A Visual Attention Grounding Neural Model for Multimodal Machine Translation

Abstract

Jointly optimizing the learning of a translator and 'visual-language embedding' by leveraging visual attention grounding mechanism linking visual semantics and textual semantics.

New (multimodal multilingual) product description dataset crawled from IKEA.

Introduction

Multimodal Machine Translation

Source sentence + image → Target Sentence (translation)

“In this setting, translation is expected to be more accurate compared to purely text-based translation, as the visual context could help resolve ambiguous multi-sense words.”

Uses: multimedia news, web products with images, movie subtitles

“ However, how to effectively integrate the visual information still remains a challenging problem.”

Improvements on automatic metrics are too tiny.

Text-only (no image) models have been competetive and sometimes better.

Multitask learning mechanism

(i) Translation

(ii) constructing a vision-language joint semantic embedding

", we develop a visual attention mechanism to learn an attention vector that values the words that have closer semantic relatedness with the visual context. The attention vector is then projected to the shared embedding space to initialize the translation decoder such that the source sentence words that are more related to the visual semantics have more influence during the decoding stage."

“lack of a large-scale, realistic dataset.”

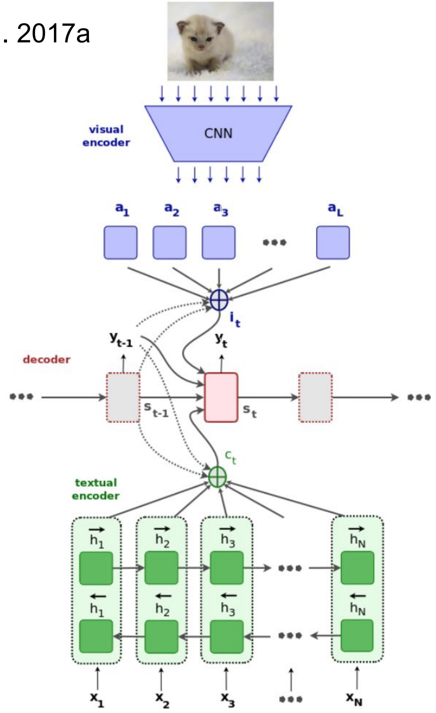
IKEA dataset

3600 products: Images + En, De, Fr descriptions.

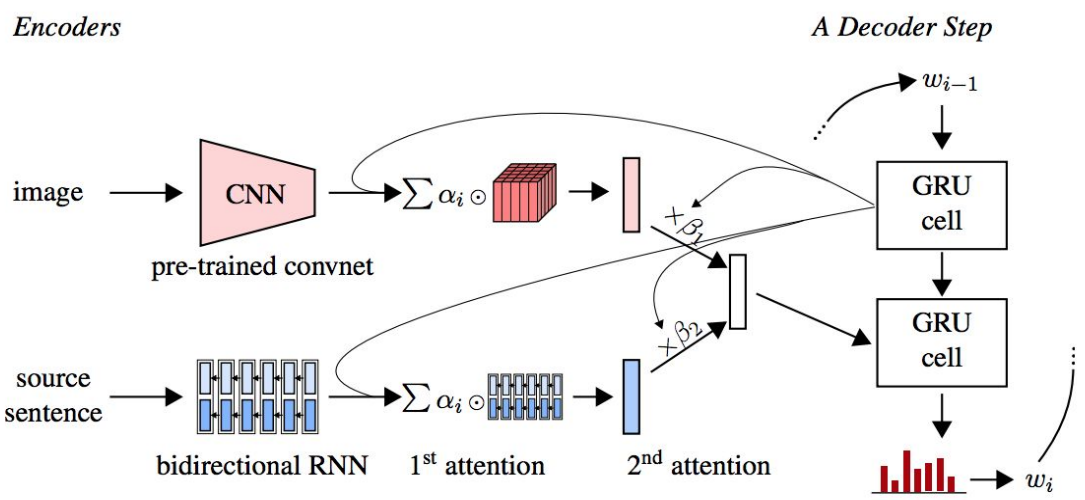
Related work

Separate attention over image and text and then merge the two.

Calixto et al. 2017a

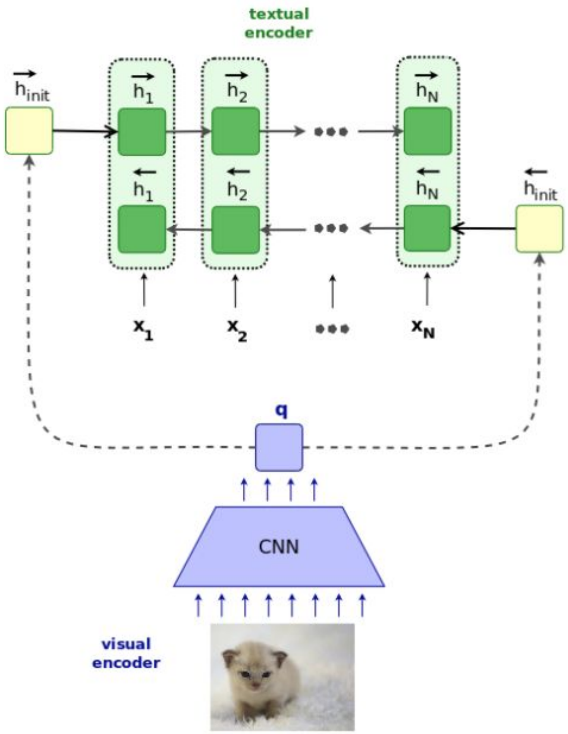


Helcl and Libovický, 2017

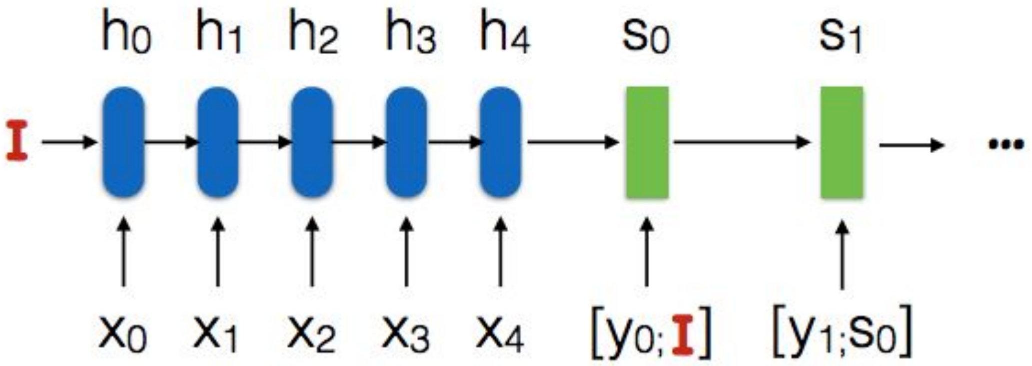


Merge image and text before attention

Calixto et al., 2017b



Ma et al., 2017

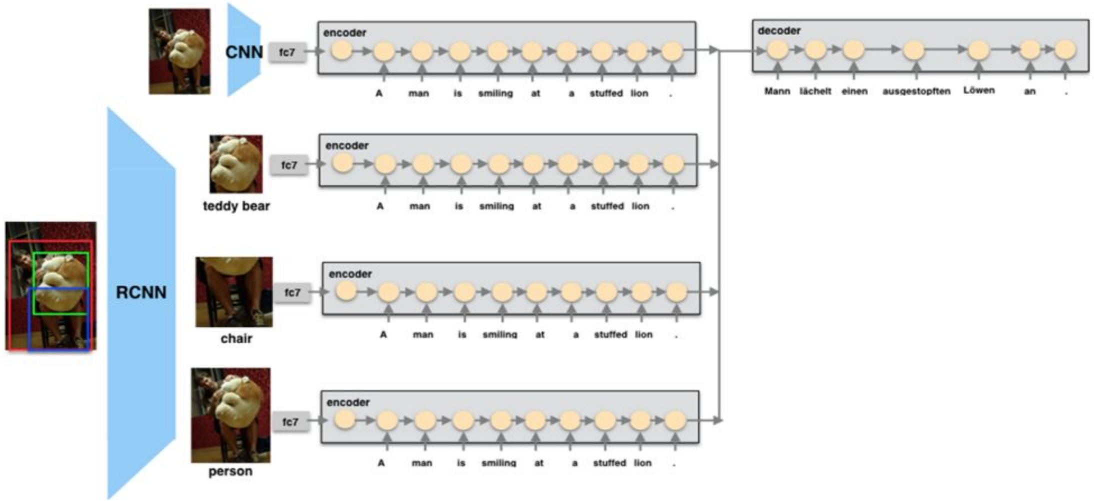


But Text-only system performed better

Zhang et al. 2017

SMT outputs were reranked by NMT

Best MMT in 2016



Huang et al., 2016

Best MMT in 2017

Element-wise product of image vector and text context vector. (Caglayan et al. 2017)

But gains too small to be conclusive.

Multimodal shared space literature

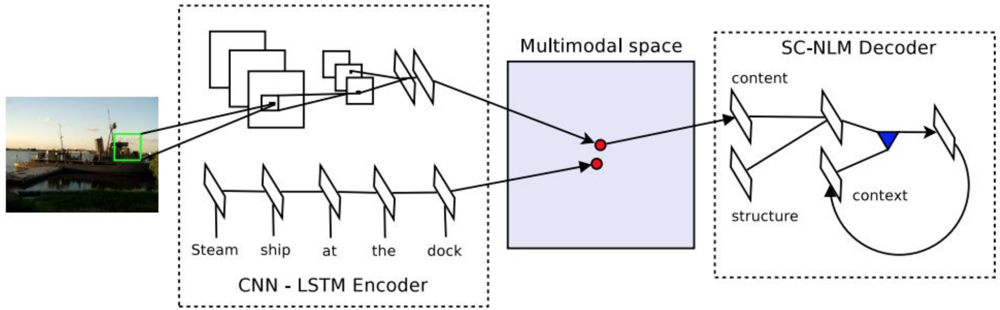


Figure 2: **Encoder:** A deep convolutional network (CNN) and long short-term memory recurrent network (LSTM) for learning a joint image-sentence embedding. **Decoder:** A new neural language model that combines structure and content vectors for generating words one at a time in sequence.

(Kiros et al., 2014)

$$S_I(\boldsymbol{v}^k, \boldsymbol{d}) = \sum_d \sum_r \max\{0, \alpha - \boldsymbol{s}_i(\boldsymbol{d}, \boldsymbol{v}^k) + \boldsymbol{s}_i(\boldsymbol{d}, \boldsymbol{v}^k_r)\} + \sum_{v^k} \sum_r \max\{0, \alpha - \boldsymbol{s}_i(\boldsymbol{v}^k, \boldsymbol{d}) + \boldsymbol{s}_i(\boldsymbol{v}^k, \boldsymbol{d}_r)\}, \quad \forall k \in K, \tag{1}$$

“a neural language model to learn a visual-semantic embedding space by optimizing a ranking objective, where the distributed representation helps generate image captions”

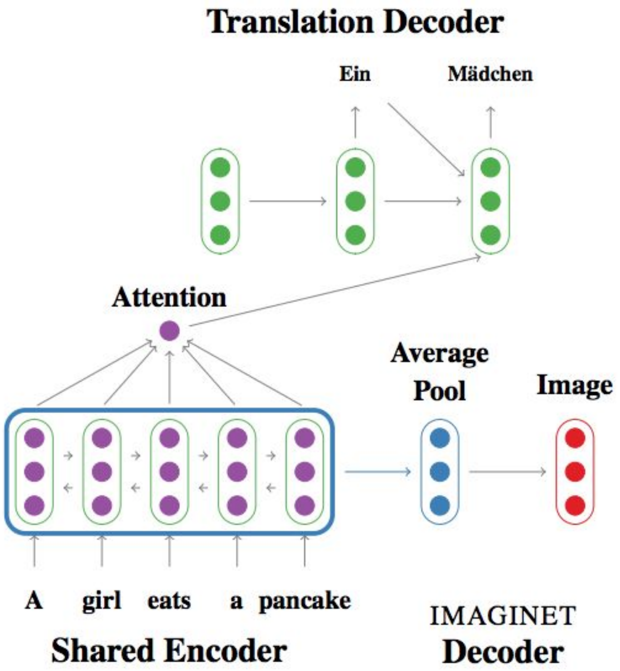
Later used by Calixto et al. 2017c and Gella et al. 2017

VAG-NMT inspired by Imagination model below

Difference is

Auxiliary task is not recreation of image feature

Visual-text attention mechanism



Elliott and Kádár, 2017

Visual Attention Grounding - NMT

Joint Objective function

$$J(\theta_T, \phi_V) = \alpha J_T(\theta_T) + (1 - \alpha) J_V(\phi_V),$$

T = Translation

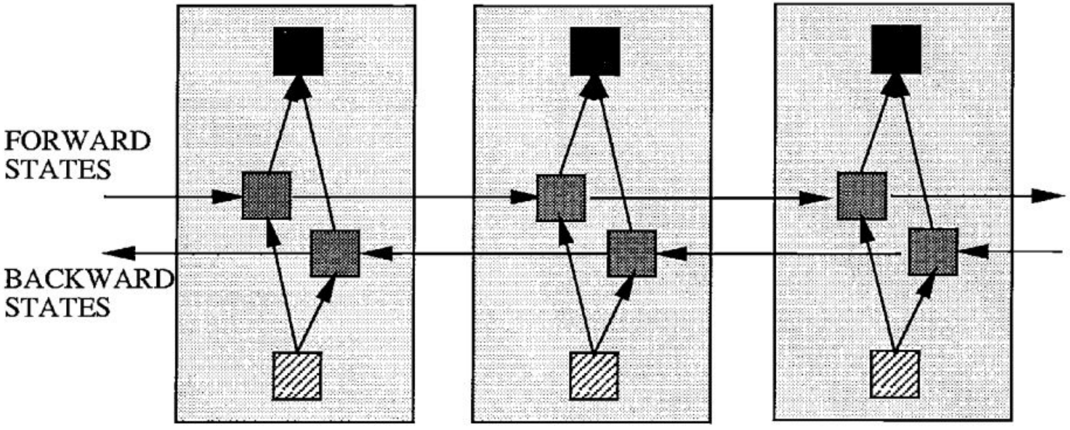
V = Joint visual-language embedding learning

Encoder

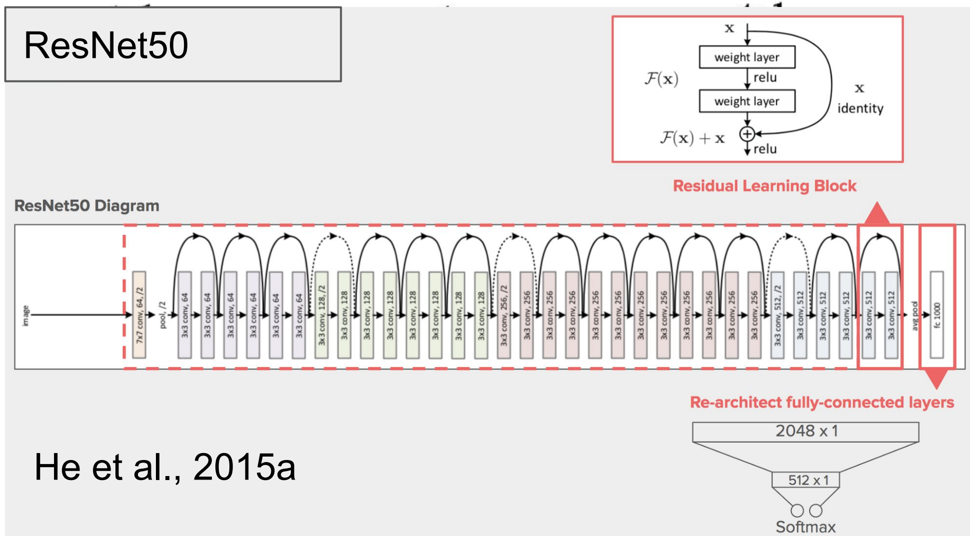
Bidirectional GRU for text

ResNet 50 for image representation

Schuster and Paliwal, 1997



ResNet50



He et al., 2015a

Visual Attention Mechanism

Specifically, we produce a set of weights $\beta = \{\beta_1, \beta_2, \ldots, \beta_n\}$ with our visual-attention mechanism, where the attention weight β_i for the i'th word is computed as:

$$\beta_i = \frac{\exp(z_i)}{\sum_{l=1}^{N} \exp(z_l)}, \qquad (2)$$

and $z_i = \tanh(W_v v) \cdot \tanh(W_h h_i)$ is computed by taking the dot product between the transformed encoder hidden state vector h_i and the transformed image feature vector v, and W_v and W_h are the association transformation parameters.

Simply take weighted average

$$t = \sum_{i=1}^{n} \beta_i h_i$$

Project t and v

$$t_{emb} = \tanh(W_{t_{emb}} t + b_{t_{emb}})$$

$$v_{emb} = \tanh(W_{v_{emb}} v + b_{v_{emb}})$$

Minimize pair-wise ranking loss

$$J_V(\phi_V) = \sum_p \sum_k \max\{0, \gamma - s(v_p, t_p) + s(v_p, t_{k \neq p})\}$$
$$+ \sum_k \sum_p \max\{0, \gamma - s(t_k, v_k) + s(t_k, v_{p \neq k})\},$$

Initializing decoder

$$s_0 = \tanh(W_{init}(\lambda t + (1 - \lambda)\frac{1}{N}\sum_i^N h_i)),$$

Translator

Standard conditional GRU decoder

Softmax Ot

Cross entropy loss function

IKEA Dataset

IKEA and UNIQLO websites

3600 products

Description in En, Fr, De are crawled

60-70 word long sentences (very long)

Pair	EN-DE		EN-FR	
Language	EN	DE	EN	FR
Tokens	256355	216892	239966	275251
Min length	6	6	6	6
Max length	343	324	334	469
Avg length	71.4	60.4	72.2	82.9
Std dev	46.3	39.1	47.2	54.7
Vocabulary	6601	10468	6442	7575

Evaluation

BLEU and METEOR

Multi30K 1000, MSCOCO 461,   IKEA 3600

[But what are the training-test splits on IKEA?

Settings: Standard stuff (including BPE)

Comparison made with LIUMCVC and Imagination and standard NMT

Not with other submissions in general.

	English → German		English → French	
Method	BLEU	METEOR	BLEU	METEOR
Imagination (Elliott and Kádár, 2017)	30.2	51.2	N/A	N/A
LIUMCVC (Caglayan et al., 2017)	31.1 ± 0.7	52.2 ± 0.4	52.7 ± 0.9	69.5 ± 0.7
Text-Only NMT	$\mathbf{31.6 \pm 0.5}$	52.2 ± 0.3	53.5 ± 0.7	70.0 ± 0.7
VAG-NMT	$\mathbf{31.6 \pm 0.3}$	52.2 ± 0.3	$\mathbf{53.8 \pm 0.3}$	$\mathbf{70.3 \pm 0.5}$

Table 1: Translation results on the Multi30K dataset

	English → German		English → French	
Method	BLEU	METEOR	BLEU	METEOR
Imagination (Elliott and Kádár, 2017)	28.0	**48.1**	N/A	N/A
LIUMCVC (Caglayan et al., 2017)	27.1 ± 0.9	47.2 ± 0.6	43.5 ± 1.2	63.2 ± 0.9
Text-Only NMT	27.9 ± 0.6	47.8 ± 0.6	44.6 ± 0.6	64.2 ± 0.5
VAG-NMT	$\mathbf{28.3 \pm 0.6}$	48.0 ± 0.5	$\mathbf{45.0 \pm 0.4}$	$\mathbf{64.7 \pm 0.4}$

Table 2: Translation results on the Ambiguous COCO dataset

	English → German		English → French	
Method	BLEU	METEOR	BLEU	METEOR
LIUMCVC-Multi	59.9 ± 1.9	63.8 ± 0.4	58.4 ± 1.6	64.6 ± 1.8
Text-Only NMT	61.9 ± 0.9	65.6 ± 0.9	65.2 ± 0.7	$\mathbf{69.0 \pm 0.2}$
VAG-NMT	$\mathbf{63.5 \pm 1.2}$	$\mathbf{65.7 \pm 0.1}$	$\mathbf{65.8 \pm 1.2}$	68.9 ± 1.4

Table 3: Translation results on the IKEA dataset

Evaluating the embedding Recall@K

Get K nearest neigbor images

Check if correct image in it or not

64% R@1, 88.6% R@5, and 93.8% R@10 Multi30K

58.13% R@1, 87.38% R@5 and 93.74% R@10 IKEA

41.35% R@1, 85.48% R@5 and 92.56% R@10 MSCOCO

Human Eval

	MSCOCO	Multi30K	IKEA
Text-Only NMT	76	**72**	75
VAG-NMT	**94**	71	**82**
Tie	30	57	43

Table 4: Human evaluation results

Discussion and Conclusion

Lets just read it out from the paper!