多模MT的集合序列级训练：OSU-百度WMT18多模机翻译系统报告

郑仁杰*[1] 杨一林*[1] 马明博†[1,2] 梁黄†[2,1]俄勒冈州立大学EECS学院，俄勒冈州科瓦利斯
[2]百度研究，桑尼维尔，加利福尼亚州
zheng@renj.me zheng@renj 我 {yilinyang721, cosmmb, liang.huang.sh} @ gmail.com

摘要

本文描述了由俄勒冈州立大学和百度研究共同开发的WMT 2018多模态翻译共享任务的多模态机器翻译系统。 在本文中，我们介绍了一种通过将图像特征馈送到解码器侧来合并图像信息的简单方法。 我们还探索了不同的序列级培训方法，包括预定的抽样和强化学习，这些都可以带来实质性我们的系统使用不同的架构和训练方法集合了几个模型，并为三个子任务实现了最佳性能：任务1中的En-De和En-C以及（En + De + Fr）-Cs任务1B。

1 介绍

近年来，神经文本生成因其令人印象深刻的生成精度和广泛的适用性而备受关注。 除了展示机器翻译的引人注目的结果（Sutskever等人。，2014; Bahdanau 等。，2014），通过简单的改编，类似的模型也被证明是成功的总结（拉什等人。，2015; Nallapati等。，2016），图像或视频字幕（Venugopalan等。，2015; 徐等人。，2015）和多模式机器翻译（艾略特等人。，2017; Caglayan等人。，2017; Calixto和刘，2017; Ma等人。，2017），旨在借助相应的图像将标题从一种语言翻译成另一种语言。 然而，传统的神经文本生成模型存在两个主要缺点。 首先，他们通常通过预测给定前一个地面真实词的下一个词来训练。 但是在测试时，模型反复将自己的预测反馈到它中。这种“暴露偏见”（Ranzato等。，2015）导致错误积累在测试时生成。 其次，通过最大化下一个地面实况词的概率来优化模型，这些概率词与期望的不可微分的评估度量不同，比如BLEU。

已经提出了几种方法来解决先前的问题。 Bengio等。 （2015）通过在训练期间以慢慢增加的概率反馈模型自身的预测，提出预定的采样以减轻“暴露偏差”。 此外，强化学习（萨顿 等。，1998）被证明有助于直接优化神经文本生成模型训练中的评估指标。 Ranzato等。 （2015）成功使用REINFORCE算法直接优化多个文本生成任务的评估指标。 Rennie等人。 （2017）; 刘等人。 （2017）使用带基线的REINFORCE实现图像字幕的最新技术，以减少训练方差。

此外，许多现有的工作表明神经文本生成模型可以通过简单地平均不同模型的输出来受益于模型集成（艾略特等人。，2017; Rennie等人。，2017）. 加马什和蒙兹 （2016）声称必须在整体中引入不同的模型。 为此，我们使用各种架构和培训方法集合模型。

本文描述了我们参与WMT 2018多模式任务的情况。 我们提交的系统包括一系列仅考虑文本信息的模型，以及还包括用于初始化解码器的图像信息的多模式模型。 我们使用预定的采样和强化学习来训练这些模型。 通过对这些模型进行整合来解码最终输出。 据我们所知，这是第一个使用序列级学习方法实现最先进技术的多模式机器翻译系统。

*平等贡献
†百度研究期间的贡献

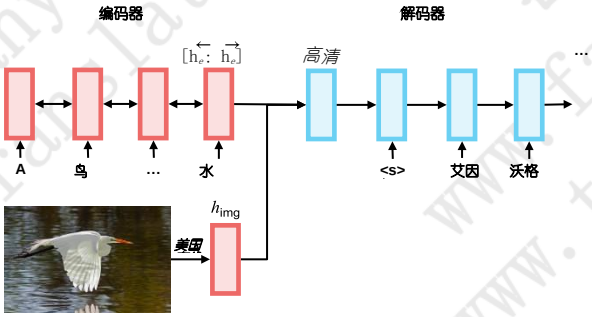

图1：多模式机器翻译模型

2 方法

我们的模型基于序列到序列的RNN架构并引起注意（Bahdanau等。，2014）。 我们将图像特征结合起来初始化解码器的隐藏状态，如图所示 1。 最初，使用隐藏状态初始化此隐藏状态
最后编码器的前向和后向隐藏状态的串联，$\overrightarrow{h_e}$和$\overleftarrow{h_e}$。 我们建议
使用编码器输出和图像特征h_{IMG}的总和来初始化解码器。 形式上，我们将最终初始化状态h_d设置为：

$$h_d = \tanh\left(W_e\left[\overrightarrow{h_e}; \overleftarrow{h_e}\right] + W_{img}\, h_{img} + b\right). \quad (1)$$

其中W_e和W_{IMG}将编码器和图像特征向量投影到解码器隐藏状态维度中，b是偏置参数。 这种方法之前已被探索过 卡利斯托
和刘 (2017).

如前所述，传统上使用交叉熵损失训练翻译系统。 为了克服训练和推理分布之间的差异，我们使用预定的采样训练我们的模型（Bengio等。，2015）将基本事实与模型预测相结合，

进一步采用带基线的REINFORCE算法直接优化翻译指标。

2.1 预定抽样

当预测标记y_t时，预定采样使用具有概率E的先前模型预测y_{t-1}或具有概率1-E的先前地面实况预测y_{t-1}。 模型预测
[illegible]的
$P(y_{t-1} | h_{t-1})$的概率分布。 在

在训练开始时，采样令牌可以是非常随机的。因此，概率E最初设定得非常低并且随时间增加。

不正确，因为它们是从地面实况数据或模型预测中随机选择的，无论选择的输入方式如何（Ranzato等。，2015）。 因此，我们使用强化学习技术直接进一步优化翻译指标模型。

2.2 强化学习

以下 Ranzato等。 (2015）和 Rennie等人。 (2017），我们使用REINFORCE和基线来直接优化评估指标。

根据强化学习文献（萨顿等人。，1998），神经网络θ定义策略p_θ，其产生作为下一个词的预测的“动作”。 在生成序列结束项（EOS）之后，模型将获得奖励r，其可以是黄金和生成序列之间的评估度量，例如BLEU得分。 培训的目标是尽量减少负面的预期奖励。

$$L(\theta) = -E_{w^s \sim p_\theta}[r^s(w)]。$$

(2）其中句子$w^s = (w^s_1, \ldots, w^s_T)$。

为了计算梯度$\nabla_\theta L(\theta)$，我们

使用REINFORCE算法，该算法基于以下观察：不可微分的奖励函数的预期梯度可以如下计算：

$$\nabla_\theta L(\theta) = -E_{w^s \sim p_\theta}[r^s(w) \nabla_\theta \log p_\theta(w^s)]。 \quad (3)$$

可以推广策略梯度以计算与相对于参考奖励或基线b的动作值相关联的奖励：

$$\nabla_\theta L(\theta) = -E_{w^s \sim p_\theta}[(r^s(w) - b) \nabla_\theta \log p_\theta(w^s)]。 \quad (4)$$

基线不会改变预期的梯度，但重要的是，它可以减少梯度估计的方差。 我们使用引入的基线 Rennie等人。 (2017这是由当前模型在测试时使用贪婪解码获得的。

$$b = r(\hat{w}^s) \quad (5)$$

哪里$\hat{}$s

预定采样的一个主要限制是在每个时间步骤，目标序列可以是

w由贪婪解码生成。

对于每个训练案例，我们用单个样本$w^{\sim} p_{\theta}$近似预期的梯度：

$$\nabla_{\theta} L(\theta) \approx - (r(w) - b) \nabla_{\theta} \log p_{\theta}(w) 。 \quad (6)$$

	培养	开发。	翻译。	翻译。 在BPE之后
恩	2,900	1,014	10,212	7,633
德	2,900	1,014	18,726	5,942
神父	2,900	1,014	11,223	6,457
铯	2,900	1,014	22,400	8,459

表1：Flickr30K数据集的统计

2.3 集成

在我们使用相对较小的训练数据集的实验中，具有不同初始化的模型的翻译质量可以显着变化。 为了使性能更加稳定并提高翻译质量，我们在解码过程中集成了不同的模型以实现更好的翻译。

为了整合，我们采用所有模型输出的平均值：

$$\hat{y}_t = \sum_{i=1}^{N} \frac{Y_t^i}{N} \quad (7)$$

其中Y_t^i表示第i个的输出分布位置t处的模型。 如同 周等人。 (2017)，我们可以集合使用不同架构和训练算法训练的模型。

3 实验

3.1 数据集

我们使用Flickr30K进行实验（埃利奥特 等。，2016）由WMT组织提供。 任务1（多模式机器翻译）包括将带有英文标题的图像翻译成德语，法语和捷克语。 任务1b（多源多模式机器翻译）涉及将并行的英语，德语和法语句子与伴随的图像翻译成捷克语。

如表所示 1，这两项任务都有2900个培训和1014个验证示例。 对于预处理，我们将所有句子转换为小写，规范化标点符号并进行标记化。 我们采用字节对编码（BPE）（Sennrich等人。，2015）包括四种语言在内的整个训练数据，并将源语言和目标语言词汇量减少到总共20k。

3.2 培训细节

使用ResNet-101提取图像功能（他等人。，2016卷积神经网络

	恩德	恩神父	恩-CS
神经网络机器翻译	39.64	58.36	31.27
NMT+SS	40.19	58.67	31.38
NMT+SS+RL	40.60	58.80	31.73
金属-非金属过渡层	39.27	57.92	30.84
MNMT+SS	39.87	58.80	31.21
MNMT+SS+RL	40.39	58.78	31.36

表2：验证集上不同方法的BLEU分数。 表中描述了集合模型的细节 9.

在ImageNet数据集上接受过培训。 我们的实现改编自基于Pytorch的OpenNMT（Klein等。，2017）。 我们使用两层双LSTM（Sutskever等人。，2014）作为编码器并共享编码器和de-之间的词汇

编码器。 我们采用长度奖励（黄等人。，2017）在En-Cs任务中找到最佳句子长度。 我们使用批量大小为50，SGD优化

辍学率为0.1，学习率为1.0。 我们的单词嵌入是随机初始化的维度500。

为了训练具有预定采样的模型，我们首先将概率E设置为0，然后每5个历元逐渐增加0.05，直到它为0.25。 基于预定采样预训练的模型训练强化学习模型。

3.3 任务1的结果

为了研究不同方法的表现，我们进行了消融研究。 表 2 显示了使用不同模型和训练方法的验证集上的BLEU分数。 通常，具有预定采样的模型比基线模型执行得更好，强化学习进一步提高了性能。 与最佳单一模型相比，集合模型可实现大约+2至+3 BLEU分数的显着改善。 但是，通过包含图像信息，MNMT per-

任务	系统	NMT+SS	NMT+SS+RL	MNMT+SS	MNMT+SS+RL
	神经网络机器翻译	7	6	0	0
	混合	7	6	5	4
	神经网络机器翻译	9	5	0	0
	混合	9	0	3	0
	神经网络机器翻译	7	6	0	0
	混合	7	6	5	4

表3：用于加入的不同模型的数量。

	秩	蓝	流星	之三
NMT	1	**32.3**	50.9	49.9
-屋宇署-混合	2	32.1	50.7	**49.6**
LIUMCVC-MNMT-和	3	31.4	51.4	52.1
umons deepgru	4	31.1	**51.6**	53.4
liumcvc E -高考	5	31.1	51.5	52.6
SHEF1-ENMT	6	30.9	50.7	52.4
底线	-	27.6	47.4	55.2

表4：测试集上的En-De结果。 总共17个系统。 （仅包括约束模型）。

	秩	蓝	流星	之三
LIUMCVC-MNMT-和	1	**39.5**	59.9	41.7
UMONS	2	39.2	**60**	41.8
liumcvc E -高考	3	39.1	59.8	41.9
NMT	4	39.0	59.5	**41.2**
上肢-MLT	5	38.9	59.8	41.5
-屋宇署-混合	9	38.6	59.3	41.5
底线	-	28.6	52.2	58.8

表5：测试集上的En-Fr结果。 总共14个系统。 （仅包括约束模型）。

	秩	蓝	流星	之三
NMT	1	**30.2**	29.5	**50.7**
-屋宇署-混合	2	30.1	**29.7**	51.2
SHEF1-ENMT	3	29.0	29.4	51.1
上肢-它	4	28.3	29.1	51.7
上肢-MLT	5	28.2	29.1	51.7
MFS shef1 -	6	27.8	29.2	52.4
底线	-	26.5	27.7	54.4

表6：测试集上的En-Cs结果。 总共8个系统。 （仅包括约束模型）。

	恩-CS	FR CS	德-CS	（恩+ FR + 德）-Cs
神经网络机器翻译	**31.27**	28.48	26.96	29.47
金属-非金属过渡层	30.84	27.02	25.99	29.23

表7：任务1B的验证集上的BLEU分数

仅在具有预定采样的En-Fr任务上形成比NMT更好的形式。

表 4, 5 和 6 使用其他顶级性能模型显示我们的模型在En-De，En-Fr和En-Cs子任务上的测试集性能。 我们根据BLEU对这些模型进行排名。 我们提交的系统在BLEU和TER上的En-De和En-Cs子任务中排名第一。

3.4 任务1B的结果

表 7 显示没有序列训练的验证集上的结果。 En-Cs，Fr-Cs，De-Cs是从一种语言到另一种

	秩	蓝	流星	之三
NMT	1	**26.4**	28.0	**52.1**
-屋宇署-混合	1	**26.4**	**28.2**	52.7
shef1——	3	25.2	27.5	53.9
上肢	4	24.7	27.6	**52.1**
上肢-MLTC	5	24.5	27.5	52.5
shef1 ARF	6	24.1	27.1	54.6
底线	-	23.6	26.8	54.2

语言训练的模型。 （En + Fr + De）-Cs模型使用多个源数据进行训练。 类似于Shuffle方法dis-

表8：测试集上的任务1B多源转换结果。 总共6个系统。

任务	系统	模型 铁				R队 谢谢			
		数字†	蓝	MET。	之三	数字‡	蓝	MET。	之三
	神经网络机器翻译	11	1	4	2				
	混合	11	2	5	1				
	神经网络机器翻译	11	4	9	1				
	混合	11	9	10	3				
	神经网络机器翻译	6	1	1	1				
	混合	6	2	2	3				
	神经网络机器翻译	6	1	2	1				
	混合	6	1	1	5				

表9：我们的模型排名。 †代表模型的总数。 ‡代表团队总数。

参与多参考培训（郑等人。，2018），我们随机抽取所有语言的源数据，并在每个时代使用传统的基于注意力的神经机器翻译模型进行训练。 由于我们对整个训练数据进行BPE，我们可以在训练期间分享不同语言的词汇。 结果表明，使用单一英语到捷克语数据训练的模型比其他数据表现更好。

表 8 显示测试集的结果。 提交的系统与任务1的En-Cs任务中使用的系统相同。尽管我们在训练期间仅考虑英语源，但我们提出的系统仍然在所有提交中排名第一。

4 结论

我们描述了提交给共享WMT 2018多模式翻译任务的系统。 我们使用序列训练方法，这导致了对强基线的实质性改进。 我们的整体模型在三个子任务的BLEU分数中实现了最佳性能：En-De，任务1的En-Cs和（En + De + Fr）-Cs任务1B。

致谢

这项工作部分得到DARPA资助N66001-17-2-4030的支持，NSF授予IIS-1817231和IIS-1656051。我们感谢匿名审稿人的建议和Juneki Hong的校对。

参考

Dzmitry Bahdanau, Kyunghyun Cho和Yoshua Bengio。 通过联合学习对齐和翻译的神经机器翻译。 CORR。

Samy Bengio, Oriol Vinyals, Navdeep Jaitly和Noam Shazeer。 2015.使用递归神经网络进行序列预测的预定抽样。 在神经信息处理系统的进展，第1171-1179页。

Ozan Caglayan, Walid Aransa, Adrien Bardet, MercedesGarc'ıa-Mart'ınez, Fethi Bougares, Lo¨ıcBarrault, Marc Masana, Luis Herranz和Joost van de Weijer。 2017.Lium-cvc提交的wmt17多模式翻译任务。 在第二次机器翻译会议论文集，第432-439页。

Iacer Calixto和Qun Liu。 2017.将全局视觉特征融入基于注意力的神经机器翻译中。 在2017年自然语言处理经验方法会议论文集，第992-1003页。

D. Elliott, S。Frank, K。Sima'an和L. Specia。 2016. Multi30k：多语言英语 - 德语图像描述。第五届视觉与语言研讨会论文集，第70-74页。

Desmond Elliott, Stella Frank, Lo¨ıcBarrault, Fethi Bougares和Lucia Specia。 2017.关于多模式机器翻译和多语言图像描述的第二个共享任务的发现。 在第二次机器翻译会议记录，第215-233页。

Ekaterina Garmash和Christof Monz。 2016.多源神经机器翻译的集成学习。 在“COLING 2016年会议录”，第26届计算语言学国际会议：技术论文，第1409-1418页。

何开明，张翔宇，任少卿，孙健。 2016.图像识别的深度残差学习。 计算机视觉和模式识别会议CVPR。

黄亮，赵凯，马明波。 2017.什么时候结束？ 最佳波束搜索神经文本生成（模光束大小）。 在EMNLP 2017中。

G. Klein, Y。Kim, Y。Deng, J。Senellart和AM Rush。 2017.Opennmt：用于神经机器翻译的开源工具包。 ArXiv电子打印。

刘思奇，朱镇海，叶宁，Sergio Guadarrama和Kevin Murphy。 2017.通过蜘蛛的策略梯度优化改进了图像标题。 在Proc。 IEEE Int。CONF。 比较。 Vis，第3卷，第3页。

马明波，李大鹏，赵凯，黄亮。 2017.Osu多模机翻译系统报告。 在第二次机器翻译会议记录，第465-469页。

Ramesh Nallapati, Bowen Zhou和Mingbo Ma。 2016.分类或选择：用于提取文档摘要的神经架构。 CORR。

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli和Wojciech Zaremba。 2015.使用递归神经网络的序列水平训练。 arXiv preprint arXiv：1511.06732。

Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross和Vaibhava Goel。 2017.图像字幕的自我关键序列训练。 在CVPR中，第1卷，第3页。

Alexander M. Rush，Sumit Chopra和Jason Weston。 用于抽象句子摘要的神经注意模型。

Rico Sennrich, Barry Haddow和Alexandra Birch。 2015.带有子单元的罕见单词的神经机器翻译。 arXiv preprint arXiv：1508.07909。

Ilya Sutskever, Oriol Vinyals和Quoc V. Le。 用神经网络进行序列学习的序列。 第27届神经信息处理系统国际会议论文集。

Richard S Sutton, Andrew G Barto, et al。 1998.强化学习：介绍。 麻省理工学院出版

Subhashini Venugopalan，Marcus Rohrbach，Jeffrey Donahue，Raymond J. Mooney，Trevor Darrell和Kate Saenko。 2015.序列到序列 - 视频到文本。 在ICCV。

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel和Yoshua Bengio。 2015.展示，参与和讲述：神经图像标题生成与视觉注意。 第32届国际机器学习大会（ICML-15）会议记录。

郑仁杰，马明波，黄亮。 2018.用于神经翻译和文本生成的伪参考的多参考训练。 arXiv preprint arXiv：1808.09564。

周龙，胡文鹏，张嘉君，宗庆清。 2017.用于机器翻译的神经系统组合。 在计算语言学协会第55届年会会议录（第2卷：短文），第2卷，第378-384页。