*What Are We Transferring Anyway? : Using Attention to Interpret Transfer Learning from Different Domains in Text Classification*

*Abstract*

In the past few years, applications of Recurrent Neural Networks to Natural Language Processing have shown state-of-the-art performance across a wide range of tasks. In recent work, researchers have demonstrated the beneficial effects to performance on downstream tasks of both supervised and unsupervised pre-training for such models. However, it is often difficult to understand what exact knowledge the model acquires in one domain and how that model changes inference in the task of interest. Such understanding would be useful, both for motivating data selection for pre-training and for mitigating unintended learned bias. In this paper, we apply a self-attention mechanism on top of RNNs for text classification to identify whether the learned attention weights differ based on the data used for pre-training, and whether those differences are informative of what the pre-training in a particular domain adds to a particular model. We find that …

**The Gist🡪Comparing the Learned Attention Weights of 3 Kinds of Transfer, constructing an (observation x sequence\_token) attention weight matrix for each model.**

Two Potential Strategies for Comparison

1.

🡪For each pairwise comparison within each kind of pretraining, will partition predictions on test data by whether the classification decisions are different.

🡪Within each partition, we will rank examples by the distance between the learned attention weight vectors from each of the 2 models.

🡪Will Examine visualization of Observations with large distances to see whether the differences

2.

🡪MPQA Dataset is granularly annotated, with markers indicating the span of pieces of text that cause a particular sentence to be subjective.

🡪We can construct a vector that encodes these annotations by giving high values to elements that correspond to tokens that are marked as the subjective parts of the sentence.

🡪We can then compare this vector to the attention weight vectors to see which model produces the attention-weighting over the sentence that most corresponds to the gold-standard annotations described above.

**Different Kinds of Pretraining**

1. Word Vector Representations

🡪GloVe (50, 100, 200, 300)

🡪CharNGram

🡪Google News Word2Vec

2. Pretraining via Unsupervised Language Model

🡪GigaWord (Will Try for different size subsets) or 1 Billion Word LM (Public so reproducible)

🡪WikiText-2

🡪Gutenburg Corpus

🡪Amazon Reviews

3. Pretraining via Supervised Text Classification

🡪 IMDB Sentiment Classification Dataset

**🡪**Brown Corpus Article Topic Classification (or Reuters or 20-NewsGroups)

🡪 TREC Question-type Classification

🡪 SUBJ dataset

Evaluation Dataset

🡪MPQA Subjectivity

**Implementation Details**

Model Architecture

**🡪**LSTM RNNLM

🡪LSTM Text Classifier with Attention

🡪save model weights from RNN and implement initializer for clf

Efficiency Tricks:

🡪Implemented noise-contrastive estimation for efficient language model inference.

**🡪**Implemented weight tying between input and ouput embedding

*Literature Review*

**Semi-supervised Sequence Learning**

[UNSUPERVISED PRETRAINING FOR SEQUENCE TO SEQUENCE LEARNING](https://arxiv.org/abs/1611.02683)[SEMI-SUPERVISED SEQUENCE LEARNING](https://arxiv.org/abs/1511.01432)

[Learning to Generate Reviews and Discovering Sentiment](https://arxiv.org/abs/1704.01444)

**Transfer Learning**

[A Survey on Transfer Learning](https://www.cse.ust.hk/~qyang/Docs/2009/tkde_transfer_learning.pdf)

[A Survey of Transfer Learning](https://link.springer.com/article/10.1186/s40537-016-0043-6)

[Learning and Transferring Mid-Level Image Representations using CNNs](https://www.cv-foundation.org/openaccess/content_cvpr_2014/papers/Oquab_Learning_and_Transferring_2014_CVPR_paper.pdf)

[Domain Adaptation for Large-Scale Text Classification: A Deep Learning Approach](http://machinelearning.wustl.edu/mlpapers/paper_files/ICML2011Glorot_342.pdf)

**Attention Transfer**

[Completely Heterogenous Transfer Learning with Attention: What and What Not to Transfer](https://www.ijcai.org/proceedings/2017/0349.pdf)

[Paying More Attention to Attention: Improving the Performance of CNNs via Attention Transfer](https://arxiv.org/pdf/1612.03928.pdf)

**Language Modeling**

[Exploring the Limits of Language Modeling](https://arxiv.org/pdf/1602.02410.pdf)

[Original Recurrent Language Model](http://www.fit.vutbr.cz/research/groups/speech/servite/2010/rnnlm_mikolov.pdf)

[Improving Language Modeling with Noise-contrastive Estimation](https://arxiv.org/abs/1709.07758)

**Attention**

[A neural attention model for abstractive summarization](https://arxiv.org/pdf/1509.00685.pdf%C3%AF%C2%BC%E2%80%B0)

[Learning to Jointly Align And Translate](https://arxiv.org/pdf/1409.0473)

**Self-Attention**

[A Self-Attentive Sentence Embedding](https://arxiv.org/abs/1703.03130)

[LSTMs for Machine Reading](https://arxiv.org/abs/1601.06733)

[A Decomposable Attention Model for NLI](https://arxiv.org/abs/1606.01933)

[Attention is All You Need](https://arxiv.org/abs/1706.03762)

RNNs

A Learning Algorithm for Continually Running Fully Recurrent Neural Networks

[Generalization of Back-propagation to RNNs](http://brainmaps.org/pdf/pineda1987.pdf)

[Sequence to Sequence Learning with Neural Networks](http://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf)

[LSTM](https://dl.acm.org/citation.cfm?id=1246450)

[Bidirectional RNNs](http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.331.9441&rep=rep1&type=pdf)

[**Interpretable NLP**](https://arxiv.org/pdf/1506.01066.pdf)

*Understanding Latent Representation properties*

[Visualizing and Understanding Neural Models in NLP](https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&uact=8&ved=0ahUKEwj68r6qqLXXAhVBKWMKHVLDA0kQFgguMAA&url=https%3A%2F%2Farxiv.org%2Fabs%2F1506.01066&usg=AOvVaw0w7CPOz5voDO1088g1mGo_)

[Understanding Neural Networks Through Representation Erasure](https://arxiv.org/pdf/1612.08220.pdf)

[LSTMVis: A Tool For Visual Analysis of Hidden State Dynamics in Recurrent Neural Networks](https://arxiv.org/pdf/1606.07461.pdf)

*Providing Rationales for classification decisions*

[Rationalizing Neural Predictions](https://arxiv.org/pdf/1606.04155.pdf)

[Explaining Document’s Classifications](http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.307.5879&rep=rep1&type=pdf)

[Examples Are Not Enough, Learn to Criticize](http://people.csail.mit.edu/beenkim/papers/KIM2016NIPS_MMD.pdf)

[Rationale Augmented CNNs for text Classification](https://arxiv.org/abs/1605.04469)

[Human-Centric Justification of ML Predictions](https://www.ijcai.org/proceedings/2017/0202.pdf)

[Explanation and Justification in ML: A Survey](http://www.intelligentrobots.org/files/IJCAI2017/IJCAI-17_XAI_WS_Proceedings.pdf#page=8)

[A Causal Framework for Explaining the Predictions of Black-Box Sequence-to-Sequence Models](https://arxiv.org/pdf/1707.01943.pdf)

**Word Vectors**

[GloVe](https://www.aclweb.org/anthology/D14-1162)

[CharNGram](https://arxiv.org/pdf/1611.01587.pdf)

[GoogleNews word2vec](https://code.google.com/archive/p/word2vec/)

**Bias**

[Semantics Derived Automatically from Language Corpora Necessarily contain Human Biases](https://arxiv.org/pdf/1608.07187.pdf)  
[Equality of Opportunity in Supervised Learning](https://drive.google.com/file/d/0B-wQVEjH9yuhanpyQjUwQS1JOTQ/view)