Google Explore CSR @ Perth

Literature Reviews

Siwen Luo
School of Computer Science
The University of Western Australia
siwen.luo@uwa.edu.au

Pretrained Model - DistilBERT

Sanh, Victor, et al. "**DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter.**" *arXiv preprint arXiv:1910.01108* (2019). [Paper]

Background

- Large-scale pretrained language model : more prevalent and leads significant improvement.

Addressed Problems

- However, operating these large model requires high computational costs and huge memories.

Research Aim

- To **reach similar performances** on many downstream-tasks using **much smaller language models** pre-trained with **knowledge distillation**
  - ✔ lighter and faster at inference time.
  - ✔ a smaller computational training budget.
  - ✔ can be fine-tuned with good performances.
  - ✔ small enough to run on the edge (e.g. on mobile)

Pretrained Model - DistilBERT

Contribution

- Propose DistilBERT, a general-purpose pre-trained version of BERT
  - ✔ 40% **smaller**, 60% **faster**, that **retains** 97% of the language understanding **capabilities**.

Knowledge Distillation

- ***Distillation*** *: (def.) the extraction of the essential meaning or most important aspects of something.*

- A **compression technique** in which a compact model is trained to **reproduce the behaviour** of a larger model or an **ensemble** of models.
  - Compact model = the student
  - Larger model = the teacher

- The student is **trained with a distillation loss** over the soft target probabilities of the teacher.

Pretrained Model - DistilBERT

Methodology

- **Student Architecture**
  - Remove the token-type embeddings and the pooler.
  - **Reduce the number of layers**.
  - Optimize the linear layer and layer normalisation from Transformer architecture.

- **Student initialization**
  - Find the right initialization for the sub-network to converge.

- **Distillation**
  - Apply **best practices for training** BERT model.
  - Be distilled on very large batches **leveraging gradient accumulation**.
  - Use **dynamic masking** and **without the next sentence prediction** objective.

Pretrained Model - DistilBERT

Experiments

- **Downstream tasks**
  - Retains **97% of BERT performance** on the GLUE dataset.
  - Only 0.6% point behind BERT on the IMDb dataset while being **40% smaller.** (Table 2)

- **Size and inference speed**
  - DistilBERT has 40% fewer parameters than BERT and is 60% faster than BERT on the STS-B (Table 3).

- **On device computation**
  - Excluding the tokenization step, DistilBERT is 71% faster than BERT.

Table 2: **DistilBERT yields to comparable performance on downstream tasks.** Comparison on downstream tasks: IMDb (test accuracy) and SQuAD 1.1 (EM/F1 on dev set). D: with a second step of distillation during fine-tuning.

Model	IMDb (acc.)	SQuAD (EM/F1)
BERT-base	93.46	81.2/88.5
DistilBERT	92.82	77.7/85.8
DistilBERT (D)	-	79.1/86.9

Table 3: **DistilBERT is significantly smaller while being constantly faster.** Inference time of a full pass of GLUE task STS-B (sentiment analysis) on CPU with a batch size of 1.

Model	# param. (Millions)	Inf. time (seconds)
ELMo	180	895
BERT-base	110	668
DistilBERT	66	410

Why Should I Trust You? - LIME

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "" Why should i trust you?" Explaining the predictions of any classifier." *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016. [Paper]

Background

1. Trust in Machine Learning Classifiers

- **Human understanding of a model's behaviour aids**:
  - Trust in individual predictions
    - E.g. Models used in medical diagnosis cannot be acted upon on blind faith
  - Trust in model's reliability when deployed
    - E.g. Users need to be confident that the model will perform well on real-world data

2. Desired Characteristics of Explainers

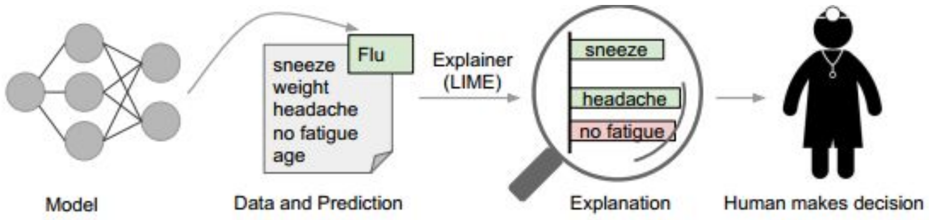
- Explanations must be interpretable
  - **Provide qualitative understanding between input and response**
- Local fidelity (*locally faithful*)
  - **Must correspond to how the model behaves** in the vicinity of the instance being predicted
- Model-agnostic : able to explain any model
- Global perspective : explain the model

Why should I trust you? - LIME

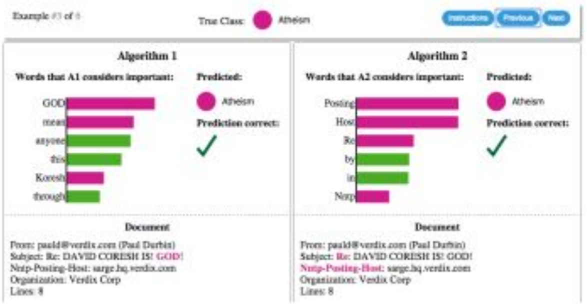
*The goal of **LIME (Local Interpretable Model-agnostic Explanations)** is to **identify an interpretable model** over the interpretable representation that is locally faithful to the classifier.*

Addressed Problems

- "**Trusting a prediction**" - provide explanations for individual predictions



- "**Trusting the model**" - select multiple representative predictions and explanations produced from the model



Contributions

- **LIME** : an algorithm that can explain the predictions of any classifier or regressor, by approximates predictions locally with an interpretable model.
- **SP-LIME** : a method that selects representative instances with explanations from the model
- **Measure impact of explanations on trust**
  - Show how understanding predictions know when and why they should not trust a model

Why should I trust you? - LIME

Methodology – LIME (Components)

1. Interpretable Data Representations

- Interpretable explanations need to be understandable to humans, regardless of the actual features used (E.g. presence of a word vs word embeddings)

2. Fidelity-Interpretability Trade-off

- Explanation produced by LIME is obtained as a trade- off.
- Balances **complexity** of the explanation (omega) with how closely it approximates an **interpretable** model within a **locality** (locality-aware loss)

3. Sampling for Local Exploration

- Sample instances around x' (uniformly at random)
- Optimise loss function to get an explanation (Fig 3)

4. Sparse Linear Explanations

- For text classification: Let interpretable representation be a bag of words and then set a limit on the number of words

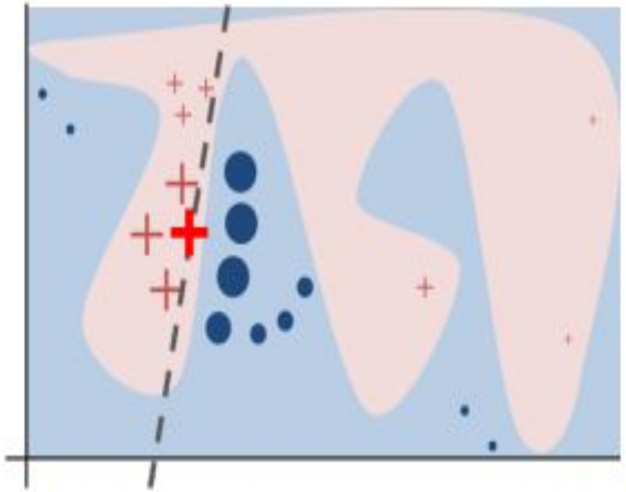


Figure 3: Toy example to present intuition for LIME. The black-box model's complex decision function f (unknown to LIME) is represented by the blue/pink background, which cannot be approximated well by a linear model. The bold red cross is the instance being explained. LIME samples instances, gets predictions using f, and weighs them by the proximity to the instance being explained (represented here by size). The dashed line is the learned explanation that is locally (but not globally) faithful.

Why should I trust you? - LIME

Simulated User Experiments

Research Questions

(1) *Are the explanations **faithful** to the model,*
(2) *Can the explanations **aid users in ascertaining trust** in predictions,*
(3) *Are the explanations **useful for evaluating** the model as a whole.*

Datasets

- 2 sentiment analysis datasets (books and DVDs) – pos/neg
- 2000 instances each (train 1,600 / test 400)

Experiment setup

- Bag of words as features
- **Decision tree, logistic regression, KNN, SVM, Random Forest**
- Baseline comparisons to LIME:
    - Random selection of K features (K : max 10)
    - Parzen: Approximates classifier globally
    - Greedy procedure
        - Removes features that contribute most to the predicted class until prediction changes

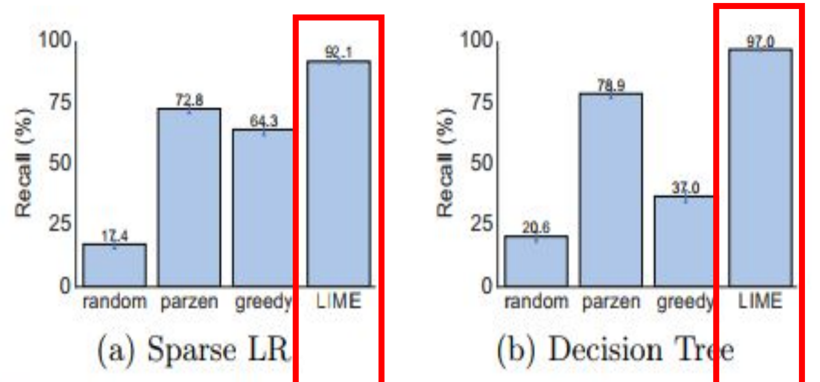


Figure 6: Recall on truly important features for two interpretable classifiers on the books dataset.

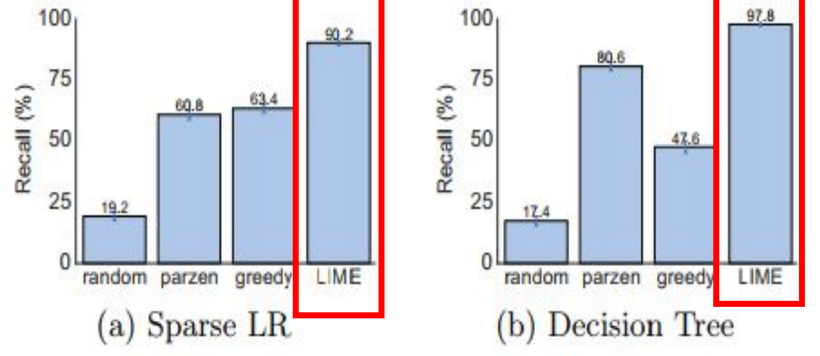


Figure 7: Recall on truly important features for two interpretable classifiers on the DVDs dataset.

Why should I trust you? - LIME

Findings

(1) Faithfulness of explanations

- Recall of model on “gold” set of features used as metric
- LIME provides **>90%** recall for both classifiers on both datasets

(2) Trustworthiness of individual predictions

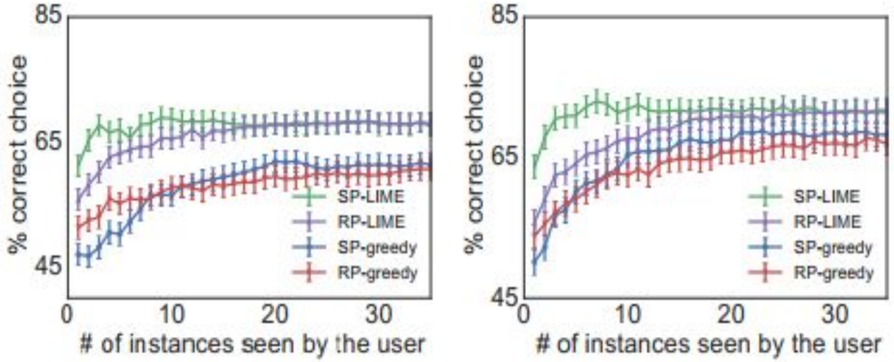
- 25% of features are randomly selected to be untrustworthy
- Prediction is untrustworthy if prediction changes when untrustworthy features are removed
- LIME dominates on both datasets (high precision and recall)

(3) Trustworthiness of model

- Add 10 noisy features, run model then simulated user classifies which explanations are untrustworthy
- SP-LIME explanations are good indicators of generalization

Table 1: Average F1 of *trustworthiness* for different explainers on a collection of classifiers and datasets.

	Books				DVDs			
	LR	NN	RF	SVM	LR	NN	RF	SVM
Random	14.6	14.8	14.7	14.7	14.2	14.3	14.5	14.4
Parzen	84.0	87.6	94.3	92.3	87.0	81.7	94.2	87.3
Greedy	53.7	47.4	45.0	53.3	52.4	58.1	46.6	55.1
LIME	**96.6**	**94.5**	**96.2**	**96.7**	**96.6**	**91.8**	**96.1**	**95.6**



(a) Books dataset (b) DVDs dataset

Figure 8: Choosing between two classifiers, as the number of instances shown to a simulated user is varied. Averages and standard errors from 800 runs.

Why should I trust you? - LIME

Evaluation with Human Subject

Research Questions

*(1) Can users choose which of two classifiers **generalizes better***
*(2) Based on the explanations, **can users perform feature engineering** to improve the model*
*(3) Are users able to identify and describe **classifier irregularities** by looking at explanations*

Datasets

- Training : Christianity and Atheism documents from 20 newsgroups dataset. Dataset contains features that do not generalise (e.g. informative header information and author names).
- Evaluation : Create a new Religion dataset to estimate real world performance / Amazon Mechanical Turk.

Findings

(1) Can users select the best classifier?

- To evaluate whether explanations can help users decide which classifier generalizes better.
- LIME outperforms greedy.
- Submodular Pick (SP) outperforms Random Pick (RP).

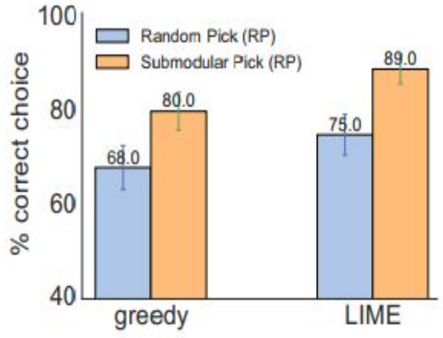


Figure 9: Average accuracy of human subject (with standard errors) in choosing between two classifiers.

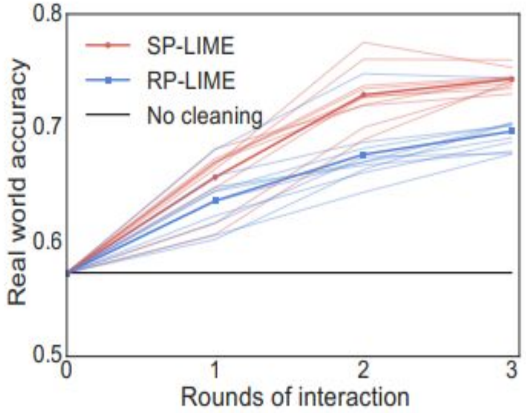


Figure 10: Feature engineering experiment. Each shaded line represents the average accuracy of subjects in a path starting from one of the initial 10 subjects. Each solid line represents the average across all paths per round of interaction.

Why should I trust you? - LIME

Evaluation with Human Subject

(2) Can non-experts improve a classifier?

- Users asked to remove words from given explanations for subsequent training
- These models trained again and given to a new set of users every round.
- The crowd workers are able to improve the model by removing features they deem unimportant for the task.

(3) Do explanations lead to insights?

- Users presented with prediction of a husky on a non-snowy background and wolf on a snowy background (reverse should be true)
- Explanation of model gave user more insight into bad model

Conclusion

- Trust in ML models is important for effective ML systems and can be assessed by explaining individual predictions
- LIME - explains predictions of any model in an interpretable manner

Future work

- Only sparse linear models are used as explanations
- Investigate other explanation families such as decision trees
- Pick step was not described for images
- Investigate potential applications of LIME in speech, video, medical domains and recommender systems
- Explore theoretical properties (approx. number of samples) and computational optimizations for accurate real-time explanations