Multi-Graph Transformer 网络

徐鹏

论文链接：https://arxiv.org/pdf/1912.11258.pdf

代码链接：https://github.com/PengBoXiangShang/multigraph_transformer

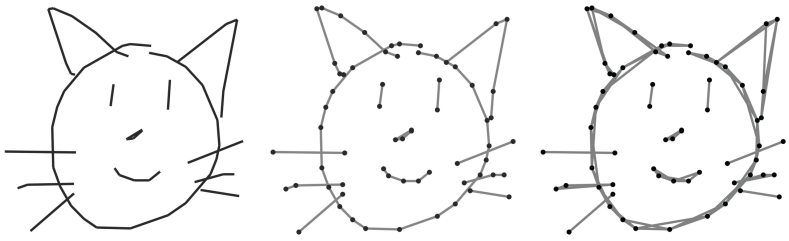
2020 年 1 月 15 日

以下将对论文《Multi-Graph Transformer for Free-Hand Sketch Recognition》进行简短且通俗地介绍，更多技术细节请大家参看我们的原文。欢迎大家通过邮件（peng.xu@ntu.edu.sg）或微信（微信号：roc56789）与我进行更深入的交流。

1 研究动机

通常，Transformer 的输入是序列化输入形式，若给定一个句子作为输入，Transformer 允许句子中的全部词之间建立相互关联的 attention 关系。所以，本质上讲，Transformer 把输入的每个句子看作一个全连接的图（fully-connected graph），Transformer 也算是一种特殊的图神经网络 (GNN)。然而，如何能为 Transformer 注入先验知识去引导它更精细化地学习图上的结构模式，是一个值得思考的问题。本论文提出以手绘草图作为一种 GNN 的实验床，探索新颖的 Transformer 网络。

手绘草图（free-hand sketch）是一种特殊数据，本质上是一种动态的序列化的数据形式。因为，手绘的过程本身就是一个“连点成线”的过程（如下图 1(b)所示）。



(a) original sketch　(b) 1-hop connected　(c) 2-hop connected

图 1: 手绘草图的离散化理解示意图

已有的手绘草图研究工作均在欧氏空间中对手绘草图进行建模，手绘草图被理解为静态图片输入到 CNN 中，或者被理解为笔画的关键点的坐标序列输入到 RNN 中。然而，在实时性要求较高的人机交互场景中，存储和传输图片会引起较大的开销，存储和传输笔画的关键点的坐标是更好的选择。文本的主要动机就是将手绘草图表示为稀疏图，将笔画的关键点理解为结点（node），且在几何空间中使用 Transformer 对其进行建模，从更具普适性的角度去理解并表示手绘草图。通过实验，我们发现且证实了，原版的 Transformer（Vanilla Transformer）并不能对手绘草图进行合理地表示。所以，文本提出了一种新颖的图神经网络，即 Multi-Graph Transformer（MGT）网络结构，将每一张手绘草图表示为多个图结构（multiple graph structure），并且这些图结构中融入了手绘草图的领域知识（domain knowledge）（如上图 1(b)和 1(c)所示）。

本文所提出的 Multi-Graph Transformer 网络也可以用于其他结构化且序列化的数据建模当中。

2 Multi-Graph Transformer （MGT）

本文所提出的网络结构可分为三个部分：(1) 网络的输入层；(2) 网络的主干，即多层的 Multi-Graph Transformer 结构；（3）网络的输出层，即分类器。

2.1 Multi-Modal Input Layer

我们采用 Google QuickDraw 数据，对每一张手绘草图都取前 100 个笔画关键点，对多于 100 个关键点或者少于 100 个关键点的手绘草图进行截断（truncation）或者补零（padding）操作。每个结点被表示为 4 维的向量，前两位是该结点在画布上的横纵坐标，第三位是用于描述画笔状态的标志位，第四位是位置编码。横纵坐标通过线性层进行升维，标志位和位置编码通过 embedding layer 进行升维，它们升维之后拼接（concatenate）起来构成 MGT 的输入。

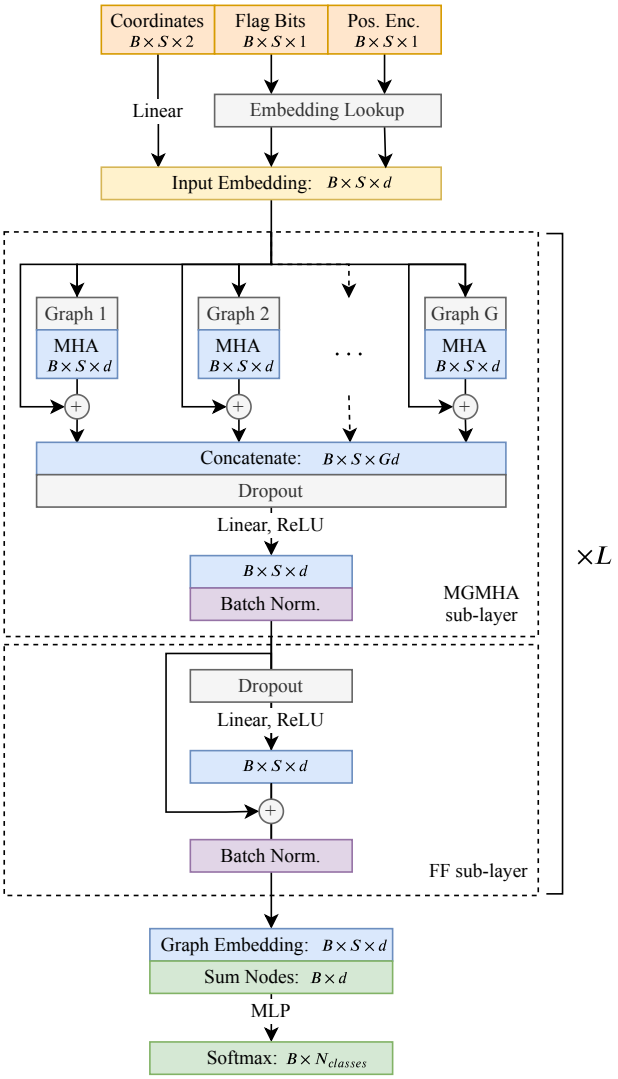


图 2: Multi-Graph Transformer 网络结构图

2.2 Multi-Graph Transformer

如图 2所示，整体上看，我们所提出的 Multi-Graph Transformer（MGT）是一个 L 层的结构，每层由两个子层构成，分别是 Multi-Graph Multi-Head Attention（MGMHA）sub-layer 和 position-wise fully connected Feed-Forward （FF）sub-layer。

我们所提出的 MGMHA 子层是一个多路并行结构，每一路都是一个基于图结构的 Multi-Head Attention 模块。这里的“图”结构是由我们基于手绘草图的领域知识所定义的图结构，也就是我们在原文中所定义的多种邻接矩阵。我们使用这些邻接矩阵来描述每张手绘草图上结点间的连通性。进而，在 Multi-Head Attention 操作中，使用邻接矩阵所描述的连通性来控制注意力分数矩阵中的连通性，允许或者屏蔽掉特定结点间的注意力关系。

FF 子层主要进行残差连接和 BN 等操作，这里不做赘述。

2.3 Sketch Embedding and Classification Layer

给定一张草图，经过 MGT 后，其每个结点都会被表示为一个向量，我们将这些结点的表示向量加起来作为该张草图的向量表示。加和过程中，不考虑数据预处理过程中 padding 操作所引入的额外结点。网络尾端的分类器由多层感知器来实现，使用 softmax 交叉熵损失函数。

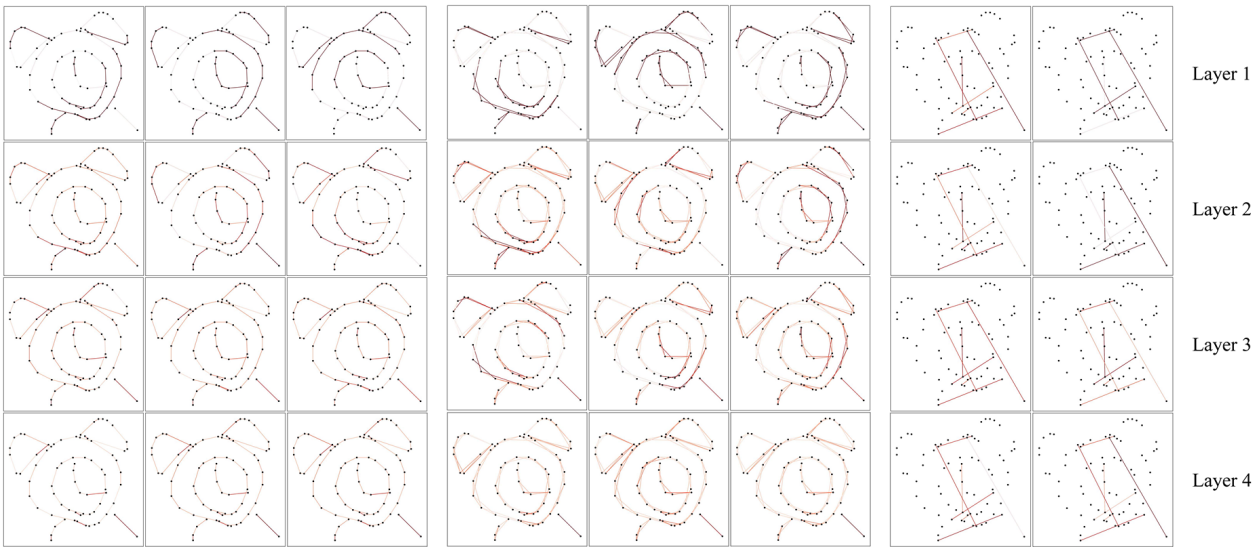
3 实验

原文中提供了 MGT 与众多经典的 RNN 结构和 CNN 网络的性能比较，同时也提供了详细的消融实验结果及可视化结果。尽管数据预处理环节的截断操作决定了 CNN 是 MGT 的性能上界，但是 MGT 所取得的识别准确率不仅远高于基于 LSTM 和 GRU 的网络，而且还超越了众多经典 CNN 网络，仅低于 Incetpion V3 和 MobileNet V2，但差距很微小。

表 1: Test set performance of MGT vs. the state-of-the-art RNN and CNN architectures. The $1^{st}/2^{nd}/3^{rd}$ best results per column are indicated in **red**/blue/magenta.

Network	Configurations	Recognition Accuracy			Parameter
		acc.@1	acc.@5	acc.@10	Amount
Bi-directional LSTM #1	4D Input, $\hat{d}=256, L=4, Dropout_{LSTM}=0.5, Dropout_{MLP}=0.15$	0.6665	0.8820	0.9189	5,553,241
Bi-directional LSTM #2	4D Input, $\hat{d}=256, L=5, Dropout_{LSTM}=0.5, Dropout_{MLP}=0.15$	0.6524	0.8697	0.9133	7,130,201
Bi-directional GRU	4D Input, $\hat{d}=256, L=5, Dropout_{GRU}=0.5, Dropout_{MLP}=0.15$	0.6768	0.8854	0.9234	5,419,097
AlexNet	Standard architecture and configurations	0.6808	0.8847	0.9203	58,417,305
VGG-11		0.6743	0.8814	0.9191	130,179,801
Inception V3		0.7422	0.9189	0.9437	25,315,474
ResNet-18		0.7031	0.9030	0.9351	11,353,497
ResNet-34		0.7009	0.9010	0.9347	21,461,657
ResNet-152		0.6924	0.8973	0.9312	58,850,713
DenseNet-201		0.7050	0.9013	0.9331	18,755,673
MobileNet V2		0.7310	0.9161	0.9429	2,665,817
Vanilla Transformer	$\hat{d}=256, L=4, I=8, Dropout=0.1$, Fully-connected graph	0.5249	0.7802	0.8486	14,029,401
MGT (Base)	$\hat{d}=128, L=4, I=24, Dropout=0.1$, $\mathbf{A}^{\text{1-hop}}, \mathbf{A}^{\text{2-hop}}, \mathbf{A}^{\text{global}}$ graphs	0.7070	0.9030	0.9351	10,096,601
MGT (Large)	$\hat{d}=256, L=4, I=24, Dropout=0.25$, $\mathbf{A}^{\text{1-hop}}, \mathbf{A}^{\text{2-hop}}, \mathbf{A}^{\text{global}}$ graphs	0.7280	0.9106	0.9387	39,984,729

下图给出了可视化的分析，将一张闹钟的草图输入到训练好的 MGT 中，其经过每一层后得到相应的注意力权重（attention heads），这里选取了其中一些有代表性的 heads。可以看到初始层的 heads 中，结点会更多地关注局部，消息传递是沿着笔画展开的，高层的 heads 中，局部的注意力在逐渐淡化，模型正在从全局地角度对图上的关系进行聚合。同时，基于全局图结构先验知识所学到的 attention heads 对跨笔画的消息传递也很重要，例如可以捕获闹钟的 body 和 feet 间的关系。



(a) $\mathbf{A}^{\text{1-hop}}$ (b) $\mathbf{A}^{\text{2-hop}}$ (c) $\mathbf{A}^{\text{global}}$

图 3: 注意力权重可视化

4 结论

本文提出了一种新颖的图神经网络，即 Multi-Graph Transformer （MGT），同时也为手绘草图提出了一种新颖的表示方法，即，把每一张手绘草图表示为多张稀疏连接的图。文本所提出的 MGT 网络的主要特性包括：（1）可以同时对手绘草图中的几何结构信息和笔画时序信息进行建模；（2）通过预定义的多种图结构为 Transformer 结构注入了领域知识；（3）充分利用了手绘草图的全局和局部图结构，即笔画内的、笔画之间的多重图结构。

希望文本可以帮助手绘草图领域的学者们从图的角度对手绘数据在更具普适性的几何空间中进行建模，同时帮助图神经网络领域的学者们把手绘数据作为一种新型的实验数据床。

最后附上我的微信二维码。