Research papers

Show and Tell

Bleu-1 score

I input
$p(S|I)$ likelihood (S given I)
of producing a target sequence words

$S = \{S_1, S_2, \ldots\}$

Encoder RNN => feature vector

Replace encoder RNN with CNN

CNN -> image encoder of last hidden layer

Ranking descriptions:
Description close to an image in a vector space

RNN starts with image processed by CNN

Joint embedding

Encode variable length input into fixed dimensional vector

$$\theta^* = \arg\max_\theta \sum_{(I,S)} \log p(S|I;\theta)$$

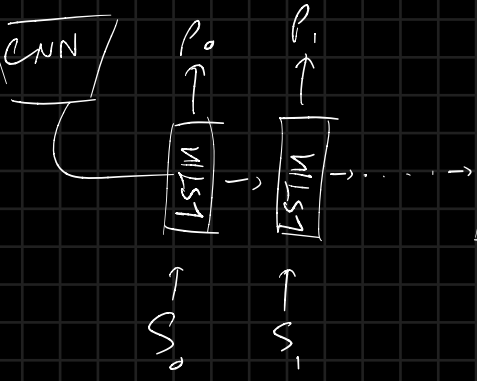
θ = parameters of model
I = image
S = correct transcription

① $\log p(S|I) = \sum_{t=0}^{N} \log p(S_t | I, S_0, \ldots, S_{t-1})$

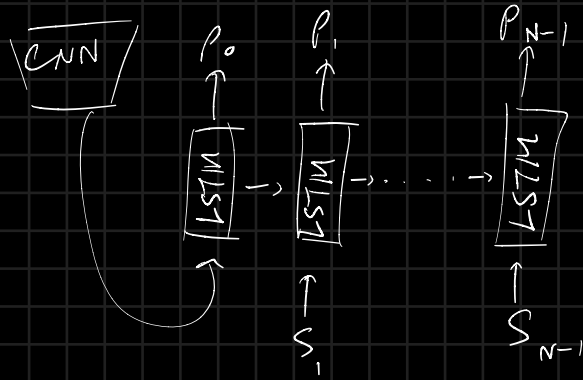
during training,

one example (S, I)
optimizer ①

LSTM net



or...



one hot vector S_t of vocab_size

start & stop word.

$t=-1$, image is input

only feed once at start -> according to results.

Loss:

$$L(I, S) = -\sum_{t=1}^{N} \log p_t(S_t)$$

Correct word at each time step

wrt all parameters

Sampling from model, until end token

BLEU Score

Use a pre-trained model on Image Net

Dropout ✓ for over fitting prevention

512 dimension embedding (probably hidden layer)

Real Time American Sign Language Recognition with Convolutional Neural Networks

Google Net

ILSVRC 2012 Dataset

Obtain video of user signing

Classify each frame to a letter

Reconstructing and displaying the most likely word from classification scores

Considerations:
- Environment (lighting & camera)
- occlusion (out of view)
- Sign Boundary Detection
- Co-articulation

What heuristics to use?

Hand cropping

Joint detection? coordinates of joints

physical component

Image processing per frame

$$\text{Loss} = \frac{1}{N}\sum_{i=1}^{N} -\log\left(\frac{e^{f_{y_i}}}{\sum_{j=1}^{c} e^{f_j}}\right)$$

$$f_j(z) = \frac{e^{z_j}}{\sum_{k=1}^{c} e^{z_k}}$$

Bottleneck: one frame per second

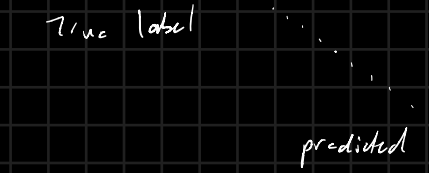
Let the user move onto the next letter

Surrey University Finger Spelling
Massey University

Datasets

Make horizontal flips of image

confusion matrix:



Reinitialize classification layer (Pre-trained model)

Can easily get confused if letters look the same.

Contrast adjustment
Background subtraction
Cropping
use CNN to locate the hand