A deep neural network for image captioning

Natural language descriptions of images

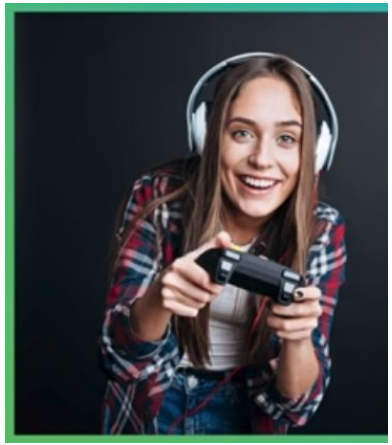
Thomas John

Overview

1. Problem description
2. Data
3. Network overview
4. Reading the image
5. Producing the caption
6. Results
7. Parallels to machine translation
8. Parallels to automatic speech recognition

Can a machine describe an image?

Input



Output

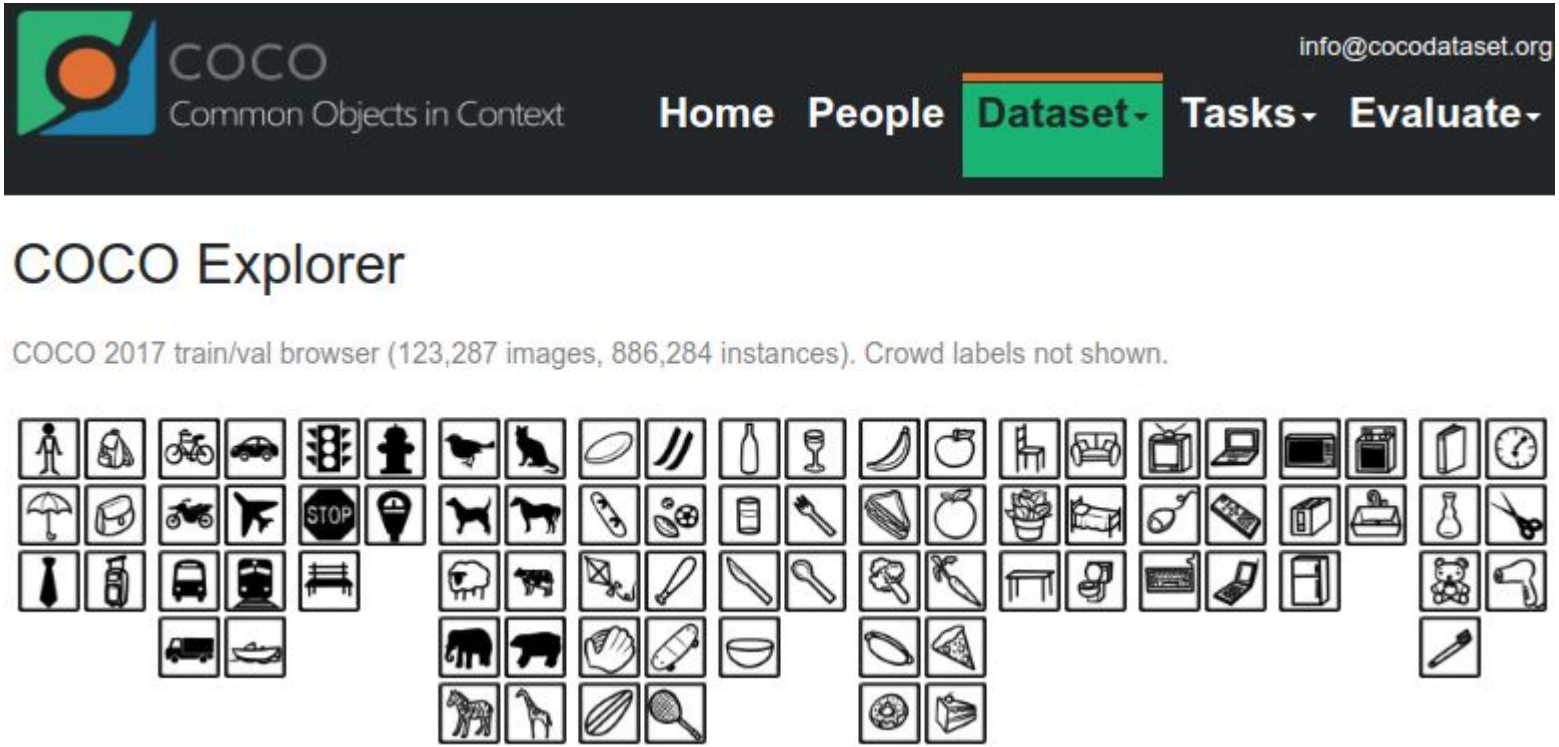
“A person playing a video game”

“Construction worker in orange safety vest working”

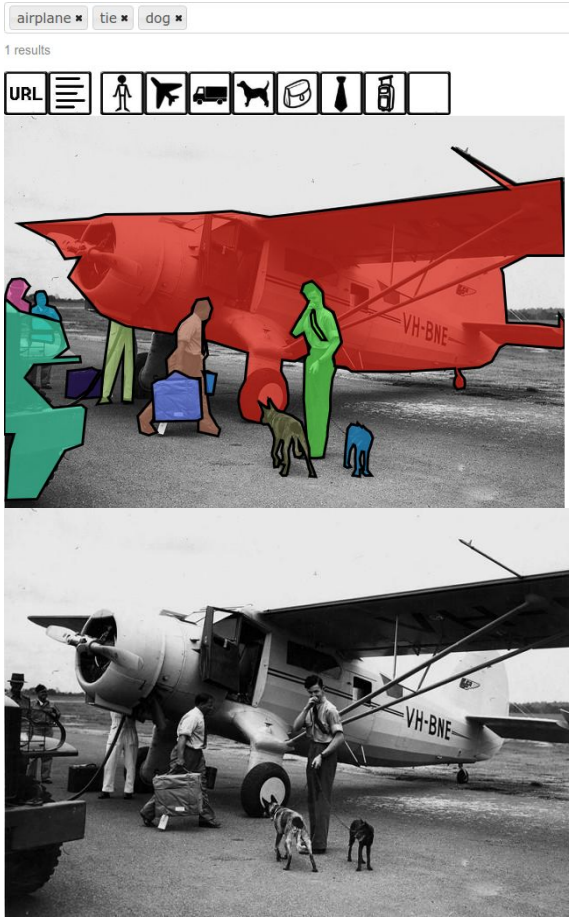
Some uses:

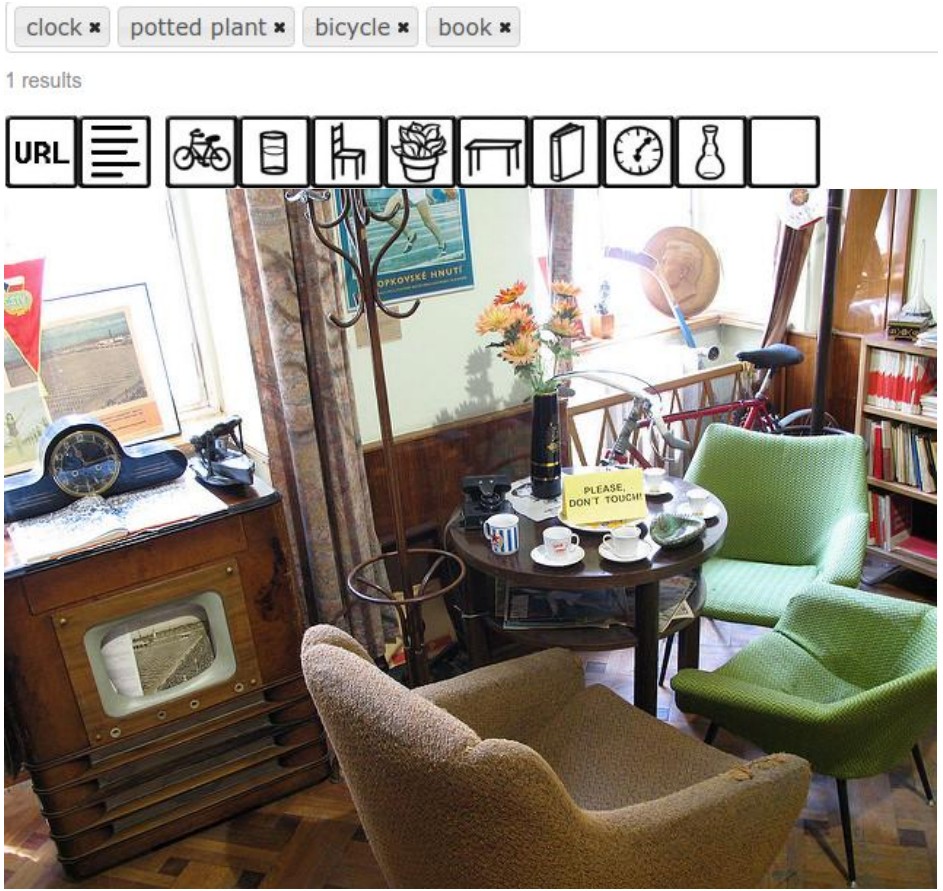
- Curate content
- Label images so they can be searched with text
- Can be used to describe images to blind persons

Dataset: MS-COCO



86 types of objects labelled





Dataset has multiple captions per image

1. men preparing an old prop plane for a trip.
2. man with suitcases preparing to board small old time antique plane.
3. a black and white photo showing a man with two dogs on leashes in front of a plane.
4. a man stands next to plane and holds two dogs on leashes.
5. a man standing next to a small airplane with two dogs.

1. a living area with a number of chairs
2. a group of chairs sitting around a table.
3. the room is crowded with many things including chairs, a bicycle, and a table with cups on it.
4. the furniture is posed in the room with a sign that says do not touch.
5. there is a small table with tea cups and three chairs around it

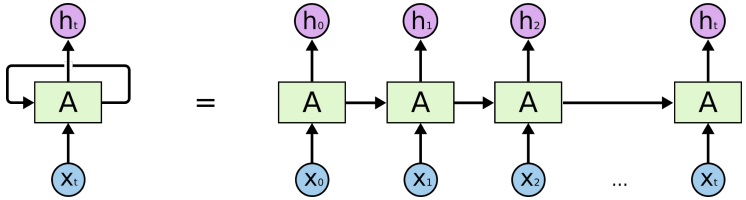
Non-unique “labels” -- appropriately design training loss and evaluation on test

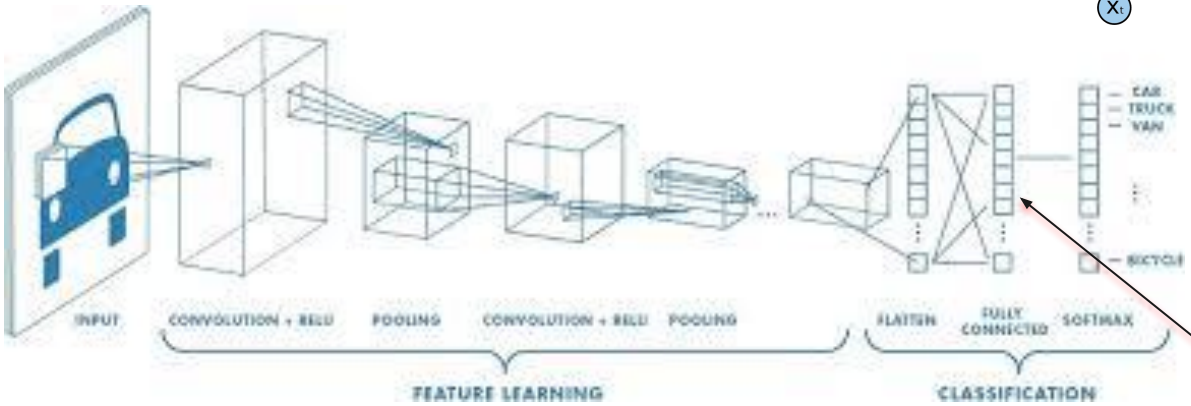
Convolutional and Recurrent Neural Networks

CNNs

- Excel in preserving the spatial structure in images
- Allow for large deep networks because of parameter sharing

RNNs Are great for ingesting or producing sequential data (*e.g.*, a sentence)





Encoder: CNN encodes the image
Decoder: RNN decodes into a sentence

This could be the connecting layer

Idea: Connect the CNN to the RNN and train on the captions

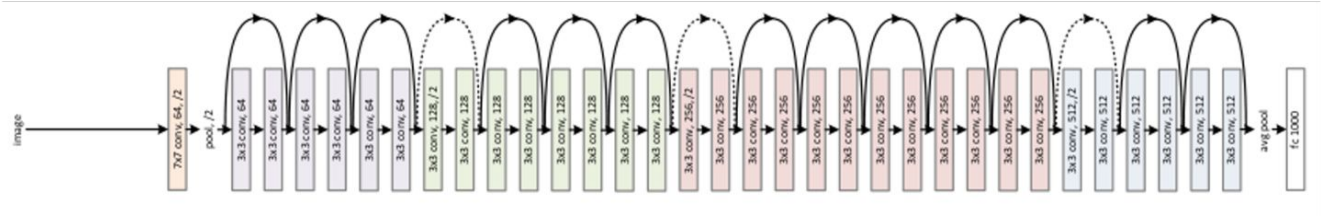
Image encoder - convolutional

Options:

A. Build a deep CNN
   a. Train from scratch - each run could take a lot of computational expense
   b. Network design choices and hyperparameter search costs can be very high
B. Use transfer learning ✔

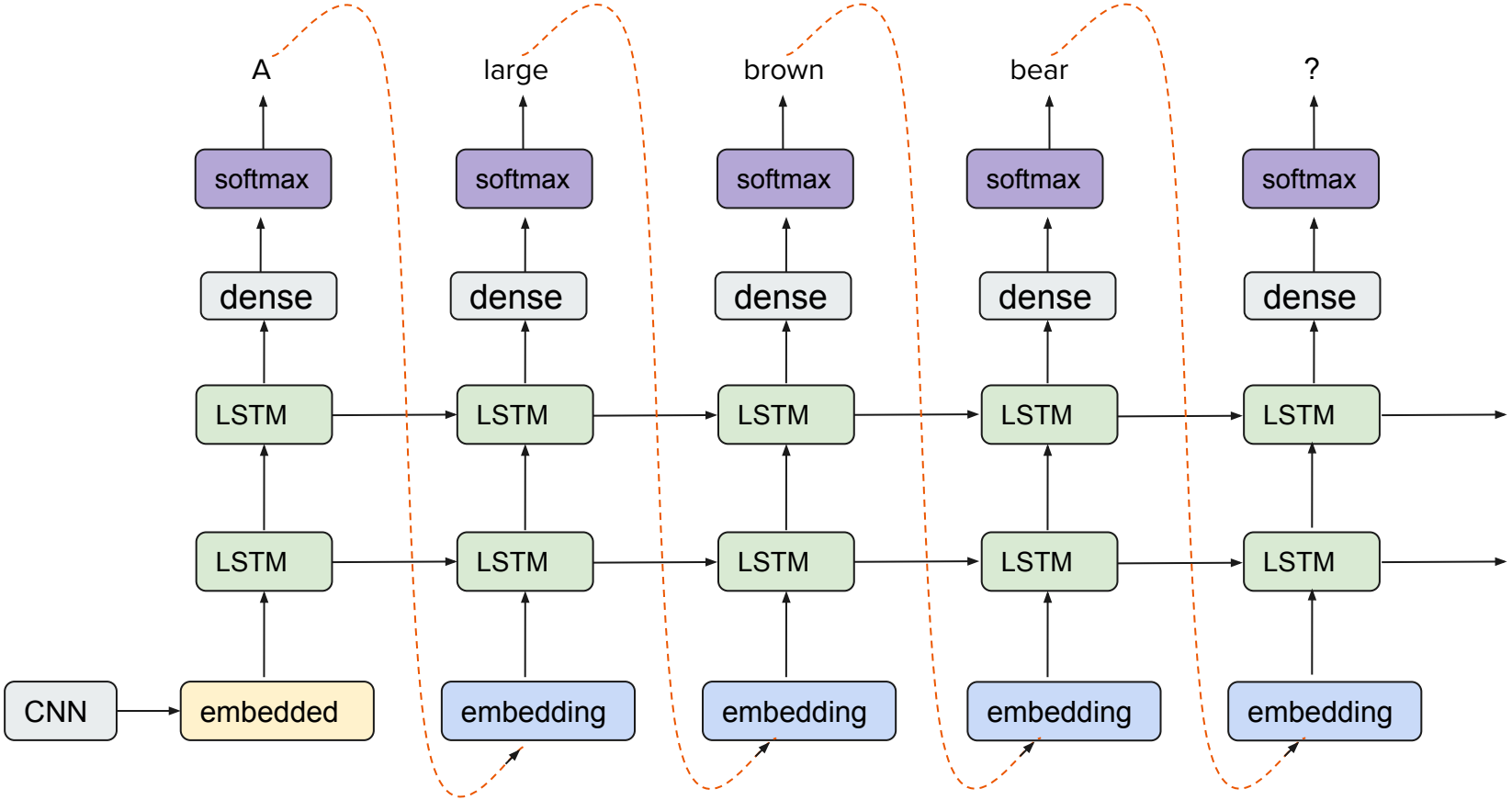
Transfer learning: Reuse parameters (weights, biases) intensively trained on a related problem to a large extent and only train a small part of the network on the problem at hand

Resnet50: A residual neural network - very deep neural net which can transmit without degradation. Full architecture visualization here.



Pre-trained on ImageNet (14+ million images), problem: classification into 1,000 classes. [Kaiming He, MIT license]

Decoder - LSTM with softmax output



Model details

Model built in PyTorch

Hyperparameters:

Embedding size: **300**
Hidden layer size (inside the LSTMs): **256**
Batch size: **512**
Num_epochs: **20**
Dropout: **20%**

Loss: **Cross Entropy Loss**

Optimizer is Adam
Learning rate: **0.001**
Beta1: **0.9**
Beta2: **0.999**
Epsilon: **1e-8**

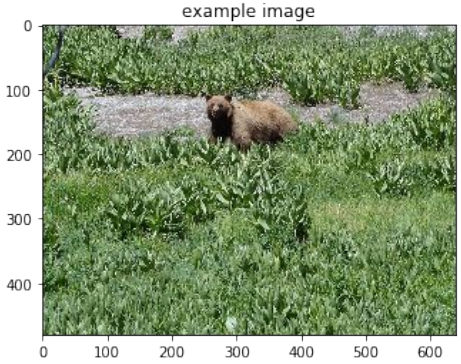
Hardware:
2 X 11 GB GPU Memory (Pascal architecture)
Clock 1569 MHz - Compute capability 6.1
32 GB memory
6-Core Intel i7-6850K CPU with 40 PCIe lanes

Batch normalization added

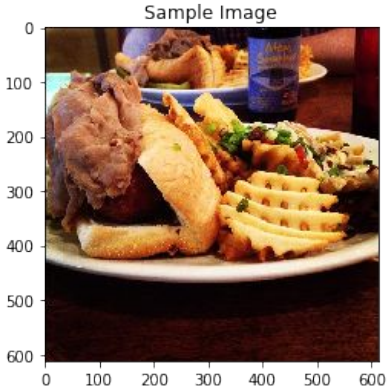
Regularization: Dropout

Code on Github:
https://github.com/gotamist/vision/tree/master/image_captioning

Successful predictions

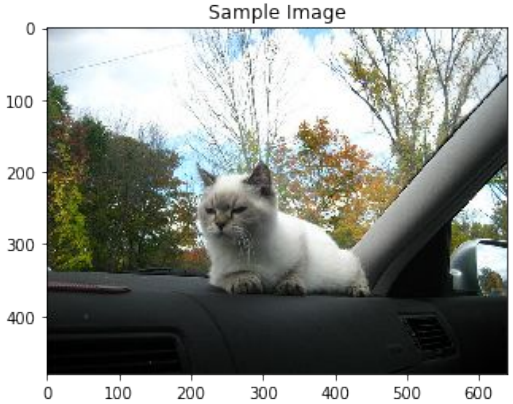


a large brown bear walking across a lush green field



a close up of a plate of food with a sandwich and a drink

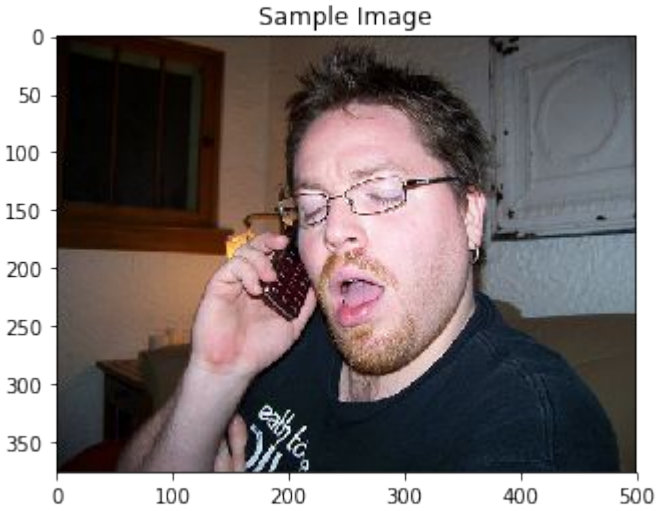
a cat sitting on the hood of a car



a group of people standing around a table filled with food

Not-so-successful predictions

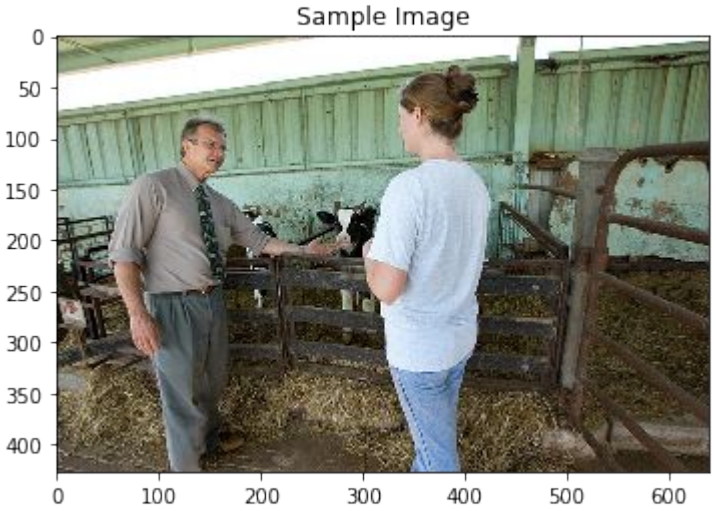
“a man brushing his teeth with a toothbrush”



Comments

- No toothbrush at all
- On the other hand:
  - Open mouth
  - About the right hand position

“a man standing next to a woman on a wooden bench”



Comments

- No bench
- Got man and woman right - obvious?
- Standing *on* a bench?

Artefacts of the training data

Scoring the generated caption with BLEU

BLEU: Bilingual Evaluation Understudy (Papineni et al, 2003) is the most widely used metric to evaluate translations

The idea is to compare outputs of a machine with human generated reference descriptions.

Bleu_1 corresponds to *adequacy* and higher order Bleu scores correspond to *fluency*.

Here, with 20 epochs of training, the 1-gram BLEU score of 0.663 on the validation set. Comparison with performance of other models from paper by Xu et al (Bengio group), 2016 below:

Full report of metrics

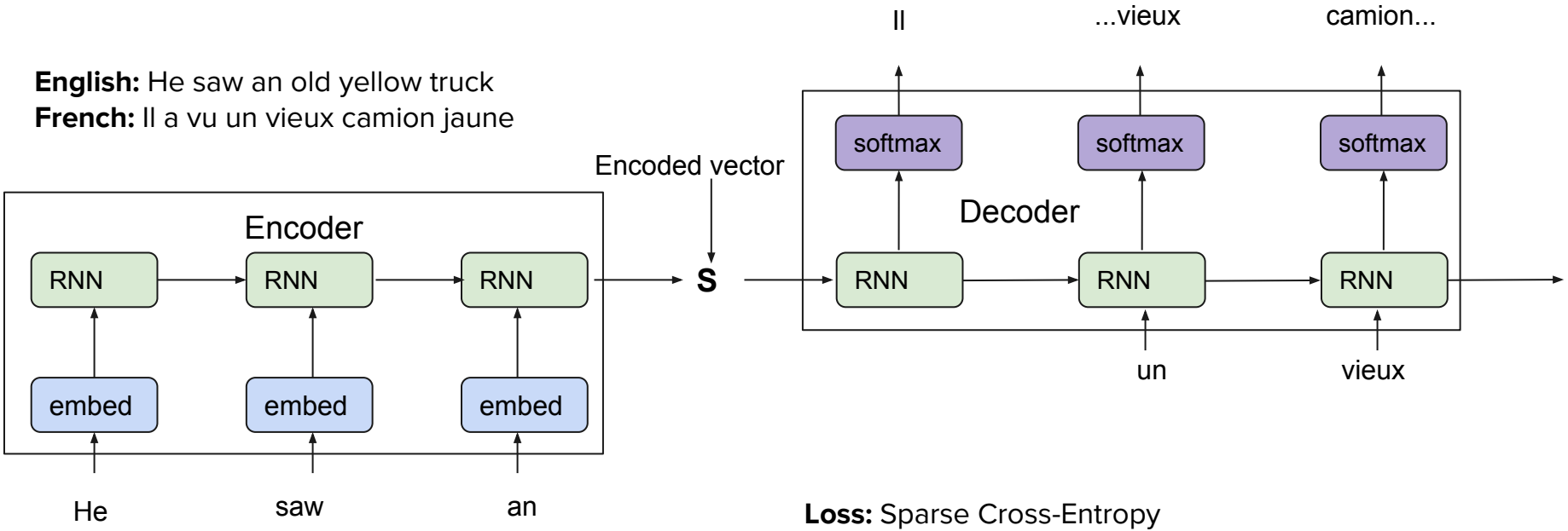
Bleu_1: 66.3
Bleu_2: 48.6
Bleu_3: 34.2
Bleu_4: 24.2
METEOR: 21.9
ROUGE_L: 49.1
CIDEr: 66.2

	BLEU				
Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
BRNN (Karpathy & Li, 2014)°	64.2	45.1	30.4	20.3	—
Google NIC†°Σ	66.6	46.1	32.9	24.6	—
Log Bilinear°	70.8	48.9	34.4	24.3	20.03
Soft-Attention	70.7	49.2	34.4	24.3	**23.90**
Hard-Attention	**71.8**	**50.4**	**35.7**	**25.0**	23.04

Similarity to the Machine Translation problem

Simple encoder-decoder architecture for translation

English: He saw an old yellow truck
French: Il a vu un vieux camion jaune

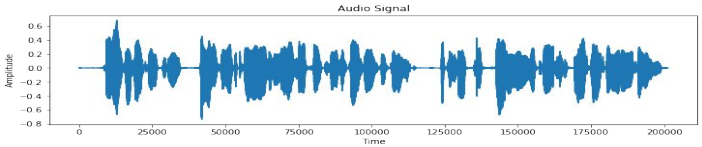


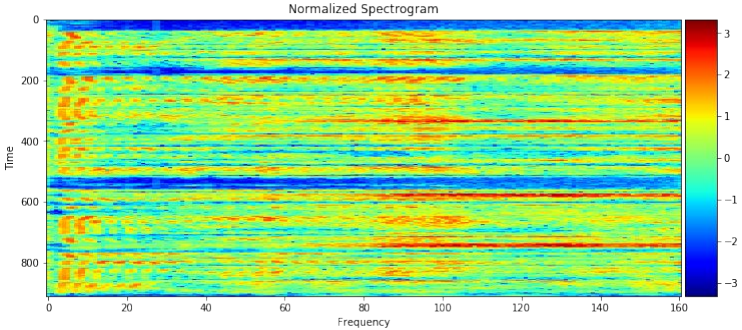
Loss: Sparse Cross-Entropy
Next level models

- Birectional
- Attention

Code on Github: **https://github.com/gotamist/nlp/tree/master/2_machine_translation**

Parallels with the speech recognition problem





Phonemes:
HH__EE_LL__OO
HHHHEL__LLOO
H_E_LL_LL__OO

Output
“HELLO”

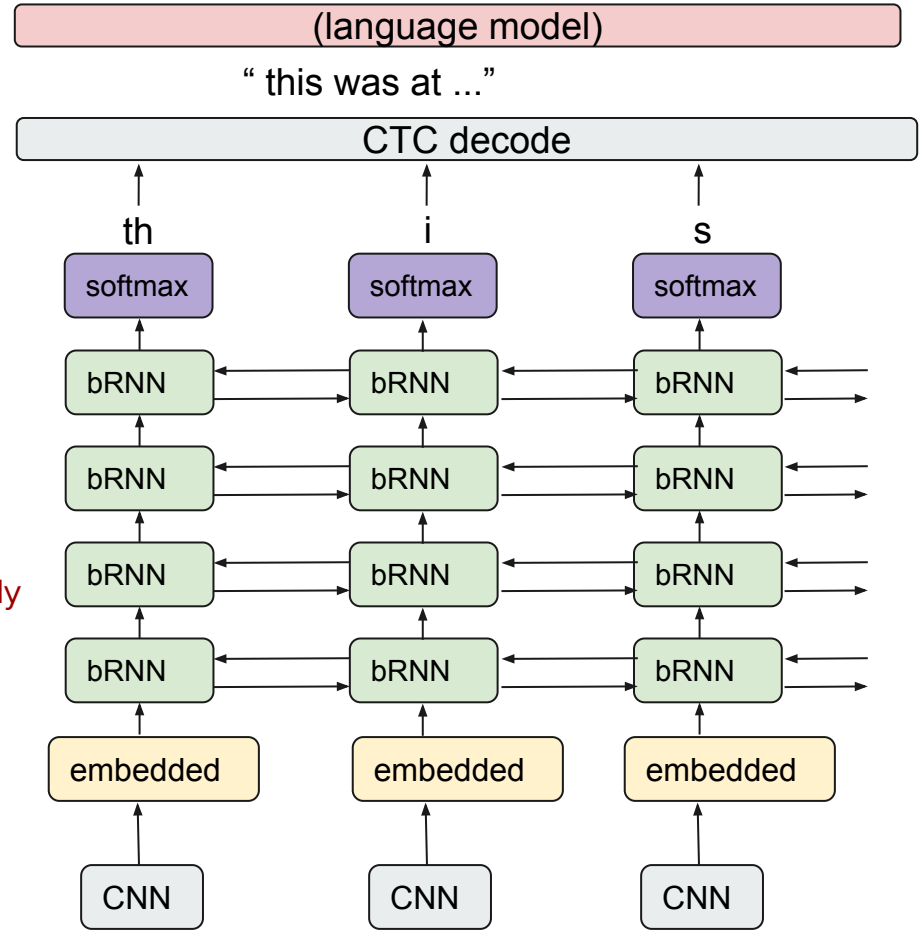
Take MP3 as input and produce the text of the sentence spoken
Data: Librispeech ASR corpus

- Voice clip (mp3) converted to spectrogram or MFCC features (Mel Frequency Cepstral Coefficients)
- Time is discretized into intervals (say 30 ms)
- Identify the phonemes that were spoken by looking at the frequency spectrum
- **CTC loss** (Connectionist Temporal Classification, Graves, ICML 2006) is used for training against the true sentences that were spoken

Code on Github: **https://github.com/gotamist/nlp/tree/master/3_vui_speech_recognizer**

Speech recognition problem - similarities & differences

- For each 30 ms interval, a 1-dimensional CNN converts the spectrogram into an embedding which is fed to the bidirectional RNNs
- Final output through a dense layer activated with a softmax over the phonemes
- CTC-decoded to get the final sentence



Sometimes, **word-boundary** issues arise in speech

Examples

True: this was at the march election eighteen fifty five
Prediction: this was at the march election aightemficty five

True: I address him in Italian and he answers very wittingly
Prediction: i adres him minitalion and he answers vering whitingly

<u>Often, a language model is still needed</u>
I’m working with **kenlm**
For matches, trying string-neighborhoods in

- Levenshtein, Dolgopolsky, Metaphone etc

Thank you

Image credits

RNN model: Colah's blog: http://colah.github.io/posts/2015-08-Understanding-LSTMs/

CNN model: Still from a mathworks video

Resnet50: https://www.kaggle.com/keras/resnet50/home

MS-COCO: http://cocodataset.org/#explore