Neural Machine Translation with Universal Visual Representation

ICLR 2020, Addis Ababa, Ethiopia

Zhuosheng Zhang[♣], Kehai Chen[♠], Rui Wang[♠,*],

Masao Utiyama[♠], Eiichiro Sumita[♠], Zuchao Li[♣], Hai Zhao[♣,*]

♣ Shanghai Jiao Tong University, China

♠ National Institute of Information and Communications Technology (NICT), Japan

Overview

TL;DR: universal visual representation for neural machine translation (NMT) using retrieved images with similar topics to source sentence, extending image applicability in NMT.

Motivation:

<u>1. Annotation Difficulty:</u>

- Parallel **sentence-image pairs**
- The **high cost** of annotation

<u>2. Limited Diversity:</u>

- A sentence is paired by only **a single image**.
- Weak in capturing the **diversity** of visual clues.

Solution:

- Apply visual representation to **text-only NMT** and **low-resource NMT**
- Propose a **universal visual representation** (VR) method

  1) relying only on **image-monolingual** instead of **image-bilingual** annotations

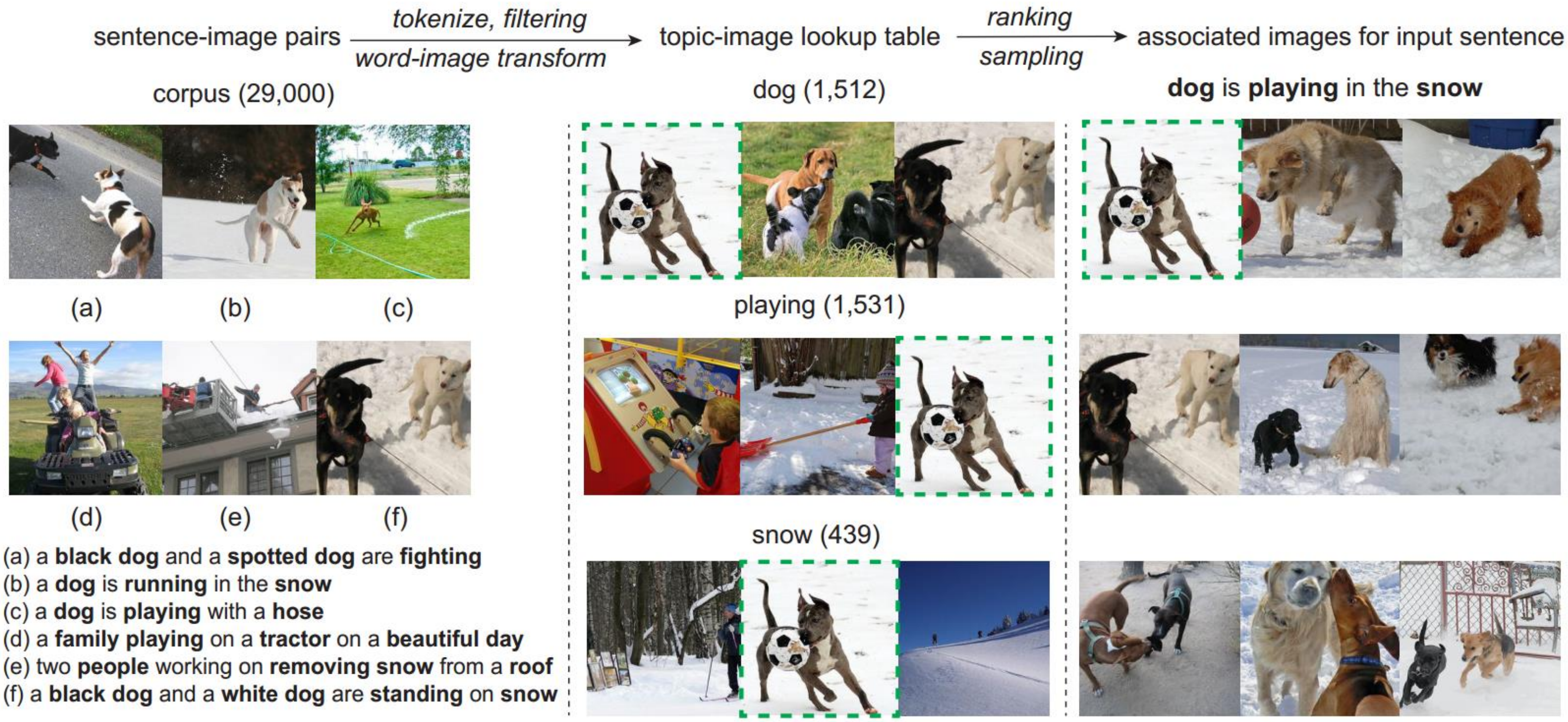
  2) breaking the bottleneck of using visual information in NMT

Paper: *https://openreview.net/forum?id=Byl8hhNYPS*

Code: *https://github.com/cooelf/UVR-NMT*

Universal Visual Retrieval

- **Lookup Table**: Transform the existing **sentence-image pairs** into **topic-image lookup table** from a small-scale multimodel dataset **Multi30K**
- **Image Retrieval**: a group of **images** with similar **topic** to the the **source sentence** will be retrieved from the topic-image lookup table learned by **TF-IDF**.

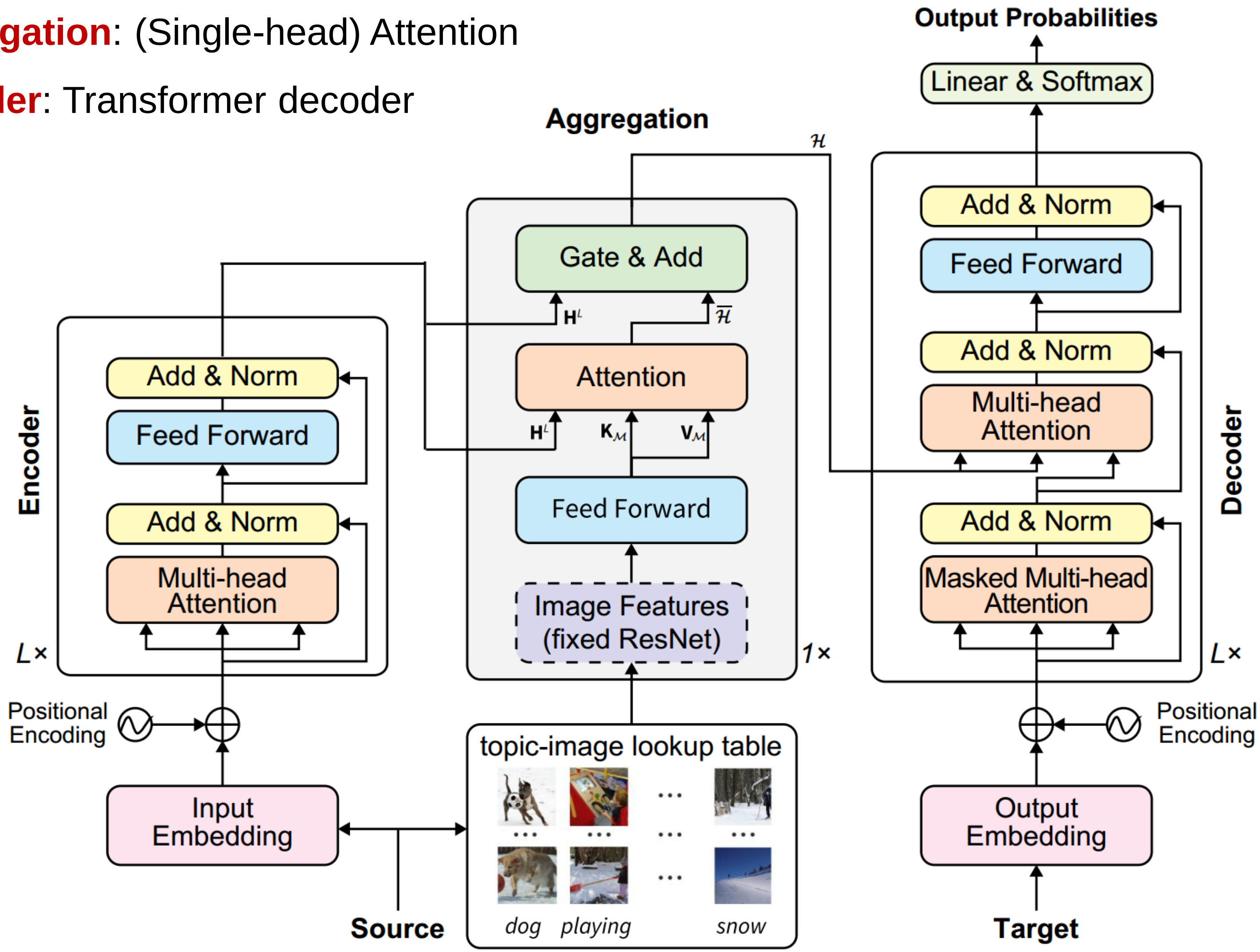


NMT With Universal Visual Representation

Encoder: Text (Transformer encoder), Image (ResNet)

Aggregation: (Single-head) Attention

Decoder: Transformer decoder



Experiments

NMT: WMT’16 EN-RO, WMT’14 EN-DE, WMT’14 EN-DE

System	Architecture	EN-RO BLEU	EN-RO #Param	EN-DE BLEU	EN-DE #Param	EN-FR BLEU	EN-FR #Param
Existing NMT systems							
Vaswani et al. (2017)	Trans. (base)	N/A	N/A	27.3	N/A	38.1	N/A
	Trans. (big)	N/A	N/A	28.4	N/A	41.0	N/A
Lee et al. (2018)	Trans. (base)	32.40	N/A	24.57	N/A	N/A	N/A
Our NMT systems							
This work	Trans. (base)	32.66	61.54M	27.31	63.44M	38.52	63.83M
	+VR	**33.78++**	63.04M	**28.14++**	64.94M	**39.64++**	65.33M
	Trans. (big)	33.85	207.02M	28.45	210.88M	41.10	211.66M
	+VR	**34.46+**	211.02M	**29.14++**	214.89M	**41.83+**	215.66M

MMT: Multi30K

System	Architecture	EN-DE Test2016	EN-DE Test2017	EN-DE #Param	EN-FR Test2016	EN-FR Test2017	EN-FR #Param
Existing NMT systems							
Calixto et al. (2017)	RNN	33.7	N/A	N/A	N/A	N/A	N/A
Elliott et al. (2017)	RNN	N/A	19.3	N/A	N/A	44.3	N/A
Elliott & Kádár (2017)	Imagination	36.8	N/A	N/A	N/A	N/A	N/A
Ive et al. (2019)	Trans. (big)	36.4	N/A	N/A	59.0	N/A	N/A
	Del	38.0	N/A	N/A	60.1	N/A	N/A
Our MMT systems							
This work	MMT. (base)	35.09	27.10	50.72M	57.40	48.02	50.65M
	MMT. (big)	35.60	28.02	190.58M	57.87	49.63	190.43M
	Trans. (base)	35.59	26.31	49.15M	57.88	48.55	49.07M
	+VR	**35.72**	**26.87**	50.72M	**58.32**	**48.69**	50.65M
	Trans. (big)	36.86	27.62	186.38M	56.97	48.17	186.23M
	+VR	**36.94**	**28.63**	190.58M	**57.53**	**48.46**	190.43M

Ablations of Hyper-parameters

Figure 4: Influence of the number of images on the BLEU score.

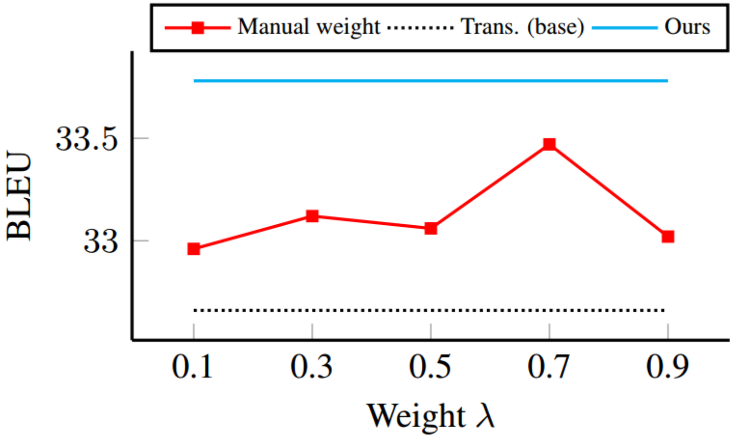


Figure 5: Quantitative study of the gating weight λ.

- A **modest** number of pairs would be beneficial.
- The degree of dependency for image information varies for each source sentence, indicating the necessity of **automatically learning** the gating weights.

Ablations of Encoders

We replace the ResNet50 feature extractor with

1)ResNet101;

2)ResNet152;

3)Caption: that adopts a standard image captioning model (Xu et al., 2015b);

4)Shuffle: shuffle the image features but keep the lookup table;

5)Random Init: randomly initialize the image embedding but keep the lookup table;

6)Random Mapping: randomly retrieve unrelated images.

Method	VR	Res101	Res152	Caption	Shuffle	Random Init	Random Mapping
BLEU	33.78	33.63	33.87	33.58	33.53	33.28	32.14

- More effective contextualized representation from the visual clue combination instead of just the single image enhancement for encoding each individual sentence or word.

Discussion

Why does it work:

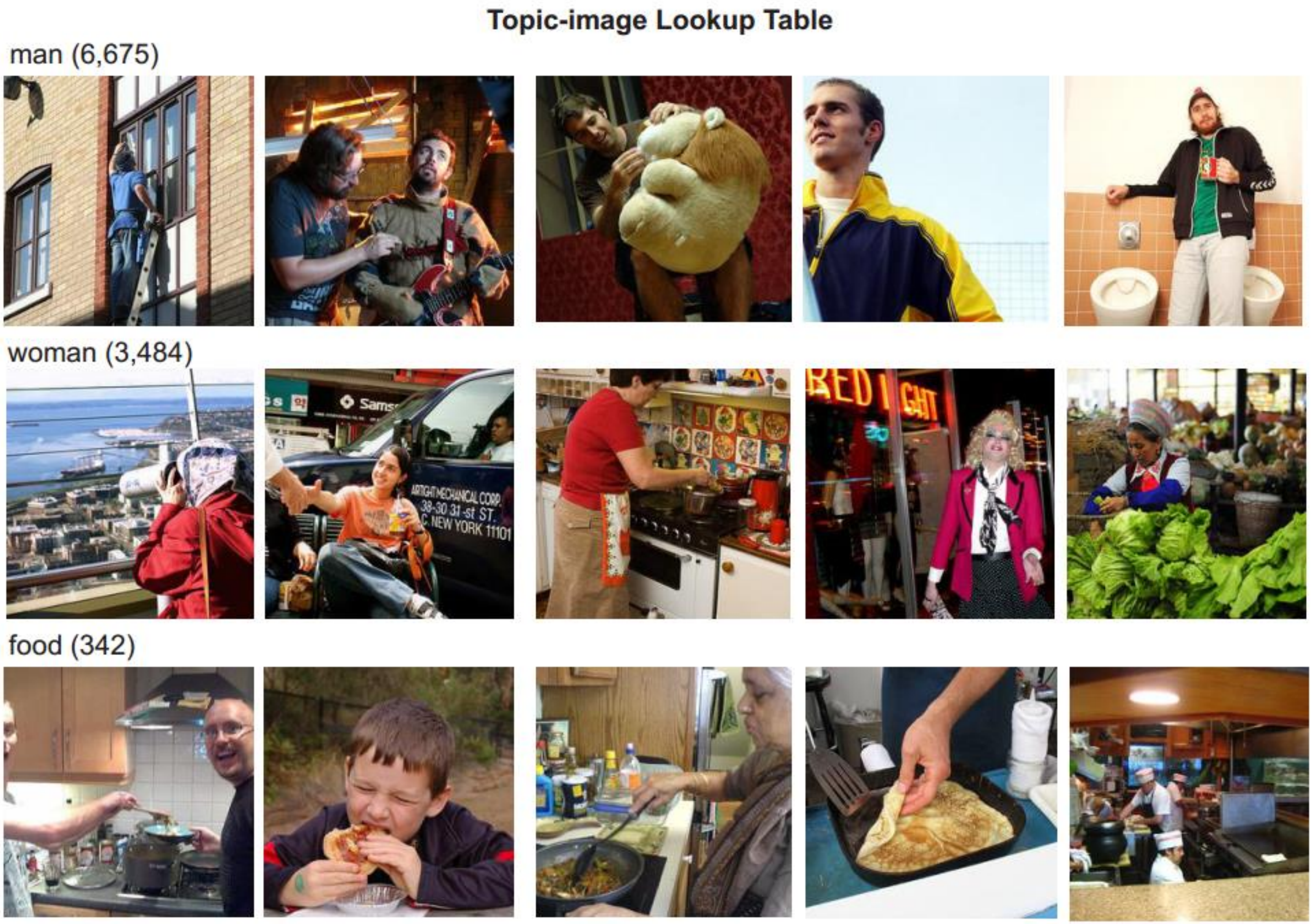
- the content connection of the sentence and images;
- the topic-aware co-occurrence of similar images and sentences.
  - *the sentences with similar meanings would be likely to pair with similar even the same images.*

A girl in a purple tutu dances in the yard.
A little girl is walking over a path of numbers.
A girl jumping rope on a sidewalk near a parking garage.
A young girl washes an automobile.

Highlights:

- Universal: potential for general text-only tasks, e.g., using the images as topic guidance.
- Diverse: diverse information entailed in the grouped images after retrieval.

Lookup Table



Retrieved Images

a **man walks** by a **silver vehicle**

an **elderly woman pan frying food** in a **kitchen**

small **boy carries** a **soccer ball** on a **field**

Thanks!
Q&A