Deep Learning: Introduction to NLP for Classification Task – Session 2

Introduction to NLP, Kaggle and Deep Learning

Presenter: Alex Dela Cruz

Ryerson University

IEEE Computational Intelligence Society

Ryerson Advanced Artificial Intelligence Laboratory

IEEE Advancing Technology for Humanity

Agenda

1) Kaggle
   - What is it?
   - What is available?
   - Ways to succeed
2) Natural Language Processing
   - Introduction
   - Tasks
   - Introduction to some algorithms
3) Deep Learning
   - Supervised, Unsupervised, semi-supervised
   - Architecture
   - Learning algorithm

Learning Objective

- How to navigate Kaggle
- Exam and analyse a dataset
- Prepare and preprocess NLP data for deep learn
- Implement and train a feedforward multilayer neural network

Kaggle

What is it?

- It is a subsidiary of Google
- Largest online community of data science
- Provides powerful tools and resources for data scientists
  - Competitions
  - Datasets
  - Notebooks
  - Courses

Kaggle

Competitions

- Price
  - Cash
  - Swag
  - Kudos (Some Form of acknowledgment)
  - Knowledge
- Types
  - Active
  - Completed
  - In Class

Kaggle

Tools and Resources

- Competition
  - Price
    - Cash
    - Swag
    - Kudos (Some Form of acknowledgment)
    - Knowledge
  - Types
    - Active
    - Completed
    - In Class
- Datasets
  - A rich resource for publicly available datasets
  - Varity of domains (medical, finance, public health, environmental, etc..)
  - Varity of Modalities (text, numerical, images, and signal)
- Notebook
  - Free computation
  - Large repository of public notebooks to learn from
- Course
  - Varity of introduction courses for data scientist

Kaggle

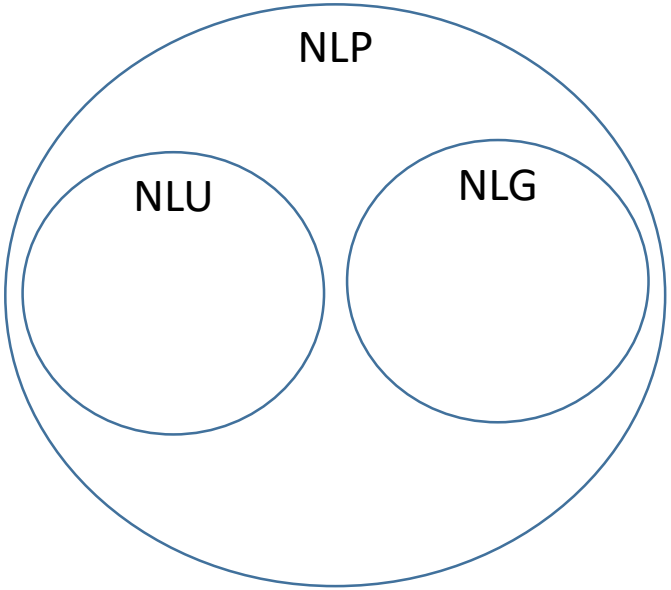
Ways to Successes

1. Understand the problem task
2. Understand the dataset
3. Learn from the community (Notebooks, Discussions)
4. Get your hands dirty
5. Persistence is key

NLP

Introduction

- Linguistics
  - “the scientific study of structure and development of language in general or of particular language” – Cambridge dictionary
- Computational Linguistics
  - “the study of computer systems for understanding and generating natural language” - [1]



[1] Grishman, Ralph. *Computational linguistics: an introduction*. Cambridge University Press, 1986.

NLP

Tasks

- Traditional Tasks
  - Tokenization
  - Name-Entity Recognition
  - Part of Speech
  - Parsing
  - Language Modelling
- Modern Task
  - Sentiment Analysis
  - Emotional Analysis
  - Summarization
  - Paraphrasing
  - Misinformation detection
  - Dialog Systems
  - Translation
  - Report Generation

Primary Focus

NLP

Algorithms:

- Bag-of-words
  - Representation of text

	it	was	the	best	of	times	worst
“it was the best of times”	1	1	1	1	1	1	0
“it was the worst of times”	1	1	1	0	1	1	1
“the best”	0	0	1	1	0	0	0

NLP

Algorithms:

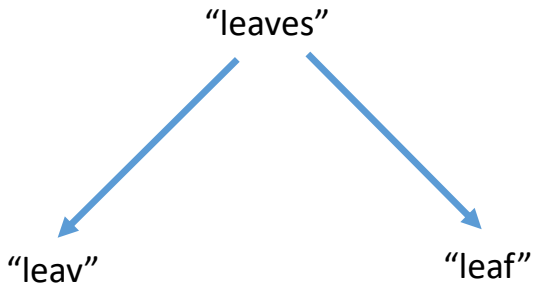
- TF-IDF
  - Similar to bag-of-words

	it	was	the	best	of	times	worst
“it was the best of times”	$\frac{1}{6} \times log(\frac{3}{2})$	$\frac{1}{6} \times log(\frac{3}{2})$	$\frac{1}{6} \times log(\frac{3}{3})$	$\frac{1}{6} \times log(\frac{3}{2})$	$\frac{1}{6} \times log(\frac{3}{2})$	$\frac{1}{6} \times log(\frac{3}{2})$	$\frac{0}{6} \times log(\frac{3}{1})$
“it was the worst of times”	$\frac{1}{6} \times log(\frac{3}{2})$	$\frac{1}{6} \times log(\frac{3}{2})$	$\frac{1}{6} \times log(\frac{3}{3})$	$\frac{0}{6} \times log(\frac{3}{2})$	$\frac{1}{6} \times log(\frac{3}{2})$	$\frac{1}{6} \times log(\frac{3}{2})$	$\frac{1}{6} \times log(\frac{3}{1})$
“the best”	$\frac{0}{2} \times \log(\frac{3}{2})$	$\frac{0}{2} \times \log(\frac{3}{2})$	$\frac{1}{2} \times \log(\frac{3}{3})$	$\frac{1}{2} \times \log(\frac{3}{2})$	$\frac{0}{2} \times \log(\frac{3}{2})$	$\frac{0}{2} \times \log(\frac{3}{2})$	$\frac{0}{2} \times \log(\frac{3}{1})$

NLP

Algorithms:

- Stemming & Lemmatization
  - Normalize words to their root word

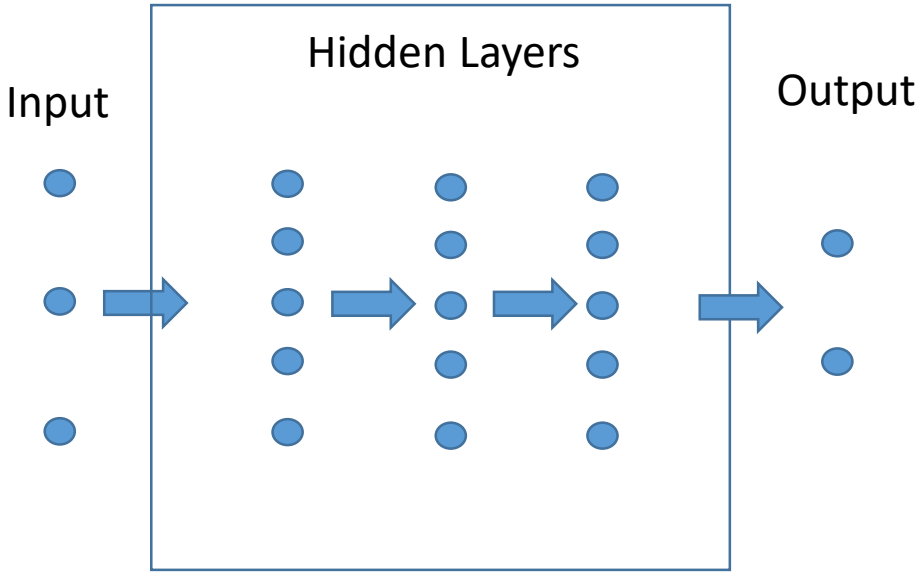


Deep Learning

- Supervised
  - Dataset contains labeled data
  - Target Output is Known during training
- Unsupervised
  - Dataset contains no labeled data
  - Target Output is not known during training
- Semi-Supervised
  - Dataset contains some labeled and non labeled data

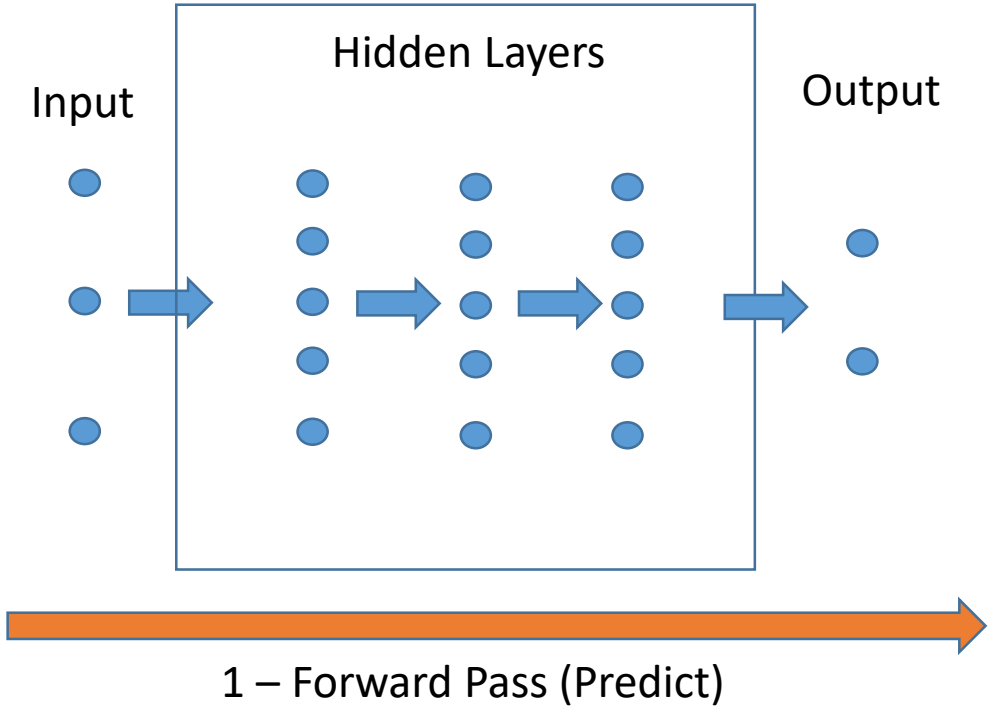
Deep Learning

Architecture



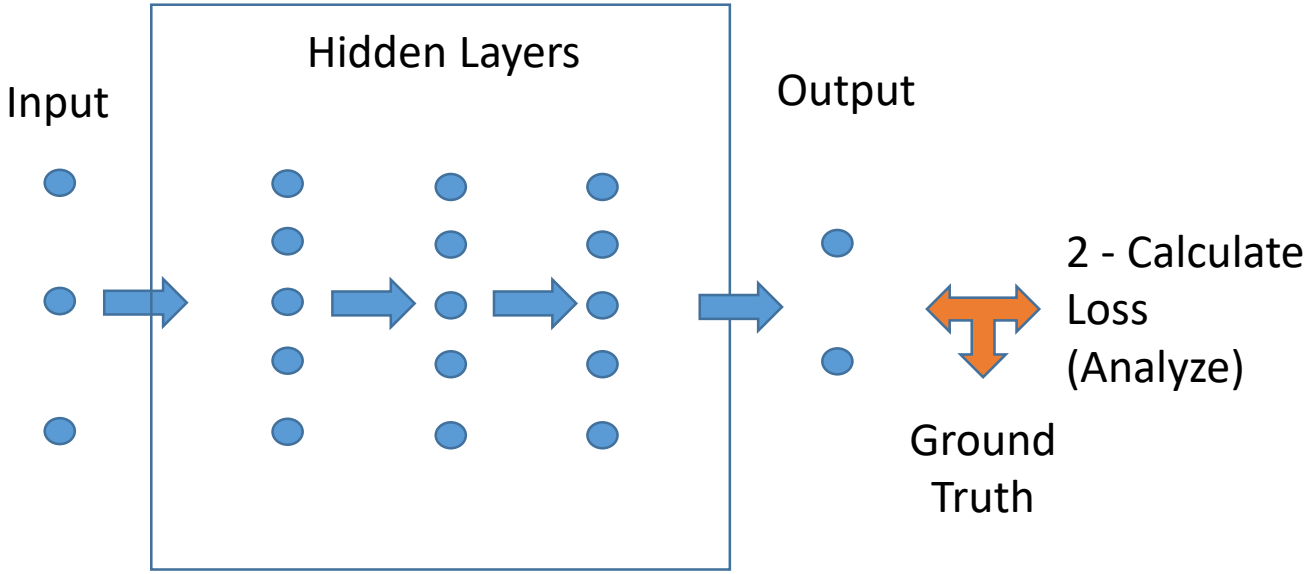
Deep Learning

Learning



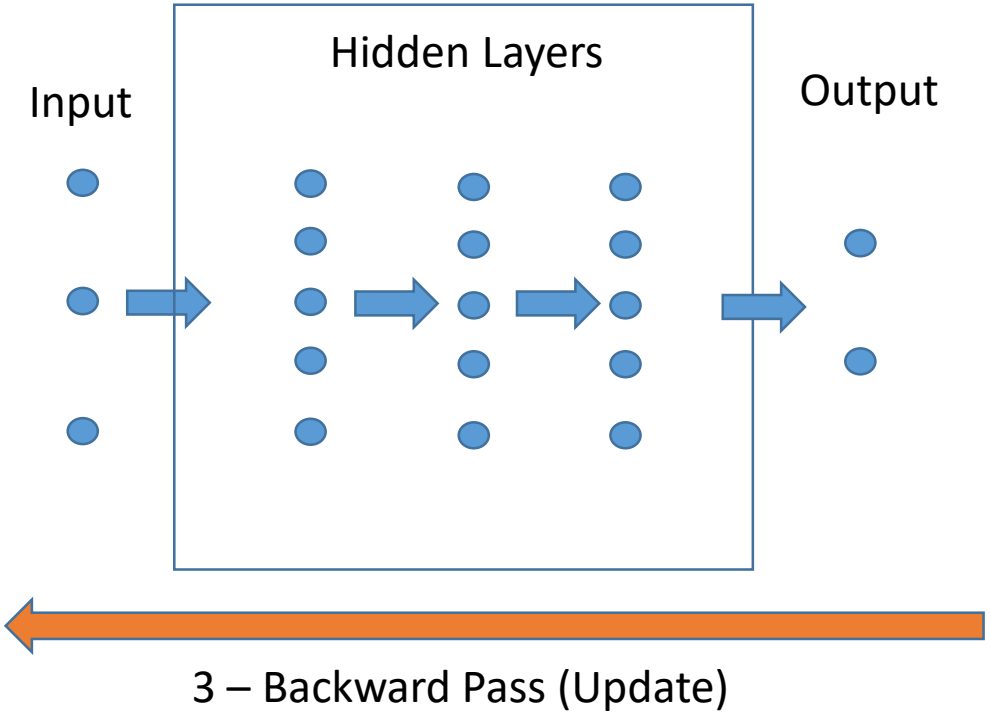
Deep Learning

Learning



Deep Learning

Learning



QA