基于多模态图的语义融合编码器的神经网络机器翻译模型

本文基于 ACL-2020 论文《A Novel Graph-based Multi-modal Fusion Encoder for Neural Machine Translation》，论文由腾讯微信 AI 团队、厦门大学、罗彻斯特大学合作完成。

导语：

多模态神经网络机器翻译的目标是将源语言和相应的图片翻译为目标语言，可以应用到多多媒体新闻翻译等场景。现有的模型都没能充分地建模不同模态语义单元（词或视觉对象）之间细粒度的语义关联，而这种关联有潜力优化多模态表示的学习。为了解决这个问题，作者提出了一种基于图的多模态语义融合编码器。具体地，首先将输入的源语言句子和对应的图片表示为一个统一的多模态图结构，此图结构包含了多种多模态语义单元之间的关系。在此图结构的基础上，作者使用多层基于图的多模态语义融合层来学习图中节点的表示。最后通过注意力机制为解码器提供源端的上下文。本模型在英德和英法多模态数据集上均取得了较好的效果。

一、 模型背景与简介

多模态机器翻译扩展了传统的基于文本的机器翻译形式，目的是将图片作为额外的输入，与源文一同翻译为译文。随着多模态机器学习的发展，多模态机器翻译也成为了机器翻译中的重要研究方向之一，它的应用有多媒体新闻翻译和网页产品信息翻译等。多模态机器翻译

期望视觉模态能为语言学习提供额外的信息来帮助解决歧义或多义词的问题，从而获得更准确的翻译。

如何充分利用视觉信息是多模态机器翻译中的核心问题，这直接影响翻译模型的性能。有很多研究工作致力于这个方面，根据使用视觉特征的方式可以大致分为以下三类：(1)将整个图片编码为一个全局的特征向量，用来初始化翻译模型的不同部件或用来学习多模态联合表示；(2)抽取基于区域的特征来补充源端序列或提供基于注意力机制的上下文；(3)将每张图片表示为基于网格的 CNN 特征，通过注意力机制为源端或目标端提供上下文。

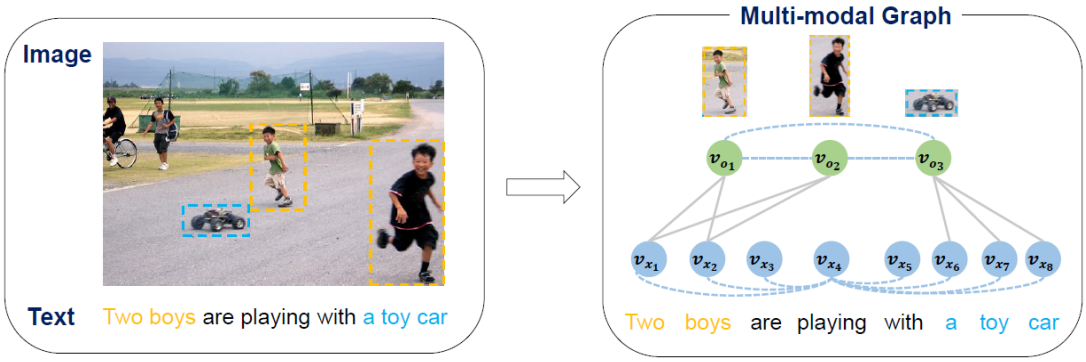

图 1：输入句子-图片对的多模态图结构

然而，以往的研究工作没有充分地利用输入的句子-图片对中细粒度的语义关联。比如如图一所示，名词词组“a toy car”和图片中玩具汽车的区域是语义对应的。忽略这一重要的语义关联可能是因为以下两个挑战：(1)如何构建一个统一的多模态表示来建立不同模态间的语义关联，和(2)如何在统一的多模态表示上进行语义交互。作者认为细粒度的语义关联可以优化多模态表示学习，因为这能够使单模态的表示通过多模态语义交互来融合跨模态信息作为补充。

在这篇论文中，作者提出了一个基于图的多模态融合编码器。首先将输入的源句子和图片构建成一个多模态图结构。图中的每个节点代表一个语义单元：文本中的词或者图片中的视觉对象，并引入不同类型的边来分别建立模态内和模态间的语义关联。在此图结构的基础

上，作者使用多层基于图的多模态语义融合层来进行图的编码，在每个融合层中顺序地执行模态内和模态间的语义融合来学习多模态图的节点表示。

总的来说，这个工作的贡献点如下：

1、 这篇工作提出了一个统一的图结构来表示输入的源句子和图片，使机器翻译模型能捕捉到多模态语义单元间的语义关联。

2、 作者提出了一个基于图的多模态语义融合编码器，是第一次在神经网络机器翻译上探索多模态图神经网络的工作。

3、 在英德和英法多模态翻译数据集上超过了有竞争力的基线模型，取得了较好的翻译效果。

二、 模型结构

本文提出了一种基于图表示的多模态融合编码器。首先，对于输入的图像和文本，使用统一的多模态图表示来建模图像中的视觉对象和文本中的名词短语的语义关联。随后，提出一个多模态语义融合编码器，在该图表示上迭代地进行多模态语义融合。

2.1 多模态图表示

多模态图表示以图的方式建模图像中的视觉对象和文本中的词之间的语义关联。图 1 给出了多模态图表示的建模过程。形式上，每个节点表示一个基本的语义单元。其中，绿色节点表示图像中的视觉对象，浅蓝色节点表示文本中的各个词。此外，引入两种类型的边分别模拟相同模态内部的语义单元之间的语义关联（蓝色虚线边）和多模态间的语义单元之间的语义关联（灰色实线边）。作者使用一个 visual grounding 技术来建立图像和文本之间的语

义关联。

2,2 基于图表示的多模态融合编码器

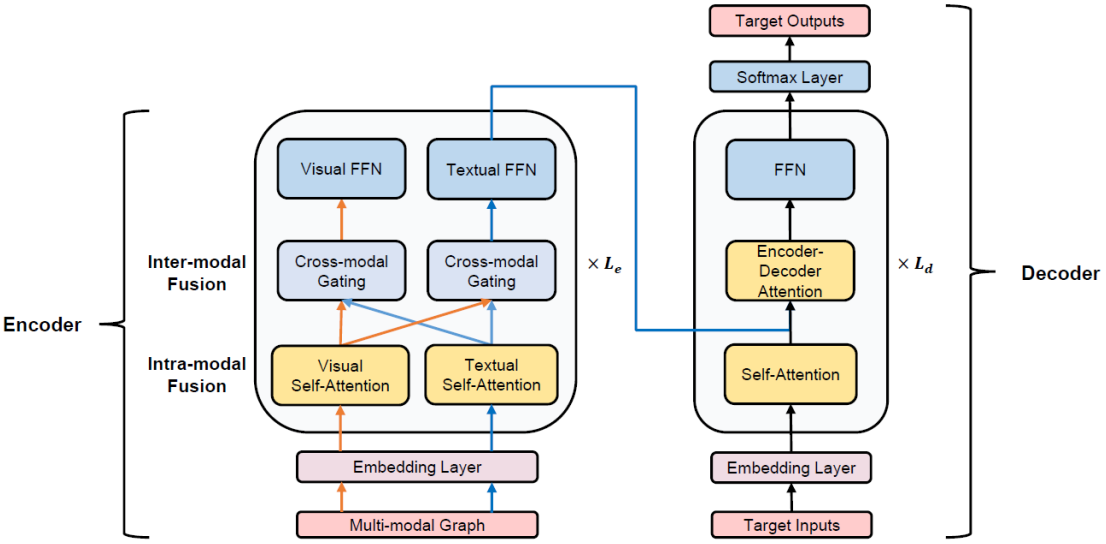

图 2：多模态神经网络机器翻译模型的整体架构

多模态融合编码器以 Transformer 为基础。如图 2 所示，整个多模态神经网络机器翻译包含L_e层基于图表示的多模态融合编码器和L_d层基于注意力机制的解码器。其中，基于图表示的多模态融合编码器包含以下多模态融合过程：

① 相同模态内部的语义融合层

作者采用自注意力机制来建模相同模态内部的语义信息。以第l层建模为例，文本句子的语义表示$\mathbf{C}_\mathbf{x}^{(l)}$的计算过程为：

$$\mathbf{C}_\mathbf{x}^{(l)} = \mathrm{MultiHead}(\mathbf{H}_\mathbf{x}^{(l-1)}, \mathbf{H}_\mathbf{x}^{(l-1)}, \mathbf{H}_\mathbf{x}^{(l-1)})$$

其中，$\mathrm{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V})$ 是多重注意力机制建模层，以三元组(Queries ，Keys ，Values) 作为输入。类似地，图像的语义表示$\mathbf{C}_\mathbf{o}^{(l)}$的计算过程为：

$$\mathbf{C}_\mathbf{o}^{(l)} = \mathrm{MultiHead}(\mathbf{H}_\mathbf{o}^{(l-1)}, \mathbf{H}_\mathbf{o}^{(l-1)}, \mathbf{H}_\mathbf{o}^{(l-1)})$$

② 多模态间的语义融合层

作者提出一种基于门机制的跨模态融合层来建模多模态间的语义融合。形式化的，

以第l层建模为例，文本模态的节点v_{x_i}的隐层表示$M_{x_i}^{(l)}$的计算过程为：

$$M_{x_i}^{(l)} = \sum_{j \in A\left(v_{x_i}\right)} \alpha_{i,j} \odot C_{o_j}^{(l)}$$

$$\alpha_{i,j} = Sigmoid(\mathbf{W}_1^{(l)} C_{x_i}^{(l)} + \mathbf{W}_2^{(l)} C_{o_j}^{(l)})$$

其中，$A(v_{x_i})$为文本节点v_{x_i}在多模态图表示的邻居节点，$W_1^{(l)}$和$W_1^{(l)}$为参数矩阵。同样地，作者以相同的计算方式建模图像节点v_{o_j}的隐层表示$M_o^{(l)}$，此处不再赘述。

经过上诉多模态融合过程后，采用前馈神经网络生成最终的隐层表示。形式化的，文本和图像的隐层表示的计算过程为：

$$\mathbf{H}_{\mathbf{x}}^{(l)} = \mathrm{FFN}\left(\mathbf{M}_{\mathbf{x}}^{(l)}\right).$$

$$\mathbf{H}_{\mathbf{o}}^{(l)} = \mathrm{FFN}\left(\mathbf{M}_{\mathbf{o}}^{(l)}\right).$$

其中，$\mathbf{M}_{\mathbf{x}}^{(l)} = \{M_{x_i}^{(l)}\}$,$\mathbf{M}_{\mathbf{o}}^{(l)} = \{M_{o_j}^{(l)}\}$。

解码器使用的是常规 Transformer 解码器，考虑到通过多层的多模态语义交互，文本节点已经充分融合了图像模态的特征，因此解码器仅考虑使用文本节点提供上下文。

三、 实验结果

作者在 Multi30k 多模态翻译数据集上进行了实验，使用 BLEU 和 METEOR 作为译文的评价指标。结果如下：

Model	En⇒De					
	Test2016		Test2017		MSCOCO	
	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR
Existing Multi-modal NMT Systems						
Doubly-att(RNN) (Calixto et al., 2017)	36.5	55.0	-	-	-	-
Soft-att(RNN) (Delbrouck and Dupont, 2017a)	37.6	55.3	-	-	-	-
Stochastic-att(RNN) (Delbrouck and Dupont, 2017a)	38.2	55.4	-	-	-	-
Fusion-conv(RNN) (Caglayan et al., 2017)	37.0	57.0	29.8	51.2	25.1	46.0
Trg-mul(RNN)(Caglayan et al., 2017)	37.8	**57.7**	30.7	**52.2**	26.4	47.4
VMMT(RNN) (Calixto et al., 2019)	37.7	56.0	30.1	49.9	25.5	44.8
Deliberation Network(TF) (Ive et al., 2019)	38.0	55.6	-	-	-	-
Our Multi-modal NMT Systems						
Transformer (Vaswani et al., 2017)	38.4	56.5	30.6	50.4	27.3	46.2
ObjectAsToken(TF) (Huang et al., 2016)	39.0	57.2	31.7	51.3	28.4	47.0
Enc-att(TF) (Delbrouck and Dupont, 2017b)	38.7	56.6	31.3	50.6	28.0	46.6
Doubly-att(TF) (Helcl et al., 2018)	38.8	56.8	31.4	50.5	27.4	46.5
Our model	**39.8**	57.6	**32.2**	51.9	**28.7**	**47.6**

图 3：英德翻译任务的实验结果

Model	En⇒De					
	Test2016		Test2017		MSCOCO	
	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR
Our model	39.8	57.6	32.2	51.9	28.7	47.6
w/o inter-modal fusion	38.7	56.7	30.7	50.6	27.0	46.7
visual grounding ⇒ fully-connected	36.4	53.4	28.3	47.0	24.4	42.9
different parameters ⇒ unified parameters	39.2	57.3	31.9	51.4	27.7	47.4
w/ attending to visual nodes	39.6	57.3	32.0	51.3	27.9	46.8
attending to textual nodes ⇒ attending to visual nodes	30.9	48.6	22.3	41.5	20.4	38.7

图 4：英德翻译任务的消融实验结果

Model	En⇒Fr			
	Test2016		Test2017	
	BLEU	METEOR	BLEU	METEOR
Existing Multi-modal NMT Systems				
Fusion-conv(RNN) (Caglayan et al., 2017)	53.5	70.4	51.6	68.6
Trg-mul(RNN)(Caglayan et al., 2017)	54.7	71.3	52.7	**69.5**
Deliberation Network(TF) (Ive et al., 2019)	59.8	74.4	-	-
Our Multi-modal NMT Systems				
Transformer (Vaswani et al., 2017)	59.5	73.7	52.0	68.0
ObjectAsToken(TF) (Huang et al., 2016)	60.0	74.3	52.9	68.6
Enc-att(TF) (Delbrouck and Dupont, 2017b)	60.0	74.3	52.8	68.3
Doubly-att(TF) (Helcl et al., 2018)	59.9	74.1	52.4	68.1
Our model	**60.9**	**74.9**	**53.9**	69.3

图 5：英法翻译任务的实验结果

在常用的英德和英法数据上，作者提出的模型均超过了几个强大的基线系统，取得了与之前最佳模型可比或者更好的翻译效果。消融实验也说明了模态间语义融合的有效性，以及验证了文本节点已经充分吸收了视觉模态的信息。

四、 总结

本文提出了一个基于图的多模态融合编码器，充分捕捉了多模态间细粒度的语义关联。在 Multi30k 数据集上的实验表明了提出方法的有效性。未来，融入视觉对象的属性和引入场景图可能是会是潜在的研究方向。