1.

(d)

Training data is very small. Because neural network usually has lots of parameters, and if there is not enough data, neural network will suffer overfitting.

(g)

i.

input gate: controls whether the input is passed on to the memory cell or ignored

forget gate: controls whether the recurrent input is passed on to the memory cell or ignored

output gate: controls whether the current activation of the memory cell is passed on to the output layer or not;

the memory cell is linear, so its gradient doesn't vanish;

ii.

cluster using the hidden states

iii.

build a lexicon to record the tagging relation with words. Final softmax should add the weight of most likely tag in the lexicon.

2.

(a)

We seek parameter values (i.e., vector representations for both

words and contexts) such that the dot product vw \_ vc

associated with \good" word-context pairs is maximized.

A(W): the loss of English context prediction

B(V): the loss of French context prediction

C (W, V): The word alignment ratio between two language embedding.

crossentropy

(b)

Balance the weight of each language. For example, in case the training data imbalance will cause model performance decreasing.

(c)

C = f(v) . g(w)

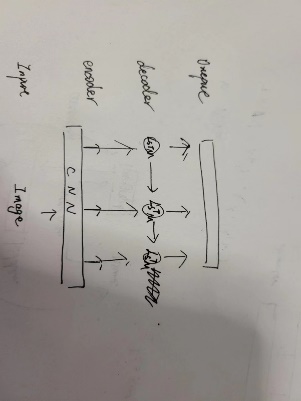
C = cos(f(v), g(w))

(d)

We could add layers between g(w) and g(v) to achieve document representation, then align the two document level representation.

3.

(a)



Encoder-decoder:

Encoder use CNN to capture the representation of image

Decoder use LSTM to generate sentence based on the decoder output and whole encoder representation.

(b)

Minimize the cross entropy loss

Stochastic gradient descent

(c)

Depending on the decoder output, we map each state by softmax function to generate the highest probability word corresponding to that representation.

The overall NLL is not the highest.

Instead of greedy search, we use beam search.

(d)

Using BLEU to measure the overlap between system output and human reference.

Ad: automatically metric, save resource, balance the adequacy and fluency.

Dis: rely highly on reference quality

Human evaluation and rating.

Ad: the evaluating result will be very reasonable

Dis: need a lot of resource (time, human), bias may occur due to the human evaluator.