国际人工智能会议
AAAI 2021 论文北京预讲会

Efficient Object-Level Visual Context Modeling for Multimodal Machine Translation

王德鑫　　熊德意
天津大学智能与计算学部
dyxiong@tju.edu.cn
https://tjunlp-lab.github.io

1

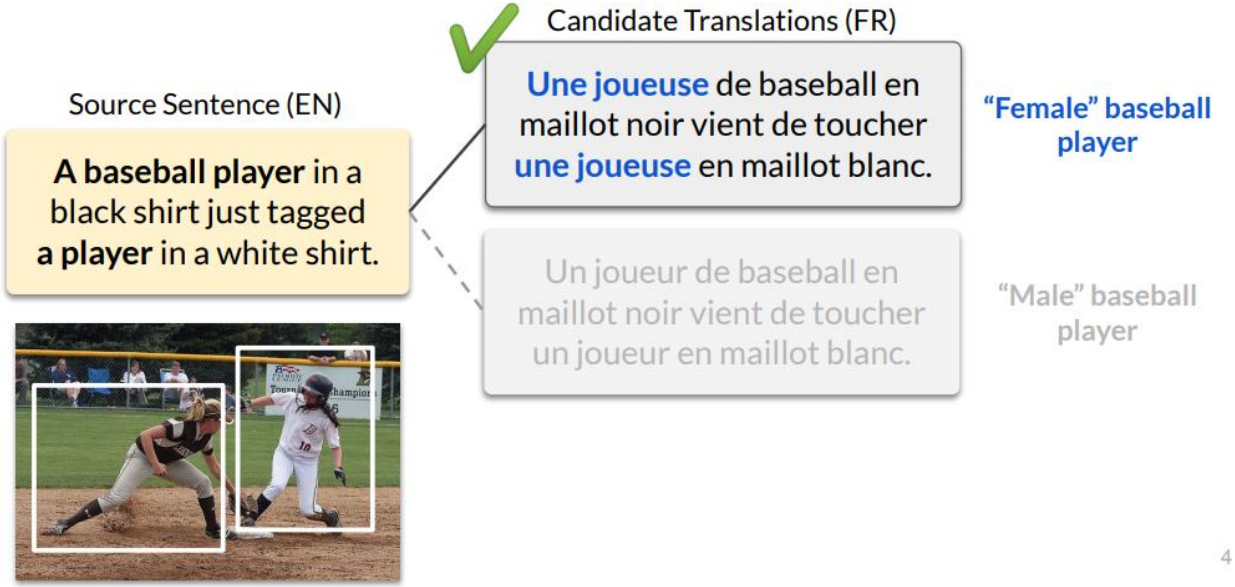
Motivation

Image-guided MMT

结合视觉信息的多模态神经机器翻译（MMT）

给定一张图片和一句相关的源语言文字，要求模型翻译该文字描述。

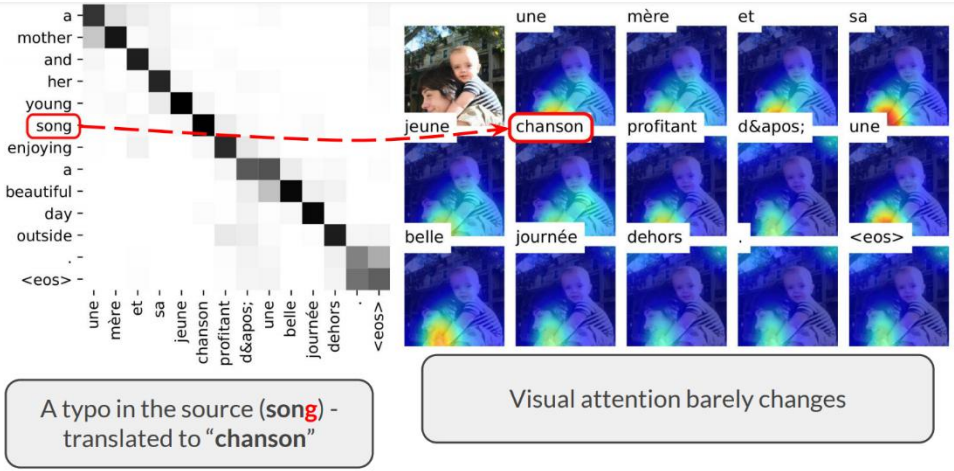
Grounding!



Is Visual Context Benefit to MMT ?

EN-DE	flickr16	flickr17	mscoco17
SUBS3M$_{LM}$ detectron	45.09	40.81	36.94
+ensemble-of-3	45.52	**41.84**	**37.49**
−visual features	**45.59**	41.75	37.43
−MS-COCO	45.11	40.52	36.47
−multi-lingual	44.95	40.09	35.28
SUBS6M$_{LM}$ detectron	45.50	41.01	36.81
SUBS3M$_{LM}$ gn2048	45.38	40.07	36.82
SUBS3M$_{LM}$ text-only	44.87	41.27	36.59
+multi-modal finetune	44.56	41.61	36.93

Results from participants of WMT 2017.



Probing the Need for Visual Context in Multimodal Machine Translation, NAACL 2019

如何更合理地**利用视觉信息**ground**文本信息**，是个棘手的问题

Construct Training Schemes Based on Visual Distilling

是否可以构造一个诱导模型ground到视觉模态的训练环境？

为此，我们总结并思考了与Visual context紧密相连的三个重要元素：

1. 源语言文本解码过程
2. 视觉信息在翻译过程是否存在冗余特征
3. 目标语言生成过程

Source-degradation Texts Settings

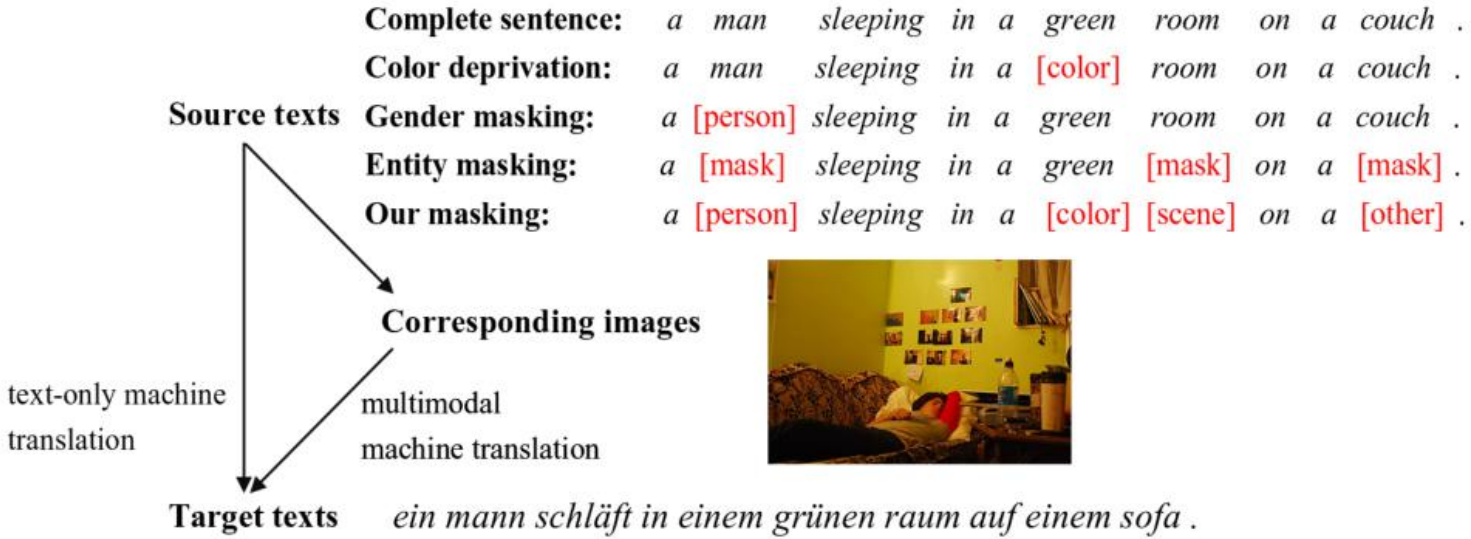
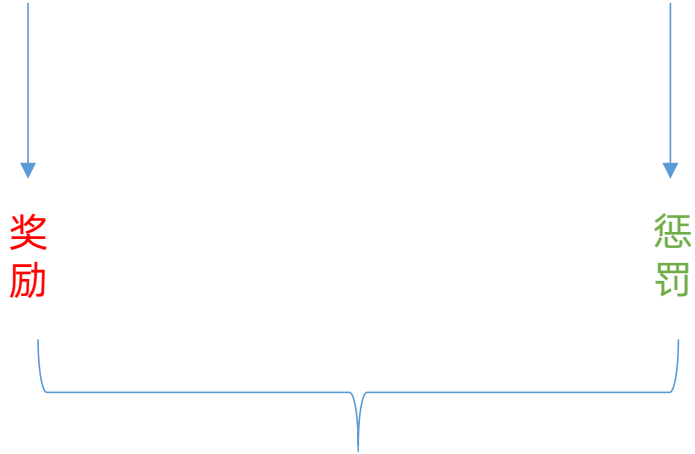


Figure 1: Word masking in multimodal machine translation.

Contribution1 Object-masking Loss for Grounding

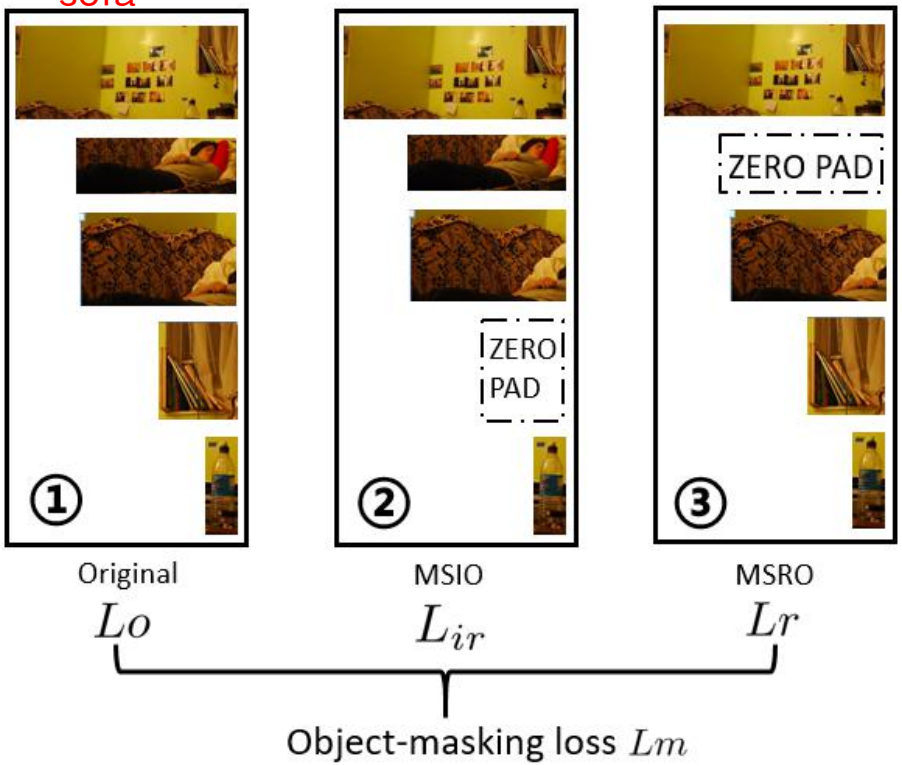
我们的目标是迫使MMT模型关注和翻译相关的视觉内容，忽略和翻译无关的视觉成分所产生的影响。



“训练方式+损失函数”的角度出发：
针对输入是否masking object设置不同程度的loss

Contribution1 Object-masking Loss for Grounding

a man sleeping in a green room on a sofa



$$L_r = L(O^m_{\emptyset_i}, W^n_s) \text{ if } d_i = 1$$

$$L_{ir} = L(O^m_{\emptyset_i}, W^n_s) \text{ if } d_i = 0$$

$$L_m = -(L_r - L_o) + (L_{ir} - L_o)^2$$

关注和翻译相关的视觉内容

降低与翻译无关的视觉成分所产生的影响

Contribution1 Object-masking Loss for Grounding

M 给定图像

Detector

O 图像中检测到的object所属类别

‘person’
‘tree’ ‘road’ ‘motorcycle’ ‘park’
‘helmet’ ‘bicycle’ ‘banner’ ‘fence’
'trees'

S 给定待翻译源文本

a group of men riding their bikes in a rac

Pretrained LM模型

S 的单词集合

‘a’ ‘group’ ‘of’ ‘men’ ‘riding’
‘their’ ‘bikes’ ‘in’ ‘a’ ‘race’

计算O-to-S的相似度矩阵

O-to-S的极大相似度

person : 1. tree : 0.79 road : 0.62 motorcycle : 0.48 fence : 0.34
park : 0.47 helmet: 0.45 bicycle : 0.4 banner : 0.35
trees : 0.33

Contribution2 Vision-weighted Loss for Grounding

相比于NMT，MMT任务的重点在于对视觉信息的准确翻译。基于这个假设，我们进一步强调了对vision words的翻译，强迫模型从图像中抽取更多信息进行grounding。

具体而言，我们借助预训练语言模型，计算target tokens和vision words in source tokens的相似度，为视觉相关的单词设置了更大的损失权重，设计了显式的视觉加权损失函数，称为vision-weighted loss。

$$\mathrm{TVS}_j = \max S'_{j,1:n}, \quad j = 1, 2, ..., r$$

→ TVS 表示目标句单词与源句视觉词汇的相关度

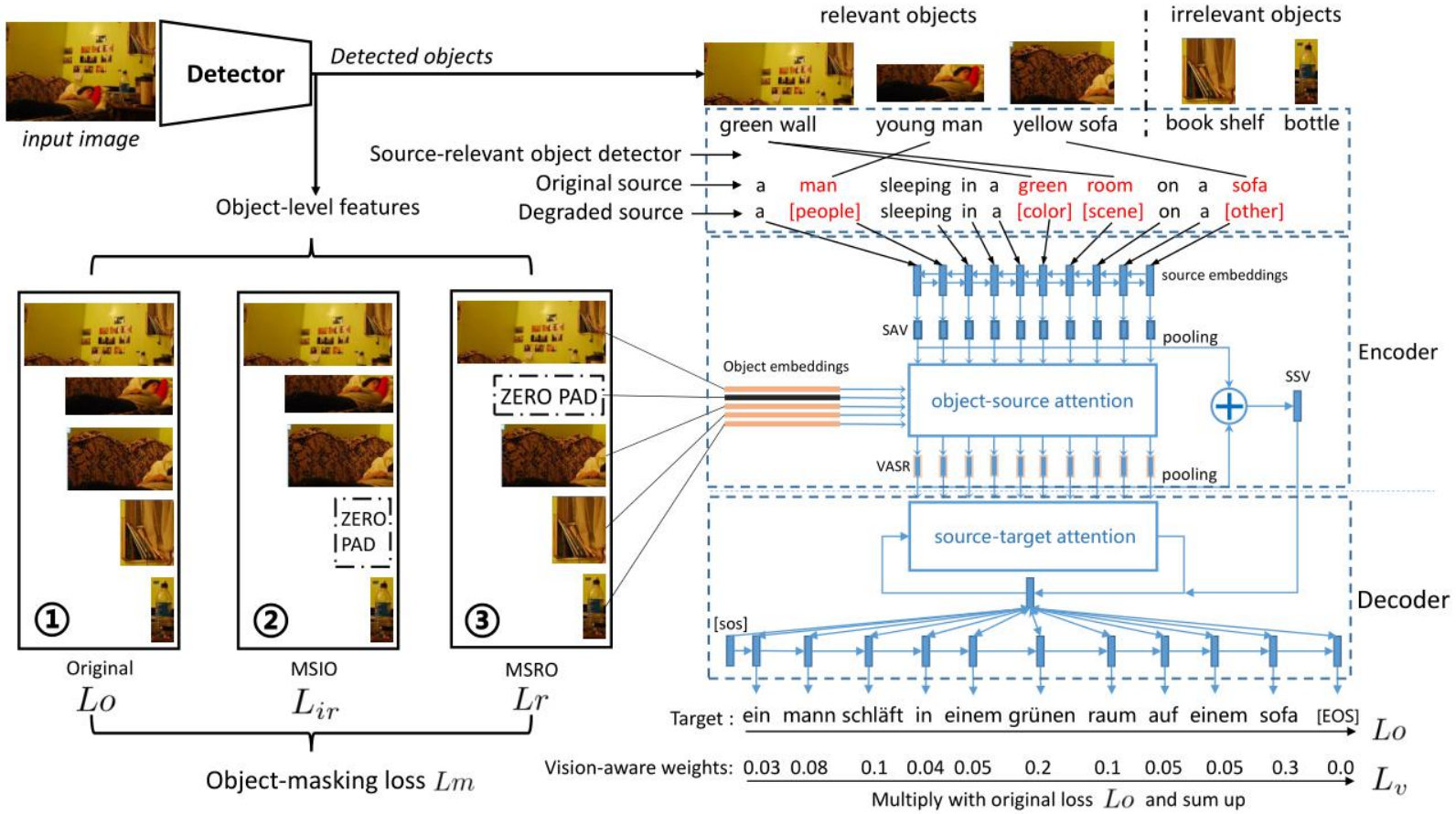
$$q_j = \frac{\mathrm{TVS}_j / f_j}{\sum_{a=1}^{r} \mathrm{TVS}_a / f_a}, \quad j = 1, 2, ..., r$$

→ 根据目标语言词频 $f_{\cdot}$ ，对TVS 进行加权归一化

$$L_v = \sum_{j=1}^{r} q_j * Lo_j$$

→ L_v 是Vision-weighted loss

Contribution3 Object-level Visual Context Modeling Framework



Contribution3 Object-level Visual Context Modeling Framework

Original loss

$$Lo$$

Object-masking loss

$$L_r = L(O^m_{\varnothing_i}, W^n_s) \text{ if } d_i = 1$$

$$L_{ir} = L(O^m_{\varnothing_i}, W^n_s) \text{ if } d_i = 0$$

$$L_m = -(L_r - L_o) + (L_{ir} - L_o)^2$$

Vision-weighted loss

$$\text{TVS}_j = \max S'_{j,1:n}, \quad j = 1, 2, ..., r$$

$$q_j = \frac{\text{TVS}_j / f_j}{\sum_{a=1}^{r} \text{TVS}_a / f_a}, \quad j = 1, 2, ..., r$$

$$L_v = \sum_{j=1}^{r} q_j * Lo_j$$

$$L_{ovc} = Lo + \alpha * L_m + \beta * L_v$$

2

Experiments

Experiment Settings

为了测试我们模型的grounding能力，采用两种实验设置，以测试OVC的翻译效果：

1. standard实验

   给定原始的源语言文本+原始图像，预测给定的翻译结果

2. source-mask实验

   给定被局部遮盖的源语言文本+原始图像，预测给定的翻译结果

Standard Experiments

Models	WMT17 MMT test set				Ambiguous COCO			
	En⇒Fr		En⇒De		En⇒Fr		En⇒De	
	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR
Existing MMT Models								
(T) Transformer‡	52.0	68.0	30.6	50.4	-	-	27.3	46.2
(R) Imagination_i	-	-	30.2	51.2	-	-	26.4	45.8
(R) VAG-NMT_i‡	53.5±0.7	70.0±0.7	31.6±0.5	52.2±0.3	44.6±0.6	64.2±0.5	27.9±0.6	47.8±0.6
(R) VAG-NMT_i	53.8±0.3	70.3±0.5	31.6±0.3	52.2±0.3	45.0±0.4	**64.7±0.4**	28.3±0.6	48.0±0.5
(R) VAR-MMT_o	52.6	69.9	29.3	51.2	-	-	-	
(T) VAR-MMT_o	53.3	70.4	29.5	50.3	-	-	-	
(R) LIUMCVC_i	52.7±0.9	69.5±0.7	30.7±1.0	52.2±0.4	43.5±1.2	63.2±0.9	26.4±0.9	47.4±0.3
(R) VMMT_i	-	-	30.1±0.3	49.9±0.3	-	-	25.5±0.5	44.8±0.2
(T) GMMT_o	53.9	69.3	32.2	51.9	-	-	28.7	47.6
Our Proposed Models								
OVC	53.5±0.2	70.2±0.3	31.7±0.3	51.9±0.4	44.7±0.6	64.1±0.3	28.5±0.5	47.8±0.3
OVC+L_m	54.1±0.7	**70.5±0.5**	32.3±0.6	**52.4±0.3**	**45.3±0.5**	64.6±0.5	**28.9±0.5**	**48.1±0.5**
OVC+L_v	**54.2±0.4**	70.5±0.5	**32.4±0.4**	52.3±0.5	45.2±0.4	64.6±0.3	28.6±0.5	48.0±0.6
OVC+L_m+L_v	54.0±0.4	70.4±0.4	32.4±0.6	52.2±0.3	45.1±0.6	64.5±0.5	28.8±0.4	48.0±0.4

Standard Experiments

Models	WMT17 MMT test set				Ambiguous COCO			
	En⇒Fr		En⇒De		En⇒Fr		En⇒De	
	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR
Existing MMT Models								
(T) Transformer‡	52.0	68.0	30.6	50.4	-	-	27.3	46.2
(R) Imagination_i	-	-	30.2	51.2	-	-	26.4	45.8
(R) VAG-NMT_i‡	53.5±0.7	70.0±0.7	31.6±0.5	52.2±0.3	44.6±0.6	64.2±0.5	27.9±0.6	47.8±0.6
(R) VAG-NMT_i	53.8±0.3	70.3±0.5	31.6±0.3	52.2±0.3	45.0±0.4	**64.7±0.4**	28.3±0.6	48.0±0.5
(R) VAR-MMT_o	52.6	69.9	29.3	51.2	-	-	-	
(T) VAR-MMT_o	53.3	70.4	29.5	50.3	-	-	-	
(R) LIUMCVC_i	52.7±0.9	69.5±0.7	30.7±1.0	52.2±0.4	43.5±1.2	63.2±0.9	26.4±0.9	47.4±0.3
(R) VMMT_i	-	-	30.1±0.3	49.9±0.3	-	-	25.5±0.5	44.8±0.2
(T) GMMT_o	53.9	69.3	32.2	51.9	-	-	28.7	47.6
Our Proposed Models								
OVC	53.5±0.2	70.2±0.3	31.7±0.3	51.9±0.4	44.7±0.6	64.1±0.3	28.5±0.5	47.8±0.3
OVC+L_m	54.1±0.7	**70.5±0.5**	32.3±0.6	**52.4±0.3**	**45.3±0.5**	64.6±0.5	**28.9±0.5**	**48.1±0.5**
OVC+L_v	**54.2±0.4**	70.5±0.5	**32.4±0.4**	52.3±0.5	45.2±0.4	64.6±0.3	28.6±0.5	48.0±0.6
OVC+L_m+L_v	54.0±0.4	70.4±0.4	32.4±0.6	52.2±0.3	45.1±0.6	64.5±0.5	28.8±0.4	48.0±0.4

Standard Experiments

Models	WMT17 MMT test set				Ambiguous COCO			
	En⇒Fr		En⇒De		En⇒Fr		En⇒De	
	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR
Existing MMT Models								
(*T*) Transformer‡	52.0	68.0	30.6	50.4	-	-	27.3	46.2
(*R*) Imagination_i	-	-	30.2	51.2	-	-	26.4	45.8
(*R*) VAG-NMT_i‡	53.5±0.7	70.0±0.7	31.6±0.5	52.2±0.3	44.6±0.6	64.2±0.5	27.9±0.6	47.8±0.6
(*R*) VAG-NMT_i	53.8±0.3	70.3±0.5	31.6±0.3	52.2±0.3	45.0±0.4	**64.7±0.4**	28.3±0.6	48.0±0.5
(*R*) VAR-MMT_o	52.6	69.9	29.3	51.2	-	-	-	
(*T*) VAR-MMT_o	53.3	70.4	29.5	50.3	-	-	-	
(*R*) LIUMCVC_i	52.7±0.9	69.5±0.7	30.7±1.0	52.2±0.4	43.5±1.2	63.2±0.9	26.4±0.9	47.4±0.3
(*R*) VMMT_i	-	-	30.1±0.3	49.9±0.3	-	-	25.5±0.5	44.8±0.2
(*T*) GMMT_o	53.9	69.3	32.2	51.9	-	-	28.7	47.6
Our Proposed Models								
OVC	53.5±0.2	70.2±0.3	31.7±0.3	51.9±0.4	44.7±0.6	64.1±0.3	28.5±0.5	47.8±0.3
OVC+L_m	54.1±0.7	**70.5±0.5**	32.3±0.6	**52.4±0.3**	**45.3±0.5**	64.6±0.5	**28.9±0.5**	**48.1±0.5**
OVC+L_v	**54.2±0.4**	70.5±0.5	**32.4±0.4**	52.3±0.5	45.2±0.4	64.6±0.3	28.6±0.5	48.0±0.6
OVC+L_m+L_v	54.0±0.4	70.4±0.4	32.4±0.6	52.2±0.3	45.1±0.6	64.5±0.5	28.8±0.4	48.0±0.4

Source-mask Experiments

En⇒De		
Metrics	BLEU	METEOR
OVC_t	21.02	40.61
OVC_i	22.02	41.91
OVC_o	21.98	41.57
OVC_o+HM	25.31	43.85
OVC_o+L_m	26.30	45.37
OVC_o+L_v	22.18	42.01
OVC_o+L_m+L_v	22.57	42.24
En⇒Fr		
OVC_t	37.01	55.35
OVC_i	37.40	55.68
OVC_o	36.94	54.92
OVC_o+HM	37.39	55.38
OVC_o+L_m	39.31	57.28
OVC_o+L_v	37.25	55.79
OVC_o+L_m+L_v	37.63	56.14

Table 2: Results for the source-degradation setting on the WMT17 MMT development set. _t denotes text-only models. HM denotes a hard masking scheme where irrelevant objects are masked in a hard way via the pretrained threshold.

实验结果显示，不使用我们所提出的损失函数时，OVC在建模中加入视觉模态信息所能带来的性能提升较为有限（见OVC_i、OVC_o所在行）。

而加入了我们所提出的损失函数，在En=>De和En=>Fr两种语言对上，翻译性能都有了0.8~3.8个BLEU值得提升，说明了我们的训练方式和损失函数有利地促进了OVC借助视觉信息来补充缺失的文本信息。

Source-mask Experiments

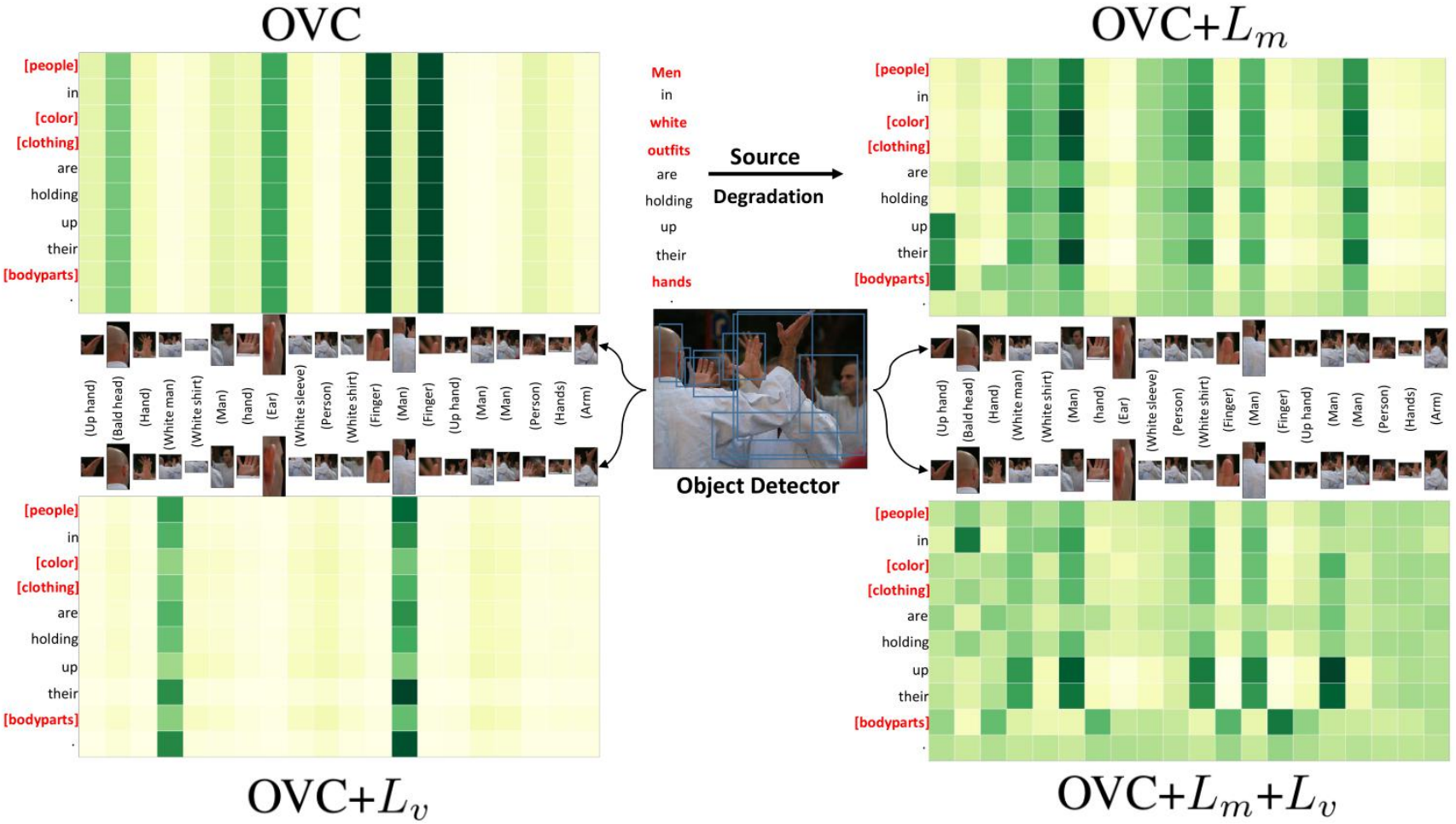
En⇒De		
Metrics	BLEU	METEOR
OVC_t	21.02	40.61
OVC_i	22.02	41.91
OVC_o	21.98	41.57
OVC_o+HM	25.31	43.85
OVC_o+L_m	26.30	45.37
OVC_o+L_v	22.18	42.01
OVC_o+L_m+L_v	22.57	42.24
En⇒Fr		
OVC_t	37.01	55.35
OVC_i	37.40	55.68
OVC_o	36.94	54.92
OVC_o+HM	37.39	55.38
OVC_o+L_m	39.31	57.28
OVC_o+L_v	37.25	55.79
OVC_o+L_m+L_v	37.63	56.14

Table 2: Results for the source-degradation setting on the WMT17 MMT development set. _t denotes text-only models. HM denotes a hard masking scheme where irrelevant objects are masked in a hard way via the pretrained threshold.

实验结果显示，不使用我们所提出的损失函数时，OVC在建模中加入视觉模态信息所能带来的性能提升较为有限（见OVC_i、OVC_o所在行）。

而加入了我们所提出的损失函数，在En=>De和En=>Fr两种语言对上，翻译性能都有了0.8~3.8个BLEU值得提升，说明了我们的训练方式和损失函数有利地促进了OVC借助视觉信息来补充缺失的文本信息。

Attention Heatmaps in Source-mask Experiments



source-mask case analysis

Images	Descriptions
	SRC: a little girl peering over a blue wall . DSRC: a little [people] peering over a [color] wall . REF: ein kleines mädchen späht über eine blaue mauer . OVC: ein kleiner junge blickt über eine grüne wand . (a little boy looks over a green wall .) OVC+L_m: ein kleiner junge guckt über eine weiße wand . (a little boy looks over a white wall .) OVC+L_v: ein kleiner **mädchen** guckt über eine weiße wand . (a little girl looks over a white wall .) OVC+L_m+L_v: ein kleines **mädchen** guckt über eine **blaue** wand . (a little girl looks over a blue wall .)
	SRC: a group of men in costume play music . DSRC: a group of [people] in [clothing] play music . REF: eine gruppe von männern in kostümen spielt musik . OVC: eine gruppe von kindern in **kost**ümen spielt musik . (a group of children in costumes play music .) OVC+L_m: eine gruppe von **männern** in uniform spielt musik . (a group of men in uniform plays music .) OVC+L_v: eine gruppe von **männern** in anzügen macht musik . (a group of men in suits makes music .) OVC+L_m+L_v: eine gruppe von **männern** in **kost**ümen spielt musik . (a group of men in costumes is playing music .)
	SRC: a group of children play in the water under a bridge . DSRC: a group of [people] play in the [scene] under a [scene] . REF: eine gruppe von kindern spielt im wasser unter einer brücke . OVC: eine gruppe von **kindern** spielt im gras unter einem berg . (a group of children play in the grass under a mountain .) OVC+L_m: eine gruppe **kinder** spielt im **wasser** unter einem wasserfall . (a group of children play in the water under a waterfall .) OVC+L_v: eine gruppe **kinder** spielt im **wasser** unter einem wasserfall . (a group of children play in the water under a waterfall .) OVC+L_m+L_v: eine gruppe von **kindern** spielt im schnee unter einem **brücke** . (a group of children play in the snow under a bridge .)

Discussion

一直以来，MMT的标注成本高昂，数据规模相对较小，仅使用少量的数据难以学习到细粒度的跨模态grounding。而视觉与文本的联系不仅仅是粗粒度的，也常常伴随细粒度的复杂关系。

如何利用少量的数据，挖掘两种模态的复杂联系是一个比较有趣的问题。这个工作是针对数据量小的情况下实现细粒度地跨模态对齐的训练方法。在一定意义上，我们的工作是从建模角度出发，将语言先验转化为跨模态信息，来更充分地利用已有数据的潜力，进行隐式的数据增强。

在未来的工作中，我们可以将这个训练方式泛化到VisLang的其他任务中，比如video-guided MT。

国际人工智能会议
AAAI 2021 论文北京预讲会

TJUNLP: https://tjunlp-lab.github.io

THANKS

2020.12.19