1.(a)

一个单词相比一个字符，其携带的信息（如词性、语义等）显然更多；一个单词由多个字符构成，字符的嵌入维度理应比单词少。

1.(b)

嵌入层 e_{word}；

卷积层参数 $\mathbf{W} \in \mathbb{R}^{f \times e_{\text{char}} \times k}$ 以及 bias $\mathbf{b} \in \mathbb{R}^{f}$；

highway 层及其 bias $\mathbf{W}_{\text{proj}}, \mathbf{W}_{\text{gate}} \in \mathbb{R}^{e_{\text{word}} \times e_{word}}$ $\mathbf{b}_{\text{proj}}, \mathbf{b}_{\text{gate}} \in \mathbb{R}^{e_{\text{word}}}$

注意：$f = e_{word}$

结果 $V_{char} \times e_{word} + f \times e_{char} \times k + f + 2 \times e_{word} \times e_{word} + 2 \times e_{word} \times e_{word} = 74156$

word-based lookup embedding model: 嵌入层：$V_{word} \times e_{word} = 12800000$

word-based lookup embedding model 含有的参数更多，倍数关系为 172 倍。

1.(c)

①1Dconvnet 的参数比 RNN 少；

②RNN 计算每个位置的上下文表示时，将（从左往右）获得的信息存储到一个定长向量中；1Dconvnet 利用多个 filter 来获取不同的特征，经过拼接可以获得不同长度的表示，可以使用不同的 attention head 以捕获特征。

1.(d)

max-pooling：优点：保留了最显著的特征；缺点：遗弃了大部分信息。mean-pooling：优点：将所有信息都保留了；缺点：稀释了较为明显的特征。

1.(h)

验证输出形状是否满足要求

```
high=Highway(5)
input=torch.randn(4,5)
pred=high(input)
assert(input.shape==pred.shape)
```

1.(i)

验证输出维度是否满足

```
cnn=CNN(50,4)
input=torch.randn(10,50,21)
assert(cnn(input).shape==(10,4))
```

3.(a)

只有 traducir:5152 和 traduce:8764 存在。

因此 word-base 时，只有它们才有对应的词向量，其他都是<UNK>进行嵌入，训练会有偏差。character-base 没有 OOV 问题，可以得到每个词的嵌入。它们的前几个字符大多是相似的。因此 CNN 的 filter

会学习到这几个字符具有相似的意思，后面几个不同的字符表示不同的形式。

3.(b)

i) ii) 使用 https://projector.tensorflow.org/，可视化地比较 Word2Vec 和该 NMT 训练出的词向量即可。

iii)Word2Vec 对词义相似性建模；CharCNN 对词形相似性建模

Explain：Word2Vec 关注语义，认为语义相似的词具有相似的上下文；而 CharCNN 通过基于 window 的特征提取建模，因此结构上相似的单词在特征空间里距离更近。