机器学习项目报告

基于 CNN 与 RNN 结合的草图分类网络

李子龙　唐　珂　裴禹乔

2022 年 6 月 19 日

摘要

本项目旨在解决 25 类草图图像分类问题。首先采用 CNN 网络（Sketch-a-Net, AlexNet, ResNet）对光栅化后的图像利用其局部信息进行识别，之后采用 RNN 的双向 LSTM 结构对笔画利用其顺序信息进行分类，最终使用 CNN 网络（Sketch-a-Net）与 RNN 网络（BiLSTM）相结合的方法得到了较好的分类结果。

1 Introduction

徒手素描是人类历史上传达信息和表达情感的一种有效方式。素描草图不仅包含目标对象生动的分类特征，而且还包含抽象、多样的视觉表现。在本项目中，我们需要针对大小为 28×28 的 25 类草图图像（如图 1 所示）构建深度学习模型以达到较好的分类效果，并对不同的网络结构性能进行比较。

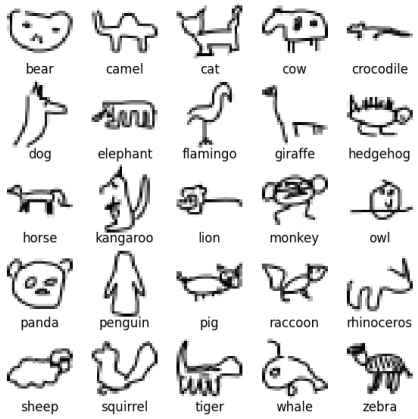


图 1: 25 个类别的草图

2 Main Ideas

本小组首先尝试从图像入手，搭建并训练了多种 CNN 网络对图像进行分类并观察其分类结果。考虑到图像由相应的笔画序列生成，我们后续尝试了通过 RNN 架构直接对笔画序列进行分类。最终我们将二者结合，通过 CNN+RNN 的方式结合图像信息和笔画序列信息对图片进行分类。

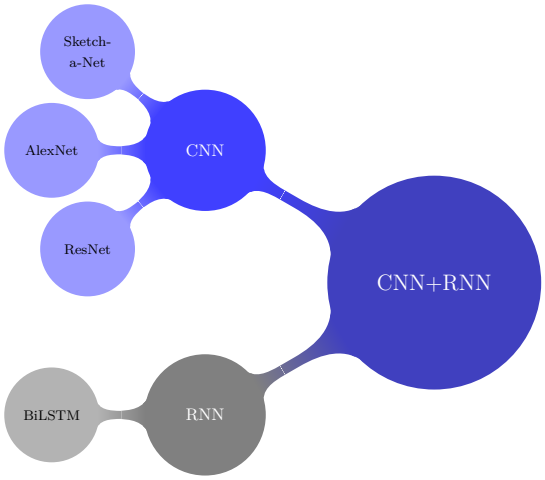


图 2: 思维导图

3 Methods and algorithms

3.1 数据集预处理

QuickDraw 数据集[1] 由数百个常见涂鸦对象类组成。每个 QuickDraw 对象类数据集包含 70000 个训练样本以及 2500 个验证样本。QuickDraw 使用了一种数据格式，将草图表示为一组笔画动作，该格式中将 0/1 笔画事件扩展为多状态事件。在此数据格式中，图形的初始绝对坐标位于原点，并将草图表示为由点构成的列表，每个点表示为由 5 个元素组成的向量：$(\Delta x, \Delta y, p_1, p_2, p_3)$。其中前两个元素是笔划在 x 和 y 方向上相对于前一点的偏移距离，后三个元素表示当前画笔的三种可能状态。第一个笔状态 p_1 表示笔当前正在接触纸张，并且将绘制一条线，将下一点与当前点连接起来。第二个笔状态 p_2 表示笔将在当前点之后从纸上提起，下一步不会画线。第三个笔状态 p_3 表示图形已结束，后续点（包括当前点）将不会渲染。

由于我们使用的 QuickDraw 数据集中的原始草图是以矢量化序列呈现的，为使用基于图片的 CNN 网络架构我们需要首先将其进一步转化为草图图像。为此我们参考了 [2] 提供的通过矢量化序列创建草图图像的方法，将 QuickDraw 数据集的笔画格式首先转换为 SVG 图像，后转化为相应的图片形式，如图 3 所示。

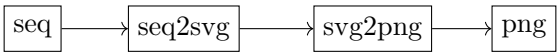


图 3: 将笔画格式转换为图像

3.2 CNN

3.2.1 动机

首先我们尝试以图片作为数据集，因此选择了适用于图像处理的 CNN 网络结构。卷积神经网络（CNN）主要是用于图像识别领域，CNN 的结构通常可以分为 3 层：卷积层 (Convolutional Layer) — 主要作用是提取特征；池化层 (Max Pooling Layer) — 主要作用是下采样 (downsampling)，却不会损坏识别结果；全连接层 (Fully Connected Layer) — 主要作用是进行分类。在图像数据集上使用 CNN 网络通常可以取得较好的分类结果，为此我们分别使用了代表性的 Sketch-a-Net、AlexNet、ResNet-18 网络架构，选择 QuickDraw 图片数据集作为我们的训练集，并在测试集对不同网络架构的分类准确性进行了测试。

3.2.2 Sketch-a-Net

Sketch-a-Net[3] 是 Qian Yu 等人针对手绘草图识别问题提出的多通道的深度神经网络框架，Sketch-a-Net 使得计算机对手绘草图的识别能力首次超过了人类。同时，Sketch-a-Net 使用了 15×15 的卷积核，由于手绘草图缺少纹理信息，较大的卷积核可以更好的体现草图的结构信息。我们复现了其代码，并在其基础上将训练数据集修改为 QuickDraw 数据集。

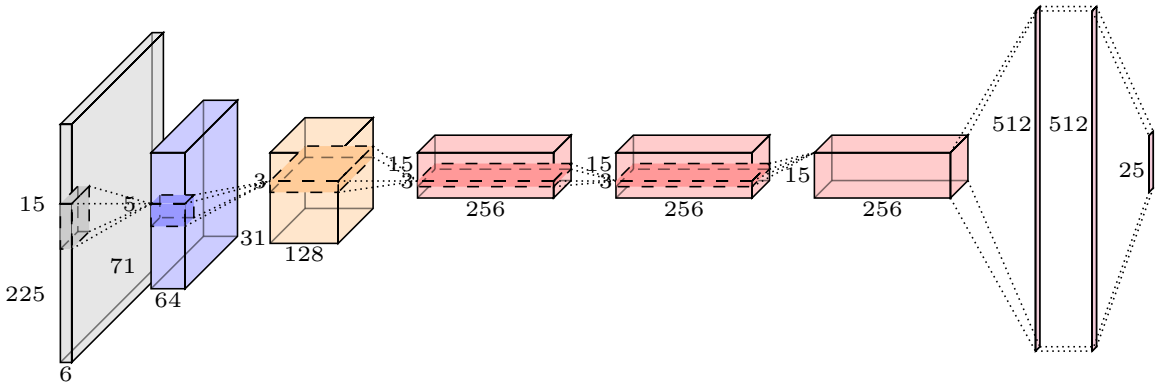


图 4: Sketch-a-Net 结构

3.2.3 AlexNet

AlexNet[4] 卷积神经网络模型由 5 个卷积层和 3 个池化 Pooling 层以及 3 个全连接层构成，如图 5 所示。其特点在于：使用 ReLU 作为 CNN 的激活函数，ReLU 函数的效果在较深的网络中超过常规的 Sigmoid 函数，解决了 Sigmoid 在网络较深时的梯度弥散问题；在训练时使用 Dropout 随机忽略一部分神经元，以避免模型过拟合；并且在 CNN 中使用重叠的最大池化 (步长小于卷积核)，此前 CNN 中普遍使用平均池化，使用最大池化可以避免平均池化的模糊效果，同时重叠效果可以提升特征的丰富性；并且 AlexNet 使用了 LRN 层（Local Response Normalization，即局部响应归一化），对局部神经元的活动创建竞争机制，使得其中响应比较大的值变得相对更大，并抑制其他反馈较小的神经元，增强了模型的泛化能力。在 AlexNet 的基础上我们将 QuickDraw 数据集中 28×28 的图片 resize 为 225×225 进而作为 AlexNet 的输入，最终通过训练得到了较好的分类结果。

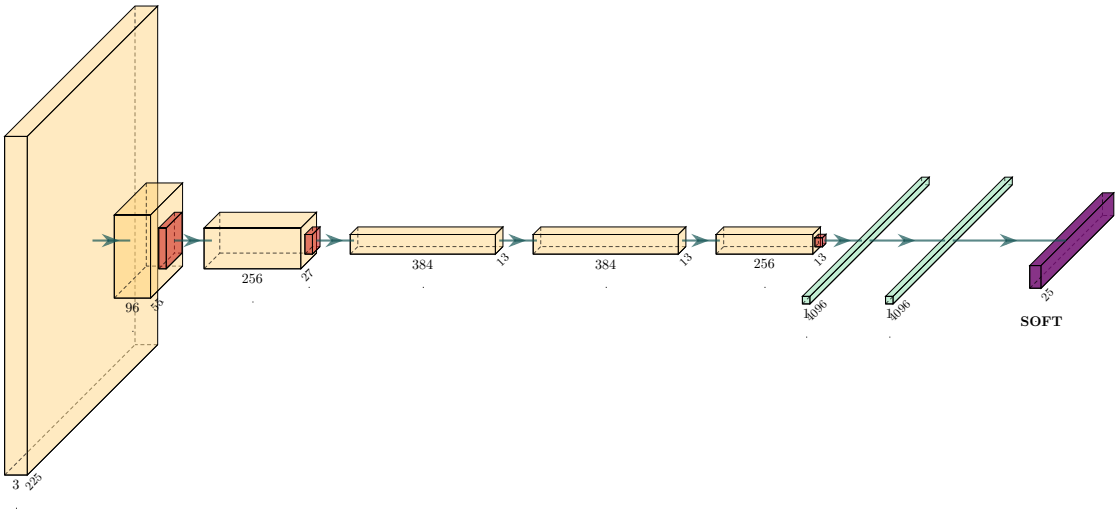


图 5: AlexNet 结构

3.2.4 ResNet-18

ResNet[5] 网络参考了 VGG19 网络，沿用了 VGG 完整的 3×3 卷积层设计，并在其基础上通过短路机制加入了残差单元。残差单元里首先有 2 个有相同输出通道数的 3×3 卷积层，每个卷积层后接一个批量规范化层和 ReLU 激活函数，残差单元通过跨层数据通路，跳过这 2 个卷积运算，将输入直接加在最后的 ReLU 激活函数中。通过以上方式，ResNet 很好的处理了深度卷积网络在图像分类中的退化问题。

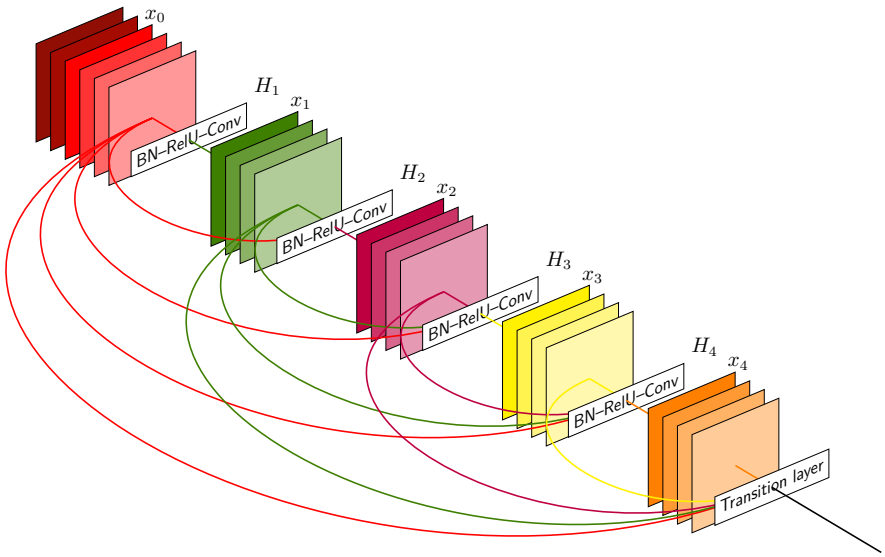


图 6: ResNet-18 结构

3.3 RNN

3.3.1 动机

考虑到原 QucikDraw 数据集由笔画序列构成，我们尝试了对序列处理较为高效的 RNN 架构。循环神经网络（RNN），同经典的前馈神经网络相比较（如多层感知器、深度置信网络、卷积神经网络等），RNN 允许网络隐藏层 (hidden layer) 的输出再以输入的形式作用于该隐藏层自己。

由于 RNN 具有记忆性并且具有参数共享的特征，在对序列的非线性特征进行学习时具有一定优势，而 RNN 单元在面对长序列数据时，很容易便遭遇梯度弥散，使得 RNN 只具备短期记忆。

而长短期记忆网络（Long-Short Term Memory, LSTM）通过引入遗忘门（forget gate）、输入门（input gate）、输出门（output gate）等结构控制信息的保留与丢弃，如图 7 所示。可以有效地解决 RNN 中的梯度爆炸和梯度消失问题，并有效地处理长距离的依赖。

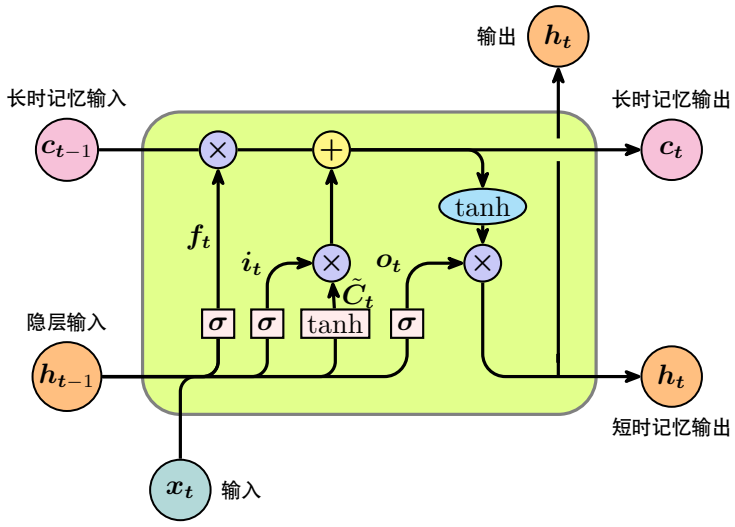


图 7: LSTM 元胞结构

3.3.2 双向 LSTM

SketchRNN[1] 采用了可变自编码器（Variational Auto Encoder, VAE）结构生成其他的笔画图像。本项目为分类问题，我们只采用编码器（Encoder）部分的双向 LSTM 结构[6] 用于分类识别，隐藏层元胞数目为 256×2，没有对 latent space 的 N_z 进行生成，而是最后直接添加线性层全连接用于图像的 25 分类。

双向 LSTM 结构（如图 8）相较于单向的 LSTM 增加了后向传播层，不仅可以考虑之前的状态，还会考虑未来的状态，这样可以更好地识别笔画信息。

3.4 CNN+RNN

Peng Xu 等人在 SketchMate[7] 文章中提出了一种将 CNN 与 RNN 结合的网络结构，分别将图片信息输入 CNN 网络分支并通过 CNN 提取抽象的视觉概念，将笔画序列信息输入 RNN 网络分支建模素描的时间顺序，之后再通过后期融合层以及量化编码层将二者进行结合。

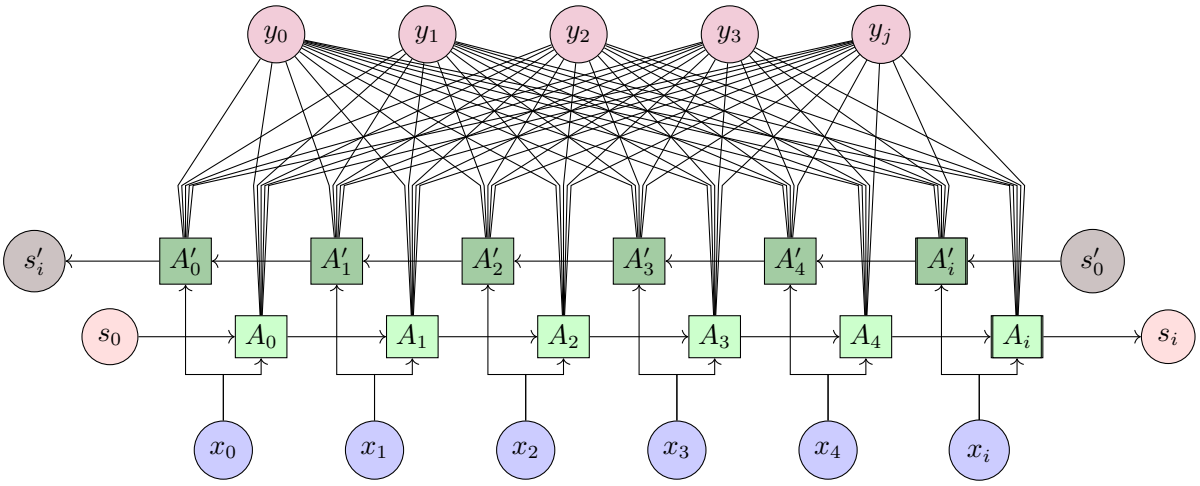


图 8: BiLSTM 结构（$i = \max \#seq, j = 25$）

我们又参考了 LiveSketch[8] 的网络结构，将 CNN 和 RNN 分支直接连接到全连接层，最后用于分类，如图 9。其中 CNN 分支使用了 Sketch-a-Net 结构，而 RNN 分支使用了 BiLSTM 结构，全连接层将会同时接收两个分支各 512 个输出（共 1024 个输出）连接于 512 个神经元上，之后全连接于后层用来进行 25 个分类的判别。

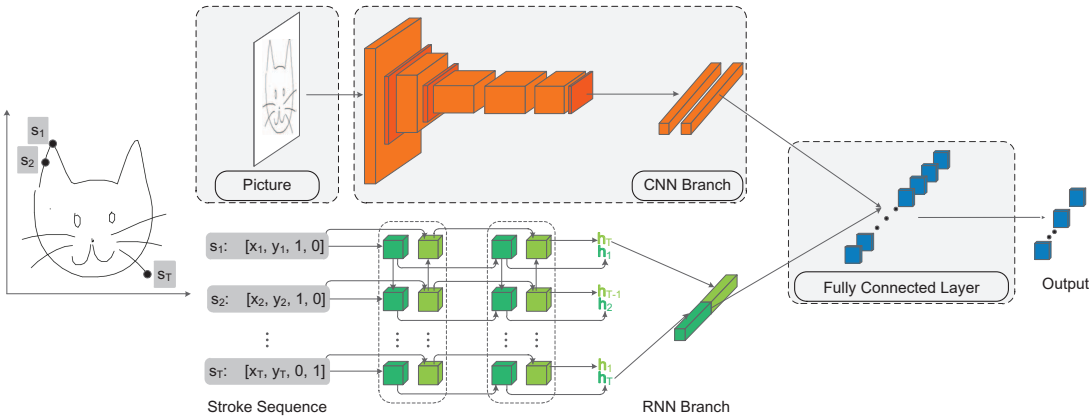


图 9: CNN+RNN 结构（分支各 512 个输出，全连接层 512 个神经元，输出层为 25 分类器）

4 Experimental settings

4.1 训练参数

网络超参数为第 3 节所言的默认参数。训练时参数如表 1 所示。

表 1: 训练参数

参数	Sketch-a-Net	AlexNet	ResNet	BiLSTM	CNN+RNN
lr			0.001		
batch size			64		
weight decay	0.001	0.001	0.001	0.0006	0.0006

4.2 优化训练方法

4.2.1 权重衰减

权重衰减（weight decay），也可称为 L2 正则化，其目的是让权重衰减到更小的值，在一定程度上缓解模型过拟合的问题。其之所以有效，是因为 L2 范数对于权重向量的大分量施加了巨大的惩罚，使学习算法会偏向在大量特征上均匀分布权重的模型，对单个变量上的观测误差更加稳定，而非将权重集中在小部分特征上。

本小组通过设置 Adam 优化器中 `weight_decay`，以抑制模型深度可能过深带来的过拟合现象，增强模型的泛化能力。

4.2.2 可变的学习率

调整学习率也是优化模型十分重要的一环。如果学习率过大，会造成结果难以收敛；而如果学习率太小，既可能导致结果收敛过慢，训练时间过长，也可能导致结果只处于一个局部最优而非更好的结果。因此，在训练的过程当中，我们需要动态衰减学习率。

我们使用了 PyTorch 包中自带的 scheduler 类来进行学习率衰减，如下面代码所示:

```
scheduler = ReduceLROnPlateau(optimizer, 'min', factor=0.3, patience=2)
train_model(model, dataloaders_dict, criterion, optimizer, scheduler, num_epochs=5)
```

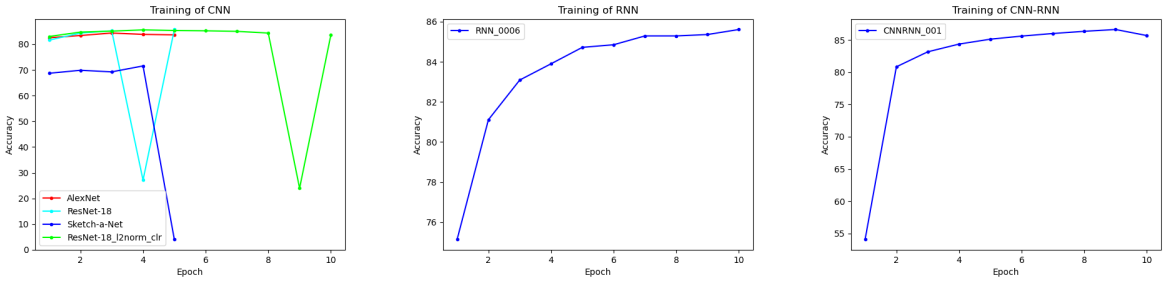
解释:

- **patience**: 这个参数是容忍程度的意思，意味着在训练数量为 patience 的轮数之后，如果 validation 集上的 loss 一直则没有下降，则要求开始降低学习率。本小组的代码中将这设置成 2。
- **factor**：这个参数是学习率衰减时的乘以的因子，本小组将此设置成 0.3，意味着每一次衰减为原来的 0.3 倍。

5 Experimental results

训练时验证集准确率的收敛情况如图 10 所示。由图可见，CNN 的收敛速度要比 RNN 要快，但是单个批次的训练时间 CNN 会更长一些。在 CNN 中，ResNet 的最终正确率会更高一些。将 CNN 与 RNN 结合的网络最终验证集正确率会比两者单独都要高一些。

最后将最佳验证准确率对应的网络在 25 个类别所有的测试集进行测试，结果如表 2 所示。



(a) CNN (b) RNN (c) CNN+RNN

图 10: 验证集收敛情况

表 2: 每一类模型最佳的测试准确率

模型	测试准确率
Sketch-a-Net	0.6948
AlexNet	0.8462
ResNet18	0.8557
BiLSTM	0.8561
CNN+RNN	**0.8567**

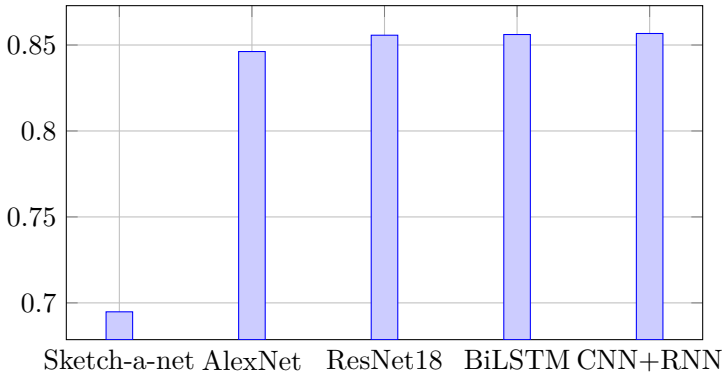


图 11: 测试准确率统计图

6 Conclusion

6.1 结果分析

1. 从 CNN 的曲线当中可以看到，验证集的 accuracy 会随着训练轮数的增加在某一个 epoch 突然下降，我们猜想这可能跟模型过拟合有关。

2. 事实上 weight_decay 参数对 RNN 模型的收敛有很大的影响，只有当 weight_decay 设置得足够小，比如低于 0.001 时，RNN 的训练才会收敛。这说明仔细调节超参的重要性。

3. alexnet 和 resnet18 都是在图像识别领域取得巨大成功的模型，在我们的实验中它们的发挥比 sketch-a-net 要好很多，这说明即使简笔画不同于一般的照片，良好的 CNN 模型也是有不错的预测能力。但是与正常图片高达 95% 以上的预测准确度相比，这些优异的 CNN 模型在简笔画上的发挥要低上十个百分点，这说明简笔画这种单单仅有几个线条的卡通图片的确会给分类造成困难。

4. BiLSTM 的结果和 AlexNet 和 ResNet18 的结果差不多，但是 BiLSTM 的训练所耗费的资源更少，在给定笔画的情况下，采用这种方法可能会更好。

5. CNN+RNN 上下两路结合的方法在我们的实验当中取得最佳的预测准确度。这种方法同时兼顾了图片的视觉信息以及笔画的序列信息，通过两种神经网路提取出不同维度的 feature，并将这两种 feature 给结合在一起，使得不同的模型可以学习到数据的不同特征，经过融合后的结果往往能有更好的表现，大有取长补短的意思。

6.2 未来的改进

1. 我们使用 RNN 的方法是直接将 decoder 给剔除，然后在 encoder 最后添加线性层，这种方法可能没有充分使用笔画的序列信息。或许使用 VAE，得到每一种种类简笔画背后的 latent space 能够对分类有很大的帮助。

2. CNN 模型表现不佳可能是因为简笔画不像一般的图片，简笔画中仅仅只有寥寥数笔，其余是大片的空白，如果能够将这些空白给填充起来，或许 CNN 就能够发挥出更大的威力。

6.3 小结

本项目从 CNN 入手，比较了 AlexNet, ResNet, Sketch-A-Net 对图像信息的分类结果，测试结果是 ResNet 略胜一筹。接着使用 RNN 中的双向 LSTM 对笔画数据进行分类，结果略有提升。最后将两者结合起来，最后通过多层神经网络进行分类，结果进一步有所提升。

CNN 优点在于准确率相对较高，收敛速度快。但缺点就是训练时间较长，从笔画转换为图像数据也需要时间。RNN 优点在于训练速度快，但缺点就在于参数不当的情况下，收敛速度会很慢。将两者结合，可以很好地利用图像的局部信息与笔画的产生顺序，更为高效地获得较好的分类模型。

Contribution

项目地址：https://github.com/LogCreative/quickdraw-classifier

表 3: 团队贡献

姓名	贡献	百分比
李子龙	RNN 模型的完善，CNN+RNN 模型的初步构造，报告	33.3%
唐　珂	CNN 模型构造，RNN 模型的初步构造，训练模型，报告	33.3%
裴禹乔	报告，CNN+RNN 模型的完善	33.3%

参考文献

[1] HA D, ECK D. A Neural Representation of Sketch Drawings[J]., 2017. eprint: 1704.03477 (cs.NE).

[2] ZANG S, TU S, XU L. Controllable stroke-based sketch synthesis from a self-organized latent space[J]. Neural Networks, 2021, 137: 138-150. DOI: 10.1016/j.neunet.2021.01.006.

[3] YU Q, YANG Y, SONG Y Z, et al. Sketch-a-Net that Beats Humans[J]., 2015. arXiv: 1501.07873 [cs.CV].

[4] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90. DOI: 10.1145/3065386.

[5] HE K, ZHANG X, REN S, et al. Deep Residual Learning for Image Recognition[J]., 2015. arXiv: 1512.03385 [cs.CV].

[6] SCHUSTER M, PALIWAL K. Bidirectional recurrent neural networks[J]. IEEE Transactions on Signal Processing, 1997, 45(11): 2673-2681. DOI: 10.1109/78.650093.

[7] XU P, HUANG Y, YUAN T, et al. SketchMate: Deep Hashing for Million-Scale Human Sketch Retrieval[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2018. DOI: 10.1109/cvpr.2018.00844.

[8] COLLOMOSSE J, BUI T, JIN H. LiveSketch: Query Perturbations for Guided Sketch-Based Visual Search[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). [S.l.]: IEEE, 2019. DOI: 10.1109/cvpr.2019.00299.