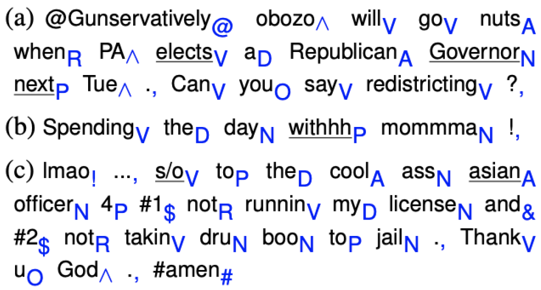
Homework 3: Energy-Based Models

This is the third assignment for DS-GA 1008 Deep Learning. Out: October 19, 2020. Due: October 30, 2020; 11:55 pm. Per the course syllabus, the solutions must be submitted on or before the deadline, and they must be typeset; handwritten answers will not be accepted. You will need to submit a zip folder which contains two things to NYU classes: (1) the typeset pdf containing solutions to problems 1 through 5 (no need to copy down the problem statements); (2) the downloaded notebook corresponding to problem 6. Please name your zip folder `lastname-firstname-netid-hw3`.

The first question is on feed-forward models. The rest of the assignment is about energy-based models.

Problem 1. (5/100 points.)



(a) @Gunservatively$_{@}$ obozo$_{\wedge}$ will$_{V}$ go$_{V}$ nuts$_{A}$ when$_{R}$ PA$_{\wedge}$ elects$_{V}$ a$_{D}$ Republican$_{A}$ Governor$_{N}$ next$_{P}$ Tue$_{\wedge}$.$_{,}$ Can$_{V}$ you$_{O}$ say$_{V}$ redistricting$_{V}$?$_{,}$

(b) Spending$_{V}$ the$_{D}$ day$_{N}$ withhh$_{P}$ mommma$_{N}$!$_{,}$

(c) lmao$_{!}$ ...$_{,}$ s/o$_{V}$ to$_{P}$ the$_{D}$ cool$_{A}$ ass$_{N}$ asian$_{A}$ officer$_{N}$ 4$_{P}$ #1$_{\$}$ not$_{R}$ runnin$_{V}$ my$_{D}$ license$_{N}$ and$_{\&}$ #2$_{\$}$ not$_{R}$ takin$_{V}$ dru$_{N}$ boo$_{N}$ to$_{P}$ jail$_{N}$.$_{,}$ Thank$_{V}$ u$_{O}$ God$_{\wedge}$.$_{,}$ #amen$_{\#}$

Figure 1: Examples of Twitter POS tagging. The black-font tweet is the input, and the blue-font tag sequence is the output. This is a screenshot of the paper in footnote 1.

(a) Consider the task Twitter part-of-speech (POS) tagging. See Table 1 in the paper[1] for explanations of the tags. The input is a sentence (a tweet), and the output is a sequence of POS tags. The output sequence length equals input sequence length given that each input token corresponds to one output POS tag. Suppose there are L possible tags. Suppose during training, an input sentence is $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T)$, and the correct sequence of POS tags is $\boldsymbol{y} = (y_1, y_2, \ldots, y_T)$ where $y_i \in \{1, 2, \ldots, L\}$ for all i. During training, we obtain $\tilde{\boldsymbol{y}} = f_\theta(\boldsymbol{x})$ where f_θ is a three-layer bidirectional LSTM (BiLSTM) and $\tilde{\boldsymbol{y}} \in [0,1]^{T \times L}$ where $\tilde{\boldsymbol{y}}_t$ is the categorical probability distribution (over all possible tags) for the POS tag at time-step t. Write down the negative log-likelihood loss (which you would use to train f_θ) using $\tilde{\boldsymbol{y}}$ and $\boldsymbol{y}$.

(b) Recall that in Lab 4[2], we were doing (left-to-right) language modeling. During training, the input is a sentence, and the gold output is the same

[1] https://www.aclweb.org/anthology/P11-2008.pdf

[2] Lab on September 28, 2020: language modeling using recurrent networks and transformer.

sentence shifted by one time-step. For both POS tagging and language modeling, both the input and the output are sequences. Why is it okay to use BiLSTM for POS tagging, but not in Lab 4, when we were doing language modeling (and we needed to use unidirectional LSTM instead)? Two or three sentences would suffice.

For the rest of the assignment, we will focus on energy-based models.

Problem 2. (5/100 points.) Open-ended question. Based on the lecture, discuss the pros and cons of energy-based models. You may want to touch upon modeling capabilities, inference speed, etc. Feel free to use any specific examples from the lecture. Write down at least two pros (in bullet points) and one cons.

Problem 3. (5/100 points.) Exact inference.

(a) Suppose $\mathcal{V} = \{\texttt{really}, \texttt{good}, \texttt{food}\}$. Suppose the input is an exactly-three-word-long sentence. Suppose for each word in a sentence, the POS tag could be one of the following $\{\texttt{noun}, \texttt{adjective}, \texttt{adverb}, \texttt{interjection}\}$. There are $3^3 = 27$ different sentences (fluent and disfluent). There are $4^3 = 64$ different sequences of POS tags. Given a perfect-quality energy function $E(\boldsymbol{x}, \boldsymbol{y})$ (input to E: a sentence and a sequence of POS tags; output of E: the corresponding scalar energy) that has 27×64 possible $(\boldsymbol{x}, \boldsymbol{y})$ input pairs, given a sentence $\boldsymbol{x}$ (e.g., `good food really`, or, `really really good`), write down the equation to obtain $\boldsymbol{y}$, and explain the equation in words. Moreover, how many times do you need to look up E in order to obtain the best output sequence of POS tags? (Suppose the lookup table corresponding to E is not ordered in any way.)

(b) (No need to explain reasoning for this part.) Suppose the vocabulary size $|\mathcal{V}| = 50000$, suppose the number of possible POS tags is 20. Suppose the input is exactly-15-word-long sentences. Similar to (a), E stores an energy for each possible $(\boldsymbol{x}, \boldsymbol{y})$ pair. (i) How many different $(\boldsymbol{x}, \boldsymbol{y})$ pairs are there? Note that again, we do not consider the fluency of the sentence. No need to compute the exact number given it is a huge number. (ii) Given a sentence, how many times do you need to look up E in order to obtain the best output sequence of POS tags? (Suppose the lookup table corresponding to E is not ordered in any way.)

In class, we discussed many strategies to shape the energy function. They can be grouped into two classes of learning methods: contrastive methods and architectural/regularized methods. Problem 4 deals with the contrastive methods. In particular, we investigate the contrastive methods using a specific example.

Problem 4. (25/100 points.) Let's keeping using the POS tagging example. Suppose there are L possible POS tags. Suppose there are N examples in the dataset; each example contains an input sentence and an output sequence of tags. Here, let $\boldsymbol{x}_i = (\boldsymbol{x}_{i,1}, \ldots, \boldsymbol{x}_{i,T_i})$ be the input sentence in the i-th example in the dataset, and let $\boldsymbol{y}_i = (y_{i,1}, \dots, y_{i,T_i})$ be the sequence of gold-standard tags in the i-th example in the dataset. Look at the following objective for training the energy function.

$$\min_{\theta} \sum_{i=1}^{N} \left[m + E_\theta(\boldsymbol{x}_i, \boldsymbol{y}_i) - \min_{\boldsymbol{y} \neq \boldsymbol{y}_i} E_\theta(\boldsymbol{x}_i, \boldsymbol{y}) \right]_+ , \tag{1}$$

The energy function can be parametrized arbitrarily; but we usually integrate "domain knowledge" into the energy function. For sequence labeling tasks like POS tagging, an energy function could capture the dependency between a part of sentence and a part of tag sequence, and it could also capture the dependency between a part of the tag sequence and other parts of the tag sequence (perhaps inspired by conditional random field). No need to understand the exact formulations for energy function today, but if you are interested, check out Section 3 of this recent paper[3] or Section 6 of this paper[4].

(a) (5 pt) Explain in words: what is $E_\theta(\boldsymbol{x}_i, \boldsymbol{y}_i)$; what is $\min_{\boldsymbol{y} \neq \boldsymbol{y}_i} E_\theta(\boldsymbol{x}_i, \boldsymbol{y})$?

(b) (5 pt) Explain the objective such that your response covers the following three questions. Why do we want to minimize such a function? Why is there a margin m? Why is there a $[\cdot]_+$?

(c) (5 pt) Given an energy function E_θ, given $\hat{\boldsymbol{y}}_i = \arg\min_{\boldsymbol{y} \neq \boldsymbol{y}_i} E_\theta(\boldsymbol{x}_i, \boldsymbol{y})$, plot the curve of the loss for this one example (from Objective (1)) as a *function*[5] of $E_\theta(\boldsymbol{x}_i, \hat{\boldsymbol{y}}_i) - E_\theta(\boldsymbol{x}_i, \boldsymbol{y}_i)$ and explain your reasoning thoroughly.

(d) (5 pt) In class, we mentioned "gradient descent for inference." In our case, given a trained energy function E_θ, given an input $\boldsymbol{x}$, suppose we want to do gradient descent inference. Write down the gradient step, in the form of $\diamond \leftarrow \circ - \eta \cdot \triangleleft$, where $\eta \in (0,1)$ is an coefficient.

(e) (5 pt) Now we do some slight modification to the objective function. See Objective (2). Define Δ to be some distance function between the two input arguments. Define $\mathcal{Y}(\boldsymbol{x}_i)$ to be the output space corresponding to input $\boldsymbol{x}_i$ (so the output space is a set of sequences of tags with length equal to length of $\boldsymbol{x}_i$).

  (i) $\Delta(\boldsymbol{y}, \boldsymbol{y}')$ could be the number of mismatches between the tag sequence $\boldsymbol{y}$ and the tag sequence $\boldsymbol{y}'$. Give another example (perhaps "better") of Δ that's not a constant function. You can check out Table 1 of footnote 1 for inspiration.

  (ii) Explain in detail what the maximization-step does.

  (iii) Explain what the following objective may achieve (to the energy function) but Objective (1) may not.

[3] https://arxiv.org/pdf/2010.02789.pdf#page=3

[4] https://arxiv.org/pdf/1703.05667.pdf#page=5

[5] Suppose $y = f(x)$ where f is a function. Each x can only correspond to at most one y.

$$\min_{\theta} \sum_{i=1}^{N} \max_{\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{x}_i)} [\Delta(\boldsymbol{y}, \boldsymbol{y}_i) + E_\theta(\boldsymbol{x}_i, \boldsymbol{y}_i) - E_\theta(\boldsymbol{x}_i, \boldsymbol{y})]_+ , \tag{2}$$

In class, we discussed many strategies to shape the energy function. They can be grouped into two classes of learning methods: contrastive methods and architectural/regularized methods. Problem 4 deals with the contrastive methods. Now, we will investigate architectural/regularized methods. We will also discuss architectural/regularized methods more in depth in the next few lectures as well as in the next assignment.

Problem 5. (20/100 points.)

(a) (4 pt) What are contrastive, architectural, and regularized methods, respectively?

(b) (4 pt) Sparse coding is a type of energy-based model. Write down the free energy where we use L_1 sparsity penalty. Suppose the code (i.e., latent) is $\boldsymbol{z}$. Suppose the decoder is called `Dec`. Define other variables as you wish. Make sure to clearly explain every symbol you're using in the equation.

(c) (4 pt) (Continuing the previous part.) If we are using images as examples, why does the model fail if we remove the regularizer in the objective function? *Hint: what happens to image reconstruction; what happens to energy function?*

(d) (4 pt) (Continuing the previous part.) What is the role of the regularizer? What does the regularizer do to the energy function? Also mention: with the regularizer, what images will have large energy and what images will have small energy, using a well-trained energy function ? (Feel free to use part of the previous part's answer in this part.)

(e) (4 pt) Finally, explain why k-means clustering is an architectural method. What is the free energy function in this case? Explain each variable in your equation; please also include the shapes of each variable. (Regarding shapes: you can make appropriate assumptions but please explain your assumptions.)

Problem 6. (40/100 points.) Implementation: denoising convolutional auto-encoder related. All the write-up for this problem should go into the colab. In your homework submission, you will need to submit the downloaded version of the notebook (*not* a URL).

If you need help, don't forget about the six weekly office hours. Best of luck with the assignment, and thanks for your hard work!