BIG DATA ANALYSIS

TA class IV

TA : Lee Chi-Hsuan

Word2Vec

1-of-N Encoding

apple = [1 0 0 0 0]

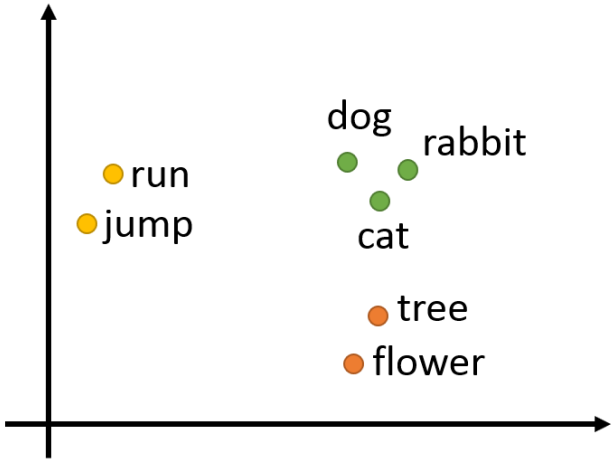
bag = [0 1 0 0 0]

cat = [0 0 1 0 0]

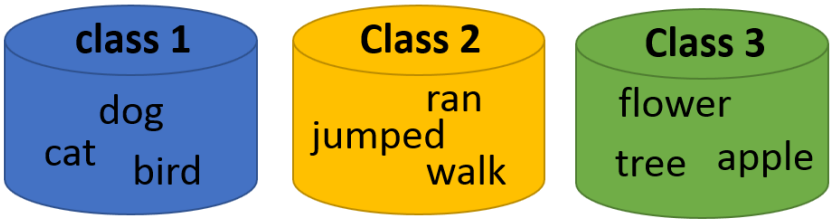
dog = [0 0 0 1 0]

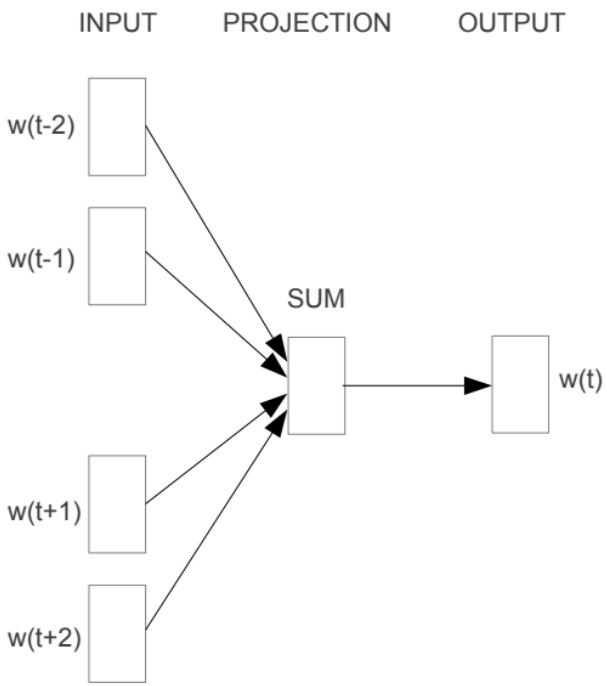
elephant = [0 0 0 0 1]

Word Embedding

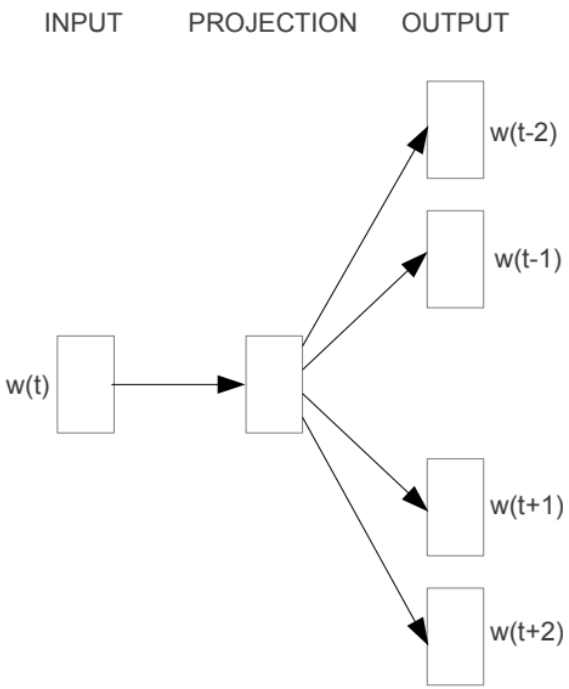


Word Class





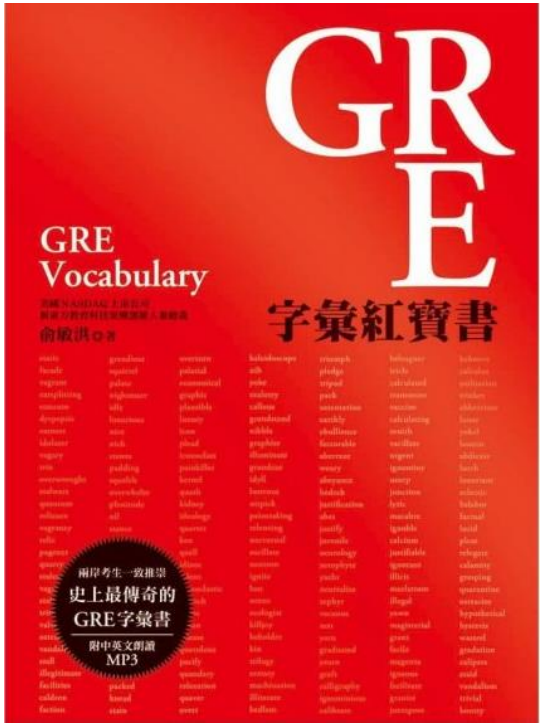
CBOW



Skip-gram

Break down words into pieces

- prefix / root / suffix



?

Gensim


Gensim is a FREE Python library
Topic modelling for humans
✓ Train large-scale semantic NLP models
✓ Represent text as semantic vectors
✓ Find semantically related documents

Why Gensim?

Super fast

The fastest library for training of vector embeddings – Python or otherwise. The core algorithms in Gensim use battle-hardened, highly optimized & parallelized C routines.

Data Streaming

Gensim can process arbitrarily large corpora, using data-streamed algorithms. There are no "dataset must fit in RAM" limitations.

Platform independent

Gensim runs on Linux, Windows and OS X, as well as any other platform that supports Python and NumPy.

Proven

With thousands of companies using Gensim every day, over 2600 academic citations and 1M downloads per week, Gensim is one of the most mature ML libraries.

Open source

All Gensim source code is hosted on Github under the GNU LGPL license, maintained by its open source community. For commercial arrangements, see Business Support.

Ready-to-use models and corpora

The Gensim community also publishes pretrained models for specific domains like legal or health, via the Gensim-data project.

Hands-on examples

jupyter Gensim Last Checkpoint: a day ago (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Run Markdown

Imports

reference

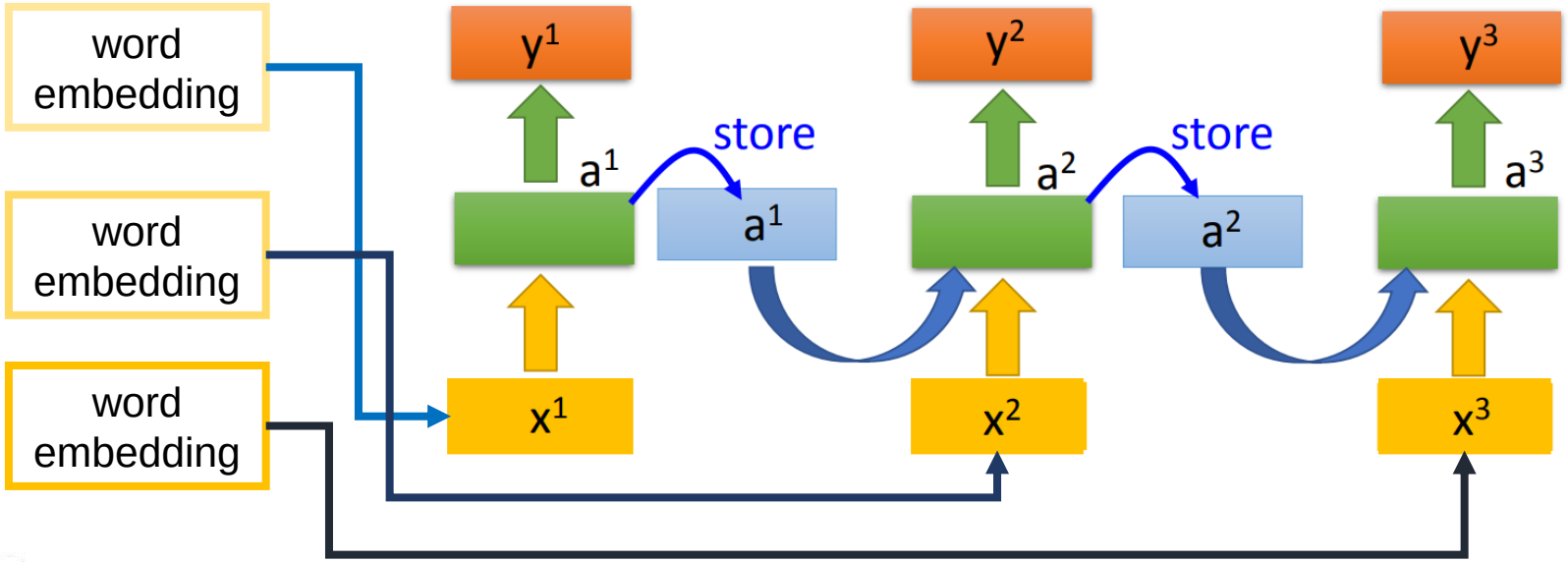
In [26]:

```
import pandas as pd
import gensim
from gensim.models import Word2Vec
import re
import nltk
import time
```

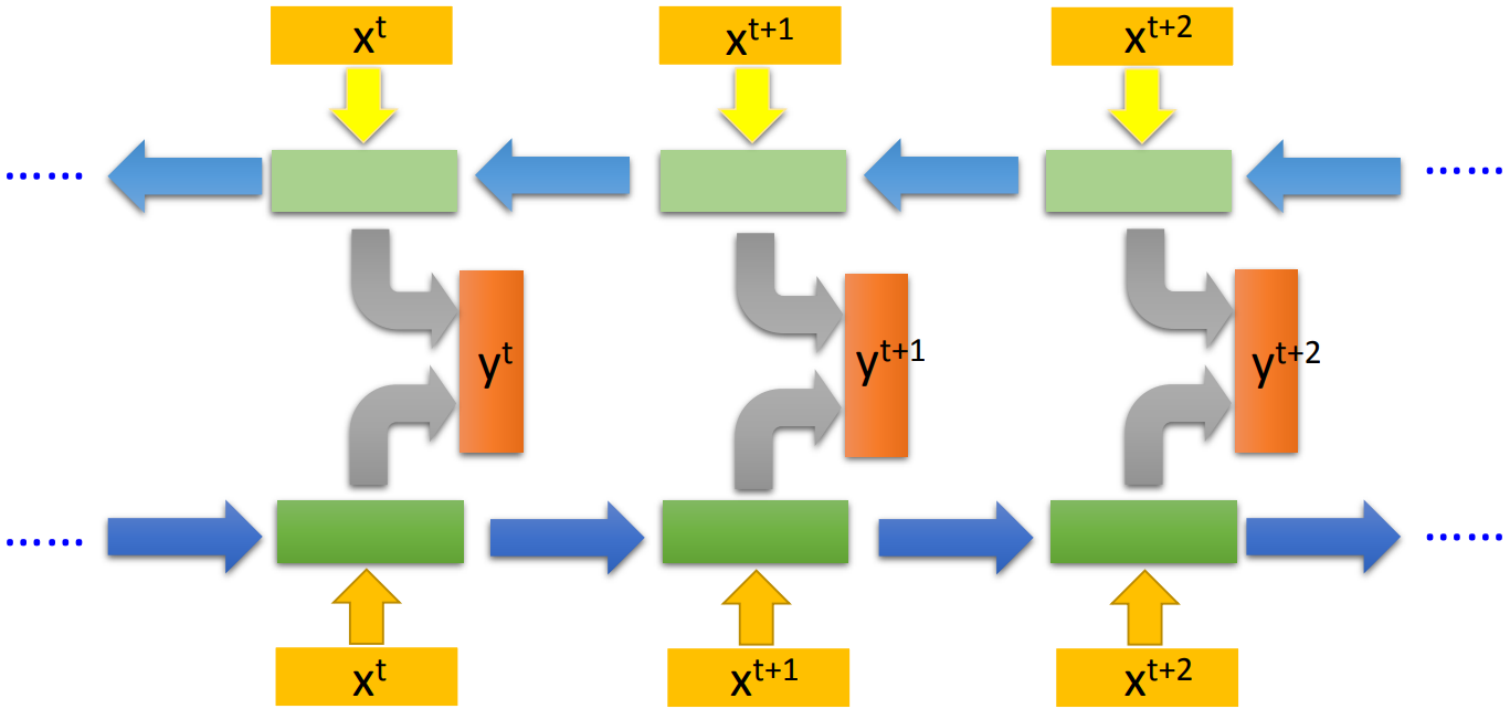
```
df_train = pd.read_csv("Train.csv")
```

More Applications

- use word embeddings to train RNN / LSTM / GRU models



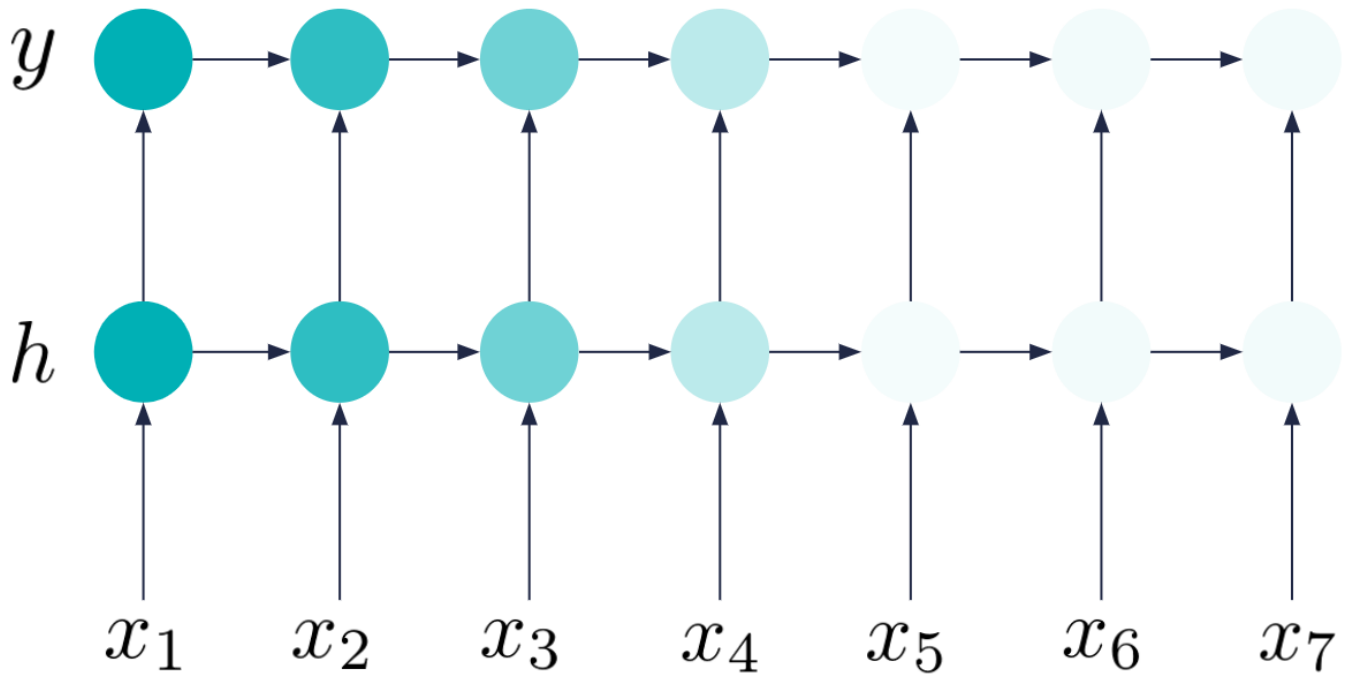
Bidirectional RNN



- problems of RNN ?

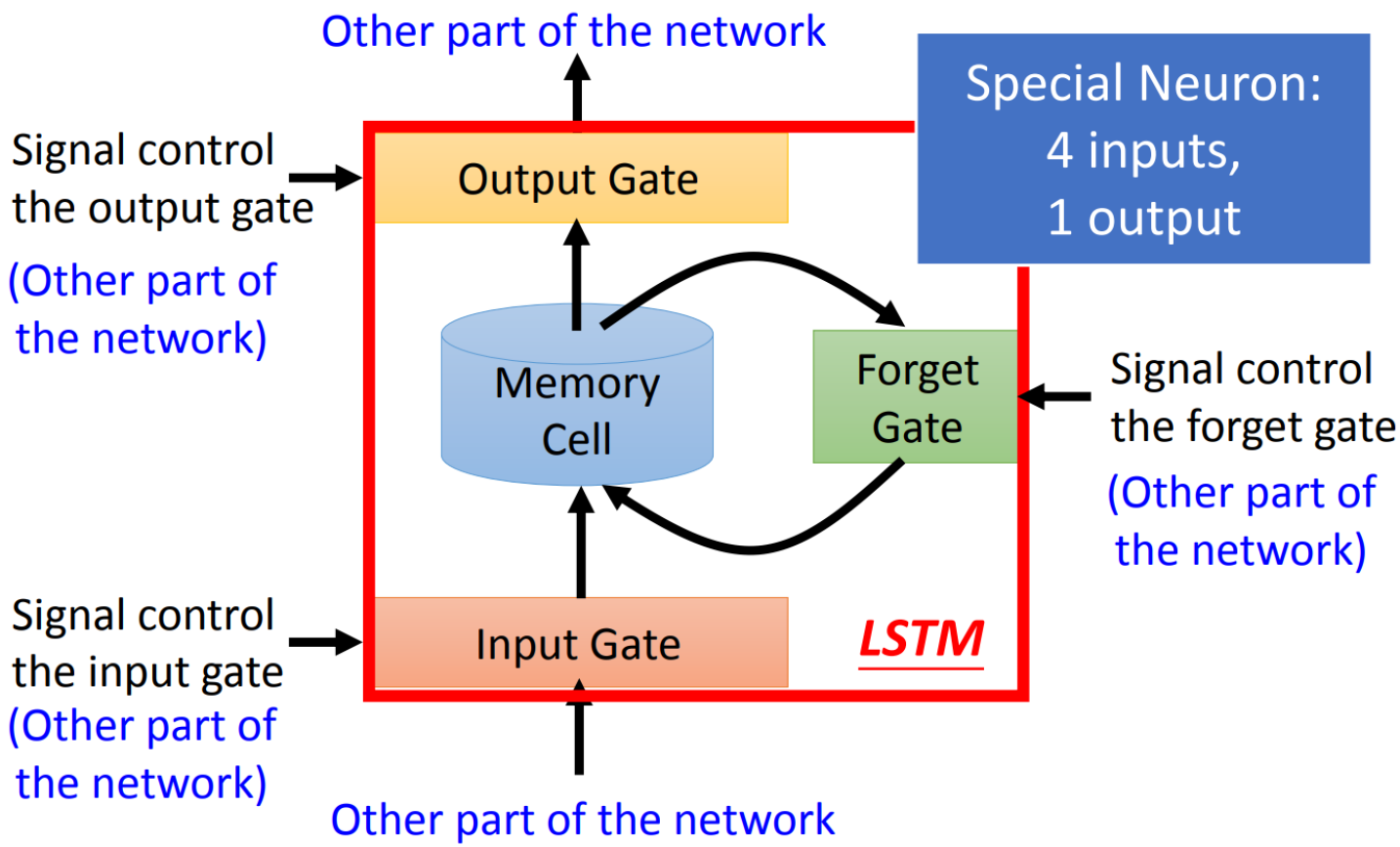
  - it forgets information

"Harry, my best friend and classmate from my childhood back in Oklahoma, is here."
Who is here? Network: Oklahoma



- LSTM and GRU

Long Short-term Memory (LSTM)



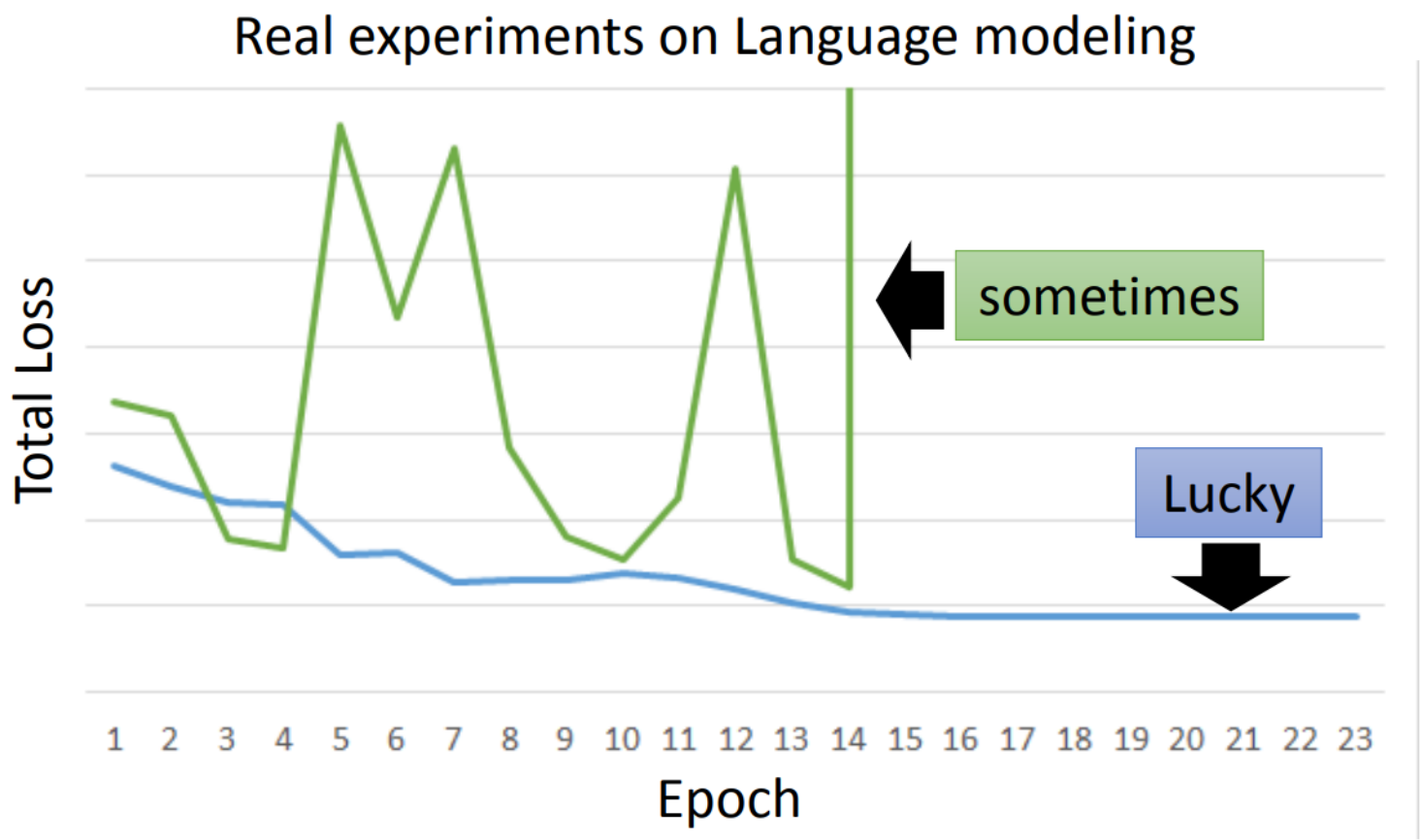
- Gated Recurrent Unit (GRU) :

  - simpler than LSTM

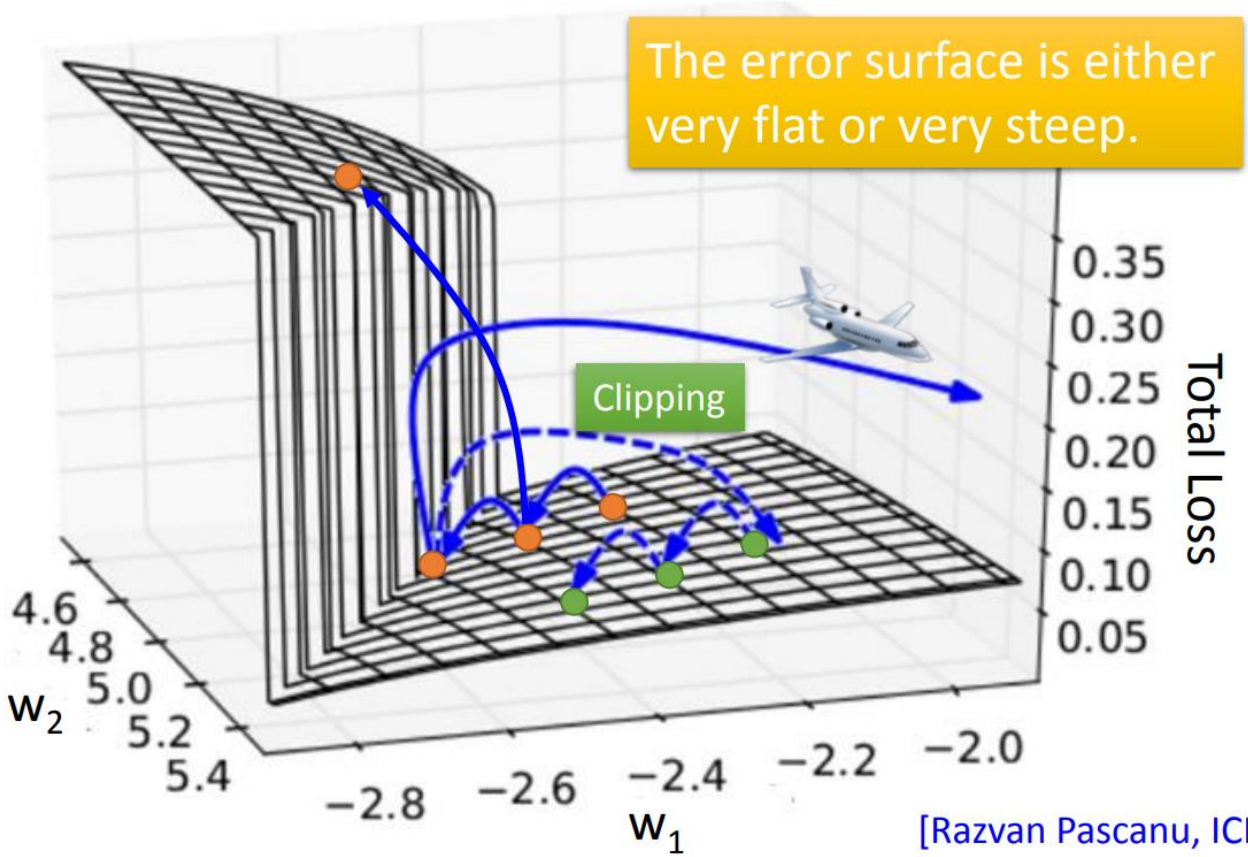
  - only contains Update Gate

    and Reset Gate

- RNN-based network is not always easy to learn



The error surface is rough.



[Razvan Pascanu, ICML'13]

Hands-on examples

jupyter Word2Vec on Bi-LSTM Last Checkpoint: Yesterday at 10:09 PM (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Run Code

Imports

In [1]:

```
import torch
import torchvision
import torch.nn as nn
import torch.optim as optim
import torch.nn.functional as F
from torch.utils.data import Dataset
from torch.utils.data import DataLoader
import torchvision.transforms as transforms
from torch.utils.data import DataLoader
from torch.nn.utils.rnn import pad_sequence
import pandas as pd
import numpy as np
import nltk
from nltk.corpus import stopwords
import time
import fasttext
```

Next Step ?

- Transformer :

  Seq2seq model with "self-attention"

FLAC 44.1kHz 16bit, MQA 44.1kHz

59 minutes

Soundtrack

Transformers: The Score

Steve Jablonsky

Play Now | Focus On Similar

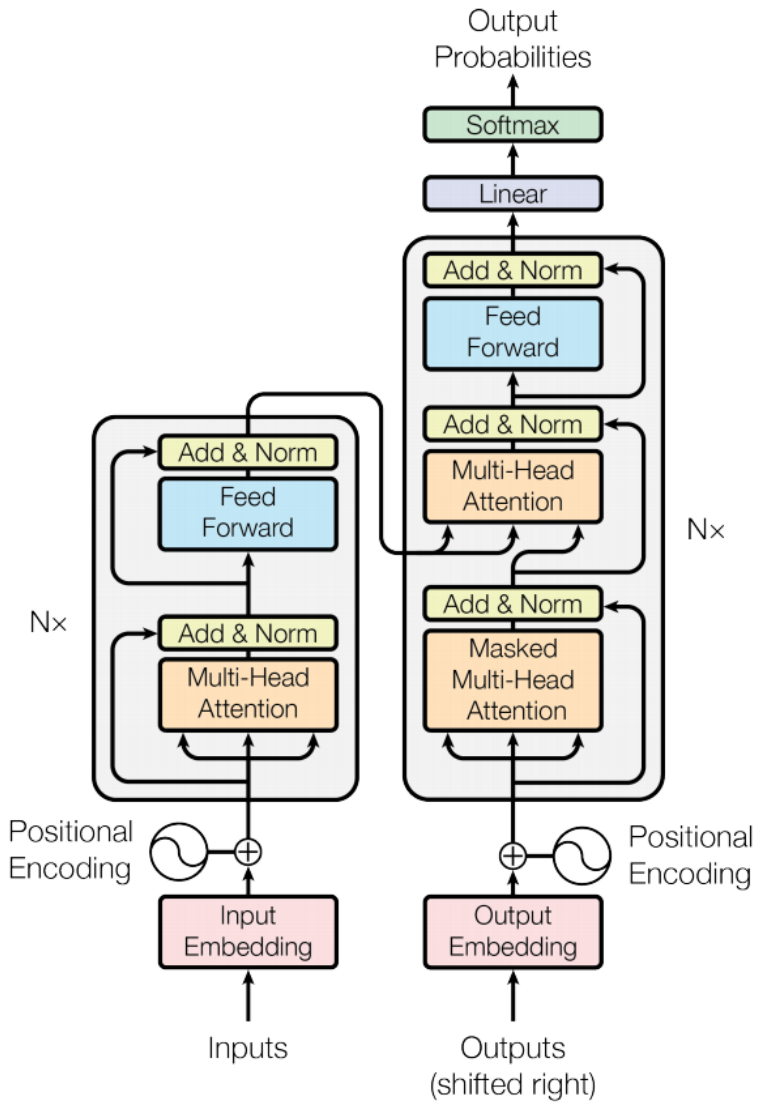
Released
9 Oct 2007

Added
6 Nov 2020

Steve Jablonsky's dynamic score to the 2007 big-budget, big-screen version of the Transformers cartoon show (and classic toy line), proved so appealing to fans of the film that an online petition was actually put into place to push for release of a CD version of the music. The voice of Transformer Nation was apparently heard loud and clear, and a CD was released in due time. Featuring a bold mix of electronic music and more traditional orchestral pieces, the score strikes a nice balance between the film's sci-fi bombast and its more terrestrial concerns.

A word can have multiple senses.

- Have you paid that money to the bank yet ?
- It is safest to deposit your money in the bank.
- They stood on the riverbank to fish.
- The hospital has its own blood bank.

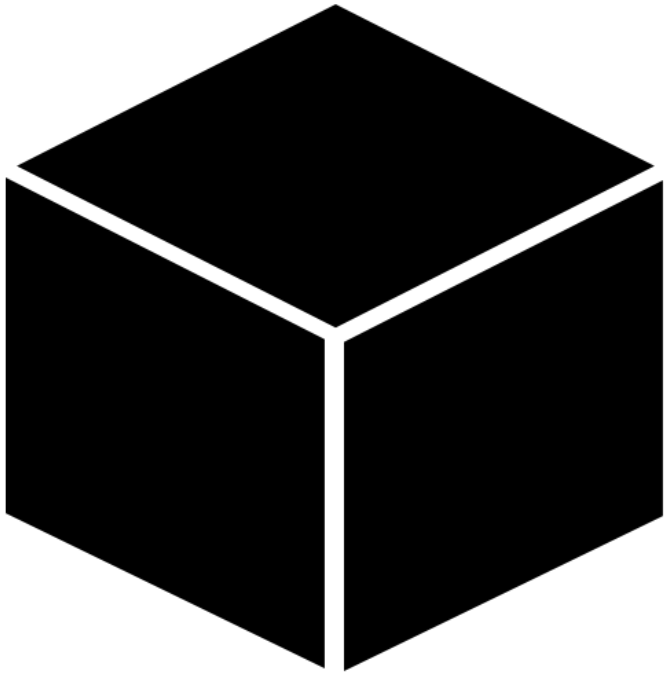

Output
Probabilities
Softmax
Linear
Add & Norm
Feed
Forward
Add & Norm
Multi-Head
Attention
Add & Norm
Masked
Multi-Head
Attention
Nx
Add & Norm
Feed
Forward
Add & Norm
Multi-Head
Attention
Nx
Positional
Encoding
Positional
Encoding
Input
Embedding
Output
Embedding
Inputs
Outputs
(shifted right)

Embeddings from Language Model (ELMO)

- RNN-based language models (trained from lots of sentences)

Bidirectional Encoder Representations from Transformers (BERT)

- BERT = Encoder of Transformer



Reference

- http://speech.ee.ntu.edu.tw/~tlkagk/index.html
- https://fasttext.cc/docs/en/english-vectors.html
- https://radimrehurek.com/gensim/models/word2vec.html
- https://github.com/facebookresearch/fastText/issues/475
- https://docs.google.com/presentation/d/1zyuwCx7knqnP-LJswlDfWSmk5FhFgFmYJGqdEZn8yhc/edit#slide=id.g33c734b7fb_0_7