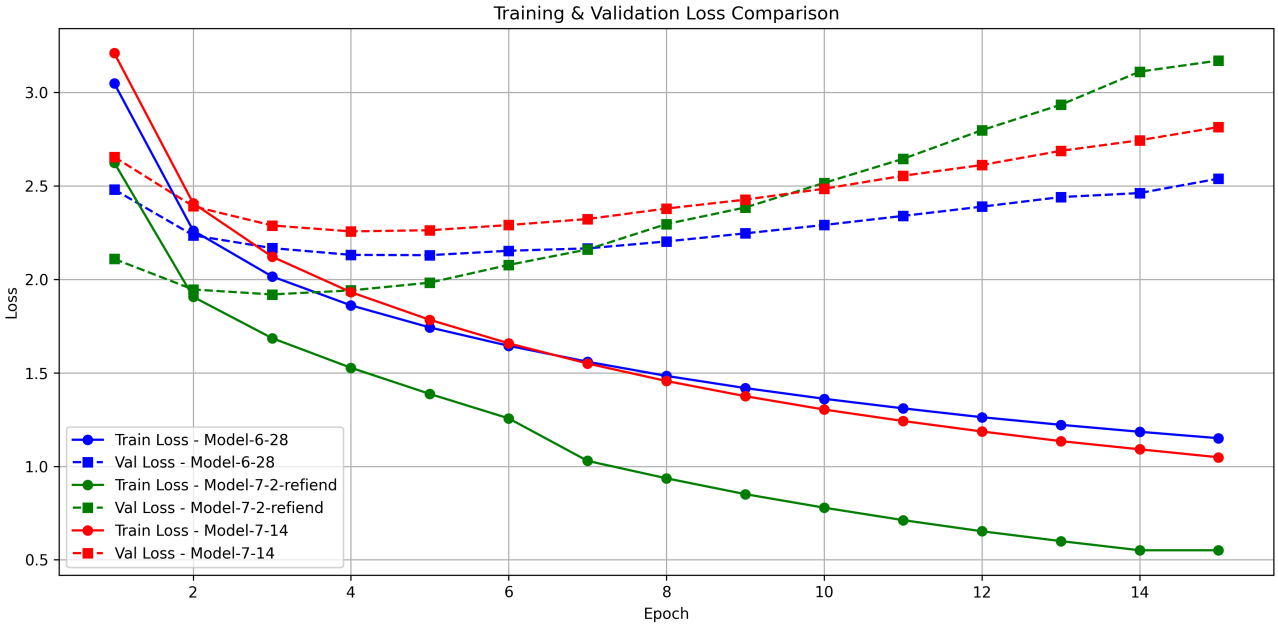
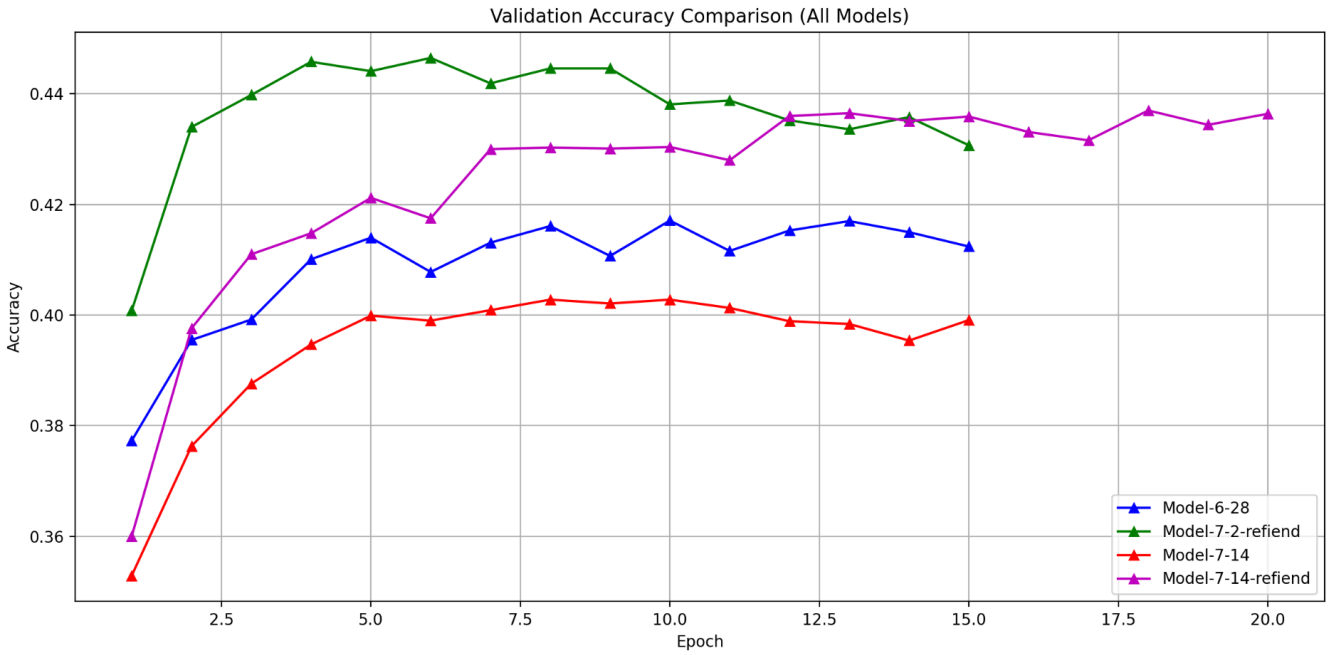
组会汇报week11

vqa模型比较

在过去一周，我把模型全部重新跑了一遍，使用相同的数据集、采取相同的评估方法(loss, acc)，采取相同的一套接口， 区别仅仅在于**模型的内部结构不一样**

- 训练数据集：vqa-v2 train
- 评估数据集： vqa-v2 val
- 训练轮次：15 epochs
- 单卡（batch_size= 512, num_workers=16) 训练时长从2-3h，每一个模型训练实际占用显存约5GB
  结果如下：





下面是这四个模型的不同结构：

> 下面是让AI总结的

模型一：平均词嵌入（Mean Pooling）

```
question_features =
self.embed(question).mean(dim=1)
```

将每个词嵌入后，取平均作为问题表示。

优点

- 实现简单，计算速度快
- 适合短文本

缺点

- 忽略词序与上下文关系
- 无法突出关键信息

模型二：LSTM 编码

```
self.q_lstm =
nn.LSTM(input_size=512,
hidden_size=512)
question_features =
self.embed(question)
question_features, _ =
self.q_lstm(question_features)
```

优点

- 考虑词序信息
- 较强语义表达能力

缺点

- 仅用最后一个时间步，可能信息损失
- 对长句不稳定

```
question_features =
question_features[:, -1, :]
```

用 LSTM 编码问题，取最后一个时间步的 hidden state。

模型三：LSTM + 注意力机制

```
self.q_lstm = nn.LSTM(input_size=512,
hidden_size=512)
self.question_attention_linear =
nn.Linear(hidden_size, 1)
question_features = self.embed(question)
question_features, _ =
self.q_lstm(question_features)
attention_scores =
self.question_attention_linear(question_features)
attention_weights =
torch.softmax(attention_scores, dim=1)
question_features = torch.sum(question_features *
attention_weights, dim=1)
```

LSTM 输出后，加权聚合每个时间步的隐藏状态。

优点

- 可关注问题关键词
- 提升语义建模能力

缺点

- 模型结构略复杂
- 对注意力参数敏感

> 模型三和模型四 是参考了论文SANs的注意力机制，均只使用了一层堆叠注意力；但是模型三的图像编码使用的方法是resnet18删去了最后一层，图形送入，得到512维度张量； 而模型四的图像编码采用的方法是删除resnet18的最后两层，借助的是其预训练好的CNN部分，和SANs结构更加贴近。

模型四：图文融合注意力模型（基于ResNet + 文本引导）

```
# 图像特征提取
self.cnn = ResNet18（去除全连接层）
v_I = CNN(image) → shape: (B, 512,
14, 14) # 问题编码
v_Q =
self.embed(question).mean(dim=1)#
注意力融合
h_A = fc1(v_I) + (fc2(v_Q) + b_A)
```

优点

- 图文信息融合充分
- 能动态关注图像关键区域

缺点

- 结构复杂，调参要求高
- 训练成本高于前三个模型

```
p_I = softmax(fc3(h_A))
v_I_a = 加权图像区域  # 融合表示
u_1 = v_I_a + v_Q → 分类器输出答案
```

使用视觉注意力机制结合问题内容对图像区域加权。

长句的vqa数据集

fsvqa数据集

论文： https://arxiv.org/pdf/1609.06657

Name： The Color of the Cat is Gray: 1 Million Full-Sentences Visual Question Answering (FSVQA)

图像依旧使用的是**COCO**，数据集提供了original和augmented两个版本

其中original版本的train data约为23w对QA对，augmented的train有66w对QA对（貌似augmented是origianl的超集？）

整理为统一格式后：

```
question_id:4870251
question:"Is there a shadow?"
answer:"Yes, there is a shadow."
image_id:487025
image_path:"/project/vqa/resources/vqa-
v1/images/train2014/COCO_train2014_000000487025.jpg"
```

lfvqa数据集

论文： https://arxiv.org/pdf/2408.06303

Name: Long-Form Answers to Visual Questions from Blind and Low Vision People

图像使用的是**VizWiz**（收集自视障人士实际拍摄的图片）的一部分（只选择了600张图像），每一个图像的对应的问题由多个大模型（LLava、BLIP、QWEN、GEMINI等）进行回复，

原数据集中大模型的回复通常比较长，~~貌似不太适合我们模型~~，我先对数据进行了整理，从大语言模型的多句回答中，选择直接回复的答案的句子，并且统一为统一格式，最后筛选得到了3129个QA对。

```
{'image_id': 16783,
 'question_id': 16783,
 'question': 'Can you tell me what kind of vehicle this is?',
 'answer': 'The vehicle in the image appears to be a utility service truck,
commonly used by contractors or maintenance departments',
 'image_path': 'image/16783.jpg'
}
```

Generative Model:

使用数据集： fsvqa

训练到第40个epoch，出现了过拟合特征：

```
--- Sample 1/5 ---
Image Path:          /project/vqa/resources/vqa-
v1/images/train2014/COCO_train2014_000000539562.jpg
Question:            Has this bed been made?
Ground Truth Answer: Yes, this bed has been made.
Model Generated Answer:  yes this bed has been made
------------------
--- Sample 2/5 ---
Image Path:          /project/vqa/resources/vqa-
v1/images/train2014/COCO_train2014_000000191639.jpg
Question:            Is it cold out?
Ground Truth Answer: Yes, it is cold out.
Model Generated Answer:  yes it is cold out
------------------
--- Sample 3/5 ---
Image Path:          /project/vqa/resources/vqa-
v1/images/train2014/COCO_train2014_000000289004.jpg
Question:            Are there clouds in the sky?
Ground Truth Answer: No, there are not clouds in the sky.
Model Generated Answer:  yes there are clouds in the sky
------------------
```

其他补充：

1. 线路二的模型（Start From Scratch）使用的数据集的统一格式和接口：
   - 格式为
   - {'image_id': 16783,
     'question_id': 16783,
     'question': 'Can you tell me what kind of vehicle this is?',
     'answer': 'The vehicle in the image appears to be a utility service truck, commonly used

by contractors or maintenance departments',
'image_path': 'image/16783.jpg' ‘question_type’: ''

}

- 接口为： loadData(train=False, num=-1, dataset_type= '')
- 使用方法： 1. 使用loadData函数导入数据，得到的数据即为上述格式的一个数组 2. 导入词表 3. 传入VQA_Dataset，在这一步会对question和image进行预处理（tokenize、编码等等） 4. 实例化Dataloader，并作为参数传入训练的函数

2. 吃了个西瓜，忘记要写什么了

Failed Attempt On Long Answer VQA

让AI写了一个模型，尝试使用上述整理好的数据集，进行训练：

训练了30个epoch的模型：（在train上的表现，过拟合，背住了答案）

```
--- Sample 4/10 ---
```

```
Question:            What color is the sea?
Ground Truth Answer: The sea is blue.
Model Generated Answer:  the <UNK> is black
-------------------
```

在val上的表现：已经完全是放飞自我，回答牛头不对马嘴了，只是学会了基本的语言规则，并没有学会看图和回答

```
--- Sample 2/10 ---
```

```
Question:            Is the pic taken from below?
Ground Truth Answer: No, the pic is not taken from below.
Model Generated Answer:  yes the time has it s been <UNK>
-------------------
```

中文vqa尝试

既然英文的成功跑起来了，那么理论上我只需要把词表换成中文词表，不就有了中文vqa模型?

理论存在，实践开始:

我用google的开源翻译模型，将fsvqa翻译成中文并且按照新的数据集重新构建了词表，训练了中文的vqa模型:

结果只能说 和上面那个英文的vqa模型 算是**卧龙凤雏、不相上下**

组会： 尝试使用预训练好的词嵌入模型和动态卷积