Natural Language Processing with Representation Learning

DS-GA 1011

Overview

How should human languages be understood and analyzed? In this course, we will examine modern computational approaches based on representation learning for understanding, processing and using human language. These include neural network-based deep learning methods and vector-space models of word meaning, and together will give you the tools to build state-of-the-art models for hard language understanding tasks like translation.

General Information

Registration

If you are a student at the NYU Center for Data Science, you can already register for this course. If not, please contact Kathryn Angeles (kangeles@nyu.edu) at the NYU Center for Data Science. Please do not email me, as I will not hand out registration code.

Lectures

Wednesdays 6:45–9:25 PM, Meyer 121 (combined lecture–lab)

Instructors

- Kyunghyun Cho: kyunghyun.cho@nyu.edu
    - Office hours Tues 10-11, 60 Fifth Ave., Rm. 616

Section Leaders and Graders

- Elman Mansimov <elman.mansimov@gmail.com>
    - Office hours Fridays 4.00-5.00pm, 60 Fifth Ave., Rm. 660
- Phu Mon Htut <pmh330@nyu.edu>
    - Office hours Fridays 4.00-5.00pm, 60 Fifth Ave., Rm. 660
- Jason Phang <jasonphang@nyu.edu>
    - Office hours Tues 1.30-2.30pm, 60 Fifth Ave., Rm. 606
- Anant Gupta <ag4508@nyu.edu>
    - Office hours Tues 1.30-2.30pm, 60 Fifth Ave., Rm. 606

Course Site

- Piazza: https://piazza.com/nyu/fall2018/dsga1011
- NYU Classes: https://newclasses.nyu.edu/portal/site/d92ab97f-91ca-467b-9fd8-6a4b79887861
- NYU Calendar: https://calendar.google.com/calendar/b/1?cid=bnl1LmVkdV9vb24xZWJudWZvcjlmZjgxb2FscDljcWs3OEBncm91cC5jYWxlbmRhci5nb29nbGUuY29t

Grade

Assignments (30%) + Project (30%) + Exam (40%)

Prerequisites

A student is expected to be familiar with the following topics:

- Undergraduate-level Probability and Statistics
- Undergraduate-level Linear Algebra
- Undergraduate-level Calculus
- Machine Learning at the level of DS-GA 1003 or CSCI-UA 0480-007

Data Science students are expected to have taken the following courses before taking this course. Students from programs other than Data Science are expected to demonstrate comparable skillsets.

- DS-GA-1002: Statistical and Mathematical Methods
- DS-GA-1003: Machine Learning and Computational Statistics

We strongly encourage students to install Python 3, Jupyter Notebook and Labs, and the PyTorch toolkit and familiarize themselves with its basic operations before the start of the course.

Materials

The main material of the course comes from my own lecture note available at:

- https://github.com/nyu-dl/NLP_DL_Lecture_Note/blob/master/lecture_note.pdf

The following (e)books are highly recommended during the course:

- Goldberg. Neural Network Methods for Natural Language Processing. 2017. (free download when on the NYU network)
- Jurafsky and Martin. Speech and Language Processing, 2nd edition. 2009. (Some free draft chapters from the 3rd edition are available here.)
- Goodfellow, Courville and Bengio. Deep Learning. 2016. (draft online)
- Manning and Schütze. Foundations of Statistical Natural Language Processing. 1999.
- Bender. Linguistic Fundamentals for Natural Language Processing. 2013. (free download when on the NYU network)

Schedule

Day	Lecture Topic	Lab Topic (in class)	Note
9/5	Course Logistics Machine Learning	Python, Jupyter Lab (JP, AG)	
9/12	No Lecture	PyTorch Setup and Introduction	
9/19	Machine Learning	No Lab	
9/26	Guest Lecture: Maximilian Nickel	Bag-of-words classification FastText	https://scholar.google.com/citations?user=KDqGTIUAAAAJ&hl=en

	(Facebook AI Research)	Programming Assignment 1	
10/3	Convolutional and recurrent networks for text classification	CNN classification RNN classification	
10/10	Language modeling (1)	GloVe/FastText and analogy evaluation Programming Assignment 2	
10/17	Language modeling (2)	Language modelling and evaluation with KenLM	
10/24	Conditional sequence modelling: neural machine translation	Language modelling and evaluation with RNN Final Project Announcement	
10/31	No Lecture	Neural machine translation FairSeq - morphological reinflection	
11/7	Delving deeper into attention (1) - self-attention - pointer networks	No lab	
11/14	Delving deeper into attention (2) - memory networks	Self-Attention Classifier, Transformers	
11/21	No Lecture	No Lab	Thanksgiving Break
11/28	Advanced Topics	https://www.youtube.com/watch?v=GP-gG2zFeBg https://www.youtube.com/watch?v=1JO1Pcr5rYA https://www.youtube.com/watch?v=XInbNFW2tX8	
12/5	Guest lecture: Ilia Kulikov	Exam Q&A	How to use v2.overleaf.com with ACL template
12/12	Final exam		

Assignments

There will be two programming assignments. Each is due at the start of class on the specified:

- Bag of N-Grams Document Classification, Hyperparameter Exploration and Analysis (10/10)
- Recurrent and convolutional document classification, hyperparameter exploration and analysis (10/31)

For each assignment, a student is expected to hand in a report of up to 5pp outlining the model, its implementation and its results. See the TAs, not the instructors, for help with these assignments.

Final Project

In this course, students are expected to conduct research on the topic described below.

Teams

You may work in teams of up to four on the final project. Teams must be finalized by Oct 24 2018.

Topic

The goal of this project is for you to build a neural machine translation system and experience how recent advances have made their way. Each team will build the following sequence of neural translation systems for two language pairs, Vietnamese (Vi)→English (En) and Chinese (Zh)→En (prepared corpora will be provided):

1. Recurrent neural network based encoder-decoder without attention
2. Recurrent neural network based encoder-decoder with attention
3. Replace the recurrent encoder with either convolutional or self-attention based encoder.
4. [Optional] Build either or both fully self-attention translation system or/and multilingual translation system.

You are expected to implement these on your own (if necessary), experiment them with both language pairs, report their performance (measured in terms of automatic evaluation metrics) and analyze their behaviours and properties.

Paper

The final report should include the description of the task, models, experiments and conclusion and be up to 6 pages long excluding unlimited pages reserved for references. It must be prepared in LaTeX in conformance with NAACL style (a.k.a. ACL style for 8.5x11" paper). Your paper should include a link to a *public* GitHub repository containing the code used in your experiments.

Contribution Statements

The final report must state (in one or two sentences) the contributions of each team member. Team projects which fail to include this will receive a 50% grade deduction.

Key date

- **Project submission 12/14 11.55pm**

The deadline for the final report will not be extended, and **late submissions will not be accepted**.

For PhD Students in CDS and CS

If you are interested and can form a team, you may choose to participate in the following competition instead after the lecturer's approval. Once the team is formed, please come see the lecturer during the office hour.

https://coleridgeinitiative.org/richcontextcompetition

Course Policies

Collaboration

We strongly encourage you to discuss the assignments and the paper with your peers. However, all work that you submit for the two assignments must be your own. Teams are not allowed on the two assignments.

Late Work

No late work will be considered.

Extra Credit

In the interest of fairness, we won't offer extra credit.

Pass/Fail Option

If you'd like to take this course as pass/fail instead of with a letter grade, come see the lecturer at the office hour. Note that the lecturer would not write a reference letter for a student who takes this course as pass/fail.

Participation

Students are expected to attend every meeting of the class, and to bring laptops (contact us if this presents a problem) for use in short in-class programming exercises. Students are also expected to participate actively in course discussion, either in person or on Piazza.

Plagiarism

If you use any text or figures at all from an outside source—yes, this includes descriptions of your baseline model and your data—you must make it clear (through quotation marks, block quotes, or an explicit statement) that you are using the work of others, and include an appropriate citation. Failure to do so will result in a zero grade for the submitted work and, except in cases of obvious mistakes, a University investigation.

Applicable University Policies

Academic Integrity

Work you submit should be your own. Please consult the CAS academic integrity policy for more information: https://cas.nyu.edu/content/nyu-as/cas/academic-integrity.html. Penalties for violations of academic integrity may include failure of the course, suspension from the University, or even expulsion.

Religious Observance

As a nonsectarian, inclusive institution, NYU policy permits members of any religious group to absent themselves from classes without penalty when required for compliance with their religious obligations. The policy and principles to be followed by students and faculty may be found here: The University Calendar Policy on Religious Holidays (http://www.nyu.edu/about/policies-guidelines-compliance/policies-and-guidelines/university-calendar-policy-on-religious-holidays.html)

Disability Disclosure Statement

Academic accommodations are available to any student with a chronic, psychological, visual, mobility, learning disability, or who is deaf or hard of hearing. Students should please register with the Moses Center for Students with Disabilities at 212-998-4980.

NYU's Henry and Lucy Moses Center for Students with Disabilities
726 Broadway, 2nd Floor
New York, NY 10003-6675
Telephone: 212-998-4980
Voice/TTY Fax: 212-995-4114
Web site: http://www.nyu.edu/csd