An Embarrassingly Simple Approach for Transfer Learning from Pretrained Language Models

Alexandra Chronopoulou[1], Christos Baziotis[1], Alexandros Potamianos[1,2]

[1]School of ECE, National Technical University of Athens, Athens, Greece
[2] Signal Analysis and Interpretation Laboratory (SAIL), USC, Los Angeles, USA

el12068@central.ntua.gr, cbaziotis@mail.ntua.gr
potam@central.ntua.gr

Abstract

A growing number of state-of-the-art transfer learning methods employ language models pretrained on large generic corpora. In this paper we present a conceptually simple and effective transfer learning approach that addresses the problem of catastrophic forgetting. Specifically, we combine the task-specific optimization function with an auxiliary language model objective, which is adjusted during the training process. This preserves language regularities captured by language models, while enabling sufficient adaptation for solving the target task. Our method does not require pretraining or finetuning separate components of the network and we train our models end-to-end in a single step. We present results on a variety of challenging affective and text classification tasks, surpassing well established transfer learning methods with greater level of complexity.

1 Introduction

Pretrained word representations captured by Language Models (LMs) have recently become popular in Natural Language Processing (NLP). Pretrained LMs encode contextual information and high-level features of language, modeling syntax and semantics, producing state-of-the-art results across a wide range of tasks, such as named entity recognition (Peters et al., 2017), machine translation (Ramachandran et al., 2017) and text classification (Howard and Ruder, 2018).

However, in cases where contextual embeddings from language models are used as additional features (e.g. ELMo (Peters et al., 2018)), results come at a high computational cost and require task-specific architectures. At the same time, approaches that rely on fine-tuning a LM to the task at hand (e.g. ULMFiT (Howard and Ruder, 2018)) depend on pretraining the model on an extensive vocabulary and on employing a sophisticated

slanted triangular learning rate scheme to adapt the parameters of the LM to the target dataset.

We propose a simple and effective transfer learning approach, that leverages LM contextual representations and does not require any elaborate scheduling schemes during training. We initially train a LM on a Twitter corpus and then transfer its weights. We add a task-specific recurrent layer and a classification layer. The transferred model is trained end-to-end using an auxiliary LM loss, which allows us to explicitly control the weighting of the pre-trained part of the model and ensure that the distilled knowledge it encodes is preserved.

Our contributions are summarized as follows: 1) We show that transfer learning from language models can achieve competitive results, while also being intuitively simple and computationally effective. 2) We address the problem of catastrophic forgetting, by adding an auxiliary LM objective and using an unfreezing method. 3) Our results show that our approach is competitive with more sophisticated transfer learning methods. Our code is publicly available (will be released in future version).

2 Related Work

Unsupervised pretraining has played a key role in deep neural networks, building on the premise that representations learned for one task can be useful for another task. In NLP, pretrained word vectors (Mikolov et al., 2013; Pennington et al., 2014) are widely used, improving performance in various downstream tasks, such as part-of-speech tagging (Collobert et al., 2011) and question answering (Xiong et al., 2016).

Aiming to learn from unlabeled data, Dai and Le (2015) use unsupervised objectives such as sequence autoencoding and language modeling as pretraining representations. Ramachandran et al.

(2017) also pretrain encoder-decoder pairs using language models and fine-tune them to a specific task. ELMo embeddings (Peters et al., 2018) are obtained from character-based bidirectional language models improving the results in a variety of tasks as additional contextual representations.

Towards the same direction, ULMFiT (Howard and Ruder, 2018) shows impressive results on a variety of tasks by employing pretrained LMs. The proposed pipeline requires three distinct steps, that include pretraining the LM, fine-tuning it on a target dataset with an elaborate scheduling procedure and transferring it to a classification model.

Multi-Task Learning (MTL) via hard parameter sharing (Caruana, 1993) in neural networks has proven to be effective in many NLP problems (Collobert and Weston, 2008). More recently, alternative approaches have been suggested that only share parameters across lower layers (Søgaard and Goldberg, 2016). By introducing part-of-speech tags at the lower levels of the network, the proposed model achieves competitive results on chunking and CCG super tagging. Our auxiliary language model objective follows this line of thought and intends to boost the performance of the higher classification layer.

3 Our Model

We introduce **SiATL**, which stands for **Si**ngle-step **A**uxiliary loss **T**ransfer **L**earning. In our proposed approach, we first train a LM. We then transfer its weights and add a task-specific recurrent layer to the final classifier. We also employ an auxiliary LM loss to avoid catastrophic forgetting.

LM Pretraining. We train a word-level language model, which consists of an embedding LSTM layer (Hochreiter and Schmidhuber, 1997), 2 hidden LSTM layers and a linear layer. We want to minimize the negative log-likelihood of the LM:

$$L(\hat{p}) = -\frac{1}{N} \sum_{n=1}^{N} \sum_{t=1}^{T^n} \log \hat{p}(x_t^n | x_1^n, ..., x_{t-1}^n) \quad (1)$$

where $\hat{p}(x_t^n | x_1^n, ..., x_{t-1}^n)$ is the distibution of the tth word in the nth sentence given the $t-1$ words preceding it and N is total number of sentences.

Transfer & auxiliary loss. We transfer the weights of the pretrained model and add one LSTM with a self-attention mechanism (Lin et al., 2017; Bahdanau et al., 2015).
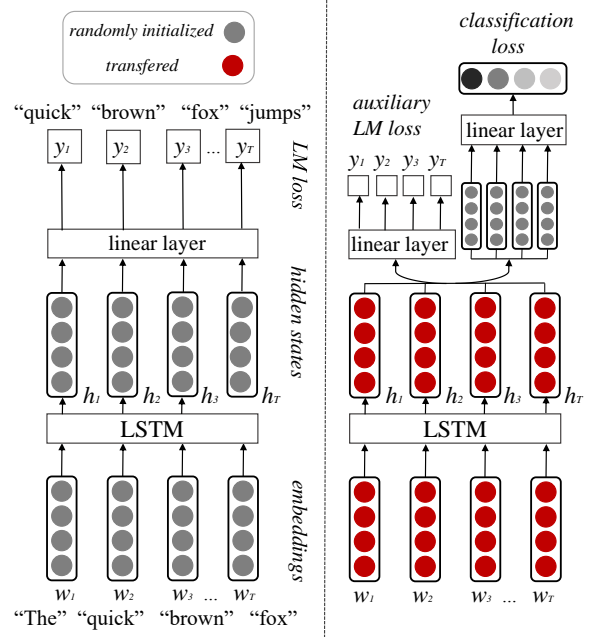

Figure 1: High-level overview of our proposed TL architecture. We transfer the pretrained LM add an extra recurrent layer and an auxiliary LM loss.

In order to adapt the contribution of the pretrained model to the task at hand, we introduce an auxiliary LM loss during training. The joint loss is the weighted sum of the task-specific loss L_{task} and the auxiliary LM loss L_{LM}, where γ is a weighting parameter to enable adaptation to the target task but at the same time keep the useful knowledge from the source task. Specifically:

$$L = L_{task} + \gamma L_{LM} \quad (2)$$

Exponential decay of γ. An advantage of the proposed TL method is that the contribution of the LM can be explicitly controlled in each training epoch. In the first few epochs, the LM should contribute more to the joint loss of SiATL so that the task-specific layers adapt to the new linguistic distribution. After the knowledge of the pretrained LM is transferred to the new domain, the task-specific component of the loss function is more important and γ should become smaller. In this paper, we use an exponential decay for γ over the training epochs.

Sequential Unfreezing. Instead of fine-tuning all the layers simultaneously, we propose unfreezing them sequentially, according to Chronopoulou et al. (2018). We first fine-tune only the extra, randomly initialized LSTM and the output layer for $n-1$ epochs. At the nth epoch, we unfreeze the

	Irony18	Sent17	SCv2	SCv1	PsychExp
BoW	43.7	61.0	65.1	60.9	25.8
NBoW	45.2	63.0	61.1	51.9	20.3
P-LM	42.7 ± 0.6	61.2 ± 0.7	69.4 ± 0.4	48.5 ± 1.5	38.3 ± 0.3
P-LM + su	41.8 ± 1.2	62.1 ± 0.8	69.9 ± 1.0	48.4 ± 1.7	38.7 ± 1.0
P-LM + aux	45.5 ± 0.9	65.1 ± 0.6	72.6 ± 0.7	55.8 ± 1.0	40.9 ± 0.5
SiATL (P-LM + aux + su)	**47.0** ± 1.1	**66.5** ± 0.2	**75.0** ± 0.7	**56.8** ± 2.0	**45.8** ± 1.6
ULMFiT (Wiki-103)	23.6 ± 1.6	60.5 ± 0.5	**68.7** ± 0.6	**56.6** ± 0.5	21.8 ± 0.3
ULMFiT (Twitter)	**41.6** ± 0.7	**65.6** ± 0.4	67.2 ± 0.9	44.0 ± 0.7	**40.2** ± 1.1
State of the art	53.6	68.5	76.0	69.0	57.0
	(Baziotis et al., 2018)	(Cliche, 2017)	(Ilic et al., 2018)	(Felbo et al., 2017)	

Table 1: Ablation study on various downstream datasets. Average over five runs with standard deviation. *BoW* stands for Bag of Words, *NBoW* for Neural Bag of Words. *P-LM* stands for a classifier initialized with our pretrained LM, *su* for sequential unfreezing and *aux* for the auxiliary LM loss. In all cases, F_1 is employed.

pretrained hidden layers. We let the model fine-tune, until epoch $k-1$. Finally, at epoch k, we also unfreeze the embedding layer and let the network train until convergence. The values of n and k are obtained through hyperparameter tuning. We find the sequential unfreezing scheme important, as it minimizes the risk of overfitting to small datasets.

Optimizers. We use Stochastic Gradient Descent (SGD) for the pretrained LM with a small learning rate, in order to preserve its contextual information and avoid catastrophic forgetting. However, we want the extra LSTM and softmax layer to train fast and adapt to the target task, so in that case Adam (Kingma and Ba, 2015) is employed.

Dataset	Domain	# classes	# examples
Irony18	Tweets	4	4618
Sent17	Tweets	3	61854
SCv2	Debate Forums	2	3260
SCv1	Debate Forums	2	1995
PsychExp	Experiences	7	7480

Table 2: Datasets used for the downstream tasks.

4 Experiments and Results

4.1 Datasets

To pretrain the language model, we collect a dataset of 20 million English Twitter messages, including approximately 2M unique tokens. We use the 70K most frequent tokens as vocabulary. We evaluate our model on five datasets: *Sent17* for sentiment analysis (Rosenthal et al., 2017), *PsychExp* for emotion recognition (Wallbott and Scherer, 1986), *Irony18* for irony detection (Van Hee et al., 2018), *SCv1* and *SCv2* for sarcasm detection (Oraby et al., 2016; Lukin and Walker, 2013). More details about the datasets can be found in Table 2.

4.2 Experimental Setup

To preprocess the tweets, we use *Ekphrasis* (Baziotis et al., 2017). For the generic datasets, we use NLTK (Loper and Bird, 2002). For the *NBoW* baseline, we use *word2vec* (Mikolov et al., 2013) 300-dimensional embeddings as features. For the neural models, we use an LM with an embedding size of 400, 2 hidden layers, 1000 neurons per layer, embedding dropout 0.1, hidden dropout 0.3 and batch size 32. We add Gaussian noise of size 0.01 to the embedding layer. A clip norm of 5 is applied, as an extra safety measure against exploding gradients. For each text classification neural network, we add on top of the transferred LM an LSTM layer of size 100 with self-attention and a softmax classification layer. For developing our models, we use PyTorch (Paszke et al., 2017) and Scikit-learn (Pedregosa et al., 2011).

5 Results & Discussion

Baselines and Comparison. Table 1 summarizes our results. The top two rows detail the baseline performance of the BoW and NBoW models. We observe that when enough data is available (e.g. *Sent17*), baselines provide decent results. Next, the results for the generic classifier initialized from a pretrained LM (P-LM) are shown with and without sequential unfreezing, followed by the results of the proposed model SiATL. SiATL is also directly compared with its close relative ULMFiT (trained on Wiki-103 or Twitter) and the state-of-the-art for each task; ULMFiT also fine-tunes a LM for classification tasks. The proposed SiATL method consistently outperforms the baselines, the P-LM method and ULMFiT in all datasets. Even though we do not perform any elaborate learning rate scheduling and we limit ourselves to pre-
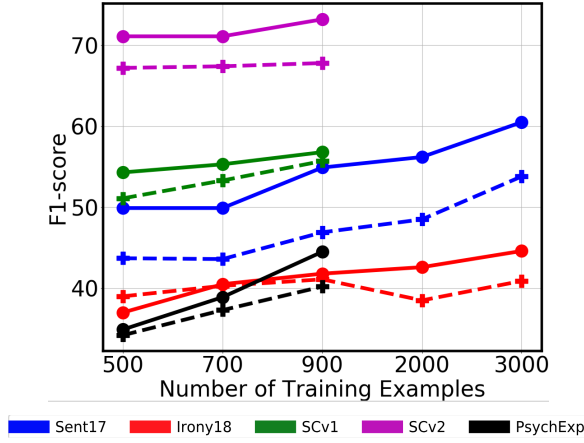
Figure 2: Results of our proposed approach (SiATL) (o) and ULMFiT (+) for different datasets as a function of the number of training examples.

Figure 3: Heatmap of the effect of γ to F_1-score, evaluated on *SCv2*. The horizontal axis depicts the initial value of γ and the vertical axis the final value of γ.

training in Twitter, we obtain higher results in two Twitter datasets and three generic.

Auxiliary LM objective. The effect of the auxiliary objective is highlighted in very small datasets, such as *SCv1*, where it results in an impressive boost in performance (7%). We hypothesize that when the classifier is simply initialized with the pretrained LM, it overfits quickly, as the target vocabulary is very limited. The auxiliary LM loss, however, permits refined adjustments to the model and fine-grained adaptation to the target task.

Exponential decay of γ. For the optimal γ interval, we empirically find that exponentially decaying γ to half of its initial value over the number of training epochs provides best results for our classification tasks. A heatmap of γ is depicted in Figure 3. We observe that small values of γ should be employed, in order to scale the LM loss in the same order of magnitude as the classification loss over the training period.

Sequential Unfreezing. Results show that sequential unfreezing is crucial to the proposed method, as it allows the pretrained LM to adapt to the target word distribution. The performance improvement is more pronounced when there is a mismatch between the LM and task domains, i.e., the non-Twitter domain tasks. Specifically for the *PsychExp* and *SCv2* datasets, sequentially unfreezing yields significant improvement in F_1 building upon our intuition.

Number of training examples. Transfer learning is particularly useful when limited training data are available. We notice that for our largest dataset *Sent17*, SiATL outperforms ULMFiT only by a
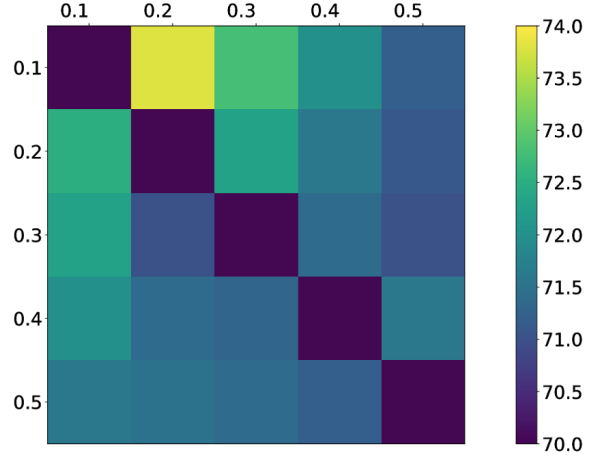
small margin when trained on all the training examples available (see Table 1), while for the small *SCv2* dataset, SiATL outperforms ULMFiT by a large margin and ranks very close to the state-of-the-art model (Ilic et al., 2018). Moreover, the performance of SiATL vs ULMFiT as a function of the training dataset size is shown in Figure 2. Note that the proposed model achieves competitive results on less than 1000 training examples for the *Irony18, SCv2, SCv1* and *PsychExp* datasets, demonstrating the robustness of SiATL even when trained on a handful of training examples.

6 Conclusions and Future Work

We introduce SiATL, a simple and efficient transfer learning method for text classification tasks. Our approach is based on pretraining a LM and transferring its weights to a classifier with a task-specific layer. The model is trained using a task-specific functional with an auxiliary LM loss. SiATL avoids catastrophic forgetting of the language distribution learned by the pretrained LM. Experiments on various text classification tasks yield results competitive to the state-of-the-art, demonstrating the efficacy of our approach. Furthermore, our method outperforms more sophisticated transfer learning approaches, such as ULMFiT in all tasks.

In future work, we plan to incorporate subword information in our LMs and experiment with bidirectional architectures. Furthermore, we plan to pretrain LMs on generic corpora. Finally, we will investigate adaptive schemes for sequential unfreezing and γ decay over training epochs.

7 Acknowledgements

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations*, San Diego, California.

Christos Baziotis, Athanasiou Nikolaos, Pinelopi Papalampidi, Athanasia Kolovou, Georgios Paraskevopoulos, Nikolaos Ellinas, and Alexandros Potamianos. 2018. Ntua-slp at semeval-2018 task 3: Tracking ironic tweets using ensembles of word and character level attentive rnns. In *Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018)*, pages 613–621, New Orleans, Louisiana.

Christos Baziotis, Nikos Pelekis, and Christos Doulkeridis. 2017. Datastories at semeval-2017 task 4: Deep lstm with attention for message-level and topic-based sentiment analysis. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 747–754, Vancouver, Canada.

Rich Caruana. 1993. Multitask learning: A knowledge-based source of inductive bias. In *Machine Learning: Proceedings of the Tenth International Conference*, pages 41–48.

Alexandra Chronopoulou, Aikaterini Margatina, Christos Baziotis, and Alexandros Potamianos. 2018. Ntua-slp at iest 2018: Ensemble of neural transfer methods for implicit emotion classification. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 57–64, Brussels, Belgium.

Mathieu Cliche. 2017. Bb_twtr at semeval-2017 task 4: Twitter sentiment analysis with cnns and lstms. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 573–580, Vancouver, Canada.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the International Conference on Machine learning*, pages 160–167.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, pages 2493–2537.

Andrew M Dai and Quoc V Le. 2015. Semi-supervised sequence learning. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 3079–3087.

Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1615–1625, Copenhagen, Denmark.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, (8):1735–1780.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the Annual Meeting of the ACL*, pages 328–339, Melbourne, Australia.

Suzana Ilic, Edison Marrese-Taylor, Jorge A. Balazs, and Yutaka Matsuo. 2018. Deep contextualized word representations for detecting sarcasm and irony. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 2–7.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations*.

Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130*.

Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, pages 63–70.

Stephanie Lukin and Marilyn Walker. 2013. Really? well. apparently bootstrapping improves the performance of sarcasm and nastiness classifiers for online dialogue. In *Proceedings of the Workshop on Language Analysis in Social Media*, pages 30–40, Atlanta, Georgia.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 3111–3119.

Shereen Oraby, Vrindavan Harrison, Lena Reed, Ernesto Hernandez, Ellen Riloff, and Marilyn A. Walker. 2016. Creating and characterizing a diverse corpus of sarcasm in dialogue. In *Proceedings of the SIGDIAL 2016 Conference, The 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 31–41.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *Journal of machine learning research*, pages 2825–2830.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543, Doha, Qatar.

Matthew Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power. 2017. Semi-supervised sequence tagging with bidirectional language models. In *Proceedings of the Annual Meeting of the ACL*, pages 1756–1765, Vancouver, Canada.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the Conference of the NAACL:HLT*, pages 2227–2237, New Orleans, Louisiana.

Prajit Ramachandran, Peter Liu, and Quoc Le. 2017. Unsupervised pretraining for sequence to sequence learning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 383–391, Copenhagen, Denmark.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. Semeval-2017 task 4: Sentiment analysis in twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518, Vancouver, Canada.

Anders Søgaard and Yoav Goldberg. 2016. Deep multi-task learning with low level tasks supervised at lower layers. In *Proceedings of the Annual Meeting of the ACL*, pages 231–235, Berlin, Germany.

Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2018. Semeval-2018 task 3: Irony detection in english tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation (SemEval-2018)*, pages 39–50, New Orleans, Louisiana.

Harald G. Wallbott and Klaus R. Scherer. 1986. How universal and specific is emotional experience? evidence from 27 countries on five continents. *Information (International Social Science Council)*, (4):763–795.

Caiming Xiong, Victor Zhong, and Richard Socher. 2016. Dynamic coattention networks for question answering.