Multimodal Machine Translation with Embedding Prediction

Tosho Hirasawa, Hayahide Yamagishi, Yukio Matsumura,
Mamoru Komachi

hirasawa-tosho@ed.tmu.ac.jp

Tokyo Metropolitan University

NAACL SRW 2019

Multimodal Machine Translation

TOKYO METROPOLITAN UNIVERSITY

- Practical application of machine translation
- Translate a source sentence along with related nonlinguistic information
  - Visual information

deux jeunes filles sont assises dans la rue , mangeant du maïs .

two young girls are sitting on the street eating corn .

Issue of MMT

TOKYO METROPOLITAN UNIVERSITY

- Multi30k [Elliott et al., 2016] has only small mount of data
  - Statistic of training data

	Sentences	Tokens	Types
English	29,000	377,534	10,210
French		409,845	11,219

- Hard to train rare word translation
  - Tend to output synonyms guided by language model

Source	deux jeunes filles sont assises dans la rue , mangeant du maïs .
Reference	two young girls are sitting on the street eating corn .
NMT	two young girls are sitting on the street eating food .

Previous Solutions

TOKYO METROPOLITAN UNIVERSITY

- Parallel corpus without images [Elliott and Kádár, 2017; Grönroos et al., 2018]
  - Out-of-domain data
  - Pseudo in-domain data by filtering general domain data
- Pseudo-parallel corpus [Sennrich et al., 2016; Helcl et al., 2018]
  - Back-translation of caption/monolingual data
- Monolingual data
  - Pretrained Word Embedding
    - Seldomly studied

Motivation

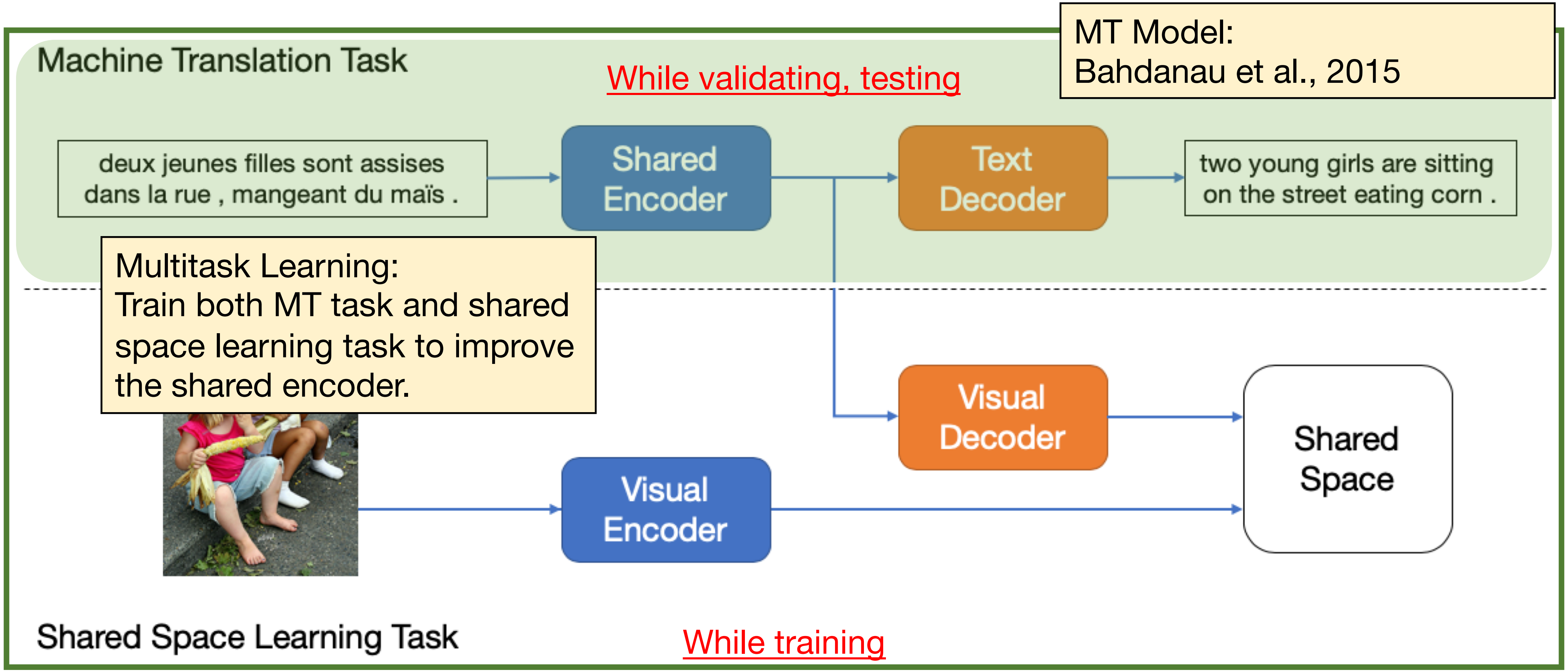
TOKYO METROPOLITAN UNIVERSITY

- Introduce pretrained word embedding to MMT
  - Improve rare word translation in MMT
  - Pretrained word embeddings with conventional MMT?
    - See our paper on MT Summit 2019 (https://arxiv.org/abs/1905.10464) !
- Pretrained Word Embedding in text-only NMT
  - Initialize embedding layers in encoder/decoder [Qi et al., 2018]
    - ✓ Improve overall performance in low-resource domain
  - Search-based decoder with continuous output [Kumar and Tsvetkov, 2019]
    - ✓ Improve rare word translation

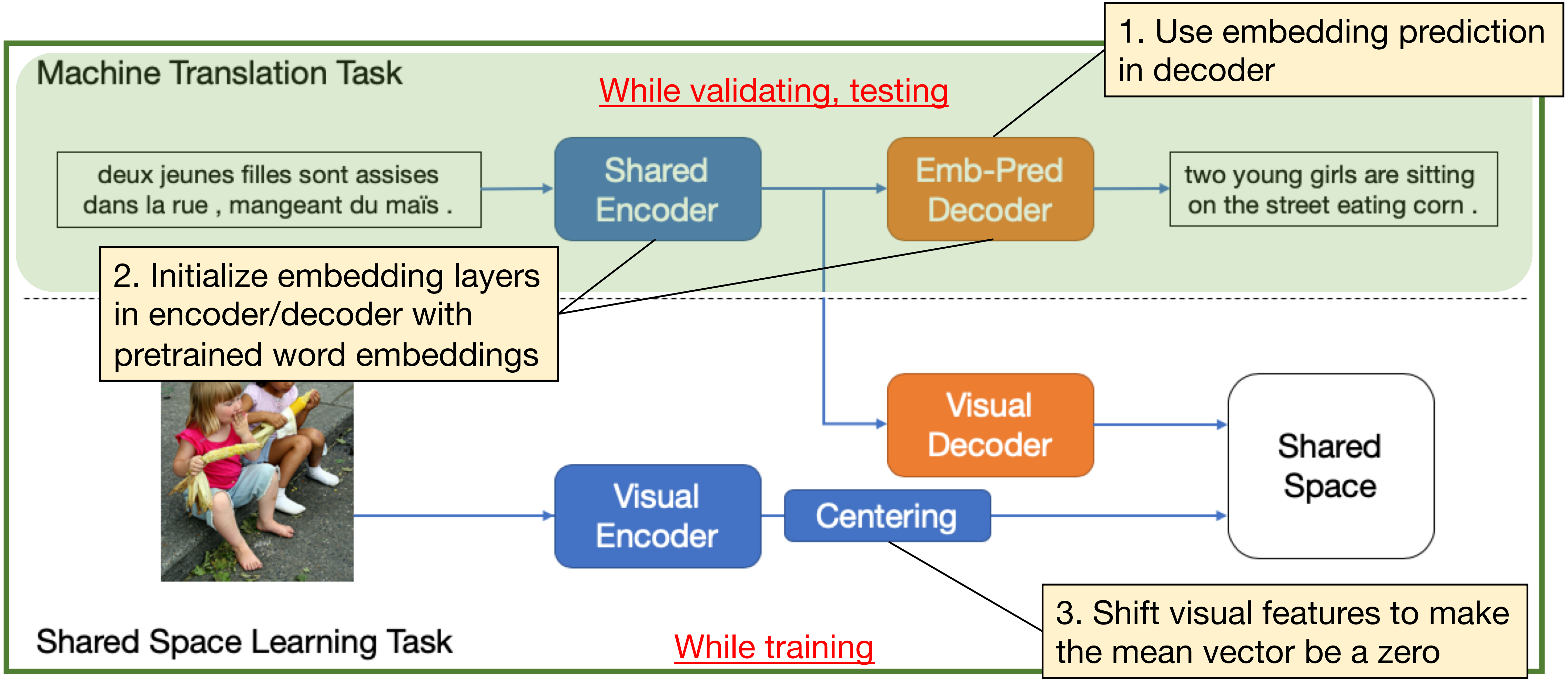
1. Multimodal Machine Translation
2. MMT with Embedding Prediction
3. Pretrained Word Embedding
4. Result & Conclusion

Baseline: IMAGINATION [Elliot and Kádáar, 2017]

TOKYO METROPOLITAN UNIVERSITY



MMT with Embedding Prediction



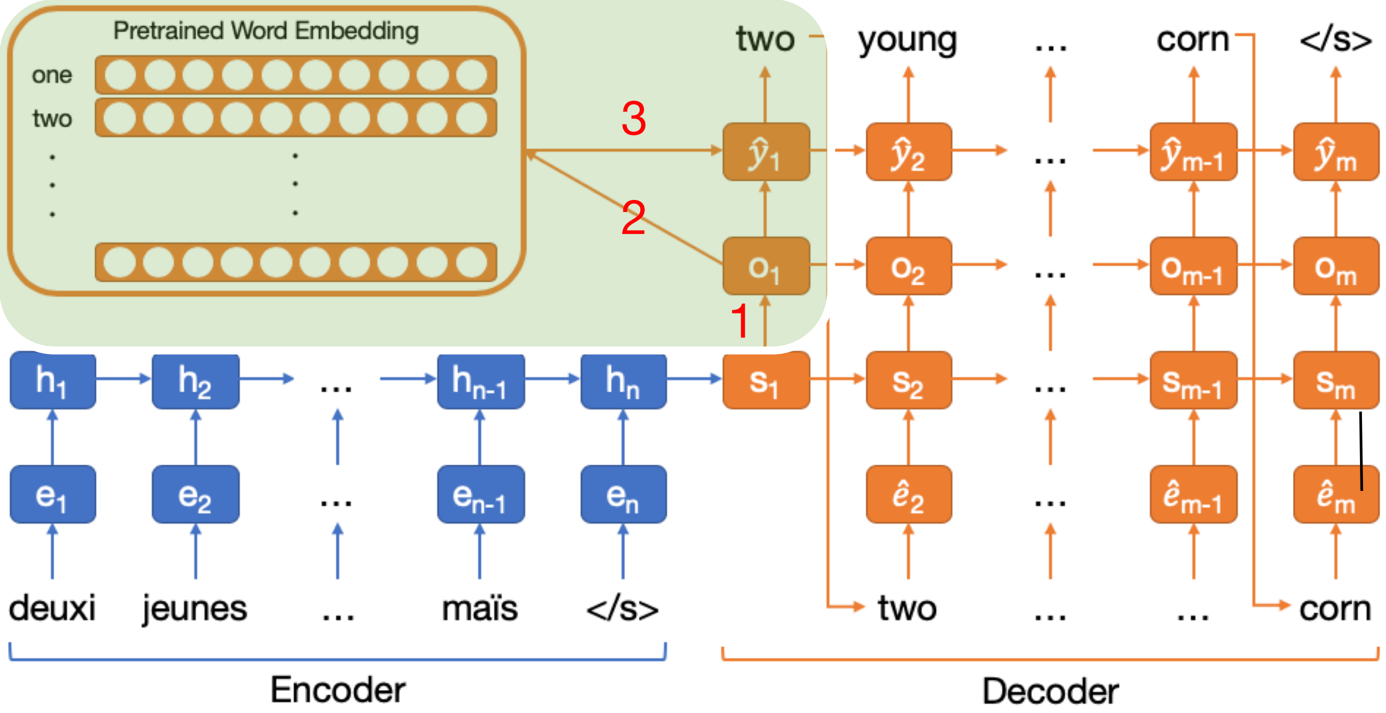
Embedding Prediction (Continuous Output)

TOKYO METROPOLITAN UNIVERSITY

- i.e. Continuous Output [Kumar and Tsvetkov, 2019]
- Predict a word embedding and search for the nearest word

1. Predict a word embedding of next word.
2. Compute cosine similarities with each word in pretrained word embedding.
3. Find and output the most similar word as system output.

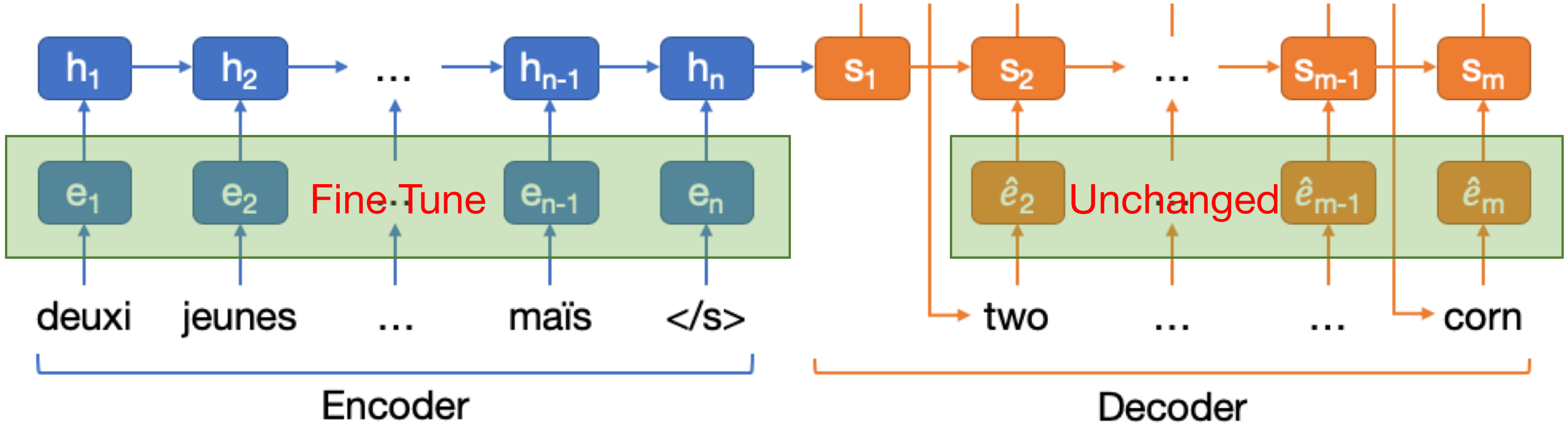
Keep unchanged:
Pretrained word embedding will NOT be updated during training.



Embedding Layer Initialization

[Qi et al., 2018]

- Initialize embedding layer with pretrained word embedding
- Fine-tune the embedding layer in encoder
- DO NOT update the embedding layer in decoder



Loss Function

TOKYO METROPOLITAN UNIVERSITY

- Model loss: Interpolation of each loss [Elliot and Kádáar, 2017]

$$J = \lambda J_{\mathrm{T}}(\theta, \phi_{\mathrm{T}}) + (1-\lambda) J_{\mathrm{V}}(\theta, \phi_{\mathrm{V}})$$

  - MT task: Max-margin with negative sampling [Lazaridou et al., 2015]

$$J_{\mathrm{T}}(\theta, \phi_{\mathrm{T}}) = \sum_{j}^{M} \max\{0, \gamma + d(\hat{\boldsymbol{e}}_{\boldsymbol{j}}, \mathbf{e}(w_j^-)) - d(\hat{\boldsymbol{e}}_{\boldsymbol{j}}, \mathbf{e}(y_j))$$

    - negative sampling

$$w_j^- = \underset{w \in \mathcal{V}}{\operatorname{argmax}}\{d(\hat{\boldsymbol{e}}_{\boldsymbol{j}}, \mathbf{e}(w)) - d(\hat{\boldsymbol{e}}_{\boldsymbol{j}}, \mathbf{e}(y_j))$$

  - Shared space learning task: Max-margin [Elliot and Kádáar, 2017]

$$J_{\mathrm{V}}(\theta, \phi_{\mathrm{V}}) = \sum_{\boldsymbol{v}' \neq \boldsymbol{v}} \max\{0, \alpha + d(\hat{\boldsymbol{v}}, \boldsymbol{v}') - d(\hat{\boldsymbol{v}}, \boldsymbol{v})\}$$

TOKYO METROPOLITAN UNIVERSITY

1. Multimodal Machine Translation
2. MMT with Embedding Prediction
3. Pretrained Word Embedding
4. Result & Conclusion

Hubness Problem [Lazaridou et al., 2015]

TOKYO METROPOLITAN UNIVERSITY

- Certain words (hubs) appear frequently in the neighbors of other words
  - Even of the word that has entirely no relationship with hubs
- Prevent the embedding prediction model from searching for correct output words
  - Incorrectly output the hub word

All-but-the-Top [Mu and Viswanath, 2018]

TOKYO METROPOLITAN UNIVERSITY

- Address hubness problem in other NLP tasks
- Debias a pretrained word embedding based on its global bias
  1. Shift all word embeddings to make their mean vector into a zero vector
  2. Subtract top 5 PCA components from each shifted word embedding
- Applied to pretrained word embeddings for encoder/decoder

TOKYO METROPOLITAN UNIVERSITY

1. Multimodal Machine Translation
2. MMT with Embedding Prediction
3. Pretrained Word Embedding
4. Result & Conclusion

Implementation & Dataset

TOKYO METROPOLITAN UNIVERSITY

- Implementation
  - Based on nmtpytorch v3.0.0 [Caglayan et al., 2017]
- Dataset
  - Multi30k (French to English)
  - Pretrained ResNet50 for visual encoder
- Pretrained Word Embedding
  - FastText
  - Trained on Common Crawl and Wikipedia
    - https://fasttext.cc/docs/en/crawl-vectors.html

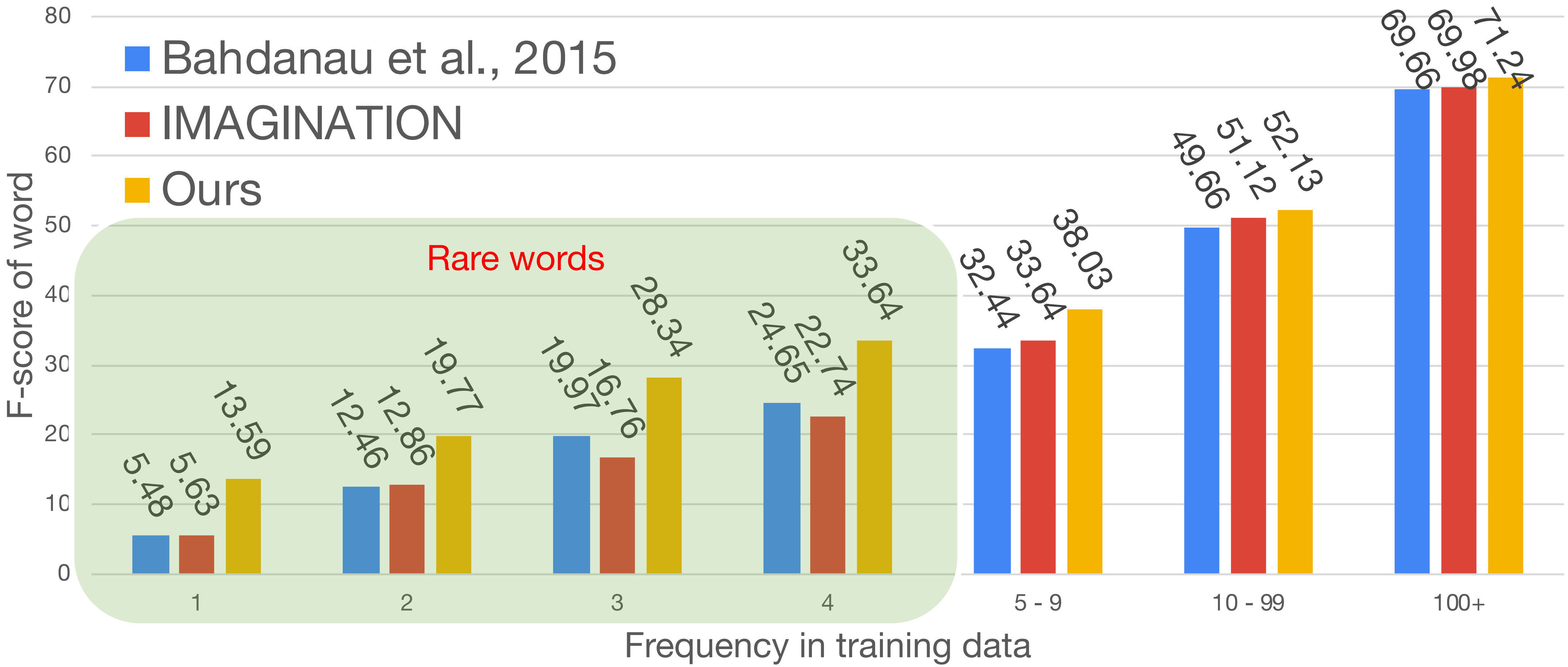
Our code is here: https://github.com/toshohirasawa/nmtpytorch-emb-pred

Hyper Parameters

TOKYO METROPOLITAN UNIVERSITY

- Model
  - dimension of hidden state: 256
  - RNN type: GRU
  - dimension of word embedding: 300
  - dimension of shared space: 2048
  - Vocabulary size (French, English): 10,000
- Training
  - $\lambda = 0.99$
  - Optimizer: Adam
  - Learning rate: 0.0004
  - Dropout rate: 0.3

Word-level F_1-score



Ablation w.r.t. Embedding Layers

TOKYO METROPOLITAN UNIVERSITY

Encoder	Decoder	Fixed	BLEU	METEOR
FastText	FastText	Yes	**53.49**	**43.89**
random	FastText	Yes	53.22	43.83
FastText	random	No	51.53	43.07
random	random	No	51.42	42.77
FastText	FastText	No	51.42	42.88
random	FastText	No	50.72	42.52

Encoder/Decoder: Initialize embedding layer with **random** values or **FastText** word embedding.
Fixed (Yes/No): Whether fix the embedding layer in decoder or fine-tune that while training.

- Fixing the embedding layer in decoder is essential
  - Keep word embeddings in input/output layers consistent

Overall Performance

TOKYO METROPOLITAN UNIVERSITY

Model	Validation BLEU	Test BLEU	Test METEOR
Bahdanau et al. 2015	50.83	51.00 ± .37	42.65 ± .12
+ pretrained	52.05	52.33 ± .66	43.42 ± .13
IMAGINATION	51.03	51.18 ± .16	42.80 ± .19
+ pretrained	52.40	52.75 ± .25	43.56 ± .04
Ours	**53.14**	**53.49 ± .20**	**43.89 ± .14**

Model (+ pretrained): Apply embedding layer initialization and All-but-the-Top debiasing.

- Our model performs better than baselines
  - Even those with embedding layer initialization

Ablation w.r.t. Visual Features

TOKYO METROPOLITAN UNIVERSITY

	Validation	Test	
Visual Features	BLEU	BLEU	METEOR
Centered	**53.14**	**53.49**	43.89
Raw	52.65	53.27	**43.91**
No	52.97	53.25	**43.91**

<u>Visual Features (Centered/Raw/No):</u> Use centered visual features or raw visual features to train model. ”No” show the result of text-only NMT with embedding prediction model.

- Centering visual features is required to train our model

Conclusion & Future Works

TOKYO METROPOLITAN UNIVERSITY

- MMT with embedding prediction improves ...
  - Rare word translation
  - Overall performance
- It is essential for embedding prediction model to ...
  - Fix the embedding in decoder
  - Debias the pretrained word embedding
  - Center the visual feature for multitask learning
- Future works
  - Better training corpora for embedding learning in MMT domain
  - Incorporate visual features into contextualized word embeddings

Thank you!

Translation Example

TOKYO
METROPOLITAN
UNIVERSITY

Source	un homme en vélo pédale devant une voûte .
Reference	a man on a bicycle pedals through an archway .
Text-only NMT	a man on a bicycle pedal past an arch .
IMAGINATION	a man on a bicycle pedals outside a monument .
Ours	a man on a bicycle pedals in front of a archway .

Translation Example (long)

Source	quatre hommes , dont trois portent des kippas , sont assis sur un tapis à motifs bleu et vert olive .
Reference	four men , three of whom are wearing prayer caps , are sitting on a blue and olive green patterned mat .
Text-only NMT	four men , three of whom are wearing aprons , are sitting on a blue and green speedo carpet .
IMAGINATION	four men , three of them are wearing alaska , are sitting on a blue patterned carpet and green green seating .
Ours	four men , three are wearing these are wearing these are sitting on a blue and green patterned mat .