gobbli: A uniform interface to deep learning for text in Python

Jason Nance¹ and Peter Baumgartner¹

¹ RTI International

Summary

Machine learning has long been used to address natural language processing (NLP) tasks like sentiment analysis (Pang & Lee, 2008) and document classification (Aggarwal & Zhai, 2012). Traditional approaches to these tasks require numerous labeled examples from the specific domains in which they will be applied. Such algorithms can only use the available training data, which is often limited in size and diversity, to learn to understand natural language. In the last few years, transfer learning (Weiss et al., 2016) has caused a paradigm shift in NLP. Rather than training distinct task-specific models from scratch, a transfer learning model first learns the rules of language from a large, diverse text corpus during an extensive self-supervised training regimen. Self-supervised tasks are formulated such that unlabeled data can be used to train supervised models (Raina et al., 2007); an example is masked language modeling, where a subset of words in a document are masked out, and the model predicts the masked words (Devlin et al., 2018). The transfer learning model thus learns a rich representation of language, which can be fine-tuned to solve specific problems. According to Torrey & Shavlik (2010), this approach mimics how humans reuse their general understanding of language across tasks. Transfer learning has not only rapidly advanced the state of the art in classification but has enabled near-human performance on more advanced tasks like question answering (Raffel et al., 2019; Rajpurkar et al., 2016) and natural language inference (Bowman et al., 2015; Rajpurkar et al., 2016; Williams et al., 2018).

While the performance gains on benchmark tasks are undeniable, applied researchers face challenges using transfer learning models to solve new problems. A wide variety of models are being developed by disparate research teams using different technologies (Liu et al., 2019; Raffel et al., 2019; Sun et al., 2019). A practitioner may therefore be required to learn a new programming language, a deep learning library, a containerization technology, and a model interface whenever they want to evaluate the feasibility of a new model on a custom task. gobbli was developed to address this problem.

gobbli is a Python library intended to bridge state-of-the-art research in natural language processing and application to real-world problems. The library defines a simple interface for training classification models, producing predictions, and generating embeddings. Several models implementing the interface are available using programmatically-created Docker containers to abstract away differences in underlying deep learning libraries and model hyperparameters. This approach allows users to easily evaluate models and compare performance across model types without spending time adapting their dataset and use case to each model. Compared to other deep learning libraries used for NLP like transformers (Wolf et al., 2019) and fastai (Howard & others, 2018), gobbli is designed to emphasize simplicity and interoperability rather than customization and performance in order to make deep learning more accessible to applied researchers.

Beyond its model wrappers, gobbli provides several helpful utilities for NLP practitioners. Data augmentation has emerged as a popular technique to improve model performance when
training data is limited. Multiple methods for data augmentation are implemented in gobbl, including backtranslation (Shleifer, 2019), word replacement (Wei & Zou, 2019), and contextual augmentation (Kobayashi, 2018). These methods can be used independently of gobbl models for interoperability with other modeling libraries. gobbl also bundles a set of interactive web applications built using Streamlit (Teixeira et al., 2018) which can be used to explore a dataset, visualize embeddings, evaluate model performance, and explain model predictions without writing any code.

gobbl was developed from experiences on client contracts and internal projects at RTI International. It is intended for use by anyone solving problems using applied NLP, including researchers, students, and industry practitioners.

Acknowledgments

Work on gobbl was funded by RTI International.

References


