Sentence-level Multilingual Multi-modal Embedding for Natural Language Processing

Iacer Calixto, Qun Liu

August 14, 2018

ADAPT Centre
School of Computing
Dublin City University
{*FirstName.LastName*}*@adaptcentre.ie*

Outline

Introduction

Introduction

- Distributional semantic models (DSMs): compute word (sentence, paragraph) vector representations from text based on word co-occurrence patterns.

- The meaning of a word depends on the “company it keeps” (Harris, 1954):
  - word2vec (Mikolov et al., 2013);
  - skip-thought vectors (Kiros et al., 2015);

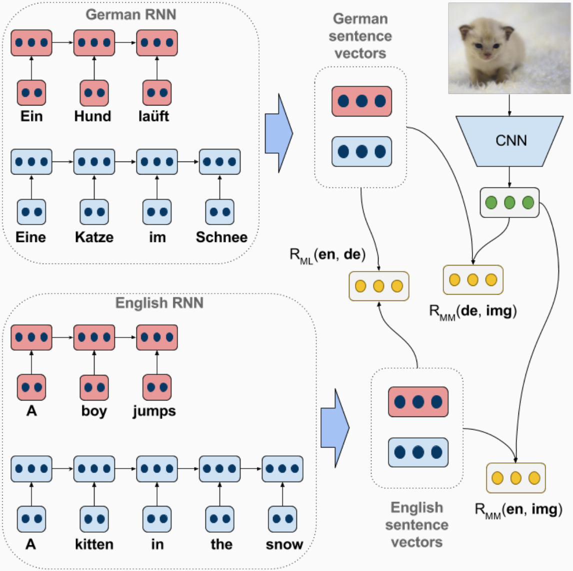
Introduction

- Distributional semantic models (DSMs): compute word (sentence, paragraph) vector representations from text based on word co-occurrence patterns.

- The meaning of a word depends on the "company it keeps" (Harris, 1954):
  - word2vec (Mikolov et al., 2013);
  - skip-thought vectors (Kiros et al., 2015);

Our Model

Multilingual Multi-modal Embedding



Model Formulation and Training

Language encoder(s):

for each language k (e.g., English, German):

- $\boldsymbol{X^k} = (\boldsymbol{x}_1^k, \boldsymbol{x}_2^k, \cdots, \boldsymbol{x}_{N_k}^k)$;
- $\boldsymbol{v}^k \in \mathbb{R}^{1024} = \text{RNN}(\boldsymbol{X^k})$;

Visual encoder(s):

for each image $\boldsymbol{i}$:

- $\boldsymbol{q} \in \mathbb{R}^{4096} = \text{CNN}(\boldsymbol{i})$;
- $\boldsymbol{d} \in \mathbb{R}^{1024} = W_I \cdot \boldsymbol{q}$;

Multi-modal ranking

$$
\begin{aligned}
\mathrm{R}_{\mathrm{MM}} = & \sum_{d} \sum_{r} \max\{0, \alpha - \boldsymbol{d}^T \cdot \boldsymbol{v}^k + \boldsymbol{d}^T \cdot \boldsymbol{v}_r^k\} + \\
& \sum_{v^k} \sum_{r} \max\{0, \alpha - (\boldsymbol{v}^k)^T \cdot \boldsymbol{d} + (\boldsymbol{v}^k)^T \cdot \boldsymbol{d}_r\}, \\
& k \in K, \qquad (1)
\end{aligned}
$$

- $\boldsymbol{d}$: correct image;
- $\boldsymbol{v}^k$: correct sentence in language k;
- $\boldsymbol{v}_r^k$: random sentence in language k;
- $\boldsymbol{d}_r$: random image;
- α: margin;

Multilingual ranking

$$
\begin{aligned}
\mathrm{R_{ML}} = \sum_{v^k}\sum_{r} &\max\{0, \alpha - (\boldsymbol{v}^k)^T \cdot \boldsymbol{v}^l + (\boldsymbol{v}^k)^T \cdot \boldsymbol{v}_r^l\} + \\
\sum_{v^l}\sum_{r} &\max\{0, \alpha - (\boldsymbol{v}^l)^T \cdot \boldsymbol{v}^k + (\boldsymbol{v}^l)^T \cdot \boldsymbol{v}_r^k\}, \\
& k \in K, l \in K, l \neq k, \qquad (2)
\end{aligned}
$$

- $\boldsymbol{v}^k$: correct sentence in language k;
- $\boldsymbol{v}^l$: correct sentence in language l;
- $\boldsymbol{v}_r^k$: random sentence in language k;
- $\boldsymbol{v}_r^l$: random sentence in language l;
- α: margin;

Final optimisation function

$$\min_{\theta_k, W_l} \beta \mathrm{R}_{\mathrm{MM}} + (1-\beta)\mathrm{R}_{\mathrm{ML}}, \forall k \in K,$$
$$0 \geq \beta \geq 1, \tag{3}$$

Results

Image–sentence ranking

	English							German				
	Skip-T.	VSE		Ours				VSE	Ours			
		paper	current	β=1	β=.75	$\beta = 0.5$	$\beta = 0.25$	current	β=1	β=.75	$\beta = 0.5$	$\beta = 0.25$
Sentence to image												
r@1	18.2	16.8	16.5	23.0 (+6.2)	**24.9** (+8.1)	22.3 (+5.5)	21.3 (+4.5)	13.5	**21.6** (+8.1)	20.3 (+6.8)	20.3 (+6.8)	19.5 (+6.0)
r@5	41.9	42.0	41.9	49.3 (+7.3)	**52.3** (+10.3)	48.3 (+6.3)	45.5 (+3.5)	36.6	**48.8** (+12.2)	45.0 (+8.4)	43.7 (+7.1)	43.0 (+6.4)
r@10	53.5	56.5	54.4	61.1 (+4.6)	**63.6** (+7.1)	58.4 (+1.9)	56.7 (+0.2)	49.0	**59.5** (+10.5)	56.6 (+7.6)	55.4 (+6.4)	54.4 (+5.4)
mrank	9	8	9	6	**5**	6	7	11	**6**	7	8	8
Image to sentence												
r@1	26.8	23.0	30.7	**33.1** (+2.4)	30.7 (+0.0)	27.4 (−3.3)	26.7 (−4.0)	30.5	**32.3** (+1.7)	24.9 (−5.6)	23.0 (−7.5)	21.8 (−8.7)
r@5	54.9	50.7	**57.8**	57.2 (−0.6)	55.4 (−2.4)	54.5 (−3.3)	51.4 (−6.4)	56.0	**58.6** (+2.6)	52.3 (−3.7)	48.4 (−7.6)	49.8 (−6.2)
r@10	67.5	62.9	**70.6**	68.7 (−1.9)	65.6 (−5.0)	64.0 (−6.6)	61.9 (−8.7)	**68.9**	68.1 (−0.8)	63.6 (−5.3)	62.8 (−6.1)	61.3 (−7.6)
mrank	5	5	**4**	4	4	4	5	**4**	4	5	6	6

- Skip-T.: skip-thought vectors (Kiros et al., 2015);
- VSE: visual-semantic embeddings (Kiros et al., 2014);
- `r@{1,5,10}`: recall-at-{1,5,10};
- `mrank`: median rank;

Semantic textual similarity

- We use our model to **compute the distance between a pair of sentences** (equivalent to **cosine similarity** and therefore lie in the $[0, 1]$ interval), where 0 **means complete dissimilarity and** 5 **complete similarity**.

Test set	VSE	Our model β=1	β=.75	β=.5	β=.25	SemEval best
in-domain data						
Image descriptions (2014)	.791	.797	.819	**.826**	.817	.821
Image descriptions (2015)	.834	.880	.882	.885	**.886**	.864

Table 1: Pearson rank correlation scores for semantic textual similarities in two different SemEval test sets.

- We train one ***weak* model**, one ***regular* model** and one ***optimised* NMT model** on the **translated Multi30k training data set (without images)** to translate from **English into German**.
    - **weak**: no regularisation;
    - **regular**: L2 = `1e-8`, dropout = `0.5`;
    - **optimised**: L2 = `0.0`, dropout = `0.2`;

Re-ranking n-best lists (n=20)

	BLEU		METEOR		TER	
Weak NMT model						
baseline	25.7		43.1		56.1	
+ VSE	25.8	(+0.1)	43.2	(+0.1)	56.1	(-0.0)
+ MLMME, $\beta = 1$	26.1	(+0.4)	**44.4**†‡	**(+1.3)**	55.5	(-0.6)
+ MLMME, $\beta = 0.75$	26.1	(+0.4)	44.3†‡	(+1.2)	55.9	(-0.2)
+ MLMME, $\beta = 0.5$	26.0	(+0.3)	43.9†‡	(+0.8)	55.9	(-0.2)
+ MLMME, $\beta = 0.25$	**26.3**†‡	**(+0.6)**	44.3†‡	(+1.2)	**55.2**†‡	**(-0.9)**
oracle	33.1		51.4		46.5	
Regular NMT model						
baseline	32.4		50.7		51.9	
+ VSE	32.2	(-0.2)	50.7	(+0.0)	52.6	(+0.7)
+ MLMME, $\beta = 1$	**33.8**†‡	**(+1.4)**	**51.4**†‡	**(+0.7)**	49.0‡	(-2.9)
+ MLMME, $\beta = 0.75$	33.5‡	(+1.1)	51.3†‡	(+0.6)	49.0‡	(-2.9)
+ MLMME, $\beta = 0.5$	**33.8**†‡	**(+1.4)**	**51.4**†‡	**(+0.7)**	**48.6**†‡	**(-3.3)**
! + MLMME, $\beta = 0.25$	33.7‡	(+1.3)	**51.4**†‡	**(+0.7)**	49.4‡	(-2.5)
oracle	41.9		59.3		41.2	
Optimised NMT model						
baseline	35.3		52.3		44.9	
+ VSE	32.3	(-3.0)	49.8	(-2.5)	46.5	(+1.6)
+ MLMME, $\beta = 1$	35.3‡	(+0.0)	**52.7**†‡	**(+0.4)**	**44.5**‡	**(-0.4)**
+ MLMME, $\beta = 0.75$	35.2‡	(-0.1)	52.6‡	(+0.3)	44.6‡	(-0.3)
+ MLMME, $\beta = 0.5$	35.1‡	(-0.2)	52.3‡	(+0.0)	44.9‡	(-0.0)
+ MLMME, $\beta = 0.25$	**35.7**‡	**(+0.4)**	**52.7**‡	**(+0.4)**	**44.5**‡	**(-0.4)**
oracle	43.2		59.7		37.8	

Table 2: Results improve significantly over the corresponding 1-best baseline (†) or over the translations obtained with the VSE re-ranker (‡) with $p = 0.05$.

Re-ranking n-best lists (n=50)

	BLEU		METEOR		TER	
Weak NMT model						
baseline	25.7		43.1		56.1	
+ VSE	25.8	(+0.1)	43.5†	(+0.4)	56.1	(-0.0)
+ MLMME, $\beta = 1$	26.2	(+0.5)	**44.6**†‡	**(+1.5)**	55.4	(-0.7)
+ MLMME, $\beta = 0.75$	**26.4**†	**(+0.7)**	44.5†‡	(+1.4)	55.6	(-0.5)
+ MLMME, $\beta = 0.5$	25.9	(+0.2)	43.9†	(+0.8)	55.9	(-0.0)
+ MLMME, $\beta = 0.25$	**26.4**†‡	**(+0.7)**	44.5†‡	(+1.4)	**55.0**†‡	**(-1.1)**
oracle	36.2		53.8		43.4	
Regular NMT model						
baseline	32.4		50.7		51.9	
+ VSE	32.7	(-0.3)	50.8	(+0.1)	51.4	(-0.5)
+ MLMME, $\beta = 1$	**34.2**†‡	**(+1.8)**	**51.6**†‡	**(+0.9)**	48.3‡	(-3.6)
+ MLMME, $\beta = 0.75$	34.1†‡	(+1.7)	**51.6**†‡	**(+0.9)**	47.6†‡	(-4.3)
+ MLMME, $\beta = 0.5$	34.0†‡	(+1.6)	51.4†‡	(+0.7)	**47.3**†‡	**(-4.6)**
+ MLMME, $\beta = 0.25$	34.1†‡	(+1.7)	**51.6**†‡	**(+0.9)**	48.5‡	(-3.4)
oracle	46.6		61.8		34.1	
Optimised NMT model						
baseline	35.3		52.3		44.9	
+ VSE	30.7	(-4.6)	47.9	(-4.4)	48.6	(+3.7)
+ MLMME, $\beta = 1$	35.4‡	(+0.1)	**52.7**†‡	**(+0.4)**	**44.4**†‡	**(-0.5)**
+ MLMME, $\beta = 0.75$	35.2‡	(-0.1)	52.5‡	(+0.2)	44.7‡	(-0.2)
+ MLMME, $\beta = 0.5$	35.1‡	(-0.2)	52.3‡	(+0.0)	44.7‡	(-0.2)
+ MLMME, $\beta = 0.25$	**35.6**‡	**(+0.3)**	52.6‡	(+0.3)	**44.4**†‡	**(-0.5)**
oracle	46.3		61.9		34.9	

Table 3: Results improve significantly over the corresponding 1-best baseline (†) or over the translations obtained with the VSE re-ranker (‡) with $p = 0.05$.

Conclusions

Conclusions

- **multilingual multimodal embedding** model trained with a **modified pairwise ranking loss objective**;
- **promising results** in **three downstream NLP tasks**:
- ISR:
  - **consistent improvements in image→sentence** ranking;
  - **sentence→image ranking**;
- STS:
  - **consistent improvements** in **in-domain tasks**;
  - **out-of-domain tasks**;
- NMT:
  - **significant improvements** in METEOR in **n-best re-ranking**;
  - 20-best lists and 50-best lists;
  - weak, regular, and optimised NMT baselines;

Conclusions

- **multilingual multimodal embedding** model trained with a **modified pairwise ranking loss objective**;
- **promising results** in **three downstream NLP tasks**:
- ISR:
  - **consistent improvements in image→sentence** ranking;
  - **sentence→image ranking**;
- STS:
  - **consistent improvements** in **in-domain tasks**;
  - **out-of-domain tasks**;
- NMT:
  - **significant improvements** in METEOR in **n-best re-ranking**;
  - 20-best lists and 50-best lists;
  - weak, regular, and optimised NMT baselines;

Conclusions

- **multilingual multimodal embedding** model trained with a **modified pairwise ranking loss objective**;
- **promising results** in **three downstream NLP tasks**:
- ISR:
  - **consistent improvements in image→sentence** ranking;
  - **sentence→image ranking**;
- STS:
  - **consistent improvements** in **in-domain tasks**;
  - **out-of-domain tasks**;
- NMT:
  - **significant improvements** in METEOR in **n-best re-ranking**;
  - 20-best lists and 50-best lists;
  - weak, regular, and optimised NMT baselines;

Conclusions

- **multilingual multimodal embedding** model trained with a **modified pairwise ranking loss objective**;
- **promising results** in **three downstream NLP tasks**:
- ISR:
  - **consistent improvements in image→sentence** ranking;
  - **sentence→image ranking**;
- STS:
  - **consistent improvements** in **in-domain tasks**;
  - **out-of-domain tasks**;
- NMT:
  - **significant improvements** in METEOR in **n-best re-ranking**;
  - 20-best lists and 50-best lists;
  - weak, regular, and optimised NMT baselines;

Conclusions

- **multilingual multimodal embedding** model trained with a **modified pairwise ranking loss objective**;
- **promising results** in **three downstream NLP tasks**:
- ISR:
  - **consistent improvements in image→sentence** ranking;
  - **sentence→image ranking**;
- STS:
  - **consistent improvements** in **in-domain tasks**;
  - **out-of-domain tasks**;
- NMT:
  - **significant improvements** in METEOR in **n-best re-ranking**;
  - 20-best lists and 50-best lists;
  - weak, regular, and optimised NMT baselines;

References

Harris, Z. (1954). Distributional structure. *Word*, 10(23):146–162.

Kiros, R., Salakhutdinov, R., and Zemel, R. S. (2014). Unifying visual-semantic embeddings with multimodal neural language models. *CoRR*, abs/1411.2539.

Kiros, R., Zhu, Y., Salakhutdinov, R., Zemel, R. S., Torralba, A., Urtasun, R., and Fidler, S. (2015). Skip-thought vectors. *CoRR*, abs/1506.06726.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *CoRR*, abs/1310.4546.

Thank you!

Questions?