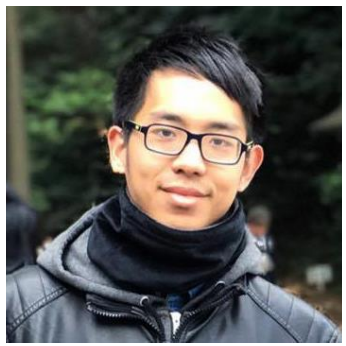
FIND: Human-in-the-Loop Debugging Deep Text Classifiers



Piyawat Lertvittayakumjorn, Lucia Specia, Francesca Toni

Department of Computing, Imperial College London

{pl1515, l.specia, ft}@imperial.ac.uk

Motivation

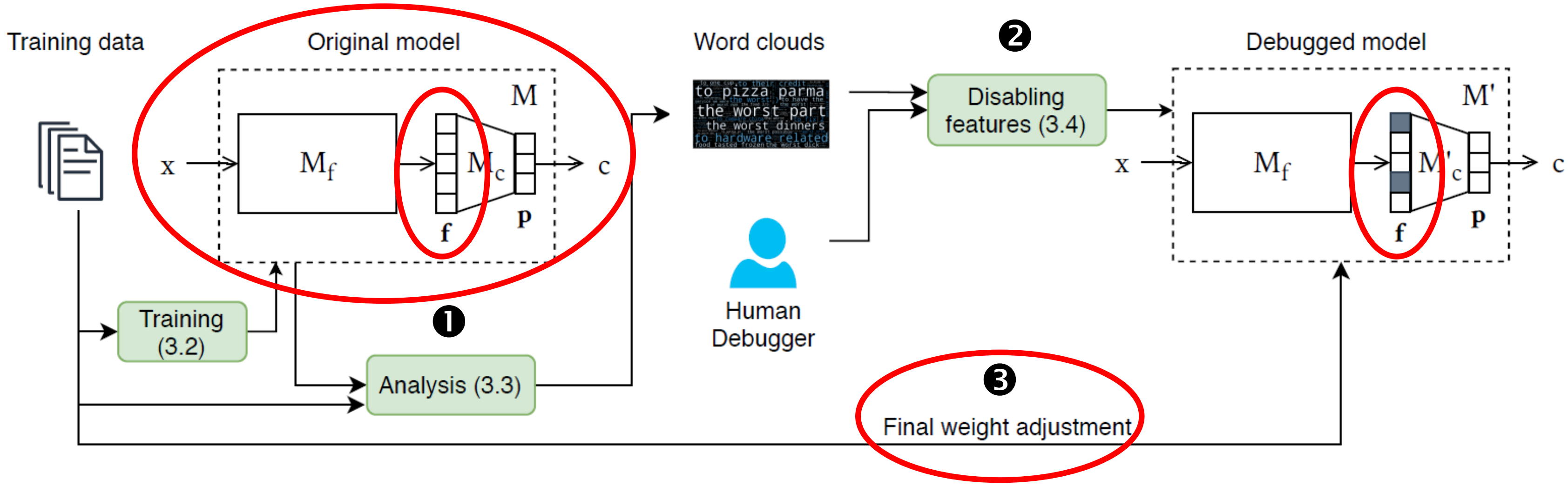
- Many real-world text classifiers are trained on available, yet imperfect, datasets.
  - Too small / Biased / Different from unseen cases
- These classifiers are thus likely to have undesirable properties.
  - They may have biases against some sub-populations
  - They may not work effectively in the wild due to overfitting
- We need to mitigate these problems !

Existing Solutions

- Preventing the anticipated problems
  - Gender swapping (Park et al., 2018; Zhao et al., 2018)
  - Adversarial training (Jaiswal et al., 2019; Zhang et al., 2018)
  - Using human rationales or prior knowledge (Zaidan et al., 2007; Bao et al., 2018; Liu and Avci, 2019)

- Fixing the problems with human in the loop (post-hoc)
  - Using interpretable models (Stumpf et al., 2009; Kulesza et al., 2015)
  - Using explanation methods (Ribeiro et al., 2016; Teso and Kersting, 2019)

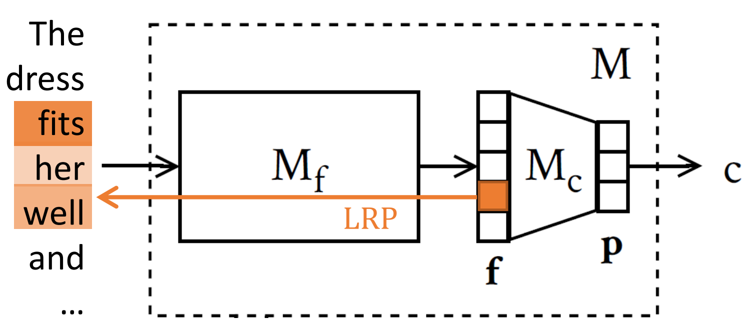
FIND:
Feature Investigation aNd Disabling

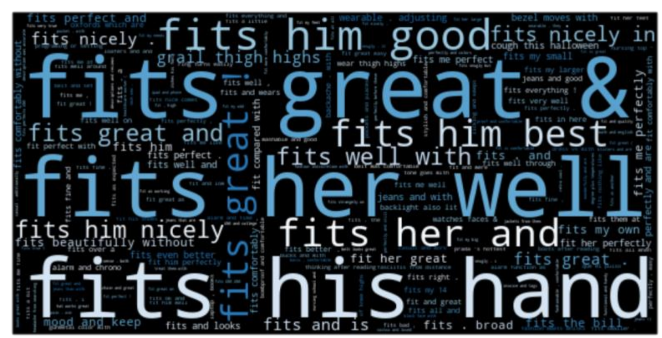


- ❶ Understand the patterns which each learned feature detects
- ❷ Disable features that are irrelevant to the classification task
- ❸ Fine-tune the model on the original dataset again to fully exploit the remaining features.

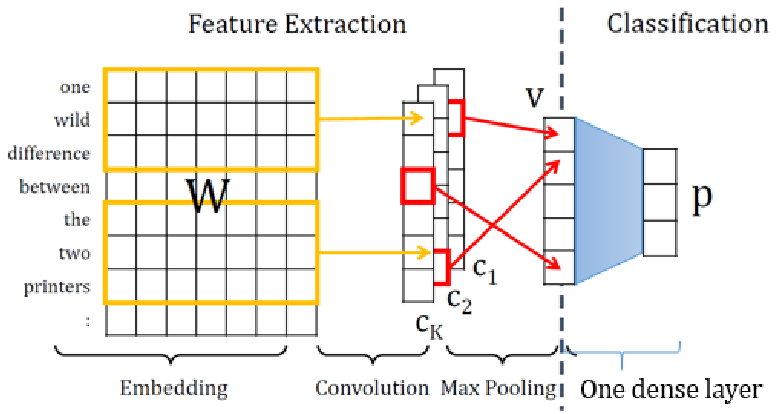
Understanding the Features

- To understand feature f_i , we consider each training example and calculate the relevance scores of input words for the value of f_i
    - Using layer-wise relevance propagation (LRP) (Bach et al., 2015)
    - Words that get higher scores are important words for f_i
- Word clouds are used to visualize important words collected from all the training examples





Understanding TextCNN Features

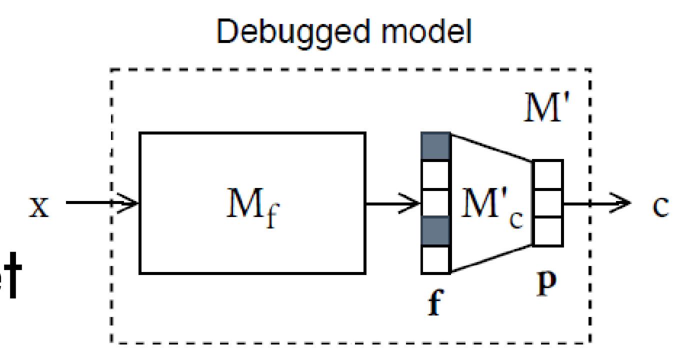


- For TextCNNs (Kim, 2014), after using LRP we crop the consecutive input words with non-zero LRP scores to show in the word clouds.
- This is equivalent to showing the n-grams, from the training examples, which were selected by the max-pooling of the CNNs.

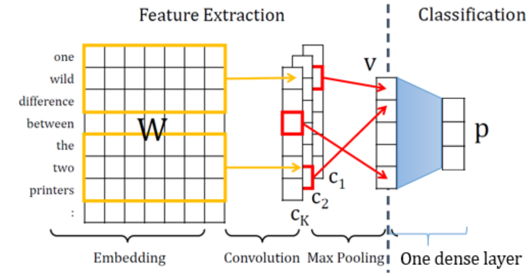
Imperial College London

Feature Disabling & Fine-tuning

- We disable features that are irrelevant to the task or that are contributing to unreasonable classes according to the weight matrix $\boldsymbol{W}$
- We modify the classification part M_c of the model
  - From $\boldsymbol{p} = M_c(\boldsymbol{f}) = \text{softmax}(\boldsymbol{W}\boldsymbol{f} + \boldsymbol{b})$
  - To $\boldsymbol{p} = M_c'(\boldsymbol{f}) = \text{softmax}((\boldsymbol{W} \odot \boldsymbol{Q})\boldsymbol{f} + \boldsymbol{b})$ where $\boldsymbol{Q}$ is a masking matrix containing ones. To disable feature $\boldsymbol{f}_i$, we set the i^{th} column of $\boldsymbol{Q}$ to be a zero vector.
- After disabling features, we then freeze the parameters of M_f and fine-tune the parameters of M'_c (except Q) with the training dataset



Experimental Setup



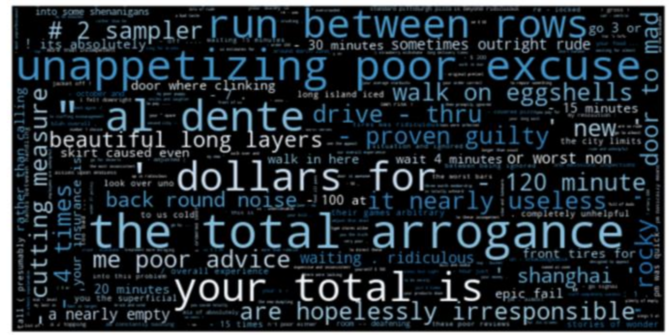
- We conducted three human experiments using TextCNNs
  - For each task, we ran and improved three models, and the reported results are the average of the three runs.
  - Filter sizes [2, 3, 4] x 10 filters for each size (30 features in total)
  - Non-trainable 300-dim GloVe vectors (Pennington et al., 2014)

- We used Amazon Mechanical Turk (MTurk) to collect crowdsourced responses for selecting features to disable.
  - Each question was answered by ten workers and the answers were aggregated using majority votes or average scores.

Experiment 1: Feasibility Study

- **Hypothesis**: Using word clouds is an effective way for humans to assess the features
- **Dataset**: Yelp (Sentiment analysis), with **limited number of training examples**
- **Human feedback**: We asked Mturk workers to consider each word cloud and answer which class the word cloud support
  - If the answer matches how the model really uses this feature (as indicated by W), the feature gets a positive score from this human response.

<u>Question 1</u>: Given this word cloud, does it convey positive or negative sentiment in the context of restaurant reviews?



○	The word cloud mostly conveys positive sentiment.	-2
○	The word cloud partially conveys positive sentiment.	-1
○	The word cloud conveys neither positive nor negative sentiment.	0
●	The word cloud partially conveys negative sentiment.	1
○	The word cloud mostly conveys negative sentiment.	2

According to W, this CNN feature is used by the model for the negative sentiment class

Imperial College London

* See the results of this dataset in our paper

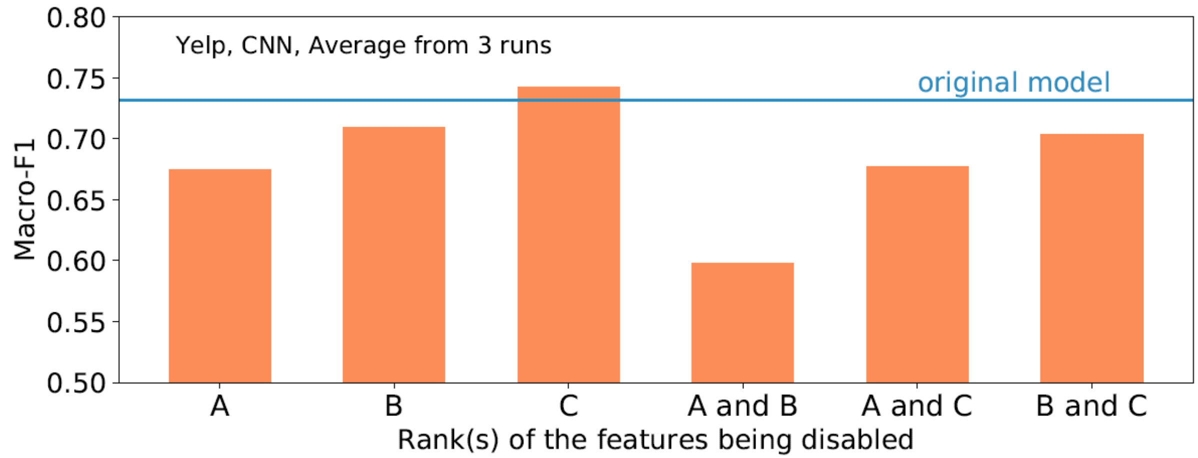
Experiment 1: Feasibility Study

Rank A - Average score = 2.0

Rank B - Average score = 1.2

Rank C - Average score = -0.7

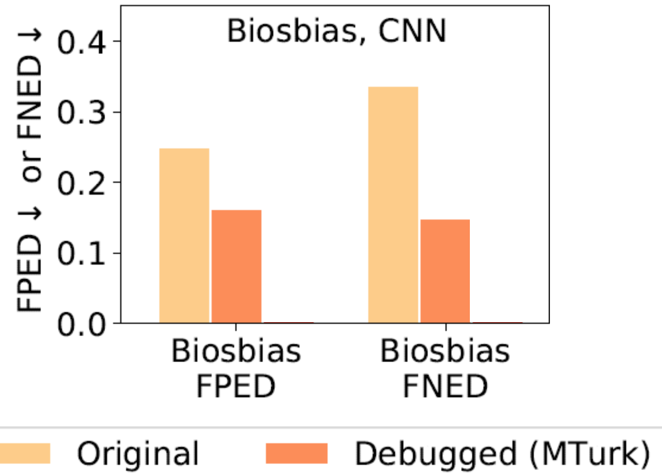
- The average macro F1, from the three runs, of all the CNN models for the Yelp dataset



Imperial College London

Experiment 2: Biased Training Data

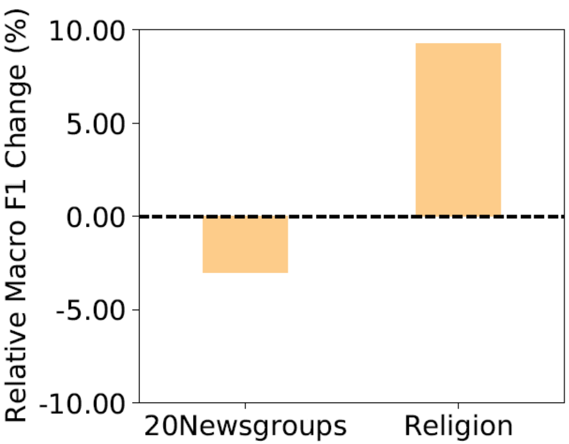
- **Hypothesis**: We can apply FIND to disable features which learn bias from the training data
- **Dataset**: Biosbias (Surgeon VS Nurse)
  - Due to gender imbalance, bios of female surgeons and male nurses are often misclassified
- **Human feedback**: For each word cloud, we asked the participants to select the relevant class from three options (Surgeon, Nurse, or it could be either). The feature will be disabled if the majority vote does not agree with the weight matrix $\boldsymbol{W}$.
- **Results**: FPED and FNED decrease after disabling the features based on human feedback



* See the results of this dataset in our paper

Experiment 3: Dataset Shift

- **Hypothesis**: FIND can disable overfitting features to increase the generalizability of the model
- **Datasets**: 20Newsgroups & Religion (Atheism VS Christian)
  - To make the models trained on the 20Newsgroups dataset work well on the Religion dataset
- **Human feedback**: For each word cloud, we asked the participants to select the relevant class from three options (Atheism, Christian, or it could be either). The feature will be disabled if the majority vote does not agree with the weight matrix $\boldsymbol{W}$.
- **Results**: The macro F1 scores
  - 20Newsgroups: 0.853 → 0.828
  - Religion: 0.731 → 0.799



* See the results of this dataset in our paper

Conclusions

- We proposed FIND, a framework which enables humans to debug deep text classifiers by disabling irrelevant or harmful features.
- Using FIND on CNN text classifiers, we found that
  - Word clouds generated by running LRP on the training data accurately revealed the behaviors of CNN features
  - Disabling the irrelevant or harmful features could improve the model predictive performance and reduce unintended biases in the model

References

- Sebastian Bach, Alexander Binder, Grοegoire Montavon, et al. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLOS ONE 2015.
- Yujia Bao, Shiyu Chang, Mo Yu, et al.. Deriving machine attention from human rationales. EMNLP 2018.
- Ayush Jaiswal, Daniel Moyer, Greg Ver Steeg, et al. Invariant representations through adversarial forgetting. AAAI 2020.
- Yoon Kim. Convolutional neural networks for sentence classification. EMNLP 2014.
- Todd Kulesza, Margaret Burnett, Weng-Keen Wong, et al.. Principles of explanatory debugging to personalize interactive machine learning. IUI 2015.
- Ji Ho Park, Jamin Shin, and Pascale Fung. Reducing gender bias in abusive language detection. EMNLP 2018.
- Jeffrey Pennington, Richard Socher, Christopher Manning. Glove: Global vectors for word representation. EMNLP 2014.
- Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. “why should i trust you?”: Explaining the predictions of any classifier. KDD 2016.
- Simone Stumpf, Vidya Rajaram, Lida Li, et al. Interacting meaningfully with machine learning systems: Three experiments. International Journal of Human-Computer Studies. 2009.
- Stefano Teso and Kristian Kersting. Explanatory interactive machine learning. AIES 2019.
- Omar Zaidan, Jason Eisner, and Christine Piatko. Using “annotator rationales” to improve machine learning for text categorization. NAACL-HLT 2007.
- Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. AIES 2018.
- Jieyu Zhao, Tianlu Wang, Mark Yatskar, et al. Gender bias in coreference resolution: Evaluation and debiasing methods. NAACL 2018.

https://github.com/plkumjorn/FIND

FIND: Human-in-the-Loop Debugging Deep Text Classifiers

Piyawat Lertvittayakumjorn, Lucia Specia, Francesca Toni

Department of Computing, Imperial College London

{pl1515, l.specia, ft}@imperial.ac.uk

@plkumjorn, @lspecia, @fra_toni