CS F425: Deep Learning

23

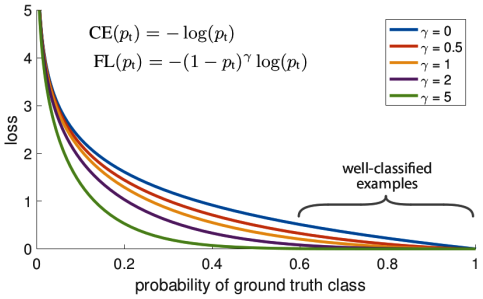
RetinaNet, 3D-CNN Sequence Modeling

Dr. Kamlesh Tiwari
Assistant Professor, Department of CSIS,
BITS Pilani, Pilani Campus, Rajasthan-333031 INDIA

Mar 21, 2023 ON-CAMPUS Campus @ BITS-Pilani [Jan-May 2023]

http://ktiwari.in/dl

RetinaNet [1]

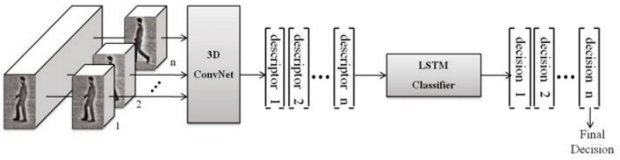


Negative examples are many

Single shot detector with better accuracy.

[1] **Cite 12436:** Lin, Tsung-Yi, et al. "Focal loss for dense object detection." Proceedings of the IEEE ICCV 2017.

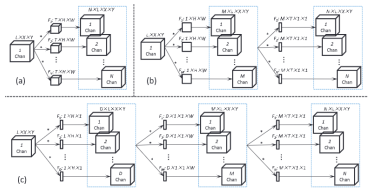
3D-CNN [2]

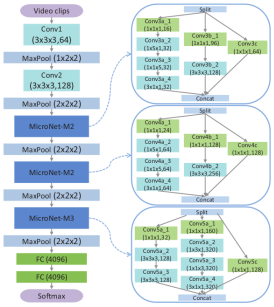


- More suitable to video
- Builds spatio temporal feature
- Human action KTH dataset with 6 actions (walking, jogging, running, bending, hand-waiving, clapping) the accuracy was 94.39%

[2] Baccouche, Moez and Mamalet, Franck and Wolf, Christian and Garcia, Christophe and Baskurt, Atilla, *Sequential deep learning for human action recognition*, In: International workshop on human behavior understanding, pp 29–39, Springer 2011

Human Action Recognition [3]





- **UCF-101** YouTube, 101 action, 13,320 videos (27 h). **HMDB-51** movies and web, 6849 clips, 51 action.
- Obtained 92.6% and 65% accuracy respectively that are better than previous state of the art

[3] Yang, Hao and Yuan, Chunfeng and Li, Bing and Du, Yang and Xing, Junliang and Hu, Weiming and Maybank, Stephen J, *Asymmetric 3d convolutional neural networks for action recognition*, In: Pattern recognition, pp 1–12, Elsevier (85) 2019

Whether I should go to see a movie?

Suppose you have a system, whom you can ask whether should you go to see a movie or NOT

- Ofcourse, the system need to consider many factors such as 1) who are the actors 2) what is IMDB rating of movie 3) your interests in type of movie 4) do you have money to purchase tickets *etc*.
- Everytime the system would give same answer. How many times you can see the same movie?

Basic ML models considers only the current input
(sometime it is useful to consider previous input/output as well)

Application of Sequence Data

There are various places we encounter Sequence Data

Sequence Prediction
Weather Forecasting
Stock Market Prediction
Product Recommendation

Sequence Classification
DNA Sequence Classification
Anomaly Detection
Sentiment Analysis

Sequence Generation
Text Generation
Music Generation
Image Captioning

Sequence-to-Sequence Prediction
Multi-Step Time Series Forecasting
Text Summarization
Language Translation

Sequences

Data is essentially a set of examples (every row represents one)

- Mostly we have fixed number of attributes in an example. However, some interesting applications (such as text, voice, video *etc*) have variable number of points
- These points could depend on each other in complicated way
- We want to do something like
  1. Given a text we want to predict sentiment or next world
     `Agra is a great place, if you go there you must visit _____`
  2. Given a voice we want to recognize the speaker
  3. Given a video we want to determine the activity
  4. Given a series (say stock values) predict next one

ISSUE: Basic ML models do **NOT** handle variable number of inputs

Modeling Sequences

There are some ideas to fix the issue (difficulty in using ML models)

- Consider only fixed length.
  (Unable to model long term dependencies)
- Bag of Words: use a vector of length equal to dictionary, and mark/count which words are present
  (since order is not preserved; following lines becomes same)
  `David is good at math but is bad in sciecne`
  `David is bad at math but is good in sciecne`

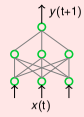
To model sequences we need

1. To deal with variable size input
2. Maintain sequence order
3. Keep track of long term dependencies
4. Share parameters across the sequences

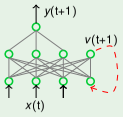
Recurrent Networks (RNN) [4]

Feedforward network cannot capture the dependence of $y(t+1)$ on earlier values of x such as $x(t-1)$

Feedforward Network



Recurrent Network

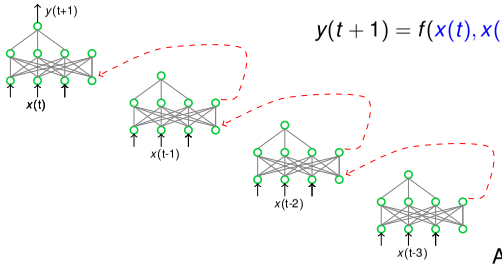


- RNN uses states (called self-state) of the network units available at time t as an input to the other units at time $t+1$
- RNN is suitable for temporal data (like time series)
- Training may involve unfolding and averaging.

[4] S. Hochreiter and J. Schmidhuber, **Long short-term memory**, Neural computation, vol. 9, no. 8, pp. 1735–1780, 1997

Unfolding RNN in Time

When input at the time t is provided, what is output at time $(t+1)$?



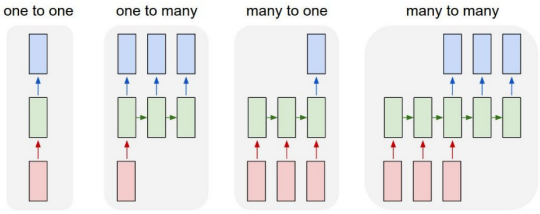
$$y(t+1) = f(x(t), x(t-1), ..., x(0))$$

And so on ...

By this way RNN incorporates history of the network in output

Various arrangements are possible based on need

- `What should be the next world?` **One-to-One**
- `Caption the given image?` **One-to-Many**
- `Segmentation or classification` **Many-to-One**
- `Translate from one language to other` **Many-to-Many**



Output may be different for same input

We are optimizing over programs not on functions

Thank You!

Thank you very much for your attention!