Neural Machine Translation with Semantically Relevant Image Regions

Yuting Zhao[1], Mamoru Komachi[1], Tomoyuki Kajiwara[2], Chenhui Chu[3]
[1]Tokyo Metropolitan University
[2]Ehime University
[3]Kyoto University
zhao-yuting@ed.tmu.ac.jp, komachi@tmu.ac.jp
kajiwara@cs.ehime-u.ac.jp, chu@i.kyoto-u.ac.jp

1 Introduction

Neural machine translation (NMT) [1] has achieved near human-level performance. However, there remain numerous situations where textual context alone is insufficient for correct translation, such as in the presence of ambiguous words and grammatical gender. Many studies [2] have increasingly focused on incorporating multimodal contents, particularly images, to improve translations. Researchers in this field have established a task called multimodal machine translation (MMT), which translates sentences paired with images into a target language.

Subsequent studies [3, 4] have started utilizing a global visual feature extracted from an entire image to initialize encoder/decoder recurrent neural network (RNN) hidden states to contextualize language representations. However, the effect of the image cannot be fully exerted because the visual features of an entire image are complex and nonspecific. To effectively use an image, some studies [5] use spatially convoluted features extracted from a convolutional neural network (CNN). Because these equally sized features are nonsemantic, the role of visual modality provides only dispensable help to the translation. [6] reported that MMT models disregard visual features because the quality of the image features or the manner in which they are integrated into the model is not satisfactory.

Consequently, current studies [7] have incorporated richer local visual features such as regional features. These studies mainly rely on object detection to automatically extract visual objects in an image. Although regional features containing semantic information can assist in generating better translations, a method to focus on only the image regions that are semantically relevant to the source words during translation has yet to be determined [8].

In this paper, we propose a model for multimodal neural machine translation (MNMT) that employs word-region alignment (WRA), called MNMT-WRA. This model is designed to focus on semantically relevant image regions during translation. We propose to generate soft/hard/entity WRA based on cosine/argmax similarity between source words and visual concepts and manual alignment of Flickr30k Entities [9]. While encoding, textual and visual modalities are represented in three aspects by leveraging WRA: (1) associating image regions with respective source words; (2) associating source words with respective image regions; and (3) crossly associating.

The main contributions of this study are as follows: (1) We propose WRA to guide the model to translate certain words based on certain image regions. (2) The proposed MNMT-WRA model outperforms competitive baselines. (3) The analysis demonstrates that MNMT-WRA utilizes visual information effectively by relating semantically relevant textual and visual information.

2 Proposed Model

2.1 WRA: Word-Region Alignment

As shown in Figure 1, we propose to create WRA. For regions, we follow [10] in detecting image regions denoted by bounding boxes on the figure. In particular, each bounding box is detected along with a visual concept consisting of an attribute class followed by an object class instead of only the object class. We take these visual concepts to represent the image regions. We set each image labeled with 36 visual concepts of image regions, which are space-separated phrases. For words, we lowercase and tokenize the source English sentences via the Moses toolkit.[1)]

1) https://github.com/moses-smt/mosesdecoder

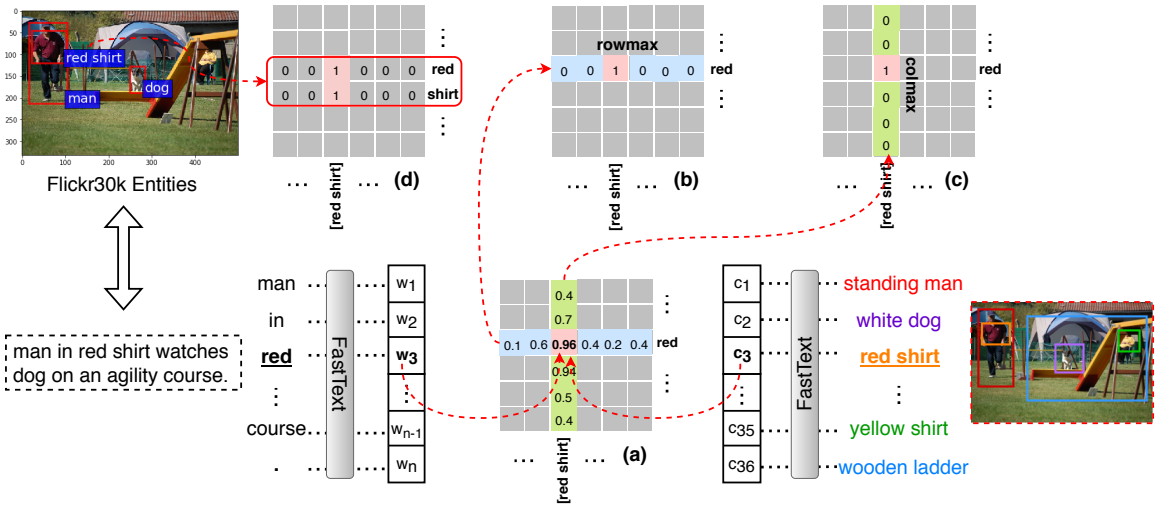

Figure 1 The WRA. (a): Soft alignment (SA). (b/c): Hard alignment (HA). (d): Entity alignment (EA).

2.1.1 Soft Alignment (SA) of Word-Region

The SA is filled with the cosine similarity between words and regions. We convert the words and concepts into subword units using the byte pair encoding (BPE) model [11]. Then we utilize fastText [12] to learn embeddings. So as to calculate a cosine similarity matrix of the word-region alignment as a SA.

2.1.2 Hard Alignment (HA) of Word-Region

The HA is based on the SA, which is a binary matrix filled with 1 in the position where the words and concepts are most similar and 0 in the remaining positions. We generate the HA from two directions: when aligning the most similar concept to a word, we use argmax function by row (rowmax). When aligning the most similar word to a concept, we employ argmax function by column (colmax).

2.1.3 Entity Alignment (EA) of Word-Region

The EA is based on Flickr30k Entities, which is a binary matrix filled with 1 in the position where the words correspond to their ground-truth regions and 0 in the remaining positions. Because the Flickr30k Entities provide a manually annotated correspondence between English entities and image regions, we use the EA as reference.

2.2 Representations with WRA

As shown in Figure 2, the textual encoder is a bi-directional RNN and the visual encoder is a object detection method [10]. All words are denoted as H and all regions are denoted as R. We represent textual annotation of n words as $\mathrm{A}^{\mathrm{txt}}$ and visual annotation of 36 regions as $\mathrm{A}^{\mathrm{img}}$ by leveraging WRA. For $\mathrm{A}^{\mathrm{txt}}$, the aligned region feature $\mathrm{R}_{\mathrm{align}}$ is calculated by the SA ($\mathrm{A}_{\mathrm{soft}}$), HA ($\mathrm{A}_{\mathrm{hard,rowmax/colmax}}$), and EA ($\mathrm{A}_{\mathrm{entity}}$) as follows.

$$
\begin{aligned}
\mathrm{A}^{\mathrm{txt}} &= \mathrm{CONCAT}(\mathrm{H}, \mathrm{R}_{\mathrm{align}}) \\
\mathrm{R}^{\mathrm{soft}}_{\mathrm{align}} &= \frac{\mathrm{A}_{\mathrm{soft}} \cdot \mathrm{R}}{|\mathrm{R}|} \\
\mathrm{R}^{\mathrm{hard}}_{\mathrm{align}} &= \mathrm{A}_{\mathrm{hard,rowmax}} \cdot \mathrm{R} \\
\mathrm{R}^{\mathrm{entity}}_{\mathrm{align}} &= \mathrm{A}_{\mathrm{entity}} \cdot \mathrm{R}
\end{aligned}
\tag{1}
$$

Similarly, $\mathrm{A}^{\mathrm{img}}$ is computed as follows:

$$
\begin{aligned}
\mathrm{A}^{\mathrm{img}} &= \mathrm{CONCAT}(\mathrm{R}, \mathrm{H}_{\mathrm{align}}) \\
\mathrm{H}^{\mathrm{soft}}_{\mathrm{align}} &= \frac{\mathrm{A}^{\mathrm{T}}_{\mathrm{soft}} \cdot \mathrm{H}}{|\mathrm{H}|} \\
\mathrm{H}^{\mathrm{hard}}_{\mathrm{align}} &= \mathrm{A}^{\mathrm{T}}_{\mathrm{hard,colmax}} \cdot \mathrm{H} \\
\mathrm{H}^{\mathrm{entity}}_{\mathrm{align}} &= \mathrm{A}^{\mathrm{T}}_{\mathrm{entity}} \cdot \mathrm{H}
\end{aligned}
\tag{2}
$$

2.3 Decoder

As shown in Figure 2, the decoder comprises double attentions and a deepGRU consisted of three cells [13].

2.3.1 Double Attentions

At time step t, the textual context vector $\mathbf{z_t}$ is generated upon a hidden state proposal $\mathbf{s}_\mathbf{t}^{(1)}$ computed by function $\mathrm{f}_{\mathrm{gru}_1}(y_{t-1}, \mathbf{s_{t-1}})$ in GRU (1) [13] and textual annotation $\mathbf{a}_\mathbf{i}^{\mathbf{txt}}$ in $\mathrm{A}^{\mathrm{txt}}$ as follows.

$$
\begin{aligned}
e^{\mathrm{text}}_{t,i} &= (V^{\mathrm{text}})^{\mathrm{T}} \tanh(U^{\mathrm{text}} \mathbf{s}_\mathbf{t}^{(1)} + W^{\mathrm{text}} \mathbf{a}_\mathbf{i}^{\mathbf{txt}}), \\
\alpha^{\mathrm{text}}_{t,i} &= \mathrm{softmax}(e^{\mathrm{text}}_{t,i}) \\
\mathbf{z_t} &= \sum_{i=1}^{n} \alpha^{\mathrm{text}}_{t,i} \mathbf{a}_\mathbf{i}^{\mathbf{txt}}
\end{aligned}
$$

where V^{text}, U^{text}, W^{text} are training parameters; $e^{\mathrm{text}}_{t,i}$ is attention energy; $\alpha^{\mathrm{text}}_{t,i}$ is attention weight matrix.

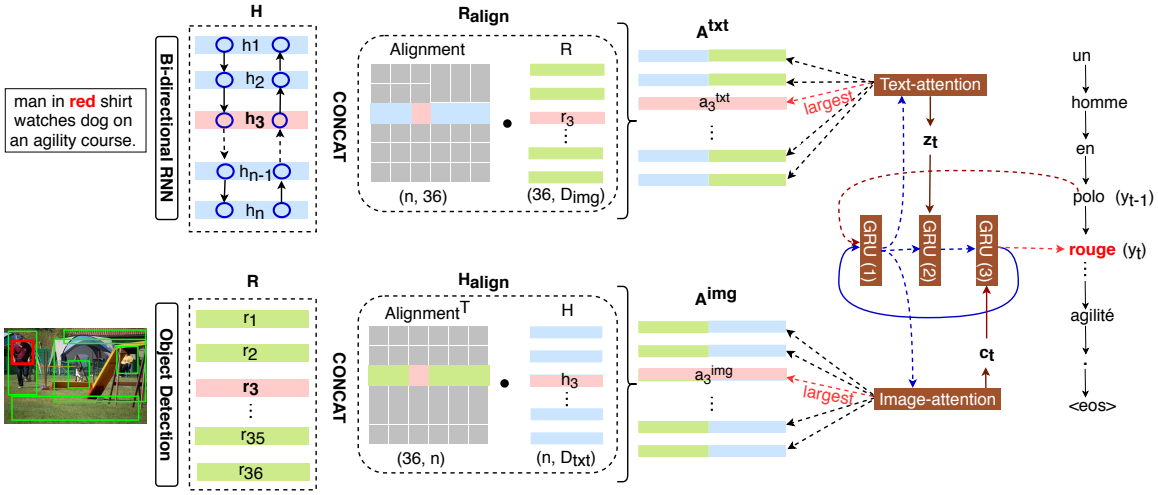

Figure 2 The proposed MNMT-WRA.

The visual context vector $\mathbf{c_t}$ is generated similarly.

$$e^{\text{img}}_{t,j} = (V^{\text{img}})^{\text{T}} \tanh(U^{\text{img}}\mathbf{s_t^{(1)}} + W^{\text{img}}\mathbf{a_j^{img}}),$$
$$\alpha^{\text{img}}_{t,j} = \text{softmax}(e^{\text{img}}_{t,j})$$
$$\mathbf{c_t} = \sum_{j=1}^{36} \alpha^{\text{img}}_{t,j}\mathbf{a_j^{img}}$$

where V^{img}, U^{img}, W^{img} are training parameters; $e^{\text{img}}_{t,j}$ is attention energy; $\alpha^{\text{img}}_{t,j}$ is attention weight matrix.

2.3.2 DeepGRU

The hidden state $\mathbf{s}_t$ is computed in GRU (3) [13]:

$$\mathbf{s}_t = \text{f}_{\text{gru}_3}([\mathbf{c}_t, y_{t-1}], \mathbf{s}_t^{(2)})$$
$$\mathbf{s}_t^{(2)} = \text{f}_{\text{gru}_2}(\mathbf{z}_t, \mathbf{s}_t^{(1)})$$

There is a textual GRU block and a visual GRU block [14] designed as below. The function f_{ght} is following [15].

$$\mathbf{b}_t^{\mathbf{v}} = \text{f}_{\text{ght}}(W_{\text{b}}^{\text{v}}\mathbf{s}_t)$$
$$\mathbf{b}_t^{\mathbf{t}} = \text{f}_{\text{ght}}(W_{\text{b}}^{\text{t}}\mathbf{s}_t^{(2)})$$
$$y_t \sim p_t = \text{softmax}(W_{\text{proj}}^{\text{t}}\mathbf{b}_t^{\mathbf{t}} + W_{\text{proj}}^{\text{v}}\mathbf{b}_t^{\mathbf{v}}),$$

where W_{b}^{v}, W_{b}^{t}, $W_{\text{proj}}^{\text{t}}$, $W_{\text{proj}}^{\text{v}}$ are training parameters.

3 Experiments

3.1 Dataset

We experimented on English→German (En→De) and English→French (En→Fr) tasks using Multi30k [16]. The dataset contains 29k train images and 1,014 valid images. We used three test sets to evaluate our models: Flickr test2016 and Flickr test2017 contain 1,000 pairs; and ambiguous MSCOCO [17] contains 461 pairs.

3.2 Settings

For baselines, we trained an NMT model [18] with the textual part of Multi30k; we trained an MNMT model [14] with 2,048-dim global visual features by ResNet-50 [19]; and we augmented [14] into a region-attentive MNMT (RAMNMT) model following [8] with 2,048-dim regional visual features extracted by [10].

For MNMT-WRA, it was implemented by three methods: associating regions with respective words (I); associating words with respective regions (T); crossly associating (C). The SA and HA were integrated in each method as two settings. The EA was integrated in the best method as a reference setting. Trainable parameters and dimensions are shown in the Appendix A.

Associating regions with respective words (I). We represented the visual annotation A^{img} by fusing R with the aligned textual features H_{align} and the textual annotation A^{txt} using textual input representation H directly. Based on equation (2), the settings were (1) MNMT-WRA (I+SA) and (2) MNMT-WRA (I+HA).

Associating words with respective regions (T). We represented textual annotation A^{txt} by fusing H with the aligned region features R_{align} and the visual annotation A^{img} using visual input representation R directly. Based on equation (1), the settings were (1) MNMT-WRA (T+SA) and (2) MNMT-WRA (T+HA).

Crossly associating (C). We cross represented textual annotation A^{txt} and visual annotation A^{img}. Based on equations (1) and (2), the settings were (1) MNMT-WRA (C+SA) and (2) MNMT-WRA (C+HA).

Model	Test2016		Test2017		MSCOCO	
	En→De	En→Fr	En→De	En→Fr	En→De	En→Fr
	B / M	B / M	B / M	B / M	B/ M	B / M
NMT	37.1 / 57.6	59.5 / 75.0	29.6 / 51.7	51.4 / 69.1	25.7 / 46.6	42.4 / 63.0
MNMT	37.6 / 57.6	59.6 / 74.8	30.1 / 51.4	52.1 / 68.8	26.8 / 47.1	43.2 / 63.2
RAMNMT	37.9 / 57.8	59.8 / 74.8	30.4 / 51.6	51.9 / 69.0	26.7 / 47.1	43.8 / 63.5
MNMT-WRA (C+SA)	35.3 / 56.0	56.6 / 73.2	26.0 / 48.4	47.2 / 65.9	23.4 / 44.7	39.3 / 61.0
MNMT-WRA (C+HA)	30.2 / 50.9	51.0 / 68.4	21.4 / 43.4	42.3 / 61.7	18.5 / 39.7	34.3 / 56.0
MNMT-WRA (T+SA)	34.9 / 55.6	57.6 / 73.4	26.0 / 48.2	48.7 / 66.4	23.6 / 44.8	41.3 / 61.8
MNMT-WRA (T+HA)	30.2 / 50.3	51.0 / 67.8	20.7 / 42.2	42.2 / 60.6	17.4 / 38.2	33.9 / 54.6
MNMT-WRA (I+SA)	38.0† / **57.9**	59.6 / 75.0	30.3† / 51.7	**52.2**† / **69.5**	26.6† / 47.2	**43.6**† / **63.9**
MNMT-WRA (I+HA)	**38.3**†‡ / 57.8	**60.2**† / **75.5**	**31.2**$^{\dagger\ddagger *}$ / **52.2**	51.8 / **69.5**	**27.4**† / **47.8**	43.5† / 63.8
MNMT-WRA (I+EA)	38.2 / 57.8	59.8 / 75.1	N/A	N/A	N/A	N/A

Table 1 †, ‡ and ∗ indicate that the result is significantly better than NMT, MNMT and RAMNMT, respectively.

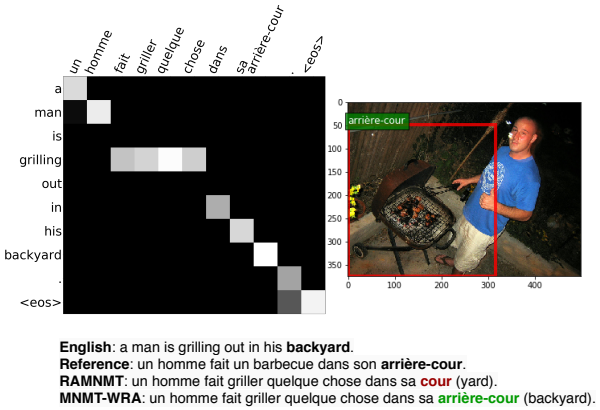

Figure 3 A good translation example of MNMT-WRA (I+HA) for En→Fr task.

Reference setting. The reference setting was MNMT-WRA (I+EA). Because the train and valid images from Flickr30k Entities were assigned to Flickr test2016 images only, we reported only the results of test2016.

3.3 Evaluation

We evaluated the translation quality according to the token level BLEU (B) [20] and METEOR (M) [21] metrics, and reported the average over three runs. We reported the statistical significance with bootstrap resampling [22] using the merger of three test results. We reported the result only if the p-value was less than 0.05.

4 Results

As shown in Table 1, the results of MNMT-WRA (I+HA/SA) outperform all the baselines on all test sets of En→De and En→Fr tasks, and the MNMT-WRA (I+HA) achieved the best performance of all. In contrast, the poor performance of MNMT-WRA (T/C) might be because of the weakened role of text and emphasized role of the image.

5 Analysis

We randomly investigated 50 examples from the En→Fr task of test2016 to do human evaluation. In this investigation, 16% of the examples is that the performance of the MNMT-WRA (I+HA) is better than the RAMNMT, and the 84% is that the performance of the two is comparable.

We show an example of MNMT-WRA (I+HA) in Figure 3 to do quality analysis. In the case, our model correctly translates "backyard" to a compound noun of "arrière-cour," which is comprised of an adverb and a noun. But the RAMNMT mistranslates it to "cour," which means "yard" in English. Through visualization, we find that the text-attention and image-attention focus on the features that are semantically relevant at that time step. It shows that translation quality improvement is due to the simultaneous attentions of semantically relevant region and word.

6 Conclusion

We presented a novel model, MNMT-WRA, that simultaneously considers semantically relevant textual and visual features during translation. Experimental results show that MNMT-WRA outperformed baselines. We also performed a human evaluation and qualitative analysis to demonstrate the specific improvements resulting from semantically relevant image regions. In the future, we plan to train supervised attention mechanisms to learn more reliable alignments, rather than external alignment.

Acknowledgments

This work was supported by Grant-in-Aid for Young Scientists #19K20343, JSPS.

References

[1] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *NIPS*, pages 3104–3112, 2014.

[2] Loïc Barrault, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank. Findings of the third shared task on multimodal machine translation. In *WMT*, pages 304–323, 2018.

[3] Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu. Are you talking to a machine? dataset and methods for multilingual image question answering. In *NIPS*, pages 2296–2304, 2015.

[4] Po-Yao Huang, Frederick Liu, Sz-Rung Shiang, Jean Oh, and Chris Dyer. Attention-based multimodal neural machine translation. In *WMT*, pages 639–645, 2016.

[5] Ozan Caglayan, Loïc Barrault, and Fethi Bougares. Multimodal attention for neural machine translation. *CoRR*, 2016.

[6] Ozan Caglayan, Pranava Madhyastha, Lucia Specia, and Loïc Barrault. Probing the need for visual context in multimodal machine translation. In *NAACL*, pages 4159–4170, 2019.

[7] Pengcheng Yang, Boxing Chen, Pei Zhang, and Xu Sun. Visual agreement regularized training for multi-modal machine translation. In *AAAI*, pages 9418–9425, 2020.

[8] Yuting Zhao, Mamoru Komachi, Tomoyuki Kajiwara, and Chenhui Chu. Double attention-based multimodal neural machine translation with semantic image regions. In *EAMT*, pages 105–114, 2020.

[9] Bryan A. Plummer, Liwei Wang, Christopher M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *IJCV*, pages 74–93, 2017.

[10] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, pages 6077–6086, 2018.

[11] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *ACL*, pages 1715–1725, 2016.

[12] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *TACL*, 5:135–146, 2017.

[13] Jean-Benoit Delbrouck and Stéphane Dupont. Bringing back simplicity and lightliness into neural image captioning. *CoRR*, 2018.

[14] Jean-Benoit Delbrouck and Stéphane Dupont. UMONS submission for WMT18 multimodal translation task. In *WMT*, pages 643–647, 2018.

[15] D. Teney, P. Anderson, X. He, and A. v. d. Hengel. Tips and tricks for visual question answering: Learnings from the 2017 challenge. In *CVPR*, pages 4223–4232, 2018.

[16] Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. Multi30k: Multilingual English-German image descriptions. In *VL*, pages 70–74, 2016.

[17] Desmond Elliott, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. Findings of the second shared task on multimodal machine translation and multilingual image description. In *WMT*, pages 215–233, 2017.

[18] Ozan Caglayan, Mercedes García-Martínez, Adrien Bardet, Walid Aransa, Fethi Bougares, and Loïc Barrault. Nmtpy: A flexible toolkit for advanced neural machine translation systems. *Prague Bull. Math. Linguistics*, 109:15–28, 2017.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[20] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318, 2002.

[21] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *WMT*, pages 376–380, 2014.

[22] Philipp Koehn. Statistical significance tests for machine translation evaluation. In *EMNLP*, pages 388–395, 2004.

[23] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, pages 1–15, 2015.

A Experimental Parameters

We ensured that the parameters in MNMT-WRA were consistent with those in the baselines. We set the encoder and decoder hidden state to 256-dim; word embedding to 128-dim; batch size to 32; beam size to 12; text dropout to 0.3; image region dropout to 0.5; and dropout of source RNN hidden states H and both blocks $\mathbf{b}_\mathbf{t}^\mathbf{t}$ and $\mathbf{b}_\mathbf{t}^\mathbf{v}$ to 0.5. We trained the model using stochastic gradient descent with ADAM [23] and a learning rate of 0.0004. We stopped training when the METEOR score did not improve for 10 evaluations on the validation set, where the maximum epoch num was set to 100.

For MNMT-WRA (I+SA/HA/EA): Between the two settings, the textual annotation A^{txt} was 512-dim, which was consistent with H. Further, the visual annotation A^{img} was 4,096-dim by a concatenation of R and H_{align}, where R was 2,048-dim and H_{align} was 2,048-dim by a linear transformation from 512-dim.

For MNMT-WRA (T+SA/HA): Between the two settings, the visual annotation A^{img} was 2,048-dim, which was consistent with R. The textual annotation A^{txt} was 2,560-dim by a concatenation of H and R_{align}, where H was 512-dim and R_{align} was 2,048-dim. The sum of the dimensions of the textual and visual annotations was consistent with that of the former method.

For MNMT-WRA (C+SA/HA): Between the two settings, the textual annotation A^{txt} was 2,560-dim by a concatenation of 512-dim H and 2,048-dim R_{align}, and the visual annotation A^{img} was 2,560-dim by a concatenation of 2,048-dim R and 512-dim H_{align}. We ensured that the dimensions of the textual and visual annotation were consistent.