Deep learning and chatbot

Le LI

iAdvize & Université d'Angers

Meetup à Pau, 03/05/2017

Outline

Outline

iAdvize

A conversational commerce platform,

Figure: logo

- object: help clients to increase their conversion rate and customer's satisfaction.
- ways: detect valuable visitors on the internet, social media etc.
- platform: integrate client's website, social media, apps etc.
- channels: chat, call, video connecting visitors and clients(consultants or experts).
- clients: Airfrance, Voyage sncf etc.

Outline

Demanding

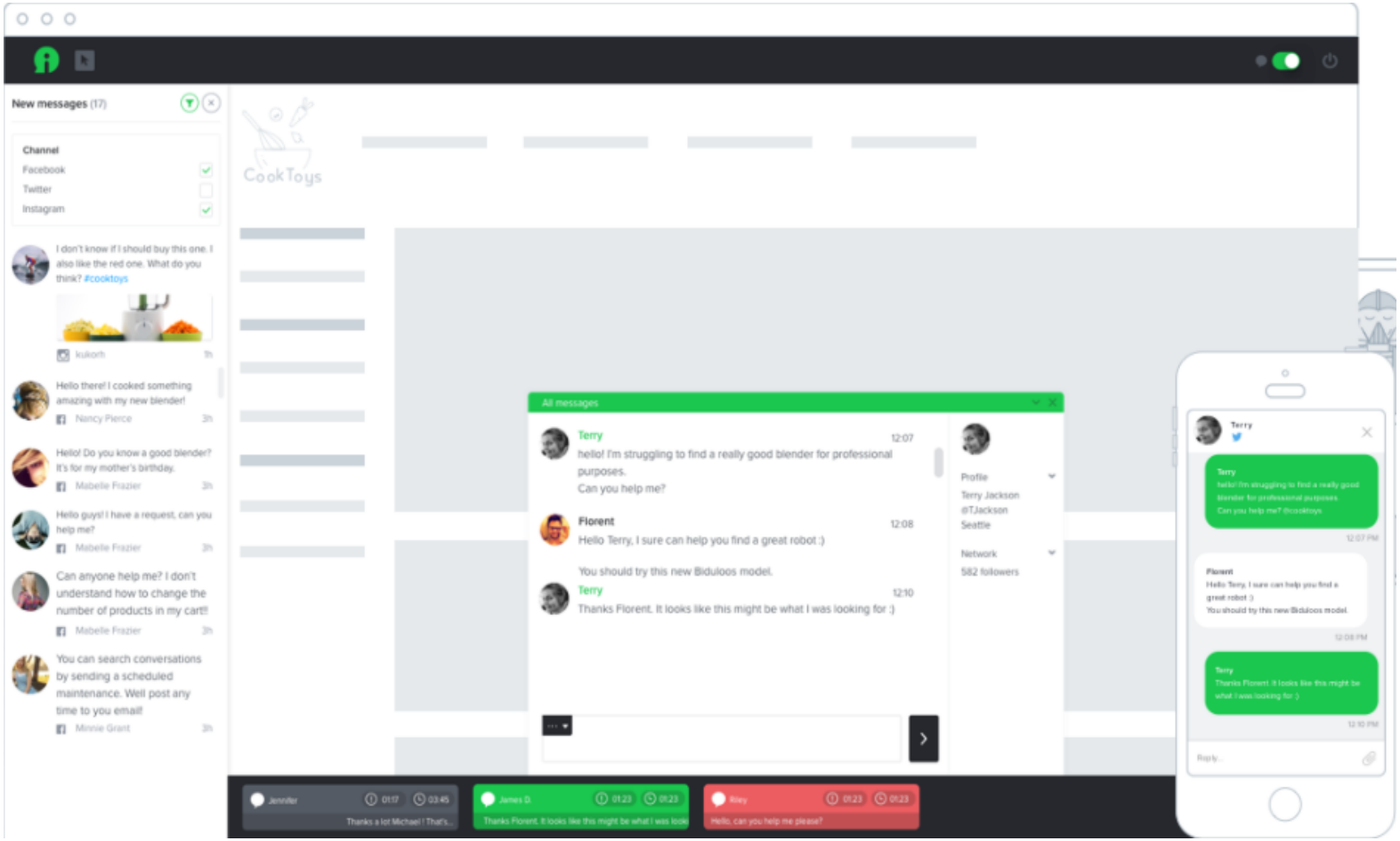


Figure: interface

Demanding

why chatbot? a big volume of visitors (6000 visitors/hour) vs a small amount of consultants.

- answer simple questions in some certain scenario.
- reduce repeated work of consultants.

Outline

Definition

chatbot: a Conversational Agent or Dialog Systems that can interact with customers by having natural conversations indistinguishable from human. eg: (Facebook (M), Apple (Siri), Google etc).

Outline

Models

- Retrieval-based model: choose a response from a prefixed set of responses.
  - expression-match, ML techniques(word2vec) etc.
  - pros: No grammar mistakes.
  - cons: work bad on unseen cases for which no pre-defined response exist. disability to refer back.
- Generative-model: generate response from scratch.
  - Recurrent Neural Network(eg:seq2seq)
  - pros: ability to refer back and cope with new cases.
  - cons: grammar mistakes, huge amount of training samples.

Text structure

- Short text conversation: one single question consecutively with one single answer.
- Long text conversation: multiple questions and responses changes(not necessarily consecutive).

Platform

- More formal platform: log of chat between a visitor and a consultant, forum.
- Less formal platform: twitter, facebook, messager: abbreviation, emoji, non-existing words.

domain

- Closed domain: the chatbot is trying to achieve a very specific goal in a certain scenario(eg:delivery).
- Open-domain: the chatbot is able to handle conversations with open subject.

Outline

A little math

Given an original message $\mathbf{x}$, find most likely response $\mathbf{y}^\star$

$$\mathbf{y}^\star = \underset{\mathbf{y} \in \mathcal{Y}}{argmin} \mathbb{P}(\mathbf{y}|\mathbf{x}).$$

Construct a model that can score responses and then find highest scoring response.

More about scoring

- original message $\mathbf{x}$: "il n'y a donc pas d'autres solutions que par carte bancaire ?"
- response $\mathbf{y}$: "C'est la seul solution je suis désolé :/"

Since we score a sequence of tokens $y_1, y_2, \ldots, y_m$ in $\mathbf{y}$, conditional on $x_1, x_2, \ldots, x_n$ in $\mathbf{x}$

$$\mathbb{P}(y_1, y_2, \ldots, y_m | x_1, x_2, \ldots, x_n) = \prod_{i=1}^{m} \mathbb{P}(y_i | y_{i-1}, \ldots, y_1, x_1, x_2, \ldots, x_n) \qquad (1)$$

Outline

Training and inference

Given a corpus of pair $(\mathbf{x}, \mathbf{y})$,

- training objective to maximize

$$\sum_{\mathbf{x},\mathbf{y}} \ln \mathbb{P}(y_1, y_2, \ldots, y_m | x_1, x_2, \ldots, x_n).$$

- inference objectives
  1. random sample from (1).
  2. greedy search: taking most likely tokens at each time.
  3. determine the likelihood of a specific response candidate.

Outline

Review Neural Network

- A simple NN with 1 layer:

$$\begin{aligned} \text{input}: \quad & \mathbf{x}_n \in \mathbb{R}^p, \\ \text{hidden layer}: \quad & \mathbf{h}_n = \sigma_1(W_i \mathbf{x}_n + B_i), \\ \text{output}: \quad & \hat{\mathbf{y}}_n = \sigma_2(W_o \mathbf{h}_n + B_o). \end{aligned}$$

  where σ_1 : activation function and σ_2 : sigmoid function.

- Eg: $\mathbf{x}_n$: pixels of images; $\mathbf{y}_n$ real character in the image, eg [cat, dog, car, house]; $\hat{\mathbf{y}}_n$: estimated character

$$[0.8, 0.2, 0.1, 0] \rightarrow \text{cat}$$
$$[0.1, 0.7, 0.1, 0.1] \rightarrow \text{dog}$$
$$[0.15, 0.05, 0.65, 0.15] \rightarrow \text{car}$$
$$[0, 0, 0.5, 0.95] \rightarrow \text{house}$$

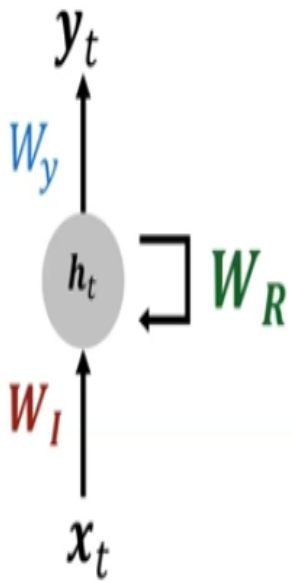
A simple RNN

- objective:

$$\text{minimize}: \sum_{n=1}^{N} (\mathbf{y}_n - \hat{\mathbf{y}}_n)^2.$$

  with respect to parameters: W_i, W_o, B_i, B_o.
- optimization: gradient descent, stochastic gradient descent etc.
- disadvantage: order of $\mathbf{x}$ is important, translation, speech recognition etc. "I am interested in Neural network".

Recurrent neuron



$$\boldsymbol{h}^{(t)} = g_h(W_I \boldsymbol{x}^{(t)} + W_R \boldsymbol{h}^{(t-1)} + \boldsymbol{b}_h)$$

$$\boldsymbol{y}^{(t)} = g_y(W_y \boldsymbol{h}^{(t)} + \boldsymbol{b}_y)$$

Figure: structure of RNN

Unrolling of RNN

Unrolling a recurrent network into a feed-forward network

$$\boldsymbol{h}^{(t)} = g_h(W_I \boldsymbol{x}^{(t)} + W_R \boldsymbol{h}^{(t-1)} + \boldsymbol{b}_h)$$

$$\boldsymbol{y}^{(t)} = g_y(W_y \boldsymbol{h}^{(t)} + \boldsymbol{b}_y)$$

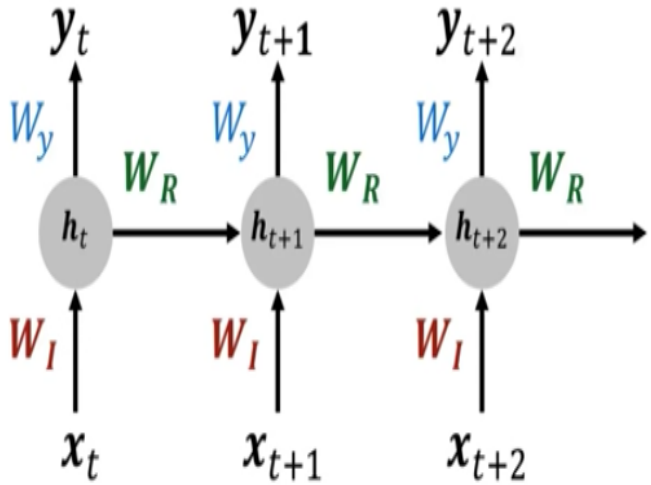


Figure: Unrolling of RNN

Long Short Term Memory

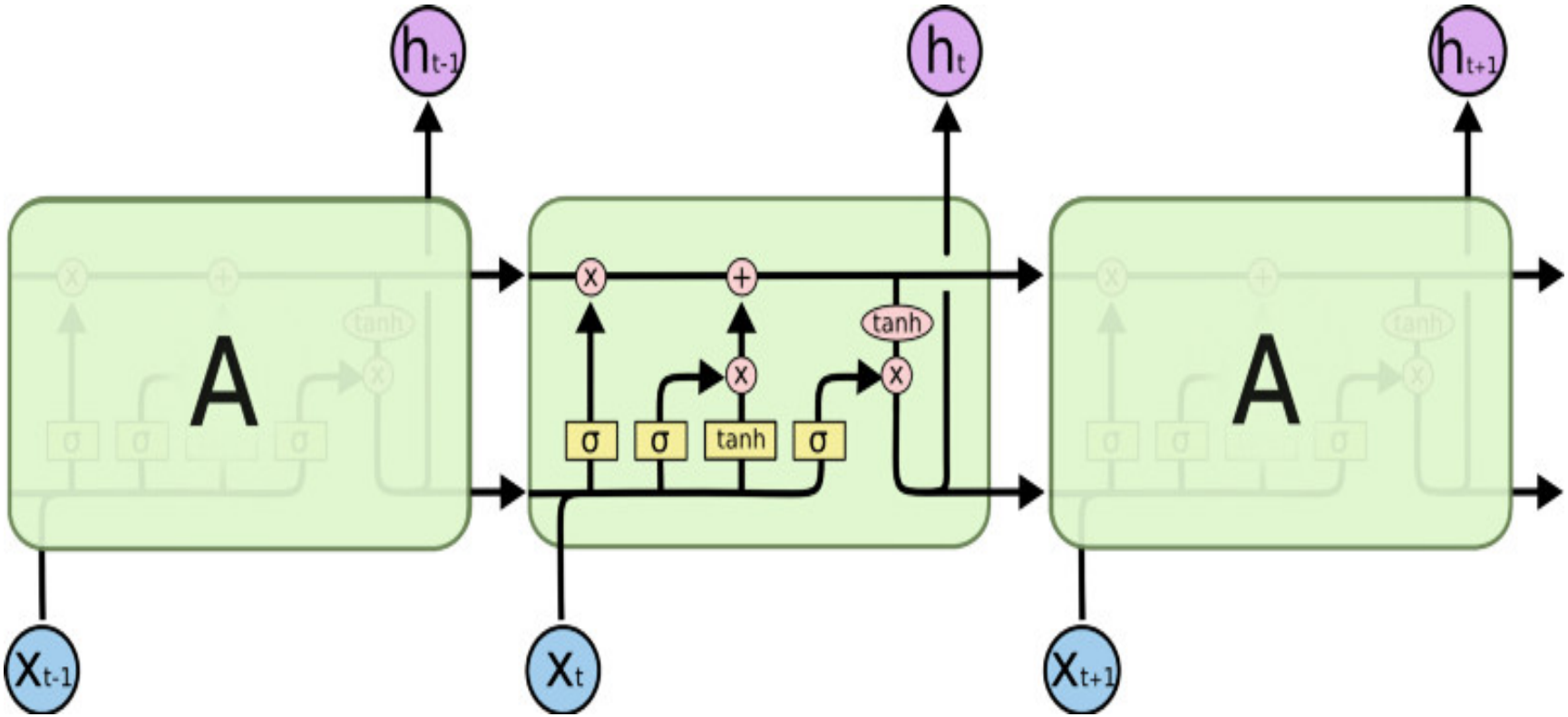


Figure: LSTM

Long short term memory

- Long short term memory(LSTM):

$$
\begin{aligned}
&\text{Forget gate layer}: f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right),\\
&\text{Input gate layer}: i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i),\\
&\text{Input}: \widetilde{c}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c),\\
&\text{Cell state}: c_t = f_t * c_{t-1} + i_t * \widetilde{c}_t,\\
&\text{Output gate layer}: o_t = \sigma\left(W_o \cdot [h_{t-1,x_t}] + b_o\right),\\
&\text{Output}: h_t = o_t * \tanh(c_t).
\end{aligned}
$$

Variants of LSTM

- Gated recurrent unit(GRU): combines the forget and input gates into a single "update gate." It also merges the cell state and hidden state.
- Coupled LSTM.
- Depth Gated RNNs().

Outline

Sequence to Sequence

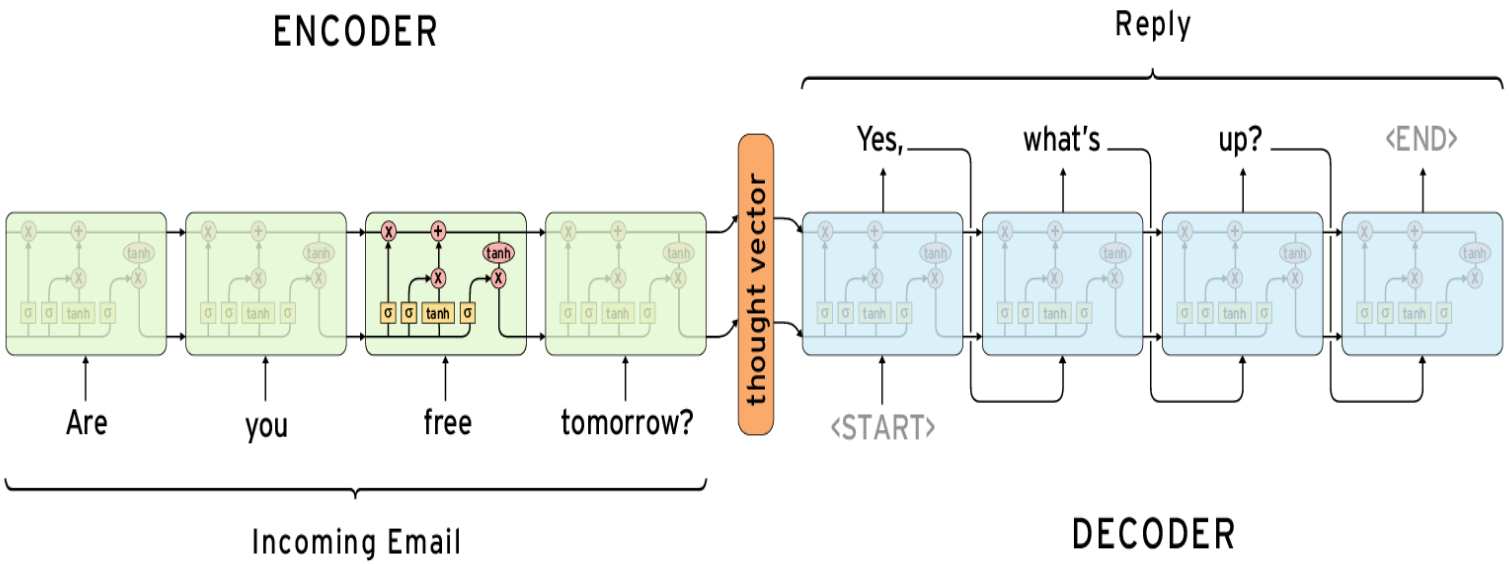


Figure: seq2seq

seq2seq

- Encoder: original message $\mathbf{x} \rightarrow$ representation of message c_n(last cell state). No output h_n in encoder step.
- Decoder: h_{t-1}, c_{t-1} + response tokens $y_t \rightarrow$ output h_t, c_t.

Outline

pre-processing

- language detection.
- remove accents: j'ai aimé → j'ai aime
- normalization: $'www.abc.com', 113$ → url, number_digit
- separate contraction: j'ai → je ai
- tokenization: Pau est beau → $[Pau, est, beau]$
- padding: $[pau, est, beau, pad, pad]$
- bucketing: grouper des questions et reponses.

Outline

Tensorflow and other framework

- Tensorflow
- Keras, TFlearn etc.

For Further Reading I


A. Kannan.
smart reply: automatied response suggestion for email.
https://arxiv.org/abs/1606.04870.


L. Sutskever.
Sequence to Sequence Learning with Neural Networks.
https://arxiv.org/abs/1409.3215


Q. Le.
A Neural Conversational Model. https://arxiv.org/abs/1506.05869