

小米AI实验室关于端到端图片翻译的研究工作入选EMNLP 2023

近日，EMNLP 2023公布录用结果，小米AI实验室联合北京理工大学郭宇航老师团队关于端到端图片翻译的研究工作被录用为Findings长文(Findings Long Paper)。EMNLP全称为Conference on Empirical Methods in Natural Language Processing，由国际计算语言学会(ACL)主办，每年举行一次，为自然语言处理和人工智能领域的最有影响力的国际会议之一，今年会议将于2023年12月6日在新加坡举行。



题目：In-Image Neural Machine Translation with Segmented Pixel Sequence-to-Sequence Model

作者：田炎智，李响，柳泽明，郭宇航*，王斌

1. 研究背景

图片翻译(In-Image Machine Translation)是一类新型的机器翻译任务，其目标是将带有源语言文本的图片翻译为带有目标语言文本的图片。这种任务有广泛的应用场景，例如翻译软件中的"拍照翻译"功能，在用户输入带有文本的图片后，软件会给出带有对应翻译的图片，能让用户更好地理解图片中文本的含义。

目前图片翻译常用的方案是级联方法，包括识别、翻译和文本嵌入三个流程。在识别过程中，调用光学字符识别模型(optical character recognition, OCR)对图片中文本进行识别；在翻译过程中，调用神经机器翻译模型(neural machine translation, NMT)对OCR识别结果进行翻译；在文本嵌入过程中，使用图像修补算法擦除原图文本区域，再通过一定特征(如字体大小、文本长度、图片背景等)将文本嵌入至图片，保证较好的视觉效果。

然而，级联方法存在的一大问题是错误传递，图1左侧展示了一个错误传递的例子。OCR模型将原图中的"STADT"错误识别为了"STAOT"，导致后续模型的表现变差，无法得到准确的翻译结果。

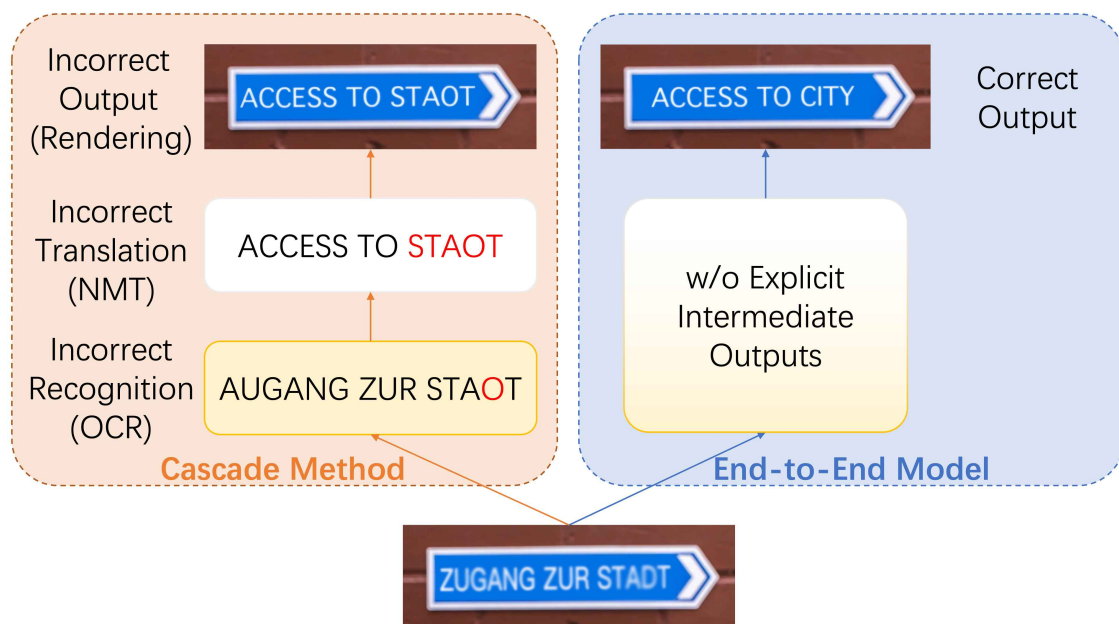


图1: 图片翻译的级联方法和端到端模型

一种缓解错误传递的方法是设计端到端模型，然而，端到端的图片翻译仍处于初步研究阶段。现有的端到端模型[1]使用构建的一行白底黑字的图片进行训练，翻译质量较低。因此，目前端到端的机器翻译存在两个难点：

- 目前并没有公开的图片翻译数据集，需要利用现有资源进行构建。
- 目前的端到端图片翻译模型的翻译质量差，需要设计更合理的模型对图片进行建模，从而得到更好的翻译质量。

2. 方法介绍

2.1 数据集构建

对于图片翻译任务，目前并没有开源的数据集，本文使用现有的平行语料文本，构造图片数据。具体来说，平行语料中的文本对<x,y>会被分别嵌入到白色背景的图片中，如图2所示。

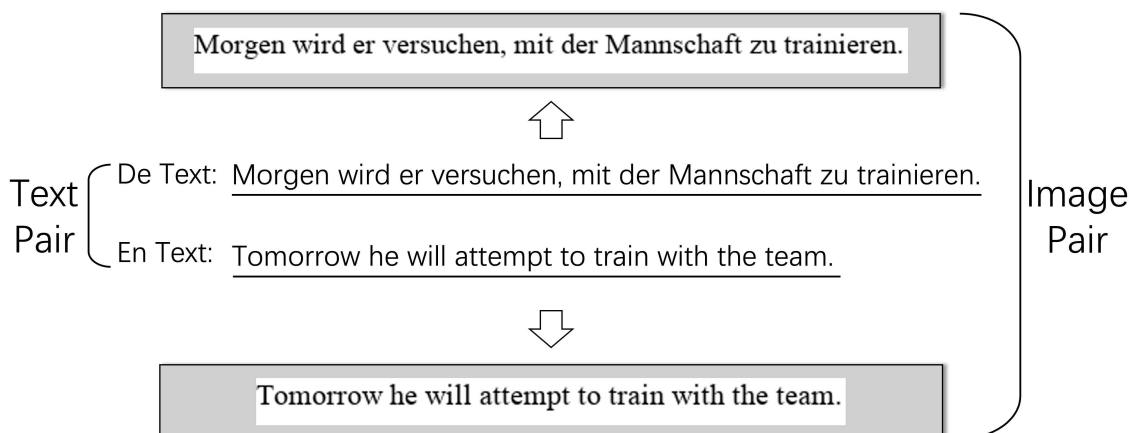


图2: 使用平行语料构建的图片翻译数据集

2.2 端到端模型设计

从数据模态层面来看，端到端图片翻译任务可以看作图像到图像的生成任务。然而，现有的图像生成模型很难生成带有清晰文本的图片[2, 3]。对于图片翻译任务，最重要的是在保证图片背景变化不大的情况下尽可能生成清晰的文本。因此，本文并没有采用常用的图像生成模型，而是将带有文本的图片视为像素序列，从而将图片到图片的转换任务转变为序列到序列的转换任务，在模型架构上可以参考常用的序列-序列模型。

由于图像的表达一般为 $x \in \mathbb{R}^{H \times W \times C}$ ，而序列的表达一般为 $x_{tok} \in Char^T \times 1$ ，两种数据形式的差别较大。本文设计了一种可以实现图像和序列相互映射转换的方法，并使用序列模型对其进行建模，如图3所示。

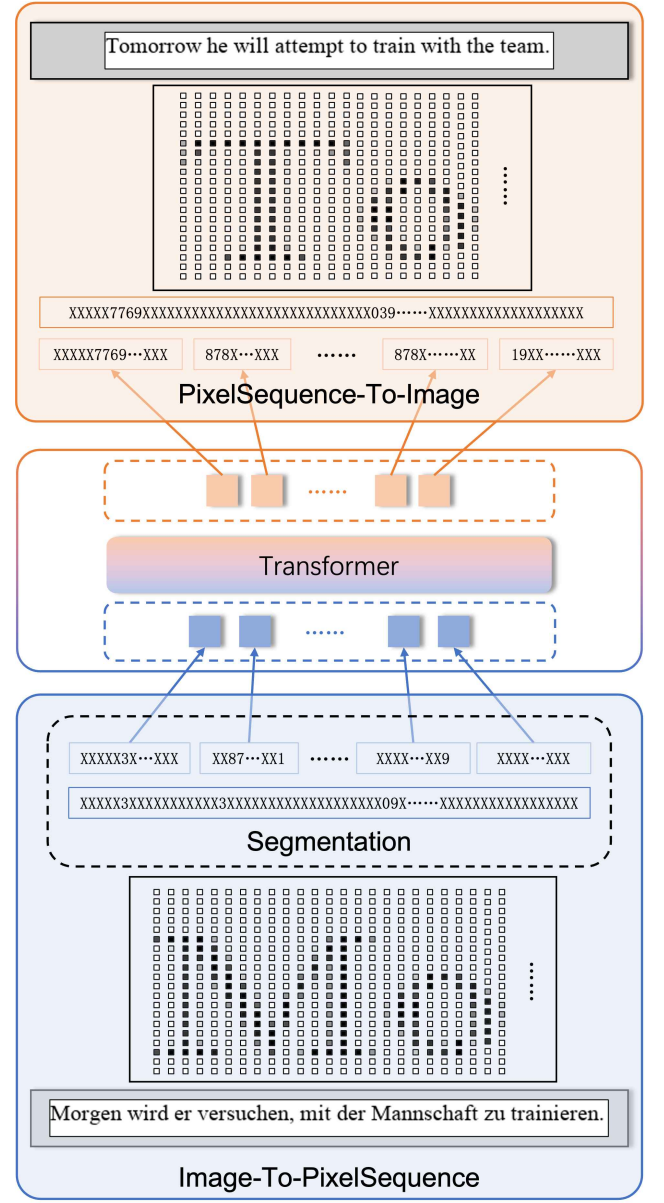


图3: 模型整体架构

在对输入图片和输出图片进行序列化表示后，端到端图片翻译的优化目标可以用下式表达：

$$P(y_{tok} | x_{tok}, \theta) = \prod_{i=1}^{|y_{tok}|} p(y_{tok}^i | x_{tok}, y_{tok}^{<i}, \theta)$$

图像转换为像素序列(Image-To-PixelSequence, ITS)

首先，需要将原始的RGB图像 $x \in \mathbb{R}^{H \times W \times C}$ 转换为对应的灰度图 $x_g \in \mathbb{R}^{H \times W \times 1}$ ，灰度图中的灰度值为0到1之间的浮点数。之后将0到1的区间进行十等分，每个区间对应一个特定的符号，一个符号对应原图中的一个像素，得到符号图。具体方式如下，v是灰度图中的灰度值，c是对应的符号，char()为定义的十等分区间和符号之间的映射关系：

$$c = char(\lfloor \frac{v}{0.1} \rfloor)$$

其次，将原图转换得到的符号图按照从上至下、从左至右的顺序重新排列为二维序列。在得到序列后，可以使用序列模型对其进行建模，本文使用的是Transformer[4]。

像素序列转换为图像(PixelSequence-To-Image, STI)

序列模型的输出也是同样形式的序列，如果要得到输出图像，需要将输出序列进行处理，使其复原为图像。首先需要将一维序列重新排列，得到形状为 $H \times W$ 的符号图。之后将符号图中的每个符号恢复为0到1之间的浮点数。具体方式如下， v 是灰度图中的灰度值， c 是对应的符号， $int()$ 为定义的符号和1到10整数之间的映射关系：

$$v = int(c) \times 0.1$$

复原得到灰度图后，对其中的每个灰度值 v ，按照 $(v \times 255, v \times 255, v \times 255)$ 的方式将其转换为RGB值，得到复原的图像。

像素序列分段(PixelSequence Segmentation)

在使用上述方法将图像转换为序列后，序列的长度为 $H \times W$ ，如果将序列中每个符号作为一个 token(序列处理中的最小单元)，会导致序列长度过长。又因为Transformer的计算复杂度是 $O(N^2)$ ，过长的序列长度会严重影响计算效率，因此本文对得到的像素序列进行分段，得到粒度更大的token，从而减小序列长度。

像素序列分段是一种迭代式算法，首先需要在训练数据集上学习分段的方式。在每次迭代中，会将出现频率最高的二元组进行合并，形成一个新的token。按照这种方式多次迭代后，序列的长度会明显减小。

3. 实验

由于图片翻译的输出是图片，为了能够利用现有的指标对该任务进行评估，首先需要使用OCR模型对图片中的文本进行识别，然后对得到的文本使用BLEU[5]，COMET[6]两个指标评测翻译质量。

本文分别在领域内和领域外的数据集上对比了提出的端到端方法和现有级联方法的翻译质量，结果如表1所示。

Systems	In-Domain				Out-Domain			
	newstest-2013		newstest-2014		tst-COMMON		Himl	
	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET
Cascade	27.1	78.3	27.3	75.8	30.1	79.7	34.3	80.7
Our E2E	28.1	81.9	28.2	80.4	30.7	83.1	36.0	84.3

表1: 本文提出的端到端模型和现有级联方法的翻译质量

此外，本文还和现有的端到端模型进行了对比，结果如表2所示。

Systems	BLEU
Conv Baseline (Mansimov et al., 2020)	0.5
AttnConv (Mansimov et al., 2020)	7.7
Our E2E	28.1

表2: 本文提出的端到端模型和现有端到端模型的翻译质量

根据实验结果可以看出，本文提出的端到端图片翻译模型的翻译质量高于现有级联方法和端到端模型。

4. 分析

本文从5个方面进行分析，分别是：

(RQ1) 不同迭代次数的像素序列分段是否会影响翻译质量？

(RQ2) 图片中有可能会有不完整的文本，这种图片会影响级联模型的效果，本文提出的端到端模型在这种情况下是否有更好的表现？

(RQ3) 图片中不同位置的文本是否会影响翻译质量？

(RQ4) 图片中不同字号和字体的文本是否会影响翻译质量？

(RQ5) 级联错误传递是怎样影响级联方法的翻译质量的？

RQ1 本文使用不同迭代次数的像素序列分段对序列进行处理，训练了不同模型，结果如图4所示。

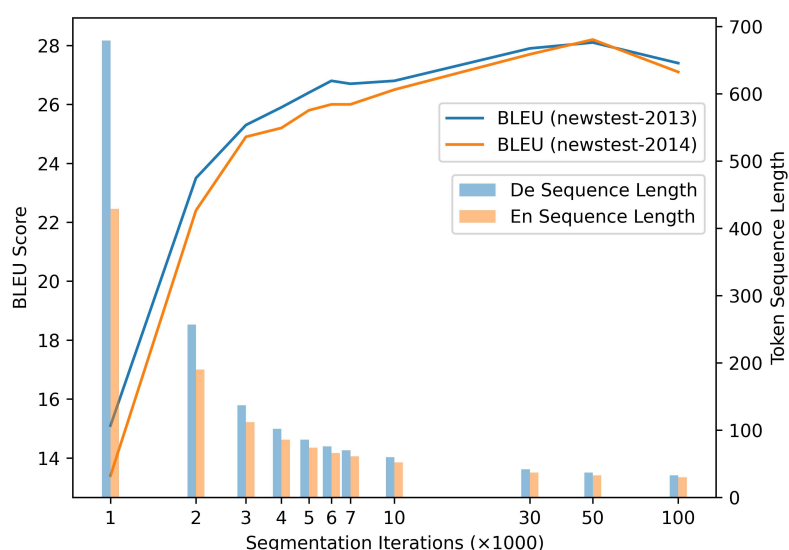


图4: 不同迭代次数的像素序列分段方法的翻译质量

根据结果可以看出，随着迭代次数的增加，序列长度减小，翻译质量增加，并且在迭代50000次时翻译质量最好。

RQ2 本文在原始数据集的基础上，对其进行遮掩，构造了带有不完整文本的图片测试集，如图5所示。

Das Problem werde an diesem Abend behoben.

Das Problem werde an diesem Abend behob

图5: 带有不完整文本的图片示例

在这类数据集上的实验结果如表3的RQ2所示。

Types	Systems	newstest-2013		newstest-2014		tst-COMON		Himl	
		BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET
Origin	Cascade	27.1	78.3	27.3	75.8	30.1	79.7	34.3	80.7
	Our E2E	28.1	81.9	28.2	80.4	30.7	83.1	36.0	84.3
RQ2	Cascade	24.2	71.2	24.5	69.1	26.4	71.7	30.5	72.4
	Our E2E	25.0	72.3	24.7	71.6	26.9	72.1	31.1	73.6
RQ3	Cascade	27.1	78.3	27.3	75.8	30.1	79.7	34.3	80.7
	Our E2E	27.8	81.5	27.8	80.0	30.5	82.7	35.5	84.1

表3: 不同形式数据集上的分析实验结果

根据实验结果可以看出，对于带有不完整文本的图片测试集，本文提出的端到端模型有更好的翻译质量，其原因是端到端模型并没有对图片中文本进行显式建模，因此受不完整的文本影响更小。

RQ3 由于级联模型包含OCR模块，OCR模块一般会先对图片中的文本位置进行检测，再对检测得到的文字部分进行识别，因此文本在图片中的位置一般不会对识别结果造成影响。然而本文提出的端到端模型并没有对文本位置进行检测，因此构造了带有偏移文本的图片数据集，如图6所示。

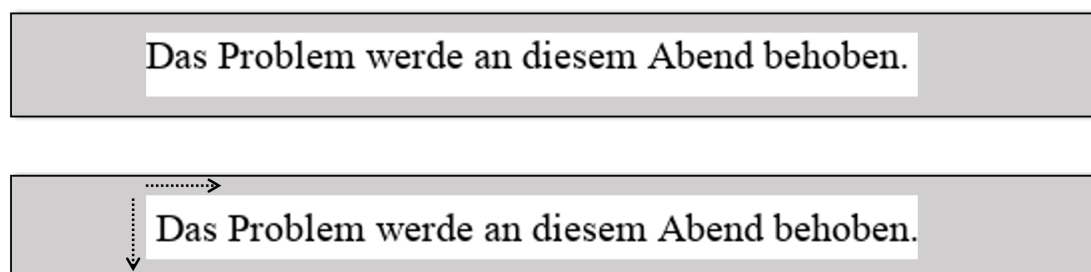


图6: 带有偏移文本的图片示例

在这类数据集上的实验结果如表3的RQ3所示。根据实验结果可以看出，对于存在偏移文本的数据集，端到端模型的翻译质量有所下降，但依然优于级联模型。这主要得益于使用像素序列的建模方式，图片中的文本存在偏移，等价于在像素序列中添加或移除部分前缀，不会影响后续的序列。

RQ4 本文在原有数据集的基础上，还使用不同字号、不同字体构造了额外的数据集，如图7所示。

Original Dataset: Times New Roman; Font Size 20

Das Problem werde an diesem Abend behoben.

TNR15: Times New Roman; Font Size **15**

Das Problem werde an diesem Abend behoben.

TNR25: Times New Roman; Font Size **25**

Das Problem werde an diesem Abend behoben.

Arial20: **Arial**; Font Size 20

Das Problem werde an diesem Abend behoben.

图7: 使用不同字号和字体构建的数据集

使用这些数据集训练模型并进行测试，结果如表4所示。从结果可以看出TNR15的实验结果明显偏低，主要原因是用于评测的OCR模型对于较小字号的文本识别错误率较高，导致评测结果较差。具体来说，对于原始测试集，WER分别为1.4%和1.6%，但是对于TNR15，WER分别为4.2%和4.0%。因此，较差的识别结果导致了评测质量较差。从TNR25和Arial20的测试结果可以看出，模型的翻译质量基本没有受到影响，说明对于不同字号或字体的数据集，模型的效果没有受到影响。

Datasets	newstest-2013	newstest-2014
TNR15	22.3	22.6
TNR25	27.5	27.7
Arial20	27.9	28.1

表4: 在不同数据集上的实验结果

RQ5 本文还对级联模型的错误传递进行了分析，在级联模型中，OCR模型输出结果中的错误会影响NMT模型的性能。本文首先使用图片中的文本，模拟不存在OCR错误的理想情况(Golden NMT)进行测试，结果如表5所示。

Systems	newstest-2013		newstest-2014	
	BLEU	COMET	BLEU	COMET
Golden NMT	29.8	83.0	29.7	81.5
Cascade	27.1 (-2.7)	78.3 (-4.7)	27.3 (-2.4)	75.8 (-5.7)
	tst-COMMON		Himl	
	BLEU	COMET	BLEU	COMET
Golden NMT	33.0	84.0	37.8	84.7
Cascade	30.1 (-2.9)	79.7 (-4.3)	34.3 (-3.5)	80.7 (-4.0)

表5: 级联模型中的错误传递

可以看出，错误传递严重影响了翻译质量。此外，本文还研究了导致NMT性能严重下降的原因，此处以一个例子进行展示，如表6所示。

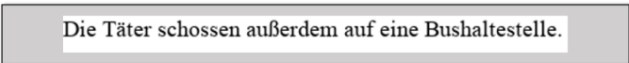

	Cascade Method (w/ OCR error)	Golden (w/o OCR error)
Input Image		
OCR	Die Täter schossen auBerdem auf eine Bushaltestelle.	Die Täter schossen auBerdem auf eine Bushaltestelle.
BPE	Die Täter scho@@ ssen au@@ Ber@@ dem auf eine Bushaltestelle .	Die Täter scho@@ ssen außerdem auf eine Bushaltestelle .
NMT	The perpetrators shot Berdem on a bus stop .	The perpetrators also shot at a bus stop .
Output Image		

表6: 级联模型中错误传递的例子

OCR模型的输出结果可能会存在错误的字符，其中的错误会对NMT模型的预处理(tokenization，如BPE[7]算法)造成影响，进而影响NMT模型的性能。例如OCR模型将"außerdem"错误识别为了"auBerdem"，从而对BPE结果造成了较大影响，因此得到了错误的翻译"Berdem"。

此外，本文还对比了级联方法和端到端模型的输出结果，如表7所示。从结果中可以看出，端到端模型能够缓解级联方法中的错误传递。

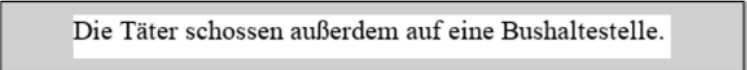
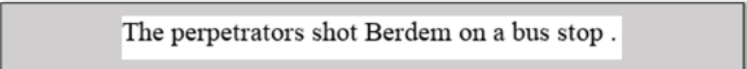
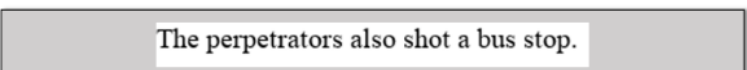
Input Image	
Output Image of Cascade Method	
Output Image of Our End-to-End Model	

表7: 级联模型和端到端模型的对比

5. 总结及未来展望

本文提出了一种端到端的图片翻译模型，在使用平行语料构建的图片数据集上实现了优于现有级联和端到端方案的翻译质量。通过实验分析，本文提出的方法在不同形式(如存在残缺和文本偏移)的测试集上同样有更好的表现。

然而，本文构建的数据集距离现实场景仍有一定距离，在后续的研究中，我们计划构建更复杂的数据集，尝试使用端到端模型达到更好的翻译质量。

主要参考文献

- [1] Mansimov E, Stern M, Chen M, et al. Towards end-to-end in-image neural machine translation[J]. arXiv preprint arXiv:2010.10648, 2020.
- [2] Liu R, Garrette D, Saharia C, et al. Character-aware models improve visual text rendering[J]. arXiv preprint arXiv:2212.10562, 2022.
- [3] Ma J, Zhao M, Chen C, et al. GlyphDraw: Learning to Draw Chinese Characters in Image Synthesis Models Coherently[J]. arXiv preprint arXiv:2303.17870, 2023.
- [4] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.
- [5] Papineni K, Roukos S, Ward T, et al. Bleu: a method for automatic evaluation of machine translation[C]//Proceedings of the 40th annual meeting of the Association for Computational Linguistics. 2002: 311-318.
- [6] Rei R, Stewart C, Farinha A C, et al. COMET: A neural framework for MT evaluation[J]. arXiv preprint arXiv:2009.09025, 2020.
- [7] Sennrich R, Haddow B, Birch A. Neural machine translation of rare words with subword units[J]. arXiv preprint arXiv:1508.07909, 2015.

