA & MANNEM	-	0.33	0.18	0.07

Table 3. Results on the Attributes Discovery dataset.

MODEL TYPE	PPL.	B-1	B-2	B-3
BACK-OFF GT2	117.7	0.163	0.033	0.009
BACK-OFF GT3	**93.4**	0.166	0.032	0.011
LBL	97.6	0.161	0.031	0.009
MLBL-B-CONV-R	154.4	0.166	0.035	0.012
MLBL-B-GIST	95.7	0.185	0.044	0.013
MLBL-F-GIST	115.1	0.182	0.042	0.013
MLBL-B-CONV	99.2	**0.189**	**0.048**	**0.017**
MLBL-F-CONV	113.2	0.175	0.042	0.014
MLBL-B-DECAF	98.3	0.186	0.045	0.014
MLBL-F-DECAF	133.0	0.178	0.041	0.012

Table 4. F-scores for retrieval on IAPR TC-12 when a text query is used to retrieve images ($T \rightarrow I$) or when an image query is used to retrieve text ($I \rightarrow T$). Each row corresponds to DeCAF, Conv and Gist features, respectively.

$T \rightarrow I$			$I \rightarrow T$		
BOW	MLBL-B	MLBL-F	BOW	MLBL-B	MLBL-F
0.890	0.889	**0.899**	**0.755**	0.731	0.568
0.726	0.788	**0.851**	0.687	0.719	**0.736**
0.832	0.799	0.792	0.599	**0.675**	0.612

Table 5. F-scores for retrieval on Attributes Discovery when a text query is used to retrieve images ($T \rightarrow I$) or when an image query is used to retrieve text ($I \rightarrow T$). Each row corresponds to DeCAF, Conv and Gist features, respectively.

$T \rightarrow I$			$I \rightarrow T$		
BOW	MLBL-B	MLBL-F	BOW	MLBL-B	MLBL-F
0.808	**0.852**	0.835	0.579	**0.580**	0.504
0.730	**0.839**	0.815	**0.607**	0.590	0.576
0.826	**0.844**	0.818	0.555	**0.621**	0.579