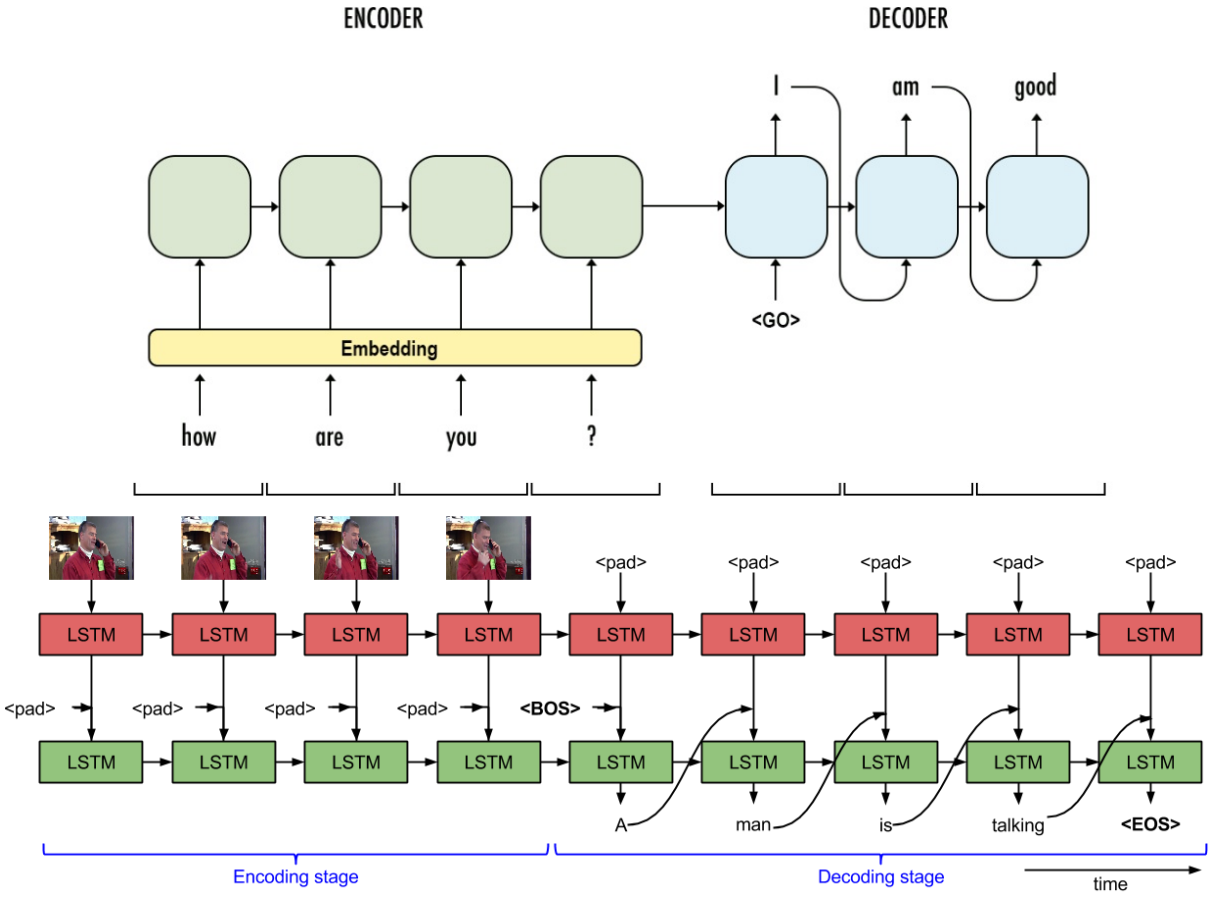
2-1

- Model description (3%)
  - Describe your seq2seq model



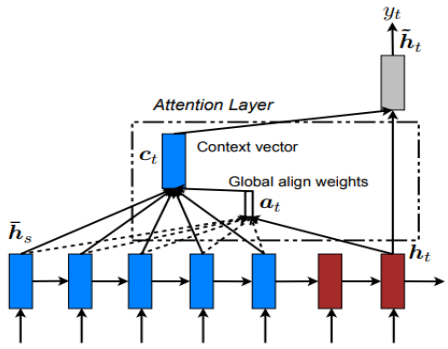
在這次作業中使用了幾個不同套件實做seq2seq model，有s2vt和其他幾種不同的架構。基本架構使用lstm做為encoder和decoder，將encoder最後輸出的context vector和state傳入decoder，幾種不同的變型是加入了attention、使用teacher forcing、把每個timestep output取argmax或masking、加入不同regularizer等，以下對於每種會有更詳細的說明。

- How to improve your performance (3%)

  (e.g. Attention, Schedule Sampling, Beamsearch...)
  - Write down the method that makes you outstanding (1%)
  - Why do you use it (1%)
  - Analysis and compare your model without the method. (1%)

以下將列出我嘗試的幾種方法並分析，然而根據實驗結果觀察，bleu無法有效呈現出文法的優劣，當bleu很高時文法也可能很糟，因此以下在test set上的實驗結果將文法和bleu分開做討論(train set都表現很好)，就這個的現象而言，我認為很可能是overfitting所造成的結果。

1. 使用 pretrained embedding
   在這個實驗中使用了word2vec(GoogleNews-vectors-negative300.bin)，然而結果和使用one-hot差不多。
2. Spelling correction
   為了使用pretrained embedding發現了文本中有不少錯字和非英文字母，因此過慮了一些字並將錯字使用edit distance 1從word2vec model中選取最有可能的字做修改，這個方法可以明顯改善錯字的情形但對seq2seq model準確率沒有明顯進步。
3. Attention
   參考自*effective approaches to attention-based neural machine translation*中的global attention，將encoder每個timestep output經過attention layer後得到alignment score。實做結果沒有發現特別的進步。



4. Teacher forcing
   使用teacher forcing是我觀察到最能夠有效提升文法，並且loss也會很快收斂到很小，若沒有teacher forcing，雖然可以達到很高的bleu(0.81)，但會有輸出的結果文法非常糟糕的情況。
5. Schedule sampling
   Schedule sampling比起只使用teacher forcing取得的bleu更好一些，並同時維持良好的文法，是實驗中看起來效果最好的(0.699)。
6. Beam search
   根據實驗結果，使用beam search會讓bleu稍微下降，但仍然會產生通順的句子。將search path提高到3條或以上的時候，bleu下降更多(0.6以下)並出現詭異文法的句子。
7. Masking
   在decoder做sampling的時候將上一步timestep sample到的結果的機率變0可以有效提升文法(可讀性)和bleu，比起postprocess把輸出句子連續相同字拿掉來的效果更好。
8. Noise layer, Dropout
   model在train set上表現很好，幾乎可以達到100正確率，然而test set上卻越來越差，因此使用這些方法防止overfitting，雖然收斂會比較慢，但最終結果都有明顯提升。
9. char-based model
   我嘗試使用char-based model發現其實是不錯的，雖然bleu偏低(~0.56)，但句子非常通順，猜測因為char-based會拼錯字，因此計算bleu會偏低，

於是加入了修改錯字之後上升到了(~0.59)，ex. a merson is riding a micycle → a person is riding a bicycle。

10. 篩選caption
    根據打聽，移除太長太複雜的句子有助於模型準確率上升，然而就自身實驗結果來看準確率反而下降，猜測可能是因為原先模型為防止overfitting中加入很多dropout和nosiy layer有關，資料數量減少影響較大，但並未做實驗證實。

- Experimental results and settings (1%)
  - parameter tuning, schedual sampling … etc

parameters setting	
LSTM hidden size	256
batch size	64
rmsprop learning rate	0.001
dropout	0.5
vocab size	6060

一開始使用64做為hidden size就可以通baseline，增加到256有稍為進步，再繼續增加hidden size就沒有觀察到進步了。

results	**bleu**
base model	~0.66
teacher forcing	~0.68
schedule sampling	~0.7
beam search	~0.63
char-based model	~0.59

這裡大約列出bleu score，特別注意到若不使用teacher forcing 不停的train ， bleu可以達到0.81的，但是文法很糟糕。

分工表：
2-1: 蔡昀達