Multimodal Neural Language Models

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

Review by: David Carlson

April 24, 2015

- 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



change is in the air, what starts out as a swinging wristlet easily transforms. to an adorable handbad with just two quick clicks of the trigger clasp . or , you can attach it to a farger bag in a snap , the zipper closure reveals and interior packet and ample room for all the essentials . NUM wrist strap . trigger clasp lets yo ...



our 14k NUM / Act round diamond stud earrings feature two fiery , perfectly proportioned & meticulously matched diamonds, these earrings measure NUM NUM NUM Rmm in diameter, and the total carat weight of the diamonds mounted in these carrings is NUM / 4ct , the diamonds and flirty tear silhouette to glam up your are [/] / k in color and [2 in clarity , these earrings are made out ...

this product contains a slip resistant and mesh upper is fully designed for breathable durability , the detachable leather footbed is the high , they feature a lady - like footbed that light sophistication feet , style to help your thing , with traditional support

FOOTIOY



this product contains a variety of strategically placed peter stripes . multi - haed silk opper in pore waven red silk bow tie in newy messculine first. width , this hand - based silk ties from forzieri offers success to any wardrobe . decidedly blue , imported clean . NOM % silk .





'nine children , some of them waving at the cameral others are a bit shy and are looking away ; one person is pointing to the camera :

a room with white walls , a red tiled floor . a blue window. Two double beds with reddish bed covers , two lamos , a telephone and a nicture :



in this picture there is another grey pavement on the right ; three grey clouds and a blue sky in the background the houses and on the left before it : a dark green , wooded slopes behind it ; grey clouds in a light blue sky in the background - snow covered mountains

in this picture there is another wall in the background ; a man with a white chequered waistcoat and a small books there is food and a white train table and a window with black waistcoat and green trousers (tables in the background another man in a classroom with salt bedraver and

(日) (同) (三) (三)

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.

April 24, 2015 3/18

- 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

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

The Log-Bilinear (LBL) Model

- Let **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 **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

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

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

- As before, we have previous tuple of words (w_1, \ldots, w_n)
- Add in a new associated vector $\pmb{x} \in \mathbb{R}^M$
 - 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 *x* to predict future words
 - Factored 3-way Log-Bilinear Model (MLBL-F): Changes word representation structure based on *x*



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



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

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

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

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

- Incorporates the features x by changing the word representation matrix
- Instead of a single word representation R, have

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

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

$$\mathbf{R}^{x} = (\mathbf{W}^{fh})^{T} \operatorname{diag}(\mathbf{W}^{fx} \mathbf{x}) \mathbf{W}^{f\hat{r}}$$
(7)

• The marginal distribution for the next word is

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

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

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

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

- 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

• A standard approach to evaluating language models is perplexity

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

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

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

- 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 (with images). Third row: neighbors from a MLBL-F model (with images).

	tranquil	sensuous	somber	bleak	cheerful	dreary
gloomy	dismal	slower	feeble	realistic	brighter	strong
	hazy	stormy	foggy	crisp	cloudless	dull
	laptop	dorm	desk	computer	canteen	darkroom
classroom	pub	cabin	library	bedroom	office	cottage
	library	desk	restroom	office	cabinet	kitchen
	bamboo	silk	gold	bark	flesh	crab
flower	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
-	championship	trophy	bowl	league	tournament	cups
cup	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

æ

イロト イポト イヨト イヨト

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

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

э

Image: A math a math

16 / 18

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 0.726 0.832	0.889 0.788 0.799	0.899 0.851 0.792	0.755 0.687 0.599	0.731 0.719 0.675	0.568 0.736 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 0.730 0.826	0.852 0.839 0.844	0.835 0.815 0.818	0.579 0.607 0.555	0.580 0.590 0.621	0.504 0.576 0.579