On Vision Features in Multimodal Machine Translation

Anonymous ACL submission

Abstract

Previous work on multimodal machine translation (MMT) has focused on the way of incorporating vision features into translation but little attention is on the quality of vision models. In this work, we investigate the impact of vision models on MMT. Given the fact that Transformer is becoming popular in computer vision, we experiment with various strong models (such as Vision Transformer) and enhanced features (such as object-detection and image captioning). We develop a selective attention model to study the patch-level contribution of an image in MMT. On detailed probing tasks, we find that stronger vision models are helpful for learning translation from the vision modality. Our results also suggest the need of carefully examining MMT models, especially when current benchmarks are small-scale and biased.

1 Introduction

Multimodal machine translation (MMT) has emerged as an active field of research which marries the worlds of computer vision (CV) and natural language processing (NLP) (Specia et al., 2016). Early models of this kind produce a translation given the fused representation of both the visual and textual inputs (Caglayan et al., 2016; Libovický and Helcl, 2017; Calixto and Liu, 2017). As expected, such a paradigm achieves promising BLEU improvements and inspires the community to follow up.

But soon researchers find that MMT systems do not act as what we would ordinarily design: the vision modality contributes to translation little. For example, it is not harmful to MMT systems when the input image is irrelevant to the text (Grönroos et al., 2018; Lala et al., 2018), or even when the vision features are absent (Elliott, 2018). More recently, Wu et al. (2021) have pointed out that the use of the vision modality is a way of regularization for training but not a complement to the text modality. As another response to the analysis of MMT, Caglayan et al. (2019) investigate how the vision features correlate to the text. They find that the input image helps translation when some of the input words are masked.

Note that previous work has for the most part focused on integrating off-the-shelf vision models (such as ResNet-50) into MMT. The underlying assumption here is that the existing vision models are powerful enough to encode the image. This implicitly ignores the quality of vision models in representing images. But computer vision is facing a new trend by moving from CNNs to Transformer as the backbone model (Dosovitskiy et al., 2021; Liu et al., 2021b; Carion et al., 2020). A natural question that arises is: *how will MMT systems behave if stronger vision models are adopted?*

In this work, we address this question by a systematic study of using various vision models in MMT, in particular using the most successful models in recent studies (such as Vision Transformer, or ViT for short). We find that the patching method used in Transformer-based vision models offers an opportunity to detail the patch-level contribution of the image. This leads us to develop a selective attention model to correlate words with image patches. Beyond this, we introduce object-detection and image captioning features into MMT for further improvements of the vision models (Carion et al., 2020; Fang et al., 2021).

Following (Caglayan et al., 2019)'s work, we design more detailed probing tasks to examine to what degree the vision modality contributes to MMT. We run an extensive set of experiments on En-De and En-Fr MMT tasks. Our findings are

- Stronger vision models help. For example, ViT can beat ResNet-50 on the probing tasks though the superiority is not significant on standard MMT data.

- Automatic evaluation on current MMT tasks

SRC:	a	man	in	green	pants	walking	down	the	road
MASK1	a	[MASK_P]	in	green	pants	walking	down	the	road
MASK2	a	man	in	[MASK_C]	[MASK_NS]	walking	down	the	road
MASK3	a	man	in	[MASK_C]	[MASK_NS]	walking	down	the	[MASK_N]
MASK4	a	[MASK_P]	in	[MASK_C]	[MASK_NS]	walking	down	the	[MASK_N]

Table 1: An example of the proposed the probing tasks. We replace the masked token by four symbols respectively.

might not be a good indicator for the effectiveness of MMT models. For example, models enhanced with object-detection and image captioning features yield good BLEU scores on the original MMT task but show modest or no contributions on the probing tasks.

We hope that the results here can inspire more research on exploring better vision models and evaluation methods for multimodal NLP.

2 Preliminary

We start with description of the probing tasks. It is followed by a design of vision features and a selective attention mechanism for introducing ViT-like representations into MMT.

2.1 Insufficient Text Generation

To know how an image contributes to translation, a way is to mask some of the input words (call this insufficient text) and force the translation model to learn from the image. Following previous design of color deprivation and entity-based masking, we present detailed probing tasks which are complementary to Caglayan et al. (2019)'s work. In preliminary experiments[1], we find that "color", "character" and "noun" are three kinds of words which could be complemented according to the vision modality once the corresponding texts are masked. The following probing tasks are designed accordingly.

Color-based Probing In training, all source words referring to a color are replaced by a special token [Mask_C]. There are $8,919$ sentences involving color words, and nearly one third of them involve more than one color. It is worth noting that each color may have two or more translations due to the rich morphology in German and French. For example, the English "green" can be translated to "grün", "grüne", "grünes", "grüner", "grünen" and "grünem" in German. We design two criteria to measure the accuracy of translation. The first criterion is strict. The correct translation requires generating the same color and the same gender as in reference translations. The second criterion is relaxed and all translations expressing the same color are correct.

Character-based Probing For character words, we choose "man", "woman", "people", "men", "girl" and "boy". Each character word has a single translation only, except for "people". Because about 60% sentences contain character words in our training data, they are reasonable indicators of assessing the ability of inferring correct translations from the input image. Here we use [MASK_P] for masking.

Noun-based Probing For more complex scenarios, a sentence can be masked with several kinds of ambiguous words, such as animals, clothing, and vehicles, provided by Flickr30K (Plummer et al., 2015). High-frequency words labeled with noun (or nouns) are more likely to be masked as [MASK_N] (or [MASK_NS])). See Table 1 for example insufficient text with different numbers of masks.

2.2 Various Vision Features

In addition to ResNet-50, we choose several Transformer-based vision models.

- General Backbone. Vision Transformer (ViT) and Swin Transformer are popular models in computer vision (Dosovitskiy et al., 2021; Liu et al., 2021b). We use ViT with various model capacities to vary from weak to strong ViT models.

- Object-detection. For pretrained object-detection vision models, we choose DETR (Carion et al., 2020) and QueryInst (Fang et al., 2021) for their strong performance.

- Image Captioning. For image captioning models, we choose CATR because it is a

[1] We choose the Multi30K En-De and En-Fr datasets for experiments.

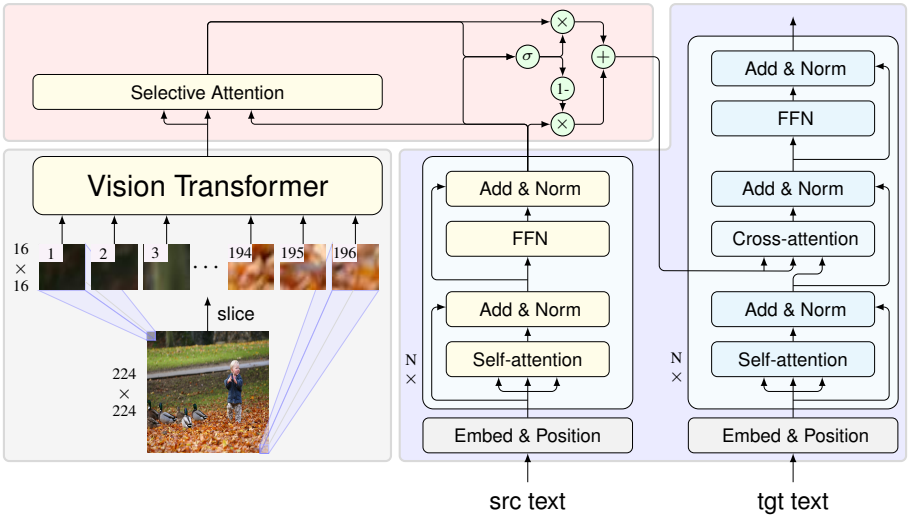

Figure 1: The overview of selective attention multimodal Transformer when using ViT as the vision feature.

Transformer-based image captioning architecture and can be easily implemented on top of ViT.

We form a number of vision features by combining the methods described above. More details are presented in Section 3.

2.3 Selective Attention

ViT and related models perform in almost the same way as Transformer in NLP (Vaswani et al., 2017). Unlike the general models in CV, ViT does not represent the image as a single vector. Instead, it generates a sequence of patches for image representation. An advantage of this design is that we can use the attention mechanism to correlate image patches to words. Thus, we present a selective attention model to model the patch-level contribution of the image. See Figure 1 for the architecture.

Text-only Transformer Transformer follows an encoder-decoder paradigm (the purple region in Figure 1) . The encoder is a stack of identical layers. Each layer consists of a self-attention (SAN) block and a feedforward network (FFN) block. The decoder shares a similar design with the encoder, but with an additional cross-attention block.

Gated Fusion Gated fusion mechanism is a popular technique for fusing representations from different sources (Wu et al., 2021; Zhang et al., 2020; Lin et al., 2020; Yin et al., 2020). Given the text input X^{text} and the image input X^{img}, the text representation H^{text} and the image feature H^{img} can be defined as:

$$H^{\text{text}} = \text{TransformerEncoder}(X^{\text{text}}) \quad (1)$$

$$H^{\text{img}} = W\ \text{ViT}(X^{\text{img}}) \quad (2)$$

where W is a projection matrix to convert the shape of $\text{ViT}(X^{\text{img}})$ into that of H^{text}. Note that $\text{ViT}(\cdot)$ can be replaced by other vision models, e.g. DETR, Swin Transformer and etc. Then, the gate $\lambda \in [0, 1]$ and the fuzed output are defined as:

$$\lambda = \text{Sigmoid}(UH^{\text{text}} + VH^{\text{img}}) \quad (3)$$

$$H^{\text{Out}} = (1 - \lambda) \cdot H^{\text{text}} + \lambda \cdot H^{\text{img}} \quad (4)$$

where U and V are trainable variables. λ controls how much visual information is kept. Then, the fusion vector H^{Out} is fed into the decoder. See the pink region in Figure 1 for an illustration of the gated fusion models.

Selective Attention After obtaining the text and image representations (or features), we use a single-head attention network to correlate words with image patches, where the query, key and value are H^{text}, H^{img} and H^{img}, respectively. Then the selective attention output H^{img}_{attn} is defined to be:

$$H^{\text{img}}_{attn} = \text{Softmax}(\frac{QK^{\text{T}}}{\sqrt{d_k}})V \quad (5)$$

where d_k is the same as the dimension of H^{text} because a single head is used. Then the fused representation could be obtained by using Eqs. 3 and 4 and replacing H^{img} with H^{img}_{attn}.

#	Model	Feature	English→German						English→French					
			Test2016		Test2017		MSCOCO		Test2016		Test2017		MSCOCO	
Text-only Transformer														
1	**Tiny**	-	41.02	68.22	33.36	62.05	29.88	56.64	61.80	81.02	53.46	75.62	44.52	69.43
Existing MMT Systems														
2	**Doubly-ATT**	ResNet	41.45	68.04	33.95	61.83	29.63	56.21	61.99	81.12	53.72	75.71	45.16	70.25
3	**Imagination**	ResNet	41.31	68.06	32.89	61.29	29.90	56.57	61.90	81.20	54.07	76.03	44.81	70.35
4	**UVR-NMT**	ResNet	40.79	-	32.16	-	29.02	-	61.00	-	53.20	-	43.71	-
5	**Gated Fusion**	ResNet	41.96	67.84	33.59	61.94	29.04	56.15	61.69	80.97	54.85	76.34	44.86	70.51
Our MMT Systems														
6	**Gated Fusion**	ViT-Large	41.55	68.34	33.49	61.67	29.27	55.64	61.93	81.08	54.98	75.12	45.65	70.81
7	**Selective Attn**	ViT-Large	41.84	68.64	34.32	62.32	30.22	56.91	62.24	81.41	54.52	76.30	44.82	70.63
8	**7 + ViT-Tiny**	ViT-Tiny	40.74	67.20	32.48	60.46	28.10	55.19	61.44	80.91	53.31	75.65	45.82	70.75
9	**7 + ViT-Small**	ViT-Small	40.86	67.64	33.62	61.61	29.72	56.94	61.78	81.30	54.21	76.04	45.28	70.89
10	**7 + ViT-Base**	ViT-Base	41.93	68.55	33.60	61.42	31.14	56.77	62.48	81.71	54.44	76.46	44.72	71.20
11	**7 + DETR**	DETR	42.23	68.94	34.14	61.57	30.13	57.01	62.14	81.45	55.17	76.40	45.10	70.38
12	**7 + QueryInst**	QueryInst	41.90	68.64	34.90	62.27	30.20	56.89	62.33	81.26	54.97	76.61	45.56	70.64
13	**7 + CATR**	CATR	42.50	68.81	34.28	61.81	29.59	56.36	62.79	81.75	55.44	76.57	45.27	70.73

Table 2: BLEU (left) and METEOR (right) scores of En→De and En→Fr tasks. Some of the results are from Wu et al. (2021)'s work.

Systems	Test2016		Test2017		MSCOCO	
	Restrict	Relaxed	Restrict	Relaxed	Restrict	Relaxed
English→German						
Text-only Transformer	25.93	34.42	22.57	35.70	18.75	23.44
Gated Fusion + ResNet	27.23 (↑ 2.30)	35.51 (↑ 1.09)	23.10 (↑ 0.53)	37.01 (↑ 2.01)	21.88 (↑ 3.13)	25.00 (↑ 1.56)
Gated Fusion + ViT	35.08 (↑ 9.15)	42.48 (↑ 8.06)	25.46 (↑ 2.89)	41.73 (↑ 6.03)	25.00 (↑ 6.25)	31.25 (↑ 7.81)
Selective Attn + ViT	**51.20** (↑ **25.27**)	**64.71** (↑ **30.29**)	**31.76** (↑ **9.19**)	**53.54** (↑ **17.84**)	**43.75** (↑ **25.00**)	**56.25** (↑ **32.81**)
English→French						
Text-only Transformer	30.72	33.12	34.91	38.85	23.44	29.69
Gated Fusion + ResNet	32.68 (↑ 1.96)	35.51 (↑ 2.39)	32.55 (↓ 2.36)	35.17 (↓ 3.68)	17.19 (↓ 6.25)	23.44 (↓ 6.25)
Gated Fusion + ViT	45.53 (↑ 14.81)	50.76 (↑ 17.64)	45.41 (↑ 10.50)	52.23 (↑ 13.38)	34.38 (↑ 10.94)	43.75 (↑ 14.06)
Selective Attn + ViT	**62.96** (↑ **32.24**)	**68.85** (↑ **35.73**)	**49.34** (↑ **14.43**)	**55.38** (↑ **16.53**)	**43.75** (↑ **20.31**)	**53.12** (↑ **23.43**)

Table 3: The accuracy of MMT systems when applied color-based probing.

3 Experiments

3.1 Datasets

We conducted experiments on the widely used Multi30K benchmark (Elliott et al., 2016). The training and validation sets consisted of 29,000 and 1,014 instances, respectively. We reported the results on the Test2016, Test2017 and MSCOCO test sets (Elliott et al., 2017). Note that MSCOCO is more challenging for MMT models due to the out-of-domain instances with ambiguous verbs. Following the setup in (Wu et al., 2021), we learned a joint BPE code for 10,000 merging operations for both the source and target languages, resulting in vocabularies of 9,716 and 9,548 entries for the En-De and En-Fr tasks.

3.2 Experimental Setups

We followed the Wu et al. (2021)'s work to conduct experiments with Transformer-Tiny configuration, which is more suited for small datasets like Multi30K. Note that smaller models even obtains higher BLEU scores than pervious MMT models. Similar observations have been discussed when building context-aware machine translation models (Li et al., 2020). The model consists of 4 encoder and decoder layers. The hidden size is 128 and the filter size of FFN is 256. There are 4 heads in the multi-head self-attention mechanism. We set the dropout as 0.3 and the label smoothing as 0.1.

Our implementation was based on Fairseq (Ott et al., 2019). For training, we used Adam Optimizer (Kingma and Ba, 2015) with $\beta_1 = 0.1$, $\beta_2 = 0.98$ and $\epsilon = 10^{-8}$. We adopted the same learning rate schedule as (Vaswani et al., 2017), where the learning rate first increased linearly for $warmup = 2000$ steps from $1e^{-7}$ to $5e^{-3}$. After the warmup, the learning rate decayed proportionally to the inverse square root of the current step. Each training batch contained 4,096 tokens. We also adopted the early-stop training strategy (Zhang et al., 2020) to avoid the overfitting issue.

For evaluation, we averaged the last 10 checkpoints for more reliable results. The width of beam

Systems	Test2016		Test2017		MSCOCO	
	Restrict	**Relaxed**	**Restrict**	**Relaxed**	**Restrict**	**Relaxed**
English→German						
Text-only Transformer	59.49	64.05	58.56	62.53	60.94	65.62
Gated Fusion + ResNet	60.06 (↑ 0.57)	64.91 (↑ 0.86)	56.08 (↓ 2.48)	59.06 (↓ 3.47)	61.72 (↑ 0.78)	65.23 (↓ 0.39)
Gated Fusion + ViT	66.33 (↑ 6.84)	70.76 (↑ 6.71)	67.00 (↑ 8.44)	71.46 (↑ 8.93)	71.09 (↑ 10.15)	75.78 (↑ 10.16)
Selective Attn + ViT	**73.04** (↑ **13.55**)	**78.89** (↑ **14.84**)	**70.97** (↑ **12.41**)	**77.17** (↑ **14.64**)	**73.44** (↑ **12.50**)	**77.73** (↑ **12.11**)
English→French						
Text-only Transformer	63.48	65.48	61.04	62.53	64.84	67.19
Gated Fusion + ResNet	61.63 (↓ 1.85)	63.62 (↓ 1.86)	63.52 (↑ 2.48)	65.01 (↑ 2.48)	64.45 (↓ 0.39)	66.80 (↓ 0.39)
Gated Fusion + ViT	73.47 (↑ 9.99)	75.89 (↑ 10.41)	76.43 (↑ 15.39)	77.92 (↑ 15.39)	**80.47** (↑ **15.63**)	**82.81** (↑ **15.62**)
Selective Attn + ViT	**78.89** (↑ **15.41**)	**81.31** (↑ **15.83**)	**78.16** (↑ **17.12**)	**79.65** (↑ **17.12**)	79.69 (↑ 14.85)	81.64 (↑ 14.45)

Table 4: The accuracy of several MMT systems on character-based probing.

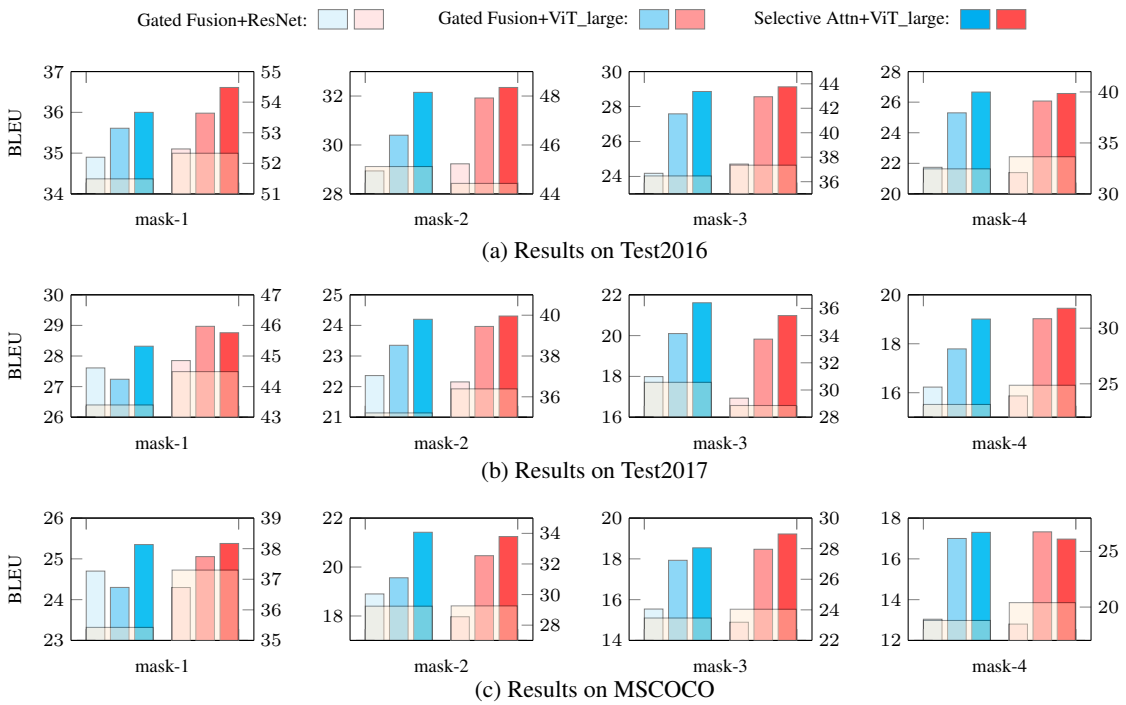

Figure 2: Comparison of systems 5-7 in Table 2 with limited textual context on Test2016. Blue/Red pillars denote the results evaluated on the En-De and En-Fr tasks, respectively. We exhibit the BLEU scores of three MMT models with different masking granularities. The shadow denotes the score obtained by text-only Transformer.

size was set to 5. The performance was measured by BLEU and METEOR for all test sets. Also we used accuracy for evaluation on the probing tasks.

3.3 Results

Table 2 summarizes the results on standard MMT data. Each model was evaluated on three test sets on two language pairs. We see, first of all, that the improvements of previous methods (Rows 2-4) over the tiny baseline are marginal in terms of both BLEU and METEOR. This confirms the assumption that the visual features is not fully used if the text is complete (Caglayan et al., 2019). When switching the vision features from ResNet (Row.5) to ViT (Row.6), there are no significant BLEU gains. Then, we test them on the proposed probing tasks to examine the "real" contribution to MMT.

Color-based Probing Table 3 shows the accuracy on the color-based probing task. We see that the accuracy improvement of the gated fusion method is marginal by both restrict and relaxed criteria. However, replacing ResNet by ViT yields gains of over 8 accuracy points across three test sets on En-De task. Similar improvements are observed on the En-Fr task. The finding here indicates that stronger vision features are helpful for representing the visual information. Moreover, selective attention can make better use of the ViT features, achieving +20 accuracy gains on three test sets. This verifies the conjecture that the selective attention can further enhance the fused representation

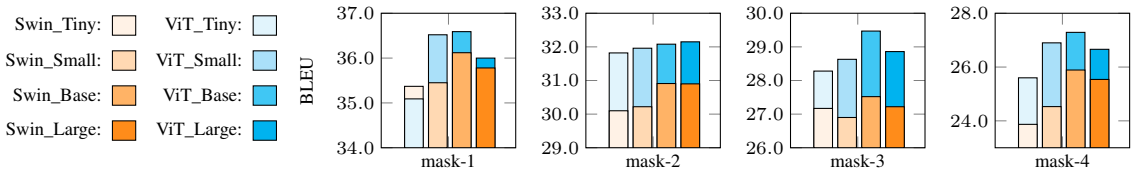

Figure 3: Comparison of BLEU[%] for MMT models with ViT/Swin in various capacities.

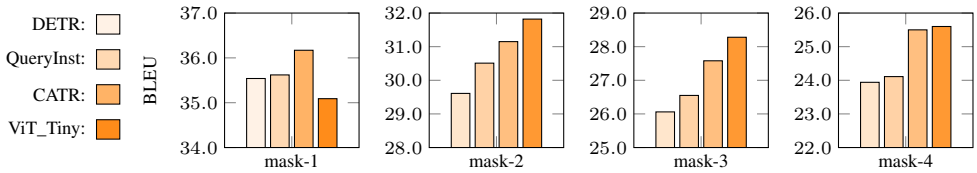

Figure 4: BLEU scores [%] of various vision features on En-De Test2016.

for the ViT features.

Character-based Probing Table 4 shows similar results as in Table 3. ViT with selective attention performs the best. While the gated fusion method with ResNet feature behaves the worst, even compared with the text-only Transformer.

Noun-based Probing Figure 2 plots the results of noun-based masking. It again verifies the above conjecture. The histograms in blue and red denote the results on the En-De and En-Fr tasks, respectively. The ViT features can significantly outperform the ResNet features across all masking methods on the two language pairs. We also observe that the gap between the ResNet and ViT features are gradually enlarged as more nouns are masked. This confirms the results in (Dosovitskiy et al., 2021).

4 Analysis

4.1 How Vision Features Improve the MMT

We further explore the impact of model capacity. Here, we report the results of ViT and Swin Transformer because they are strong models in recent studies. Our conjecture here is that larger ViT/Swin models can describe the image more accurately, which enables the text encoder to receive richer complementary information. Figure 3 depicts the BLEU scores in progressive noun masking scenarios. Intuitively, larger ViT and Swin models provide more complementary knowledge to complete the insufficient text representations.

Nevertheless, a counterintuitive phenomenon is the inferiority of Swin across all scenarios in the same configuration, though it outperforms ViT on most computer vision benchmarks. We attribute the reason to short length of the patch sequence. In patching, ViT has a length of 577 (576 sequence segments and a special token `CLS`) when the image resolution and the patching size are 384×384 and 16×16. However, Swin has a fixed sequence length (49) restricted by the shifted window operation. This leads to more fine-grained local features for ViT, which is beneficial to the selective attention mechanism for extracting more relevant pieces.

4.2 Impact of Learning Objectives

Then, we investigate the impact of the enhanced vision features on MMT. Previous studies have already attempted to leverage object-detection features (Zhao et al., 2020; Wang and Xiong, 2021) but the observation here is slightly different. Beyond the object-detection pretrained features, we also take the image captioning task into account.

Rows 11-13 in Table 2 summarize the results of the three enhanced vision features on the standard MMT data, and Figure 4 depicts the results on insufficient texts. Here we choose ViT-Tiny-based models for comparison due to the similar model capacity they own[2]. We see that not only the object-detection (DETR and QueryInst), but also the image captioning (CATR) pretrained features obtain superior performance compared with ViT-tiny (Row 8) when the text is complete. It is consistent with previous findings (Yin et al., 2020; Zhao et al., 2020). However, the advantages do not persist when switching to limited text scenarios. A possible explanation is that these methods are sensitive to the quality of the extracted objects. We leave this as future work.

[2]Only pretrained vision models in a 256 hidden-size are available

System	Patch	Reso.	Leng.	Color Probing		Character Probing		Noun Probing			
				Restrict	Relaxed	Restrict	Relaxed	$Mask_1$	$Mask_2$	$Mask_3$	$Mask_4$
ViT	16×16	384	576	49.67	64.49	74.32	79.46	36.59	32.08	29.47	27.29
ViT	16×16	224	196	50.11	61.87	68.47	74.32	36.27	31.49	29.70	26.51
ViT	32×32	384	144	49.02	63.18	70.19	76.03	35.53	30.50	28.28	26.20
ViT	32×32	224	49	48.80	61.00	68.19	73.47	35.14	30.30	28.12	25.19
Swin	4×4	224	49	43.57	54.47	70.04	75.18	36.12	30.91	27.52	25.89

Table 5: Comparison of various resolutions and patch sizes on the En-De (Test2016) probing tasks.

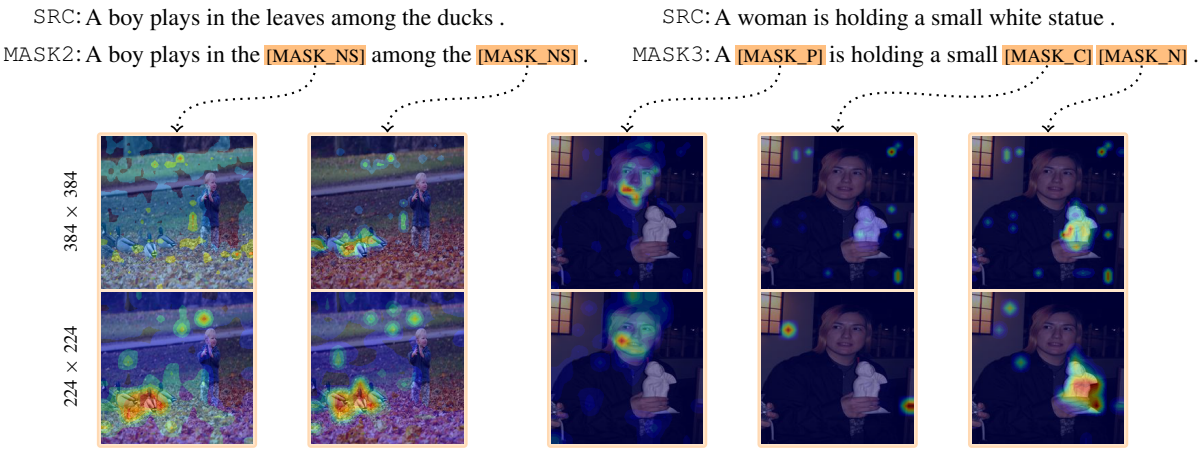

Figure 5: Attention map of ViT in 384 × 384 vs 224 × 224 resolution and 16 × 16 patching.

4.3 Impact of Resolution and Patching Size

It is well-known that higher resolutions are beneficial to the accuracy improvement in computer vision tasks (Dosovitskiy et al., 2021). Despite the success of the Transformer architecture, recent studies show that the success of ViT mainly comes from the successful use of the patching schema (Dosovitskiy et al., 2021). Here, we compare MMT systems with different resolutions and patch sizes based on ViT-Base. The results on three probing tasks (see Table 5) again confirm the above assumption that fine-grained vision features are more suited for the selective attention. Also, the attention map visualized in Figure 5 demonstrate that high resolution with fine-grained patching schema can attend to correct region of the image for each masked token. For example, both models pay the right attention to the masked character and noun, but the model with low resolution fails to detect the right region of color. The finding here may shed light to other multimodal tasks, such as VQA.

4.4 Incongruent Decoding

Incongruent decoding is a widely used manner to evaluate whether the vision modality contributes to the text (Caglayan et al., 2019, 2021). Table 6 shows that incongruent decoding causes obvious BLEU drops except for the ResNet feature. ViT beats the ResNet with gated fusion. It yields higher BLEU scores with congruent decoding and exhibits larger BLEU drop with incongruent decoding. We also find that the ViT features learned from scratch is also insensitive to the vision modality. This is reasonable that the learned vision systems are not sufficiently strong due to the data scarcity of Multi30K. Thus the vision modality acts more like noise signals. In addition, focusing on the results of pretrained selective attention + ViT, the gap between congruent and incongruent decoding gradually becomes larger. Also, the ensemble vision features perform the best. These results indicate that visual contexts help.

4.5 Case Study

Finally, we compare several real cases. We choose gated fusion (`CNN`) (Wu et al., 2021) and selective attention + ViT_Base (`ViT`) for comparison. The qualitative examples in Table 7 demonstrate that the visual modality is complementary rather than redundant if the text is insufficient. To figure out whether the German translation is right or not, we provide the human-translation results. For example, `ViT` can fill in the masked entities and generate the correct translations even four entities were masked. Unfortunately, `CNN` incorrectly judges the man as a woman. Also, it cannot distinguish the right color

System		Mask$_1$		Mask$_2$		Mask$_3$		Mask$_4$	
		Cong.	**Icong.**	**Cong.**	**Icong.**	**Cong.**	**Icong.**	**Cong.**	**Icong.**
Transformer-Tiny		34.37	-	29.12	-	24.03	-	21.64	-
Gated Fusion + ResNet	Pretrained	34.90	34.88	28.94	28.08	24.18	22.56	21.74	20.79
Gated Fusion + ViT	Pretrained	35.61	33.77	30.40	25.43	27.58	19.79	25.30	16.66
Selective Attn + ViT	Pretrained	36.59	32.88	32.08	25.58	29.47	20.42	27.29	15.80
	Scratch	34.91	34.81	28.91	28.91	23.40	23.40	19.63	19.63
Selective Attn + DETR	Pretrained	35.54	33.92	29.61	27.20	26.06	21.65	23.94	18.88
Selective Attn + CATR	Pretrained	36.17	33.13	31.15	26.40	27.58	20.72	25.50	16.98
Select. Attn + ViT + CATR	Pretrained	**36.97**	32.98	**32.45**	24.71	**30.30**	19.92	**28.14**	16.09

Table 6: The impact of incongruent decoding for the noun masking strategy. Here Cong./Icong. denotes congruent and incongruent decoding, respectively. The results (BLEU [%]) were measured on the En-De Test2016 test set.

SRC: a brown-haired [man] in a [green] [shirt] plays a [trumpet] outdoors .
REF: ein **mann** mit braunen haaren in einem **grünen** **hemd** spielt im freien **trompete** .
MK4: a brown-haired [MASK_P] in a [MASK_C] [MASK_N] plays a [MASK_N] outdoors .
CNN: eine braunhaarige ~~frau~~ in einem ~~roten~~ ~~kleid~~ spielt im freien ~~gitarre~~ .
(a brown-haired ~~woman~~ in a ~~red~~ ~~dress~~ plays a ~~guitar~~ outdoors.)
ViT: ein braunhaariger **mann** in einem **grünen** **hemd** spielt im freien **trompete** .
*(a brown-haired **man** in a **green** **shirt** plays a **trumpet** outdoors.)*

SRC: a [boy] is leaning on a [car] with [flowers] on the [hood] .
REF: ein junge lehnt sich an ein auto mit blumen auf der motorhaube .
MK4: a [MASK_P] is leaning on a [MASK_N] with [MASK_NS] on the [MASK_N] .
CNN: ein ~~mann~~ lehnt an einer ~~wand~~ mit ~~bäumen~~ auf der ~~straße~~ .
(a ~~man~~ is leaning on a ~~wall~~ with ~~trees~~ on the ~~street~~.)
ViT: ein kind lehnt sich an einem **auto** mit **blumen** auf dem ~~gehweg~~ .
*(a child is leaning on a **car** with **flowers** on the ~~sidewalk~~.)*

Table 7: Qualitative examples from two complex scenarios. ~~Strikethrough~~ and **bold** words present the incorrect and good lexical choices. Underline denotes the acceptable but not totally right translation.

of shirt due to the complex background. When given a more complex image, it is still a challenge for ViT to generate the totally right translation.

5 Related Work

Multimodal machine translation is a cross-domain task in the filed of machine translation. Early attempts mainly focused on enhancing the MMT model by better incorporation of the vision features (Calixto and Liu, 2017; Elliott and Kádár, 2017; Delbrouck and Dupont, 2017). However, directly encoding the whole image feature brings additional noise to the text (Yao and Wan, 2020; Liu et al., 2021a). To address the above issue, Yao and Wan (2020) proposed a multimodal self-attention to consider the relative difference of information between two modalities. Similarly, Liu et al. (2021a) used a Gumbel Softmax to achieve the same goal.

Researchers also realize that the vision modality maybe redundant. Irrelevant images have little impact on the translation quality, and no significant BLEU drop is observed even the image is absent (Elliott, 2018). Encouraging results appeared in Caglayan et al. (2019)'s work. They pointed out that the visual modality is still useful when the linguistic context is scarce, but is less sensitive when exposed to complete sentences. More recently, Wu et al. (2021) attributed the BLEU gain on MMT tasks to the regularization training. Caglayan et al. (2021) proposed a cross-lingual visual pretraining approach. In this work, we make a systematic study on whether stronger vision features are helpful. We also extend the research to enhanced features, such as object-detection and image captioning, which is complementary to previous work.

6 Conclusions

In this work, we show that stronger vision features (e.g. ViT-like models) strengthen MMT systems on three proposed probing tasks. We present a selective attention method for ViT-based models to make better use of the patch-level representation. The result here shows a promising line of research on developing better vision models for multimodal tasks. Our code and metrics for probing tasks will be open source soon.

References

Ozan Caglayan, Loïc Barrault, and Fethi Bougares. 2016. Multimodal attention for neural machine translation. *CoRR*, abs/1609.03976.

Ozan Caglayan, Menekse Kuyu, Mustafa Sercan Amac, Pranava Madhyastha, Erkut Erdem, Aykut Erdem, and Lucia Specia. 2021. Cross-lingual visual pre-training for multimodal machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1317–1324, Online. Association for Computational Linguistics.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

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, Copenhagen, Denmark. Association for Computational Linguistics.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I*, volume 12346 of *Lecture Notes in Computer Science*, pages 213–229. Springer.

Jean-Benoit Delbrouck and Stéphane Dupont. 2017. An empirical study on the effectiveness of images in multimodal neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 910–919, Copenhagen, Denmark. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Desmond Elliott. 2018. Adversarial evaluation of multimodal machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2974–2978, Brussels, Belgium. Association for Computational Linguistics.

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, Copenhagen, Denmark. Association for Computational Linguistics.

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, hosted by the 54th Annual Meeting of the Association for Computational Linguistics, VL@ACL 2016, August 12, Berlin, Germany*. The Association for Computer Linguistics.

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, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Yuxin Fang, Shusheng Yang, Xinggang Wang, Yu Li, Chen Fang, Ying Shan, Bin Feng, and Wenyu Liu. 2021. Instances as queries. *CoRR*, abs/2105.01928.

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, Belgium, Brussels. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

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, Belgium, Brussels. Association for Computational Linguistics.

Bei Li, Hui Liu, Ziyang Wang, Yufan Jiang, Tong Xiao, Jingbo Zhu, Tongran Liu, and Changliang Li. 2020. Does multi-encoder help? a case study on context-aware neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3512–3518, Online. Association for Computational Linguistics.

Jindřich Libovický and Jindřich Helcl. 2017. Attention strategies for multi-source sequence-to-sequence learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*

(Volume 2: Short Papers), pages 196–202, Vancouver, Canada. Association for Computational Linguistics.

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 *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*, pages 1320–1329. ACM.

Pengbo Liu, Hailong Cao, and Tiejun Zhao. 2021a. Gumbel-attention for multi-modal machine translation. *CoRR*, abs/2103.08862.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021b. Swin transformer: Hierarchical vision transformer using shifted windows. *CoRR*, abs/2103.14030.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

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 *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2641–2649. IEEE Computer Society.

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, Berlin, Germany. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 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 *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 2720–2728. AAAI Press.

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. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6153–6166, Online. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Zhuosheng Zhang, Kehai Chen, Rui Wang, Masao Utiyama, Eiichiro Sumita, Zuchao Li, and Hai Zhao. 2020. Neural machine translation with universal visual representation. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Yuting Zhao, Mamoru Komachi, Tomoyuki Kajiwara, and Chenhui Chu. 2020. Double attention-based multimodal neural machine translation with semantic image regions. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 105–114, Lisboa, Portugal. European Association for Machine Translation.