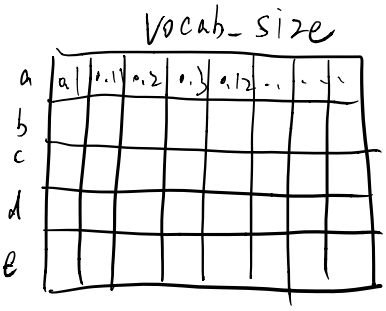
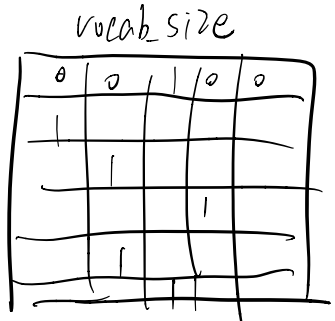
概率.



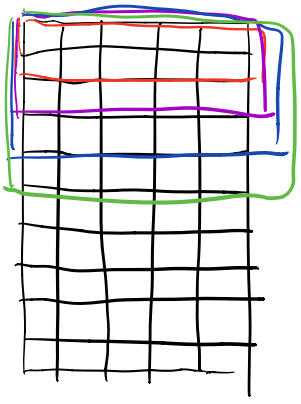
简单示意图



G oweput

seq_size

16



20

不同filter size

这里不考虑batch，假设矩阵shape为(20,16)

seq_size　　vocab_size

DIM=25 为feature数

Discrimintor

Input (20,16)

(20,25)

Conv(1)

filter_size=1

filter_input_dim=16

filter_oweput_dim=25

将概率垂数转换到语义特征矩阵上来.

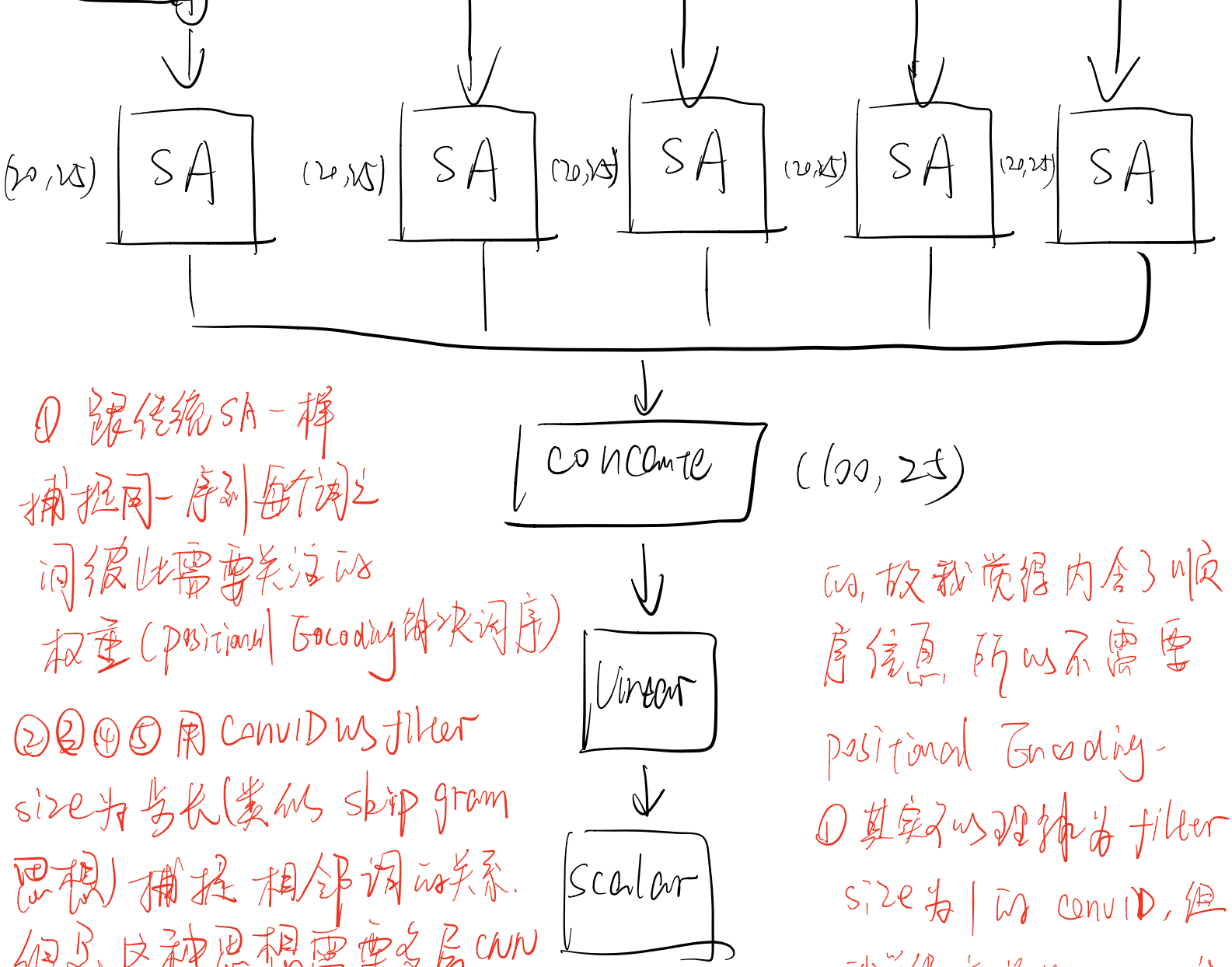
① positional Encoding

② Conv1D filter_size=2 filter in=out =16

③ Conv1D filter_size=3 filter in=out =16

④ Conv1D filter_size=6 filter in=out =16

⑤ Conv1D filter_size=5 filter in=out =16



① 跟传统SA一样捕捉同一序列每个词之间彼此需要关注的权重(positional Encoding给次词序)

②③④⑤用conv1D ws filter size为当长(类似skip gram思想)捕捉相邻词的关系.但是这种思想需要多层CNN叠加才能捕捉远距离特征.在后面加SA层用来弥补这个缺陷,可以捕捉类似消息帧中位置i到位置(i+filter size)部分的数据与位置i到位置(i+filter_size)的关系.这里filter取2,3,4,5涵盖大部消息头的长度. conv1D是根据stride大小滑动的,故我觉得内含了顺序信息,所以不需要positional Encoding.

① 其实这里的理解为filter size为1的conv1D,但我觉得应该比conv1D效果好.最后concate起来,保证filter size 1-5每个占的比例一致.

linear之后得到一个scalar表示概率.

Q1. SA是否需要multi-head机制?
head数似乎会影响长距离特征捕捉能力.但是我们的数据长度其实不长.