What can we learn from Semantic Tagging?

Anonymous EMNLP submission

Abstract

We investigate the effects of multi-task learning using the recently introduced task of semantic tagging. We employ semantic tagging as an auxiliary task for three different NLP tasks: part-of-speech tagging, Universal Dependency parsing, and Natural Language Inference. We compare full neural network sharing, partial neural network sharing, and what we term the learning what to share setting where negative transfer between tasks is less likely. Our findings show considerable improvements for all tasks, particularly in the learning what to share setting which shows consistent gains across all tasks.

1 Introduction

Multi-task learning (MTL) is a recently resurgent approach to machine learning in which multiple tasks are simultaneously learned. By optimising the multiple loss functions of related tasks at once, multi-task learning models can achieve superior results compared to models trained on a single task. The key principle is summarized by Caruana (1998) as “MTL improves generalization by leveraging the domain-specific information contained in the training signals of related tasks”. Neural MTL has become an increasingly successful approach by exploiting similarities between Natural Language Processing (NLP) tasks (Collobert and Weston, 2008; Søgaard and Goldberg, 2016; Plank et al., 2016). Our work builds upon Bjerva et al. (2016), who demonstrate that employing semantic tagging as an auxiliary task for Universal Dependency (McDonald et al., 2013) part-of-speech tagging can lead to improved performance.

The objective of this paper is to investigate whether learning to predict lexical semantic categories can be beneficial to other NLP tasks. To achieve this we augment single-task models (ST)\(^1\) with an additional classifier to predict semantic tags and jointly optimize for both the original task and the auxiliary semantic tagging task. Our hypothesis is that learning to predict semantic tags as an auxiliary task can improve performance of single-task systems. We believe that this is, among other factors, due to the following:

- Providing the main task’s model with a useful inductive bias, encouraging it to prefer representations that lead to semantically plausible hypotheses over those that are not.
- Putting the focus of the main task model’s attention on features that actually matter by providing additional evidence for the relevance or irrelevance of those features.
- Reducing the risk of overfitting by minimizing the model’s Rademacher Complexity.\(^2\) Representations which are learned for multiple tasks have been shown to generalize better (Baxter et al., 2000).

We test our hypothesis on three disparate NLP tasks: (i) Universal Dependency part-of-speech tagging (UPOS), (ii) Universal Dependency parsing (UD DEP), a complex syntactic task; and (iii) Natural Language Inference (NLI), a complex task requiring deep natural language understanding.

2 Background and Related work

2.1 Semantic Tagging

Semantic tagging (Bjerva et al., 2016; Abzianidze and Bos, 2017) is the task of assigning language-neutral semantic categories to words. It is designed to overcome a lack of semantic information syntax-oriented part-of-speech tagsets, such as the state-of-the-art. Our choice of models is based on replicability.\(^2\)

\(^1\)We replicate models which perform at or close to the

\(^2\)The ability to fit random noise.
The semantic tagset is language-neutral, abstracts over part-of-speech and named-entity classes, and includes fine-grained semantic information. The tagset consists of 80 semantic tags grouped in 13 coarse-grained classes. The tagset originated in the Parallel Meaning Bank (PMB) project (Abzianidze et al., 2017), where it contributes to compositional semantics and cross-lingual projection of semantic representations. Recent work has highlighted the utility of the tagset as a conduit for evaluating the semantics captured by vector representations (Belinkov et al., 2018), or employed it in an auxiliary tagging task (Bjerva et al., 2016), as we do in this work.

2.2 Learning What to Share

Recently, there has been an increasing interest in the development of models which are trained to learn what to (and what not to) share between a set of tasks, with the general aim of preventing negative transfer when the tasks are not closely related (Meyerson and Miikkulainen, 2017; Ruder et al., 2017; Lu et al., 2017; Misra et al., 2016). Our Learning What to Share setting is based on this idea and closely related to Liu et al. (2016)’s shared layer architecture.

Specifically, a layer $\hat{h}_X$ which is shared between the main task and the auxiliary task is split into two subspaces: a shared subspace $\hat{h}_{X_S}$ and a private subspace $\hat{h}_{X_P}$. The interaction between the shared subspaces is modulated via a sigmoidal gating unit applied to a set of learned weights, as seen in Equations (1) and (2) where $\hat{h}_{X_{S(main)}}$ and $\hat{h}_{X_{S(aux)}}$ are the main and auxiliary tasks’ shared layers, $W_{a \rightarrow m}$ and $W_{m \rightarrow a}$ are learned weights, and $\sigma$ is a sigmoidal function.

$$\hat{h}_{X_{S(main)}} = \sigma(\hat{h}_{X_{S(aux)}} W_{a \rightarrow m}) \quad (1)$$
$$\hat{h}_{X_{S(aux)}} = \sigma(\hat{h}_{X_{S(main)}} W_{m \rightarrow a}) \quad (2)$$

Unlike Liu et al. (2016), in our setup, each task has its own shared subspace rather than one common shared layer. This enables the sharing of different parameters in each direction (i.e., from main to auxiliary task and from auxiliary to main task).

3 Multi-task Learning Settings

We implement three neural MTL settings, shown in Figure 1. They differ in the way the network’s parameters are shared between the tasks:

- **Fully shared network (FSN):** All hidden layers are entirely shared among the tasks, each task has a separate output layer;
- **Partially shared network (PSN):** A subset of hidden layers is shared among the tasks; each task has at least one private hidden layer and a separate output layer;
- **Learning What to Share (LWS):** Each task has a dedicated set of hidden layers. For sharing, a hidden layer is split into a shared subspace and a private subspace. A gating unit modulates the transfer of information between the shared subspaces as shown in Equations (1) and (2).
4 Data

In the UPOS tagging experiments, we utilize the UD 2.0 English corpus (Nivre et al., 2017) for the POS tagging and the semantically tagged PMB release (sem-PMB) 3 for the MTL settings. Note that there is no overlap between the two datasets. Conversely, for the UD DEP and NLI experiments there is a complete overlap between the datasets of main and auxiliary tasks, i.e., each instance is labeled with both the main task’s labels and semantic tags. We use the Stanford POS Tagger (Toutanova et al., 2003) trained on sem-PMB to tag the UD corpus and NLI datasets with semantic tags, and then use those assigned tags for the MTL settings of our dependency parsing and NLI models. We find that this approach leads to better results when the main task is only loosely related to the auxiliary task. The UD DEP experiments use the English UD 2.0 corpus, and the NLI experiments use the SNLI (Bowman et al., 2015) and SICK-E datasets (Marelli et al., 2014). The provided train, development, and test splits are used for all datasets. For sem-PMB, the silver and gold parts are used for training and testing respectively.

5 Experiments

We run four experiments for each of the four tasks (UPOS, UD DEP, SNLI, SICK-E), one using the ST model and one for each of the three MTL settings. Each experiment is run five times, and the average of the five runs is reported. We briefly describe the ST models and refer the reader to the original work for further details due to a lack of space. 4 For reproducibility, detailed diagrams of the MTL models for each task and their hyperparameters can be found in Appendix A.

5.1 Universal Dependency POS Tagging

Our tagging model uses a basic contextual one-layer bi-LSTM (Hochreiter and Schmidhuber, 1997) that takes in word embeddings and produces a sequence of recurrent states which can be viewed as contextualized representations. The recurrent $r_n$ state from the bi-LSTM corresponding to each time-step $t_n$ is passed through a dense layer with a softmax activation to predict the token’s tag.

In each of the MTL settings a softmax classifier is added to predict a token’s semantic tag and the model is then jointly trained on the concatenation of the sem-PMB and UPOS tagging data to minimize the sum of softmax cross-entropy losses of both the main (UPOS tagging) and auxiliary (semantic tagging) tasks.

5.2 Universal Dependency Parsing

We employ a parsing model that is based on Dozat and Manning’s (2016) deep biaffine attention dependency parser. The model’s embeddings layer is a concatenation of randomly initialized word embeddings 5 and character-based word representations added to pre-trained word embeddings, which are passed through a 4-layer stacked bi-LSTM. Unlike Dozat and Manning (2016), our model jointly learns to perform UPOS tagging and parsing, instead of treating them as separate tasks. Therefore, instead of tag embeddings, we add a softmax classifier to predict UPOS tags after the first bi-LSTM layer. The outputs from that layer and the UPOS softmax prediction vectors are both concatenated to the original embedding layer and passed to the second bi-LSTM layer. The output of the last bi-LSTM is then used as input for four dense layers with a ReLU activation, producing four vector representations: a word as a dependent seeking its head; a word as a head seeking all its dependents; a word as a dependent deciding on its label; a word as head deciding on the labels of its dependents. These representations are then passed to biaffine and affine softmax classifiers to produce a fully-connected labeled probabilistic dependency graph (Dozat and Manning, 2016). Finally, a non-projective maximum spanning tree parsing algorithm (Chu, 1965; Edmonds, 1967) is used to obtain a well-formed dependency tree. 6

Similarly to UPOS tagging, an additional softmax classifier is used to predict a token’s semantic tag in each of the MTL settings, as both tasks are jointly learned. In the FSN setting, the 4-layer stacked bi-LSTM is entirely shared. In the PSN setting the semantic tags are predicted from the second layer’s hidden states, and the final two layers are devoted to the parsing task. In the LWS setting, the first two layers of the bi-LSTM are split into a private bi-LSTM private and a shared bi-LSTM shared for each of the tasks with the inter-

3http://pmb.let.rug.nl/data.php
4This applies to UD DEP and NLI only, as the POS tagger is not based on any one particular work.
5This replaces the holistic word embeddings for frequent words in Dozat and Manning (2016).
6This is recommended but not implemented by Dozat et al. (2017).
action between the shared subspaces being modulated via a gating unit. Then, two bi-LSTM layers that are devoted to parsing only are stacked on top.

5.3 Natural Language Inference
We base our NLI model on Chen et al. (2017)’s Enhanced Sequential Inference Model which uses a bi-LSTM to encode the premise and hypothesis, computes a soft-alignment between premise and hypothesis’ representations using an attention mechanism, and employs an inference composition bi-LSTM to compose local inference information sequentially.1 The MTL settings are implemented by adding a softmax classifier to predict semantic tags at the level of the encoding bi-LSTM, with rest of the model unaltered. In the FSN setting, the hidden states of the encoding bi-LSTM are directly passed as input to the softmax classifier. In the PSN setting an earlier bi-LSTM layer is used to predict the semantic tags and the output from that is passed on to the encoding bi-LSTM which is stacked on top. This follows Hashimoto et al. (2016)’s hierarchical approach. In the LWS setting, a bi-LSTM layer with private and shared subspaces is used for semantic tagging and for the ESIM model’s encoding layer. In all MTL settings, the bi-LSTM used for semantic tagging is pre-trained on the sem-PMB data.

6 Results and Discussion
Results for all tasks are shown in Table 1. In line with Bjerva et al. (2016)’s findings, the FSN setting leads to an improvement for UPOS tagging. POS tagging, a sequence labeling task, can be seen as the most closely related to semantic tagging, therefore negative transfer is minimal and the full sharing of parameters is beneficial. Surprisingly, the FSN setting also leads improvements for UD DEP. Indeed, for UD DEP, all of the MTL models outperform the ST model by increasing margins. For the NLI tasks, however, there is a clear degradation in performance.

The PSN setting shows mixed results and does not show a clear advantage over FSN for UPOS and UD DEP. This suggests that adding task-specific layers after fully-shared ones does not always enable sufficient task specialization. For the NLI tasks however, PSN is clearly preferable to FSN, especially for the small-sized SICK-E dataset where the FSN model fails to adequately learn.

<table>
<thead>
<tr>
<th>Model</th>
<th>SNLI</th>
<th>SICK-E</th>
<th>UPOS</th>
<th>UD DEP</th>
</tr>
</thead>
<tbody>
<tr>
<td>ST</td>
<td>87.01</td>
<td>81.30</td>
<td>92.12</td>
<td>80.24 / 84.87</td>
</tr>
<tr>
<td>FSN</td>
<td>84.96</td>
<td>56.69</td>
<td>92.95</td>
<td>81.03 / 85.54</td>
</tr>
<tr>
<td>PSN</td>
<td>87.08</td>
<td>77.92</td>
<td>92.34</td>
<td>80.92 / 85.81</td>
</tr>
<tr>
<td>LWS</td>
<td>87.51</td>
<td>84.57</td>
<td>95.54</td>
<td>81.39 / 86.00</td>
</tr>
</tbody>
</table>

Table 1: Results for single-task models (ST), fully-shared networks (FSN), partially-shared networks (PSN), and Learning What to Share (LWS). All scores are reported as accuracy, except UD DEP for which we report LAS/UAS $F_1$ score.

As a sentence-level task, NLI is functionally dissimilar to semantic tagging. However, it is a task which requires deep understanding of natural language semantics and can therefore conceivably benefit from the signal provided by semantic tagging. Our results demonstrate that it is possible to leverage this signal given a selective sharing setup where negative transfer can be minimized. Indeed, for the NLI tasks, only the LWS setting leads to improvements over the ST models.8 The improvement is larger for the SICK-E task which has a much smaller training set and therefore stands to learn more from the semantic tagging signal. For all tasks, it can be observed that the LWS models outperform the rest of the models. This is in line with our expectations with the findings from previous work (Ruder et al., 2017; Liu et al., 2016) that selective sharing outperforms full network and partial network sharing.

7 Conclusions
We present a comprehensive evaluation of MTL using a recently proposed task of semantic tagging as an auxiliary task. Our experiments span three types of NLP tasks and three MTL settings. The results of the experiments show that employing semantic tagging as an auxiliary task leads to improvements in performance for UPOS tagging and UD DEP in all MTL settings. For the SNLI tasks, requiring understanding of phrasal semantics, the selective sharing setup we term Learning What to Share holds a clear advantage. Our work offers a generalizable framework for the evaluation of the utility of an auxiliary task.

---

1We do not implement the additional tree-LSTM model used in Chen et al. (2017) as we focus on the effect of MTL with semantic tagging rather than on absolute performance.

8Demonstrative examples of the SNLI models’ outputs can be found in Appendix B.
References


