基于深度学习网络实现化学有机物骨架图像的有机分子式读取

1. **小组成员**

涂汉璋 邵惟至 肖铂 李京洋

GitHub: [img2inchi](https://github.com/itoshiko/img2inchi)

1. **大作业题目**

利用百万级别的有InChI标注的有机物的骨架式图片数据集，通过深度学习训练网络实现化学有机物骨架图像的有机分子式读取。

1. **研究背景**

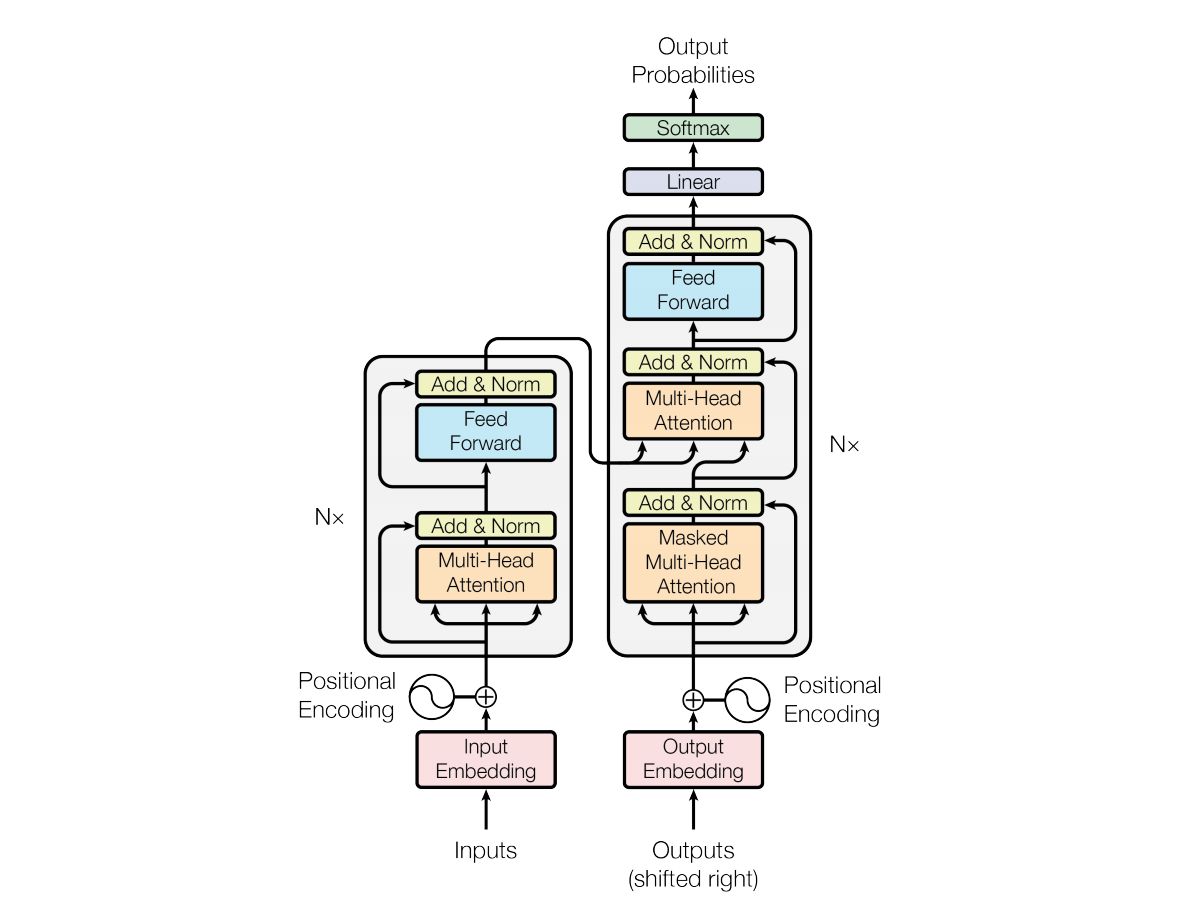
在技术进步的时代，有时最好、最简单的工具仍然是纸和笔。有机化学家经常利用骨架式（Skeletal formula），一种使用了数百年的结构符号，来绘制（有机物）分子。最近的出版物还用机器可读的化学描述（InChI）作注释，但是仍有大量扫描的文献无法用特定的化学描述来自动检索。利用机器学习，可以更快得将这些结构进行识别。

**四、基本思路**

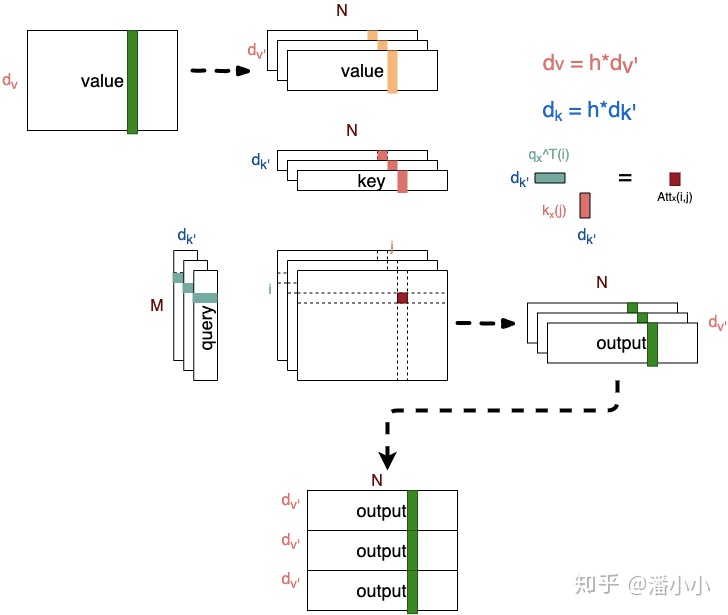
我们考虑使用两种模型来进行训练，下面将分别进行介绍：

1. 预训练ResNet + Transformer模型

**Transformer模型:**



标准的Transformer模型的框架图如上所示，由encoder和decoder组成。Encoder有6个相同的layer组成，每个结构如图所示，分别是multi-head attention mechanism和fully connected feed-forward network。

Attention是将query和key映射到同一高维空间中去计算相似度，而对应的multi-head attention把query和key映射到高维空间的不同子空间中去计算相似度。

Multi-head Attention在参数总量保持不变的情况下，将同样的query, key, value映射到原来的高维空间的不同子空间中进行attention的计算，在最后一步再合并不同子空间中的attention信息。这样降低了计算每个head的attention时每个向量的维度，在某种意义上防止了过拟合；由于Attention在不同子空间中有不同的分布，Multi-head Attention实际上是寻找了序列之间不同角度的关联关系，并在最后concat这一步骤中，将不同子空间中捕获到的关联关系再综合起来。

每一层经过attention之后，还会有一个FFN，这个FFN的作用就是空间变换。FFN包含了2层linear transformation层，中间的激活函数是ReLu。FFN的加入引入了非线性(ReLu激活函数)，变换了attention output的空间, 从而增加了模型的表现能力。

位置编码（Positional Encoding）是Transformer框架中特有的组成部分，补充了Attention机制本身不能捕捉位置信息的缺陷。Positional Embedding的成分直接叠加于Embedding之上，使得每个token的位置信息和它的语义信息(embedding)充分融合，并被传递到后续所有经过复杂变换的序列表达中去。

Decoder部分与encoder基本类似，多了一个attention的sub-layer。

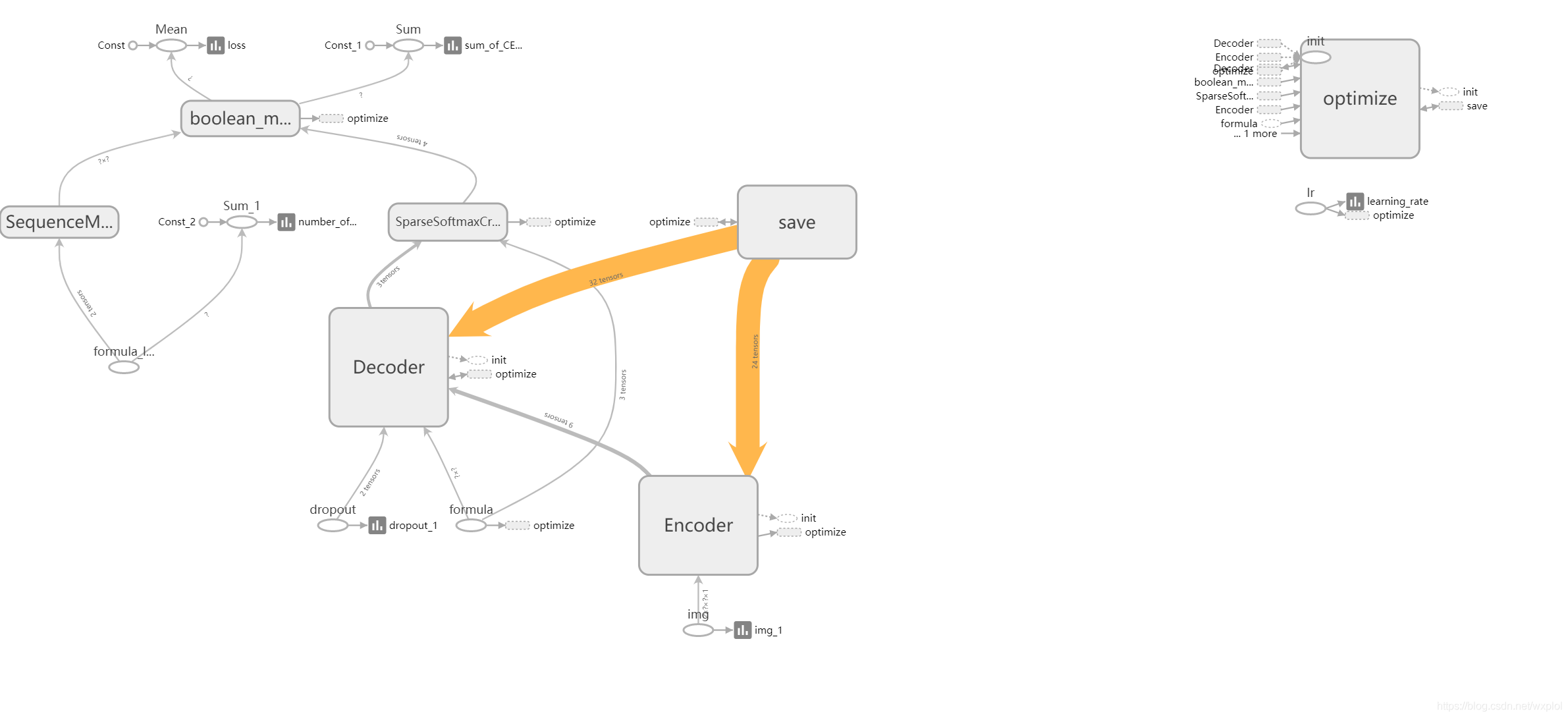
**ResNet:**

即在CNN的基础上增加残差连接以提高模型的稳定性，获得更好的表达能力。在本问题中，我们通过调用预训练的ResNet34模型作为图片的特征提取器，然后将提取得到的特征扁平化处理为一个序列输入到Transformer的Encoder中。ResNet34共有34层，提取的特征数为512，因此我们选定Transformer模型的维数为512.

1. Seq2Seq + Attention + Beam Search模型

为了提高训练效率和正确率，我们准备尝试这种模型，已经基本完成网络搭建，预计在中期之后进一步验证。这一部分代码我们参考了GitHub上的项目[LaTeX\_OCR\_PRO](https://github.com/LinXueyuanStdio/LaTeX_OCR_PRO/tree/master/model)。

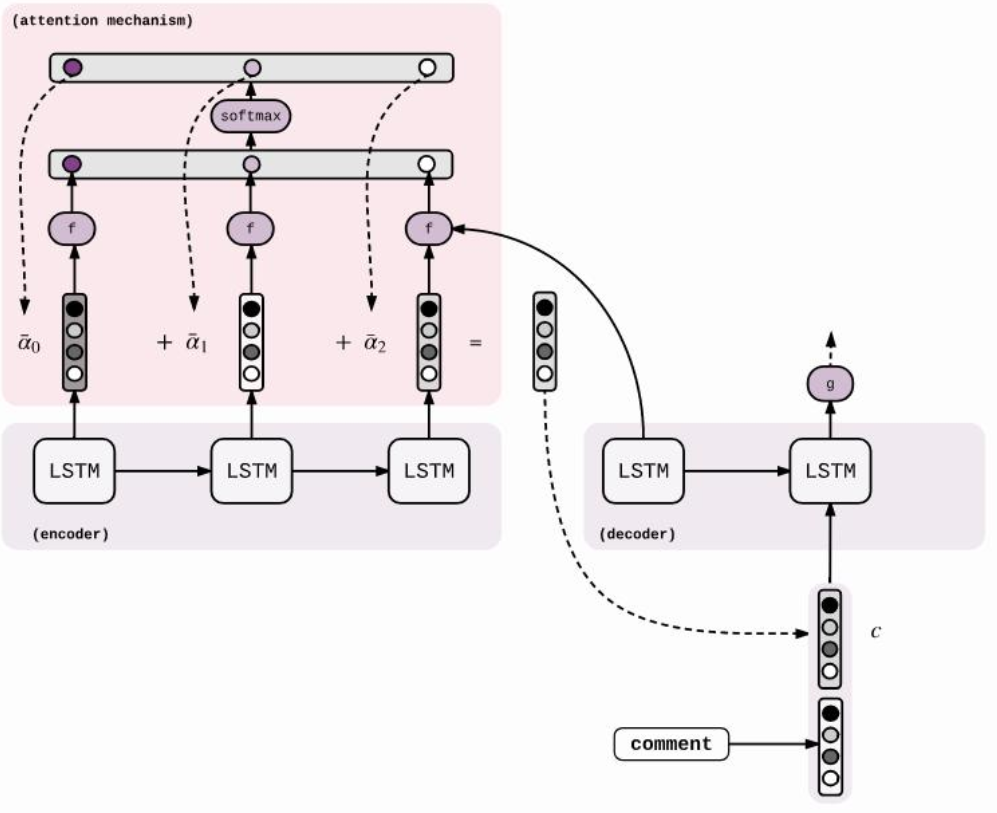
seq2seq属于encoder-decoder结构的一种，我们将CNN作为它的encoder，LSTM作为它的decoder。encoder负责将输入序列压缩成指定长度的向量，decoder则负责根据语义向量生成指定的序列。对于过长的分子式，单纯的seq2seq效果不好，所以我们引入了attention机制，在decoder阶段在每一步决定哪些分子式是重要的。在测试阶段，我们采用了beam search来决定选择最有可能的分子式。接下来进行详细介绍：



Encoder：采用传统的卷积神经网络，共6层，输出得到特征图，大小为N\*H\*W\*C，并使用reshape对其进行扁平化，得到[N,H\*W,C]的特征。

Decoder：在解码部分，首先对编码部分的最后一层进行初始化，将其转换为隐藏向量。采用全连接操作，得到新的向量：

接下来，我们将利用attention机制计算出一个上下文向量。对于encoder和decoder的隐藏状态，计算一个分数，，并根据这个分数得到权重，最后做点积相加，得到平均权重，即上下文向量。

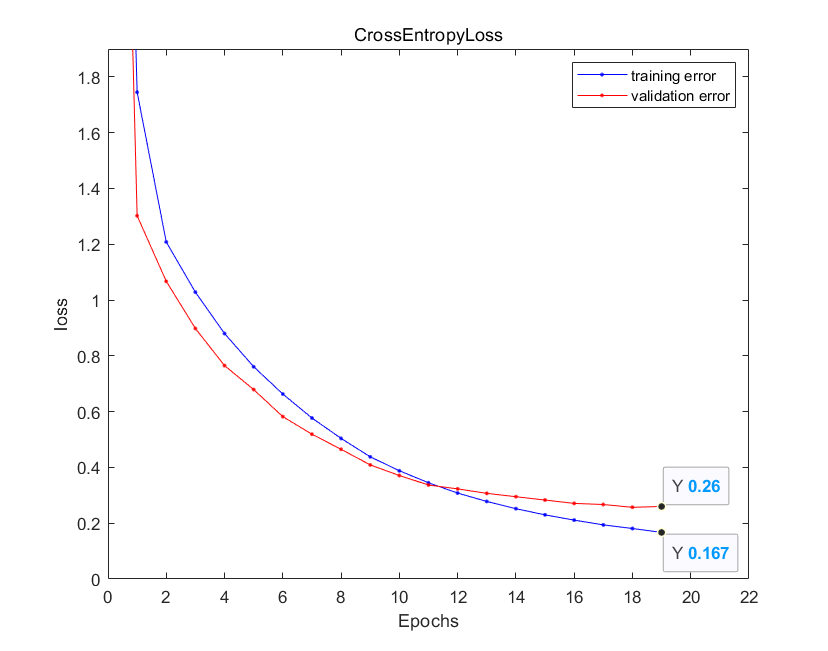


最后进行解码。将标签与上一时刻的隐藏状态传入解码器，得到当前时刻的输出，这时需要利用attention算出的上下文向量，将其与当前时刻的输出进行加权乘积，最后利用SoftMax得到概率p。

测试阶段，我们采用了beam search。这是相对于greedy search的一种改进。我们设置一种超参数beam size，记为k。每一个时间步长，都选取当前条件概率最大的k个分子式，其后的时间步长基于上一个序列，选取条件概率最大的k个分子式，最终结束后挑选出最优的结果。Beam search的搜索空间更大，结果一般也比greedy search要好。

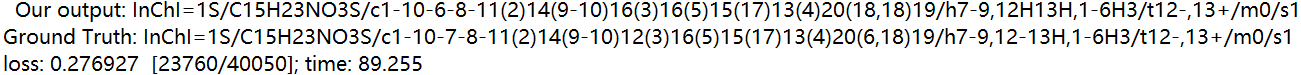
1. **训练结果**

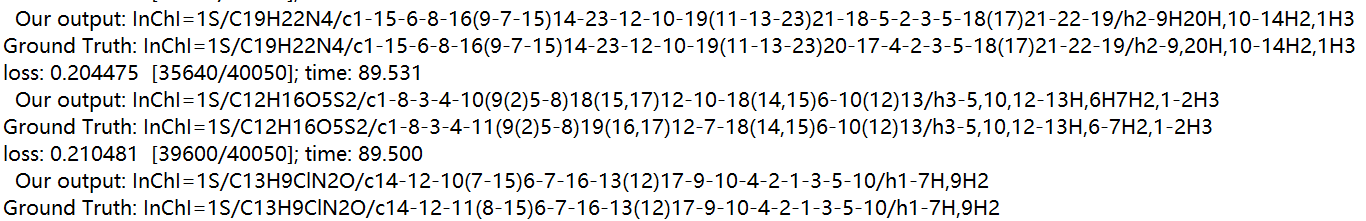
由于算力限制，我们目前只对transformer模型进行了训练。在训练中随机选择了40000个数据。在进行多个epoch之后，分别得到了training error的曲线和validation error的曲线。在19个epoch之后，它们分别下降为0.167和0.26，具体图像如图所示。从结果来说，考虑到数据的个数，这个结果还是比较理想的，如果算力提高，应该可以进一步优化。



如果epoch增加，可能会出现过拟合现象，validation error会增加。

部分训练结果如下图所示，我们可以看到现有的网络已经可以实现对部分图片十分精确的标准形式翻译。





1. **后续工作方向**
2. 完善模型搭建

目前我们准备尝试两种模型，具体已经在前述部分介绍，第二种模型仍然在搭建中。除此之外，transformer模型目前只在小样本上进行了训练，后续预计得到算法支持后，在大样本数据集上查看效果，并且进行参数的调整和一些细节的完善。

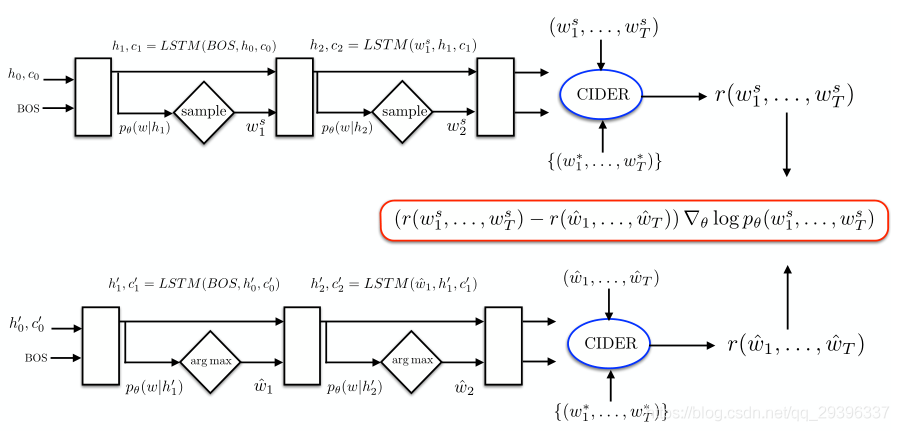
1. 对比两种模型效果

在搭建实现两种模型并调参完毕后，我们预计比较两种模型的效果，分析造成效果差异的原因，选择更加贴合分子式读取的模型，并针对差异进行优化。

附：一些可能的优化方向：

针对LSTM的优化：Self-critical sequence training(SCST)

在现有的模型中，我们还存在一些缺点，例如模型测试的时候利用的是自己生成的分子式，一旦出现缺点，就会导致误差累加，因此需要手段将它暴露出来。我们预计将采用SCST方法，即使用了自己在测试时生成的分子式作为baseline，利用测试阶段的reward signal直接优化Levenshtein Distance指标，流程如下：



在这个模型中，我们把LSTM看作智能体，每个动作之后，智能体得到一个reward，表示为r，因此训练目标就是最小化消极期望奖励：

Reinforce算法中梯度可以表示为：

这个式子可以推广到带有基线时的情况，并用单个蒙特卡洛抽样近似，得到梯度计算公式为：

如果采用SCST，就把当前模型推理得到的reward记为基线，也即在测试阶段使用greedy decoding生成的分子式得到的reward，得到梯度计算公式：

这样就避免了单独训练一个基线。未来的工作，我们决定尝试这样的方法，观察得到的结果是否会在准确性上有所提升。

1. 调参以及调研训练技巧

在我们的后期工作中，有一个很重要的环节就是调参。Transformer的超参有编码器、解码器层数，模型宽度（维数），多头注意力头数等等。Seq2Seq模型有LSTM堆叠层数，编码器维数（特征数），解码器隐藏层维数，甚至可以考虑调整和更改编码器（如换成ResNet）。

除此以外，还有训练过程中的最大学习率、warm up轮数等超参。

在训练技巧方面，我们调研到Transformer模型需要充分的warm up以及精细的学习率控制：

还有如调整batch\_size的一些技巧。(Transformer模型中，根据其他科研团队的实验，需要batch\_size尽可能大一些(到2~3k)，这样可以有效提高模型性能)

参考文献：

**[1] Xu, K., “Show, Attend and Tell: Neural Image Caption Generation with Visual Attention”, arXiv, 2015.**

**[2] Ranzato, M., Chopra, S., Auli, M., and Zaremba, W., “Sequence Level Training with Recurrent Neural Networks”, arXiv, 2015.**

**[3] Rennie, S. J., Marcheret, E., Mroueh, Y., Ross, J., and Goel, V., “Self-critical Sequence Training for Image Captioning”, arXiv, 2016.**

**[4] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., & Gomez, A. N., et al. (2017). Attention is all you need. arXiv.**

**[5] Khan, S., Naseer, M., Hayat, M., Waqas Zamir, S., Shahbaz Khan, F., and Shah, M., “Transformers in Vision: A Survey”, arXiv, 2021.**