HW3: (Neural) Language Modeling

Nicolas Drizard
nicolasdrizard@g.harvard.edu

Virgile Audi
vaudi@g.harvard.edu

March 10, 2016

1 Introduction

This assignment focuses on the task of language modeling, a crucial first-step for many natural language applications. In this report, we will present several count-based multinomial language models with different smoothing methods, an influential neural network based language model from the work of Bengio et al. (2003), and an extension to this language model which learns using noise contrastive estimation, as well as their implementation using Torch. We found this homework more challenging than the previous ones and encountered significant challenges that we will underline in this report.

2 Problem Description

The goal of the language models presented in this report is to learn a distributed representation for words as well as probability distribution for word sequences. Language models are usually represented as the probability of generating a new word conditioned on the preceeding words:

$$P(\boldsymbol{w}_{1:n}) = \prod_{i=1}^{n-1} P(w_{i+1}|w_i)$$

To simplify the analysis, it is common to make the assumption that a word is influenced only by the N words immediately preceeding it, which we call the context. Even with reasonably small values for N, building such models are extremely expensive computationally-wise as well as time-consuming if not ran on GPU. The joint proabibility of a sequence of 6 words taken from a vocabulary of 10 000 words could possibly imply training the model to fit up to $10^{4^6} - 1 = 10^{24} - 1$ parameters.

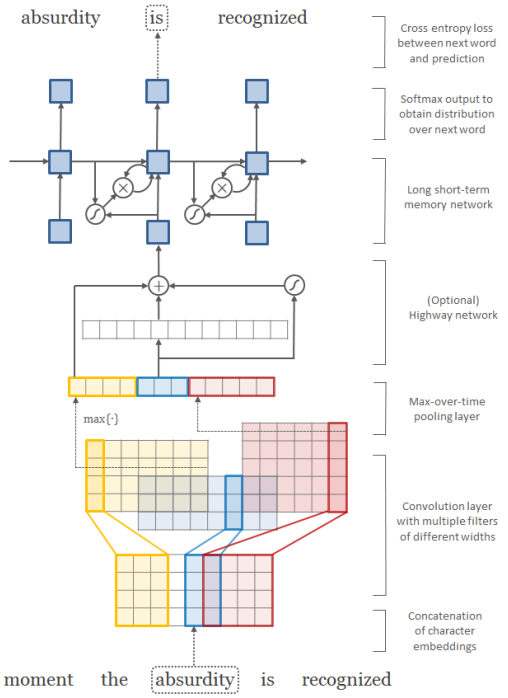
In this report, we tried implementing N-grams models in an efficient manner. Due to computational limitations (no access to GPUs...), we faced difficulties training the Neural Network and focused our efforts on building strong count-based models to solve the problem of language modeling.

3 Model and Algorithms

Here you specify the model itself. This section should formally describe the model used to solve the task proposed in the previous section. This section should try to avoid introducing new vocabulary or notation, when possible use the notation from the previous section. Feel free to use the notation from class, but try to make the note understandable as a standalone piece of text.

This section is also a great place to include other material that describes the underlying structure and choices of your model, for instance here are some example tables and algorithms from full research papers:

- diagrams of your model,



- feature tables,

Mention Features	
Feature	Value Set
Mention Head	$\mathcal{V}$
Mention First Word	$\mathcal{V}$
Mention Last Word	$\mathcal{V}$
Word Preceding Mention	$\mathcal{V}$
Word Following Mention	$\mathcal{V}$
# Words in Mention	$\{1, 2, \ldots\}$
Mention Type	$\mathcal{T}$

- pseudo-code,

procedure LINEARIZE($x_1 \dots x_N$, K, g)
  $B_0 \leftarrow \langle(\langle\rangle, \{1, \dots, N\}, 0, \boldsymbol{h}_0, \mathbf{0})\rangle$
  for $m = 0, \dots, M-1$ **do**
    for $k = 1, \dots, |B_m|$ **do**
      for $i \in \mathcal{R}$ **do**
        $(y, \mathcal{R}, s, \boldsymbol{h}) \leftarrow \text{copy}(B_m^{(k)})$
        for word w in phrase x_i **do**
          $y \leftarrow y$ append w
          $s \leftarrow s + \log q(w, \boldsymbol{h})$
          $\boldsymbol{h} \leftarrow \delta(w, \boldsymbol{h})$
        $B_{m+|w_i|} \leftarrow B_{m+|w_i|} + (y, \mathcal{R} - i, s, \boldsymbol{h})$
        keep top-K of $B_{m+|w_i|}$ by $f(x, y) + g(\mathcal{R})$
  return $B_M^{(k)}$

4 Experiments

Finally we end with the experimental section. Each assignment will make clear the main experiments and baselines that you should run. For these experiments you should present a main results table. Here we give a sample Table 1. In addition to these results you should describe in words what the table shows and the relative performance of the models.

Besides the main results we will also ask you to present other results comparing particular aspects of the models. For instance, for word embedding experiments, we may ask you to show a chart of the projected word vectors. This experiment will lead to something like Figure 1. This should also be described within the body of the text itself.

Model	Acc.
BASELINE 1	0.45
BASELINE 2	2.59
MODEL 1	10.59
MODEL 2	13.42
MODEL 3	7.49

Table 1: Table with the main results.

5 Conclusion

End the write-up with a very short recap of the main experiments and the main results. Describe any challenges you may have faced, and what could have been improved in the model.

Figure 1: Sample qualitative chart.

References

Berger, A. L., Pietra, V. J. D., and Pietra, S. A. D. (1996). A maximum entropy approach to natural language processing. *Computational linguistics*, 22(1):39–71.

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.