Distill the Image to Nowhere: Inversion Knowledge Distillation for Multimodal Machine Translation

Ru Peng[1*], **Yawen Zeng**[2*], **Junbo Zhao**[1†]
[1]Zhejiang University, Zhejiang, China
[2]Tencent WeChat, Shenzhen, China
pengru709909347@gmail.com, yawenzeng11@gmail.com, j.zhao@zju.edu.cn

Abstract

Past works on multimodal machine translation (MMT) elevate bilingual setup by incorporating additional aligned vision information. However, an *image-must* requirement of the multimodal dataset largely hinders MMT's development — namely that it demands an aligned form of [image, source text, target text]. This limitation is generally troublesome during the inference phase especially when the aligned image is not provided as in the normal NMT setup. Thus, in this work, we introduce IKD-MMT, a novel MMT framework to support the *image-free* inference phase via an *inversion knowledge distillation* scheme. In particular, a multimodal feature generator is executed with a knowledge distillation module, which directly generates the multimodal feature from (only) source texts as the input. While there have been a few prior works entertaining the possibility to support image-free inference for machine translation, their performances have yet to rival the image-must translation. In our experiments, we identify our method as the first image-free approach to comprehensively rival or even surpass (almost) *all* image-must frameworks, and achieved the state-of-the-art result on the often-used Multi30k benchmark[1].

1 Introduction

Multimodal machine translation (MMT) is an worthy task of elevating text-only translation by introducing additional image modality (Specia et al., 2016; Elliott et al., 2017; Barrault et al., 2018). Existing works mostly focus on the fusion and alignment of images and texts to improve MMT (Calixto et al., 2017; Ive et al., 2019; Yin et al., 2020), that they have managed to concept-prove the effectiveness of the aligned visual information. Nevertheless, the strict triplet data form of the dataset, in both the training and inference phases, has disabled the MMT model to generalize further. In particular, if we consider using an MMT model to conduct translation for the normal bilingual text translation as in the NMT setup, one must provide the aligned images during inference. And unfortunately, this is not often feasible. This general comparison between image-free and image-must schemes is visually illustrated in Figure 1(a). In hindsight, the quantity and quality of attached images become a bottleneck towards the development of MMT, as acquiring such resources can be scarce and expensive (e.g. Multi30K (Elliott et al., 2016)).

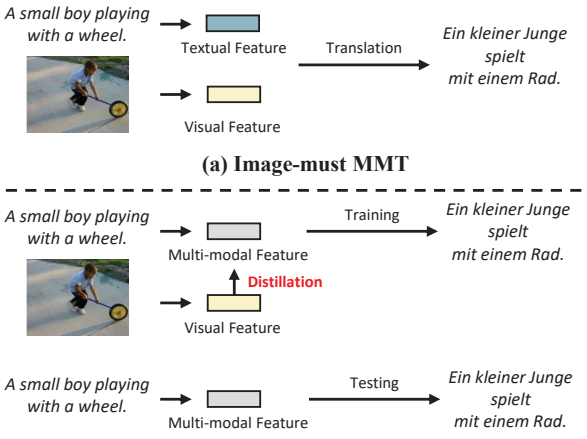

Figure 1: Examples of Image-must MMT (a), and our Image-free MMT (b). During testing, our IKD-MMT does not require the image as input.

Indeed, there have been a few attempts to resolve the image-must limitation. For instance, Elliott and Kádár (2017) present a multi-task learning model for MMT where they rely on an auxiliary visual grounding task to obtain the visual feature. Zhang et al. (2020) introduce an image retrieval paradigm to find topic-related images from a small-scale dataset. Further, Long et al. (2021) attempts to

*Both authors contributed equally to this research.
†Corresponding author.
[1]Our code and data are available at: https://github.com/pengr/IKD-mmt/tree/master.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 2379 - 2390
December 7-11, 2022

utilize a set of generative adversarial networks to obtain an imaginary vision feature. We may posit that a (nearly) common ground for such image-free frameworks is to learn and further obtain a generated visual feature representation without the actual image data provided during inference. However, none of the aforementioned works has managed to consistently reach the performance of the image-must counterpart. In this work, we hypothesise that this can be caused by the inferior representation learned, insufficient visual distribution coverage, improper multimodal fusion stage (Caglayan et al., 2017; Arslan et al., 2018; Helcl et al., 2018; Calixto and Liu, 2017), and/or lacked training stability, etc.

In this work, we intend to take a thorough exploration towards this line. As Shown in Figure 1(b), unlike prior works solely targeting visual feature generation and/or relying on later stages of fusion, our approach directly generates a **multimodal feature** using only the source text input. We enable this by proposing an inverse knowledge distillation mechanism employing pre-trained convolutional neural networks (CNN). From our experiments, we find that this architectural choice has notably enhanced the training stability as well as the final representation quality. To this end, we introduce the IKD-MMT framework, an image-free framework that systematically rivals or outperforms the image-must frameworks. To set up the inverse knowledge distillation flow, we incorporate dual CNNs with inverted data feeding flow. Of the two, the teacher network receives the pre-trained weights while the student CNN is trained from scratch aiming to provide a high-quality multimodal feature space by incorporating both inter-modal and intra-modal distillations.

Our contributions are summarized as follows:

i. IKD-MMT framework is the first method that systematically rivals or even outperforms the existing image-must frameworks, which fully demonstrates the feasibility of the image-free concept;

ii. We pioneer the exploration of knowledge-distillation combined with the pre-trained models in the regime of MMT, as well as the multimodal feature generation. We posit that these techniques have shed some light on the representation learning and training stability of MMT.

2 Related Work

2.1 Multi-modal Machine Translation

As an intersection of multimedia and neural machine translation (NMT), MMT has drawn great attention in the research community. Technically, existing methods mainly focus on how to better integrate visual information into the framework of NMT. 1) Calixto et al. (2017) propose a doubly-attentive decoder to incorporate two separate attention over the source words and visual features. 2) Ive et al. (2019) propose a translate-and-refine approach to refine draft translations by visual features. 3) Yao and Wan (2020) propose the multimodal Transformer to induce the image representations from the text under the guide of image-aware attention. 4) Yin et al. (2020) employs a unified multimodal graph to capture various semantic interactions between multimodal semantic units.

However, the quantity and quality of the annotated images limit the development of this task, which is scarce and expensive. In this work, we aim to perform the MMT in an image-free manner, which has the ability to break data constraints.

2.2 Knowledge Distillation

Knowledge distillation (KD) (Buciluǎ et al., 2006; Hinton et al., 2015) aims to use a knowledge-rich teacher network to guide the parameter learning of the student network. In fact, KD has been investigated in a wide range of fields. Romero et al. (2014) transfer knowledge through an intermediate hidden layer to extend the KD. Yim et al. (2017) define the distilled knowledge to be transferred in terms of flow between layers, which is calculated by the inner product between features from two layers. In the multimedia field, Gupta et al. (2016) first introduce the technique that transfers supervision between images from different modalities. Yuan and Peng (2018) propose the symmetric distillation networks for the text-to-image synthesis task.

Inspired by these pioneering efforts, our IKD-MMT framework is intents to take full advantage of KD to generate a multimodal feature to overcome triplet data constraints.

3 IKD-MMT Model

As illustrated in Figure 2, the proposed framework consists of two components: an image-free MMT backbone and a multimodal feature generator.

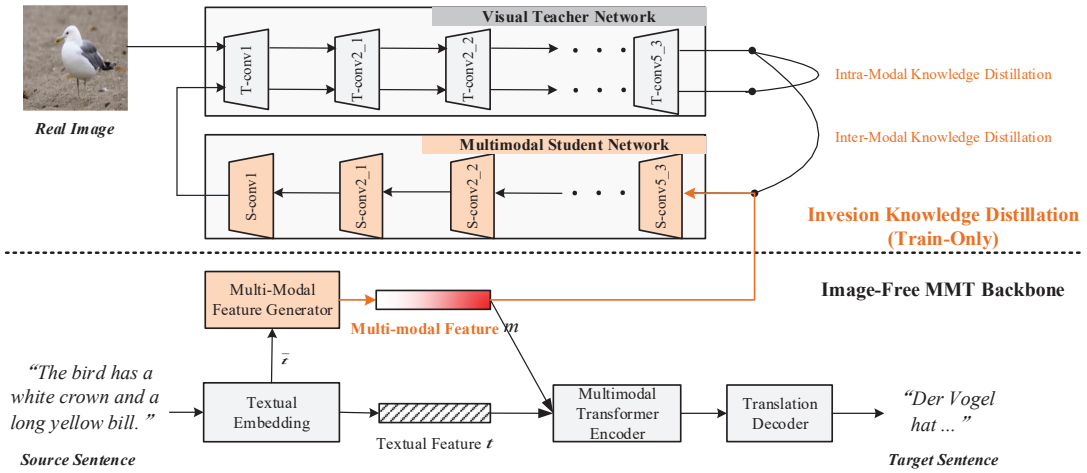

Figure 2: The framework of our IKD-MMT model. The multimodal feature generator, multimodal student network and visual teacher network are the most critical modules, which help break the dataset constraints of image-must.

3.1 Image-Free MMT Backbone

Given a source sentence $X=(x_1,\ldots,x_I)$, each token x_i is mapped into a word embedding vector $E_{x_i} \in \mathbb{R}^{d_w}$ through the textual embedding with position encoding (Gehring et al., 2017). d_w and $t=(E_{x_1},\ldots,E_{x_I})$ are the word embedding dimension and the textual feature, respectively.

Then, we feed the text feature t together with the multimodal feature m (detail in Section 3.2.1) into the multimodal transformer encoder (Yao and Wan, 2020). In the multimodal encoder layer, we cascade the multimodal feature m and the text feature t to reorganize a new multimodal feature $\widetilde{x}$ as the query vector:

$$\widetilde{x}=[t;mW^m] \in \mathbb{R}^{(I+P)*d}, \quad (1)$$

where I is the length of source sentence, and P is the size of multimodal feature. Here, we can understand this modal fusion from the perspective of nodes and graphs. If we treat each source token as a node, each region of the multimodal feature can also be regarded as a pseudo-token and added to the source token graph for modal fusion. The key and value vectors are preserved as the text feature t, and the multimodal encoder layer is calculated as follows:

$$c_k = \sum_{i=1}^{I} \tilde{\alpha}_{ki} \left(t_i W^V\right), \quad (2)$$

$$\tilde{\alpha}_{ki} = softmax\left(\frac{\left(\tilde{x}_k W^Q\right)\left(t_i W^K\right)^\top}{\sqrt{d}}\right). \quad (3)$$

In this paper, we directly adopt the Transformer decoder[2] (Vaswani et al., 2017) for translation.

[2]For details, please refer to the original paper.

Given a target sentence $Y=(y_1,\ldots,y_J)$, our framework outputs the predicted probability of the target word y_j as follow:

$$p\left(y_j|y_{<j},\mathrm{X},m\right) \propto \exp\left(W^h H_j^L + b^h\right), \quad (4)$$

where H_j^L represents the top output of the decoder at j-th decoding time step, W^h and b^h are learnable multi-layer perceptrons, and $\exp()$ is a Softmax layer.

3.2 Multimodal Feature Generation

3.2.1 Preliminaries

In this part, we introduce the frame, symbol definitions and task goal of multimodal feature generation in advance.

The frame is composed of a multimodal feature generator F, a visual teacher model T and a multimodal student model S. The detailed architecture of each module is shown in Table 7 of the appendix. The model parameters of S are denoted as θ^s. When the global text feature $\bar{t}$ is fed into S, the hidden representation produced by the l-th layer is denoted as $\varphi_l^S\left(\bar{t},\theta_l^s\right)$. The F outputs a multimodal feature m, and the S produces an inverse feature I_s after the *S-conv1* layer. The real image and the inverse feature are $\{I_s,I_r\} \in \mathbb{R}^{m*n*3}$. Given a feature I as input, the hidden representation produced by the l-th layer of T is denoted as $\varphi_l^T(I)$.

Our goal is to generate multimodal features from the source text to break the image-must restriction in testing. The visual perception of this multimodal feature is extracted from the visual distillation of the teacher-student model, while the textual

semantic of that is derived from the text translation of the input text.

3.2.2 Multimodal Feature Generator

First, we simply adopt an average pooling to transform all word embedding vectors into global textual features, which are proven to carry the overall word senses in (Zhang et al., 2010).

$$\bar{t} = {}^{1}/_{I} \sum_{i=1}^{I} E_{x_i}. \quad (5)$$

Then, the global text feature $\bar{t}$ is serially transported into the multimodal feature generator to compute a multimodal feature m:

$$m = \text{unpool}(W^t \bar{t}). \quad (6)$$

Among them, the FC layer W^t projects the global text feature $\bar{t}$ into the image space. The following average unpooling computes a high-dimensional multimodal feature map from the low-dimensional latent vector. The dimension of $m \in \mathbb{R}^{P*2048}$ is the same as that of the last convolutional activation of the teacher model. Notably, the textual semantics of multimodal features are modelled from the global textual context supervised by the text translation.

3.2.3 Inversion Knowledge Distillation

The inversion knowledge distillation transfers the visual perception from the teacher model T to the student model S, and in-depth interacts with textual semantics in the multimodal feature generator. To synthesize an information-rich multimodal feature, we formulate a novel dual distillation paradigm consisting of inter-modal (IrM-KD) and intra-modal (IaM-KD) knowledge distillations.

IrM-KD: The IrM-KD direct the student model S to extract the vital visual information from the source text, thereby bridging the inter-modal semantics of the text and the real image. Specifically, given the real image I_r, the teacher model T generates a visual representation $\varphi_l^T(I_r)$ in each layer l. Meanwhile, the S produces a inverse hidden representation $\varphi_{l+1}^S\left(\bar{t}, \theta_l^s\right)$ in next layer $l+1$. The paired representations $\varphi_l^S\left(\bar{t}, \theta_l^s\right)$ and $\varphi_l^T(I_r)$ with identical dimension entail the same-level latent concepts. We present the IrM-KD loss by the discrepancy among these two representations and an auxiliary regularization term:

$$Loss_{\text{IrM}} = \sum_{l} \left\|\varphi_l^T(I_r) - \varphi_{l+1}^S(\bar{t}; \theta_l^s)\right\|_2 + \|I_r - I_s\|_2, \quad (7)$$

where the L_2 norm $\|\|_2$ is used to measure the similarity of two vectors. The regularization term $\|I_r - I_s\|_2$ indicates the image space loss, which is the fundamental constraint for the S to learn the distribution of the real image.

IaM-KD: The IaM-KD constrains the student model S to learn the visual perception of images via the inverse feature, thus relieving the intra-modal gap between the inverse feature with the real image. Specifically, we fed the inverse feature I_s into the teacher model T to gain the teacher's cognition for it — a pseudo visual representation $\varphi_l^D(I_s)$. Then, to encourage the student model profoundly learn the distribution of images, we narrow the divergence between the $\varphi_l^D(I_s)$ and its coupled visual representation $\varphi_l^T(I_r)$. So that, the IaM-KD loss is defined as the combination of the above divergence and the image space loss:

$$Loss_{\text{IaM}} = \sum_{l} \left\|\varphi_l^T(I_r) - \varphi_l^D(I_s)\right\|_2 + \|I_r - I_s\|_2. \quad (8)$$

Compare with T2I synthesis works (Reed et al., 2016; Zhang et al., 2017; Xu et al., 2018), we are dedicated to aidding text translation through inter-modal and intra-modal bi-visual distillation. By doing so, our generated multimodal feature focuses more on the text-image alignment and fusion, but not only the authenticity of image.

3.3 Objective function

During the training phase, we optimize the proposed IKD-MMT model end-to-end by the text translation loss and the inversion distillation loss:

$$J(\theta, \theta_s) = J_{\text{trans}}(\theta, \theta_s) + Loss_{\text{IrM}} + Loss_{\text{IaM}}. \quad (9)$$

Wherein, the translation loss over the training dataset $\mathcal{D}$, not only bridges the relevance of the source and target texts, but also models the text semantics of multimodal features:

$$J_{\text{trans}}(\theta, \theta_s) = -\sum_{D} \sum_{J} \log p(y_j | y_{<j}, X, m). \quad (10)$$

In the testing phase, the trained multimodal feature generator is capable to generate rich features to embed into the MMT backbone, thus getting rid of the image-must constraints.

4 Experiment

4.1 Setup

Datasets We conduct experiments on the Multi30K benchmark (Elliott et al., 2016). The

Table 1: BLEU ("B") and METEOR ("M") scores of EN-DE and EN-FR tasks. Encouragingly, our IKD-MMT as an image-free MMT model outperforms almost all MMT systems, and even rivals the SOTA image-must systems. ‡/† mark statistically significant variations for BLEU (p-value $< 0.01/0.05$) as compared to the Transformer.

Systems	EN-DE						EN-FR			
	Test2016		Test2017		MSCOCO		Test2016		Test2017	
	B	M	B	M	B	M	B	M	B	M
Image-must MMT Systems										
$\text{NMT}_{\text{SRC+IMG}}$(Calixto et al., 2017)	36.5	55.0	-	-	-	-	-	-	-	-
IMG_D(Calixto and Liu, 2017)	37.3	55.1	-	-	-	-	-	-	-	-
Fusion-conv(Caglayan et al., 2017)	37.0	57.0	29.8	51.2	25.1	46.0	53.5	70.4	51.6	68.6
Trg-mul(Caglayan et al., 2017)	37.8	57.7	30.7	52.2	26.4	47.4	54.7	71.3	52.7	69.5
VAG-NMT(Zhou et al., 2018)	-	-	31.6	52.2	28.3	48.0	-	-	53.8	70.3
DS-SUM-L2(Caglayan, 2019)	39.4	58.7	32.6	52.9	-	-	60.7	76.0	54.2	71.0
Del+obj(Ive et al., 2019)	38.0	55.6	-	-	-	-	59.8	74.4	-	-
Multimodal(Yao and Wan, 2020)	38.7	55.7	-	-	-	-	-	-	-	-
GMNMT(Yin et al., 2020)	39.8	57.6	32.2	51.9	28.7	47.6	60.9	74.9	53.9	69.3
DCCN(Lin et al., 2020)	39.7	56.8	31.0	49.9	26.7	45.7	61.2	76.4	54.3	70.3
Gumbel-att(Liu et al., 2021)	39.2	57.8	31.4	51.2	26.9	46.0	-	-	-	-
OVC+L_m(Wang and Xiong, 2021)	-	-	32.3	52.4	28.9	48.1	-	-	54.1	70.5
Gated Fusion(Wu et al., 2021)	41.96	-	33.59	-	29.04	-	61.69	-	54.85	-
RMMT(Wu et al., 2021)	41.45	-	32.94	-	30.01	-	62.1	-	54.39	-
Image-free MMT Systems										
Transformer(Vaswani et al., 2017)	37.6	55.3	31.7	52.1	27.9	47.8	59.0	73.6	51.9	68.3
Multitask(Elliott and Kádár, 2017)	36.8	55.8	-	-	-	-	-	-	-	-
VMMT_{F}(Calixto et al., 2019)	37.7	56.0	30.1	49.9	25.5	44.8	-	-	-	-
UVR-NMT(Zhang et al., 2020)	36.94	-	28.63	-	-	-	57.53	-	48.46	-
ImagiT (Long et al., 2021)	38.5	55.7	32.1	52.4	28.7	48.8	59.7	74.0	52.4	68.3
IKD-MMT (Ours)	**41.28**‡ ±0.3	**58.93** ±0.20	**33.83**† ±0.10	**53.21** ±0.26	**30.17** ±0.14	**48.93** ±0.08	**62.53**† ±0.25	**77.20** ±0.18	**54.84**† ±0.50	**71.87** ±0.34

training and validation sets contain 29,000 and 1,014, respectively. We report the results of the Test2016, Test2017 and ambiguous MSCOCO test sets. We directly use the preprocessed sentences[3] and apply the BPE (Sennrich et al., 2016) with 10K merge operations to segment words into sub-words, which build a shared vocabulary of 9,712 and 9,544 tokens for EN-DE and EN-FR translation tasks.

Settings We follow all model settings of (Wu et al., 2021), such as the Transformer-Tiny configuration for anti-overfitting in small datasets. 4-gram case-insensitive BLEU (Papineni et al., 2002) and METEOR (Denkowski and Lavie, 2014) are used as evaluation metrics. All models are run three times and report the average results.

4.2 Main Results

EN-DE Translation Task Table 1 reports the performance of all MMT baselines on the En-DE task. Comparing all systems, we draw the following interesting conclusions:

First, the IKD-MMT significantly surpasses all image-free MMT systems on five test sets. These improvements demonstrate that a) our model can effectively embed multimodal semantics during the training and guide the translation via multimodal features among the image-free testing phase, b) benefiting from the informative richness and stable generation of multimodal features, our method is a more robust way to break data constraints.

Second, the image-must MMT systems generally exceed their image-free counterparts, showing the efficacy of additional images for translation.

Finally, encouragingly, our image-free MMT model not only overbeats almost all image-must MMT systems, but even rivals the SOTA image-must MMT. We speculate that these noticeable gains stem from the IKD-MMT's strong ability to fuse the text semantics and visual perception, and generate text-related visual representation, under the dual supervision of text translation and visual distillation.

EN-FR Translation Task We also conduct experiments on the EN-FR task. Our IKD-MMT still outperforms the compared baselines in Table 1. This verifies the robustness and generality of our model in various language scenarios.

[3]https://github.com/multi30k/dataset

Table 2: Ablation results for diverse distillation variants on the EN-DE task. The *base* row denotes the IKD-MMT in Table 1, and "-" means to retain the setting of the *base* row. Avg.B and Avg.M indicate the BLEU and METEOR scores of the three test sets

	Sim. Func.	**Dist. Gran.**	**CNN Back.**	**Dist. Loss**	Avg.B	Avg.M
base	L_2	Model	ResNet50	(IrM-KD+IaM-KD) loss	**35.09**	**53.69**
(A)	L_1	-	-	-	34.60 (-0.49)	53.39 (-0.30)
	L_∞	-	-	-	34.15 (-0.94)	53.27 (-0.42)
	Cosine	-	-	-	34.64 (-0.45)	53.27 (-0.42)
	KL-Div.	-	-	-	34.62 (-0.47)	53.56 (-0.13)
(B)	-	Block	-		34.57 (-0.52)	53.27 (-0.42)
	-	Layer	-	-	34.85 (-0.24)	53.25 (-0.44)
(C)	-	-	VGG19	-	34.38 (-0.71)	53.24 (-0.45)
	-	-	AlexNet	-	33.98 (-1.11)	52.99 (-0.70)
(D)	-	-	-	w/o (IrM-KD+IaM-KD) loss	27.30 (-7.79)	51.11 (-2.58)
	-	-	-	Image Space loss	32.91 (-2.18)	52.41 (-1.28)
	-	-	-	w/o IaM-KD loss	33.64 (-1.45)	52.81 (-0.88)
	-	-	-	w/o IrM-KD loss	34.03 (-1.06)	53.08 (-0.61)

Table 3: Validation ablation results for diverse distillation variants on the EN-DE task. The *base* row denotes the IKD-MMT in Multi30K development sets, and "-" means to retain the setting of the *base* row. Dev.B and Dev.M indicate the BLEU and METEOR scores of the development set

	Sim. Func.	**Dist. Gran.**	**CNN Back.**	**Dist. Loss**	Dev.B	Dev.M
base	L_2	Model	ResNet50	(IrM-KD+IaM-KD) loss	**42.48**	**59.20**
(A)	L_1	-	-	-	41.33(-1.15)	58.55(-0.65)
	L_∞	-	-	-	41.80(-0.68)	58.92(-0.28)
	Cosine	-	-	-	41.44(-1.04)	58.64(-0.56)
	KL-Div.	-	-	-	41.90 (-0.58)	58.89(-0.31)
(B)	-	Block	-		41.83(-0.65)	59.01(-0.19)
	-	Layer	-	-	41.67(-0.81)	58.69(-0.51)
(C)	-	-	VGG19	-	41.69(-0.79)	58.80(-0.40)
	-	-	AlexNet	-	41.20(-1.28)	58.45(-0.75)
(D)	-	-	-	w/o (IrM-KD+IaM-KD) loss	36.02(-6.46)	55.67(-3.53)
	-	-	-	Image Space loss	40.45(-2.03)	57.96(-1.24)
	-	-	-	w/o IaM-KD loss	41.14(-1.34)	58.39(-0.81)
	-	-	-	w/o IrM-KD loss	41.46(-1.02)	58.66(-0.54)

4.3 Ablation Studies

Table 2 illustrates ablation experiments on the EN-DE task to explore the impact of different collocations of distillation modules.

Similarity Function First, we explore the effect of using varied similarity functions to measure the divergence between hidden representations in our distillation module. As shown in row (A), the L_2 norm is the best option. Later, the performance order is KL Divergence (Kullback and Leibler, 1951) > L_1 norm > Cosine similarity > L_∞.

Distillation Granularity Second, in row (B), we analyze what distillation granularity would be the golden standard of our model for optimal translation performance. Specifically, the "Layer", "Block" and "Model" represents that we employ representations of each layer, each block, the last convolutional layer and the image in teacher-student models to compute the distillation loss. Based on the evaluation results, we conclude that the "Model" is optimal, and the "Block" is consistent with the "Layer" in the Meteor score, but slightly inferior in the BLEU score. The such phenomenon reflect that the initial and terminal representations in our knowledge distillation are sufficient to teach the student model to generate information-rich features. This case breaks the stereotype that KD must transmit all knowledge.

CNN Backbone Third, in row (C), we devise three variants with diverse CNN backbones to investigate their impact on the translation. The ResNet50 wins this round since the deep

residual network can derive the strongest visual representation. The VGG19 performs worse with the absence of residual connection and plenty of training samples for model convergence. Undoubtedly, the lightweight AlexNet incurs the worst translation degradation. It implies that the feature extraction capability of a small model may be difficult to undertake the heavy task of multi-supervised learning.

Distillation Loss Finally, we discuss the translation performance of different distillation loss strategies in row (D). Unsurprisingly, w/o (IrM-KD+IaM-KD) loss suffers the severest performance degradation. Removing visual distillation leads to the absence of visual perception in multimodal features, which evolves into a perturbed feature obtained by passing the global text feature into the Fc&Avg Unpool. Afterwards, w/o IrM-KD loss outperforms w/o IaM-KD loss, indicating that the capability of the IaM-KD to establish the text-image relevance that is critical for multimodal feature synthesis is stronger than the IrM-KD. We assume this event is related to that the IaM-KD covers the propagation path of the IrM-KD. Compared with Image Space loss, the improvement of our method reveals that the intermediate hidden state of the teacher model plays a vital role in teaching the student model to comprehend the text-image correlations, as also verified in preceding KD work (Romero et al., 2014; Yim et al., 2017). Overall, each distillation loss considerably improves translation.

Ablation Studies on Development set Table 3 attaches all the validation ablation results to corroborate that each distillation hyperparameter also contributes its decent gains on the model convergence rather than just the model generalization. Drawing from the tabular results, all hyperparameters can be tuned freely on the dev set. We further notice that the performances on the dev set align quite well with the testing set, in terms of tendency.

5 Analysis

In this section, we will investigate our IKD-MMT model from multiple perspectives.

5.1 Does IKD-MMT really generate multimodal features?

To explore the multimodal features generated by our distillation strategy, we test their informative

Table 4: Image retrieval tasks on the Multi30K dataset.

	R@1	R@5	R@10	R@15
Train	0	0.02	0.04	0.05
Valid	0.1	0.69	0.89	1.38
Test2016	0.1	0.7	1.0	1.5
Test2017	0.1	0.5	1.1	1.5
MSCOCO	0.22	0.65	2.39	2.82

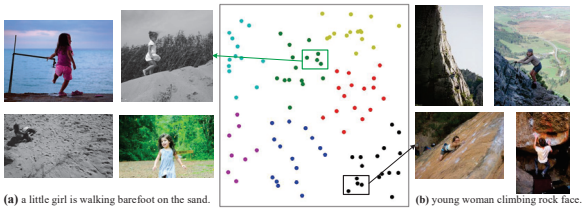

Figure 3: The cluster analysis of the learned multimodal feature, where the two colored boxes represent some representative images in the two cluster cases. The arrow points to the original image that is belonged to the current multimodal feature (i.e. cluster center).

richness from three aspects:

Image Retrieval The image retrieval task aims to analyzes the relationship between our generated multimodal feature and the visual feature. Specifically, we generate the multimodal feature from each source sentence. Further, we find the K closest visual features for each multimodal feature based on cosine similarity. Then, we measure the R@K score, which calculates the recall rate of the visual feature of current sample in these top K nearest neighborhoods. The results in Table 4 display that no matter any K, or whichever data set, the R@K scores are extremely low. These retrieval scores confirm that our model is not trying to generate the visual feature of the current image.

Cluster Visualization In Figure 3, we visualize the related pictures which retrieved by the multimodal feature at the cluster map. Here, points of different colors fall into different clusters, and the distance between points is specified by the cosine similarity between multimodal features and visual features. In the cluster case (a), the other images exist the points-of-parity with the original image, namely objects, backgrounds, and actions (girl, sand, walking). Likewise, in the cluster case (b), the other images satisfy the identical thematic content (person, rock, climbing) as the original image. Certainly, these related pictures also conform to the original text's description of the scene. So the multimodal features are confirmed to have learned commonalities between images.

Attention Weights In Figure 4, we envision the

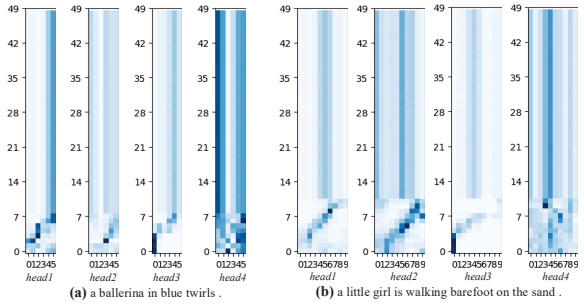

Figure 4: Visualization of attention weights for fusion of multimodal features and text features. The weight values decreasing as the color becomes lighter.

attention weights[4] for the fusion of multimodal features and text features. These weights display which text words the different regions of the multimodal feature focus on. Combining the two examples, several insights are excavated as follows: 1) Part of the multimodal feature with the size of sentence length can be regarded as "pseudo-words", and a word alignment is formed with the text feature. 2) The rest of the multimodal features pay the attention to words equally. We conjecture that these regions as non-object parts thus tend to contribute a consistent impact on text translation. 3) The attention weight of the former three attention heads are flat and presents linearization at the bottom part, while one of 4th attention head is fluctuating and presents dispersion at the bottom part. This means that the first three attention heads capture the entire sentence semantics with the "global attention" form. The 4th attention head, acts like the "local attention", and emphasizes understanding the keywords of the sentence. These findings demonstrate that the multimodal features have embedded the textual semantics.

To summarize, the above experiments can fully prove that our IKD-MMT reliably generates an information-rich multimodal feature.

5.2 Can multimodal features be directly used for translation?

Our IKD-MMT model synthesizes a multimodal feature equipped with textual and image knowledge through a multimodal generator. A natural question to ask is, can multimodal features be fed into the encoder alone, rather than being cascaded with textual features for translation?

To this end, we compare the model removing text features with the original benchmark in Table 5. We notice that w/o Text Feat. appears a cliff-like performance drop, which is explainable. In the multimodal encoding layer, the dot product of the query and key vectors is used to mark the importance of each token corresponding to other tokens in the sentence, i.e. the attention score. If we treat the multi-modal feature as the query, its fixed P regions can be regarded as a set of pseudo tokens. Considering this token set carries limited semantics and destroys the word alignment, it is difficult to obtain an available attention score alone. In addition, most studies convey that text semantics is more important than visual perception in the MMT task (Grönroos et al., 2018; Lala et al., 2018).

[4]They are computed from the 4 attention heads in the first multimodal transformer encoder layer.

Table 5: Results of IKD-MMT without text features.

	Test2016		Test2017		MSCOCO	
	B	M	B	M	B	M
IKD-MMT	41.28	58.93	33.83	53.21	30.17	48.93
w/o Text Feat.	22.06	39.44	19.35	36.67	16.00	32.54

Table 6: Results of two degraded text and original text[5].

Model	$\mathcal{D}$	$\mathcal{D}_{\mathcal{C}}$	$\mathcal{D}_{\mathcal{E}}$
Transformer	52.5	50.28 (↓2.22)	33.81 (↓ 18.61)
Multimodal	53.18	51.30 (↓1.88)	35.04 (↓ 18.14)
IKD-MMT	53.21	51.30 (↓1.91)	34.59 (↓ 18.62)

5.3 Can multimodal features recover the missing text?

In Table 6, we adopt two degradation strategies (Ive et al., 2019; Caglayan et al., 2019) for the source sentence, and feed into Transformer, Multimodal and our IKD-MMT, to probe whether multimodal features can recover the missing text. Test2017 Meteor scores are used for evaluation.

Color Deprivation We mask the source tokens that refer to colors as a special token [U], which involves 3.19% and 3.16% of the words in the training set and test set, respectively. As shown in the column $\mathcal{D}_{\mathcal{C}}$, after color deprivation, the text-only Transformer fails to align the source and target tokens, then leads to the worst performance descent. Our IKD-MMT and Image-must Multimodal hardly synthesize color information to compensate for the deterioration of color missing.

Entity Masking We tag all visually depictable entities (Plummer et al., 2015) with a special token

[5]Here, we use the result of the reproduced models.

(b)

(a)

Source text:	a ballerina in blue twirls .	a little girl is walking barefoot on the sand .
Target text:	eine ballerina in blau wirbelt herum .	ein kleines mädchen geht barfuß im sand .
Transformer:	eine ballerina in blau dreht sich .	ein kleines mädchen läuft barfuß auf dem sand .
NMT$_{SRC+IMG}$:	eine ballerina in blauer kleidung wirbelt herum .	ein kleines mädchen läuft barfuß auf dem sand .
Multitask:	eine ballerina in blauer kleidung wirbelt herum .	ein kleines mädchen läuft barfuß auf dem sand .
URA-NMT:	eine ballerina in blau dreht sich .	ein kleines mädchen läuft barfuß im sand .
Multimodal:	eine ballerina in blau dreht sich .	ein kleines mädchen läuft barfuß auf dem sand .
IKD-MMT(ours):	eine ballerina in blau wirbelt herum .	ein kleines mädchen geht barfuß im sand .
	(a)	(b)

Figure 5: Translation cases of different models. The red and blue highlight error and correct translations respectively.

[U], which affects 29.49% and 31.12% of the words in the training set and test set, respectively. In the $\mathcal{D}_{\mathcal{E}}$ column, we observe that the IKD-MMT and the text-only Transformer degrade equally in performance, which is because IKD-MMT unable to distill the visual perception of multimodal features from the entity-missed text.

Beyond these two masking experiments, we revisit such token recovery problems to pose a more common-sense insight: As per the faithfulness-first principle (Koehn, 2009) in translation, once the source sentence misses keyword information, what we need to do is translate this degraded faithfully. Re-translate back to the original target text from the degraded source text is false.

5.4 Case Study

Figure 5 depicts the 1-best translation of the two test cases generated by various systems. Other systems mistranslate and over-translate text in case (a) and distort the semantics due to mistakenly translating "geht" (walking) to "läuft" (running) in case(b). Our IKD-MMT relies on rich multimodal semantics to keep the translation fidelity.

6 Conclusion

In this work, we propose the IKD-MMT framework to address the image-must issue for multimodal machine translation (MMT) via the knowledge distillation paradigm. Under this image-free MMT system, there are three key contributions: 1) An information-rich multimodal feature is generated by the dual constraints of visual distillation and text translation to support the image-free testing stage; 2) The knowledge distillation module is flexible, and pioneers to employ of the pre-trained model to guide translation; 3) Both quantitative and qualitative results validate the feasibility of the proposed approach IKD-MMT, where it can be deemed the first framework that rivals or even surpass most (if not all) image-must frameworks.

Acknowledgements

Ru Peng and Junbo Zhao were supported by the Fundamental Research Funds for the Central Universities (No. 226-2022-00028). Junbo Zhao also wants to thank the Zhejiang University startup package. The authors would like to thank the Institute of Computer Innovation of Zhejiang University for the high-performance computing platform.

Limitations

From a representation learning perspective, this work is dedicated to introduce the visual perception pipeline and the comprehension of text-image correlations from texts, and may be limited to more complex visual descriptive text (if there exists numerous visual descriptive entities). Further, since the Multi30K is the only and most commonly used MMT benchmark, most of the experiments are centered around it. We addtionally made a "bold" attempt to move IKD-MMT onto much larger scaled NMT datasets for testing only — thanks to the image-free nature of our approach — and unfortuantely the inference results did not look decent enough. While the IKD-MMT's image-free inference pass can be fully facilitated in this scenario, we attribute the inferior results to the much simpler data distribution involved in the Multi30K. Indeed, we hope a richer or real-world MMT dataset could fully bridge this image-free performance gap between MMT and NMT. That, however, may have gone beyond the scope of this paper.

References

Hasan Sait Arslan, Mark Fishel, and Gholamreza Anbarjafari. 2018. Doubly attentive transformer machine translation. *arXiv preprint arXiv:1807.11605*.

Loïc Barrault, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank. 2018.

Findings of the third shared task on multimodal machine translation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 304–323.

Cristian Buciluco, Rich Caruana, and Alexandru Niculescu-Mizil. 2006. Model compression. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 535–541.

Ozan Caglayan. 2019. *Multimodal machine translation*. Ph.D. thesis, Université du Maine.

Ozan Caglayan, Walid Aransa, Adrien Bardet, Mercedes García-Martínez, Fethi Bougares, Loïc Barrault, Marc Masana, Luis Herranz, and Joost van de Weijer. 2017. LIUM-CVC submissions for WMT17 multimodal translation task. In *Proceedings of the Second Conference on Machine Translation*, pages 432–439.

Ozan Caglayan, Pranava Madhyastha, Lucia Specia, and Loïc Barrault. 2019. Probing the need for visual context in multimodal machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4159–4170.

Iacer Calixto and Qun Liu. 2017. Incorporating global visual features into attention-based neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 992–1003.

Iacer Calixto, Qun Liu, and Nick Campbell. 2017. Doubly-attentive decoder for multi-modal neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1913–1924.

Iacer Calixto, Miguel Rios, and Wilker Aziz. 2019. Latent variable model for multi-modal translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6392–6405.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380.

Desmond Elliott, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. 2017. Findings of the second shared task on multimodal machine translation and multilingual image description. In *Proceedings of the Second Conference on Machine Translation*, pages 215–233.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30K: Multilingual English-German image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74.

Desmond Elliott and Ákos Kádár. 2017. Imagination improves multimodal translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 130–141.

Jonas Gehring, Michael Auli, David Grangier, and Yann Dauphin. 2017. A convolutional encoder model for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 123–135.

Stig-Arne Grönroos, Benoit Huet, Mikko Kurimo, Jorma Laaksonen, Bernard Merialdo, Phu Pham, Mats Sjöberg, Umut Sulubacak, Jörg Tiedemann, Raphael Troncy, and Raúl Vázquez. 2018. The MeMAD submission to the WMT18 multimodal translation task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 603–611.

Saurabh Gupta, Judy Hoffman, and Jitendra Malik. 2016. Cross modal distillation for supervision transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2827–2836.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Jindřich Helcl, Jindřich Libovický, and Dušan Variš. 2018. CUNI system for the WMT18 multimodal translation task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 616–623.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Julia Ive, Pranava Madhyastha, and Lucia Specia. 2019. Distilling translations with visual awareness. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6525–6538.

Philipp Koehn. 2009. *Statistical machine translation*. Cambridge University Press.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105.

Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

Chiraag Lala, Pranava Swaroop Madhyastha, Carolina Scarton, and Lucia Specia. 2018. Sheffield submissions for WMT18 multimodal translation shared task. In *Proceedings of the Third Conference*

on Machine Translation: Shared Task Papers, pages 624–631.

Huan Lin, Fandong Meng, Jinsong Su, Yongjing Yin, Zhengyuan Yang, Yubin Ge, Jie Zhou, and Jiebo Luo. 2020. Dynamic context-guided capsule network for multimodal machine translation. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1320–1329.

Pengbo Liu, Hailong Cao, and Tiejun Zhao. 2021. Gumbel-attention for multi-modal machine translation. *arXiv preprint arXiv:2103.08862*.

Quanyu Long, Mingxuan Wang, and Lei Li. 2021. Generative imagination elevates machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5738–5748.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649.

Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. 2016. Generative adversarial text to image synthesis. In *International conference on machine learning*, pages 1060–1069. PMLR.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. 2014. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Lucia Specia, Stella Frank, Khalil Sima'an, and Desmond Elliott. 2016. A shared task on multimodal machine translation and crosslingual image description. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 543–553.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Dexin Wang and Deyi Xiong. 2021. Efficient object-level visual context modeling for multimodal machine translation: Masking irrelevant objects helps grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2720–2728.

Zhiyong Wu, Lingpeng Kong, Wei Bi, Xiang Li, and Ben Kao. 2021. Good for misconceived reasons: An empirical revisiting on the need for visual context in multimodal machine translation. *arXiv preprint arXiv:2105.14462*.

Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. 2018. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1316–1324.

Shaowei Yao and Xiaojun Wan. 2020. Multimodal transformer for multimodal machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4346–4350.

Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. 2017. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4133–4141.

Yongjing Yin, Fandong Meng, Jinsong Su, Chulun Zhou, Zhengyuan Yang, Jie Zhou, and Jiebo Luo. 2020. A novel graph-based multi-modal fusion encoder for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3025–3035.

Mingkuan Yuan and Yuxin Peng. 2018. Text-to-image synthesis via symmetrical distillation networks. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 1407–1415.

Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. 2017. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 5907–5915.

Yin Zhang, Rong Jin, and Zhi-Hua Zhou. 2010. Understanding bag-of-words model: a statistical framework. *International Journal of Machine Learning and Cybernetics*, 1(1-4):43–52.

Zhuosheng Zhang, Kehai Chen, Rui Wang, Masao Utiyama, Eiichiro Sumita, Zuchao Li, and Hai Zhao. 2020. Neural machine translation with universal

visual representation. In *International Conference on Learning Representations*.

Mingyang Zhou, Runxiang Cheng, Yong Jae Lee, and Zhou Yu. 2018. A visual attention grounding neural model for multimodal machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3643–3653.

A Appendix

Table 7: Architecture of each module in multimodal feature generation. The multimodal student model has inverted data flow with visual teacher model. These architectures can be easily replaced with any CNN variant (e.g. VGG19 (Simonyan and Zisserman, 2014), AlexNet (Krizhevsky et al., 2012)) with reference to ResNet50 (He et al., 2016).

Visual Teacher Model			**Multimodal Student Model**		
layer name	output size	49-layer	layer name	output size	48-layer
T-conv1	112x112	7x7, 64, stride 2	S-conv1	224x224	8x8, 3, stride 2
T-conv2_x	56x56	3x3 max pool, stride 2 $\begin{bmatrix} 1\text{x}1, 64 \\ 3\text{x}3, 64 \\ 1\text{x}1, 256 \end{bmatrix}$ x3	S-conv2_x	112x112	2x2 max unpool, stride 2 $\begin{bmatrix} 1\text{x}1, 256 \\ 3\text{x}3, 64 \\ 1\text{x}1, 64 \end{bmatrix}$ x3
T-conv3_x	28x28	$\begin{bmatrix} 1\text{x}1, 128 \\ 3\text{x}3, 128 \\ 1\text{x}1, 512 \end{bmatrix}$ x4	S-conv3_x	56x56	$\begin{bmatrix} 1\text{x}1, 512 \\ 3\text{x}3, 128 \\ 1\text{x}1, 128 \end{bmatrix}$ x4
T-conv4_x	14x14	$\begin{bmatrix} 1\text{x}1, 256 \\ 3\text{x}3, 256 \\ 1\text{x}1, 1024 \end{bmatrix}$ x6	S-conv4_x	28x28	$\begin{bmatrix} 1\text{x}1, 512 \\ 3\text{x}3, 256 \\ 1\text{x}1, 256 \end{bmatrix}$ x6
T-conv5_x	7x7	$\begin{bmatrix} 1\text{x}1, 512 \\ 3\text{x}3, 512 \\ 1\text{x}1, 2048 \end{bmatrix}$ x3	S-conv5_x	14x14	$\begin{bmatrix} 1\text{x}1, 1024 \\ 3\text{x}3, 512 \\ 1\text{x}1, 512 \end{bmatrix}$ x3
N/A	1x1	average pool			
Multimodal Feature Generator			N/A	7x7	2048-d fc, average unpool