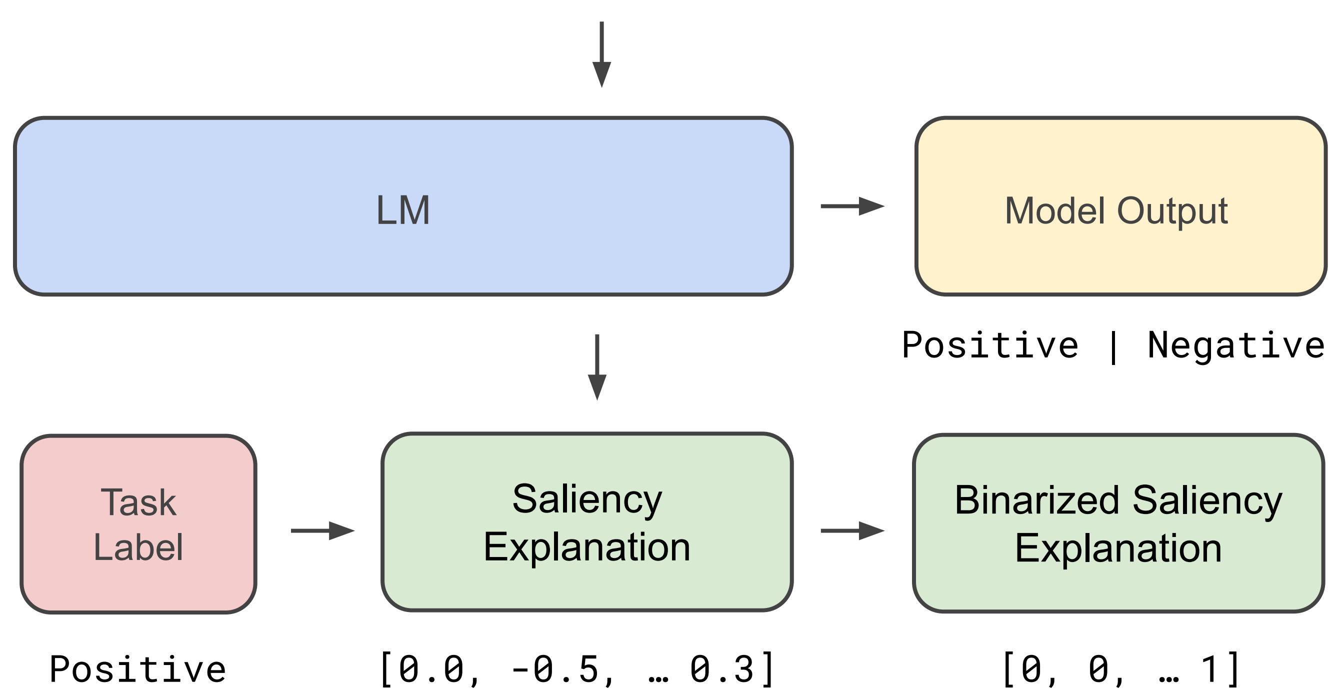
Training Language Models with Saliency Explanations

Wyatt Lake (lakewyatt@gmail.com), Harvard Westlake, Class of 2024
USC Viterbi School of Engineering, SHINE 2021
Professor Ren, USC INK Research Lab, Department of Computer Science

SHINE
Summer High School Intensive in Next-Generation Engineering

1. What are saliency explanations?

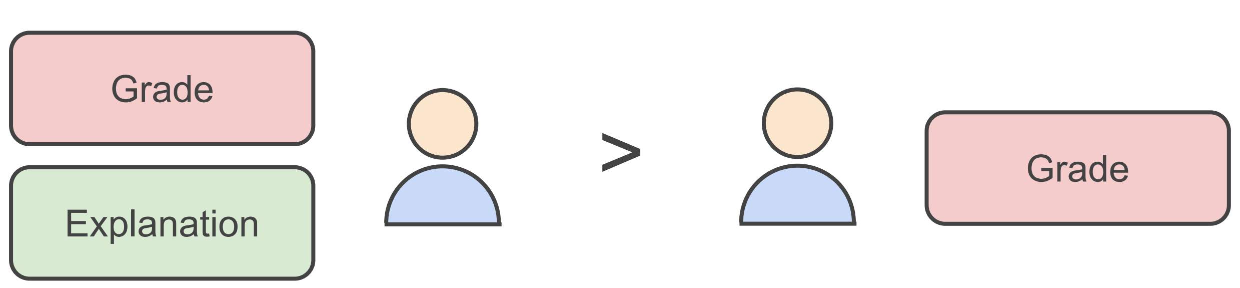
```
Still, this flick is fun, and host to some truly
excellent sequences.
```



- An **extractive explanation** highlights the most useful parts of a language model's (LM) input for solving a given task instance.
- A **saliency explanation**[1] is a type of extractive explanation that is *auto-generated* via (gradient-based) saliency methods.

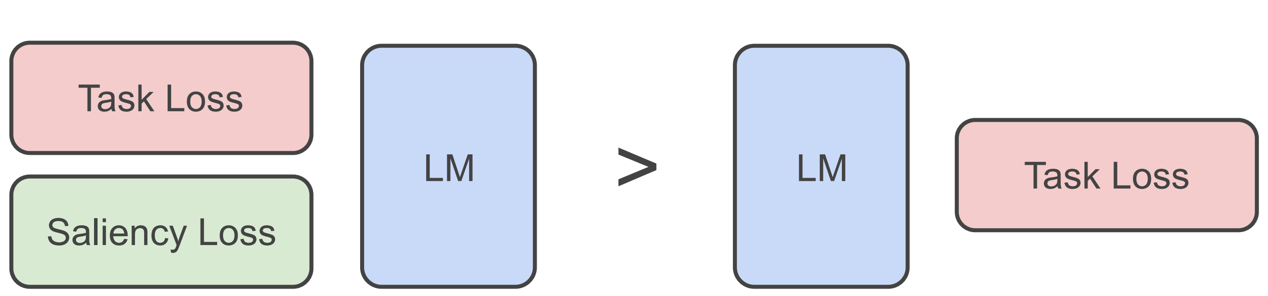
2. Motivation for Explanation-Based Learning

When a student submits a school assignment, the teacher gives them both a grade and an explanation for why they received that grade.



Students who get both grades and explanations from their teachers *perform better* than students who only get grades.

We hypothesize that a LM trained on both **task labels** and **saliency explanations** will perform better than an LM trained only on task labels.

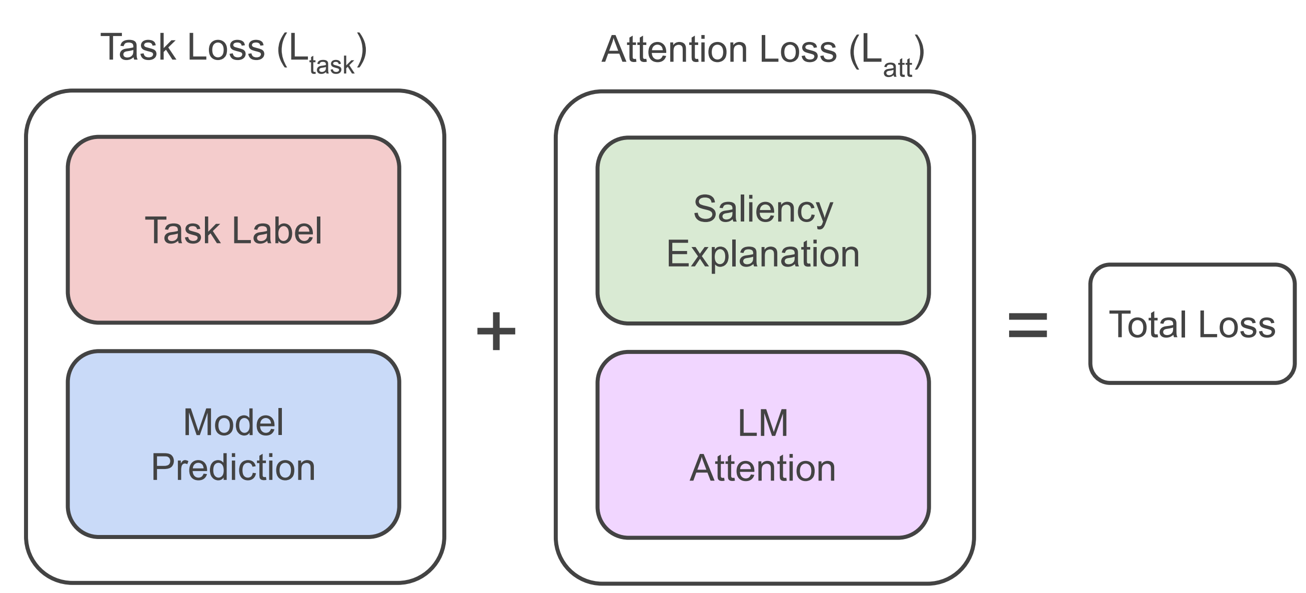


3. SalLM

- LMs use **attention** to predict which input tokens are most important[2].
- To improve LMs' attention, we propose **SalLM**, a method for regularizing LMs' attention to mimic the saliency explanations.

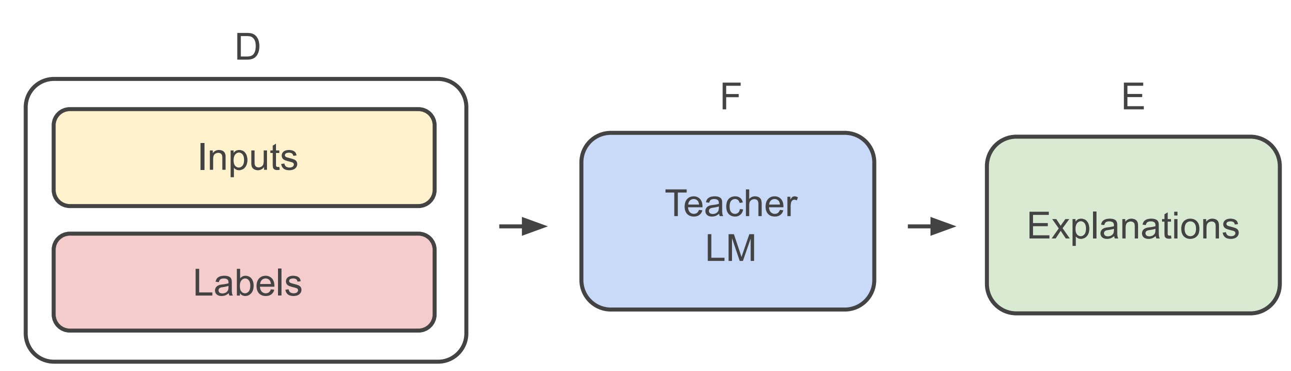
	Still,	this	flick	is	fun	
Explanation	0.0	0.0	0.0	0.0	1.0	…
Attention	0.1	0.2	0.5	0.0	1.0	…

The SalLM learning objective consists of: (1) the original **task loss** and (2) the **attention loss** for regularizing the LM's attention mechanism[3].

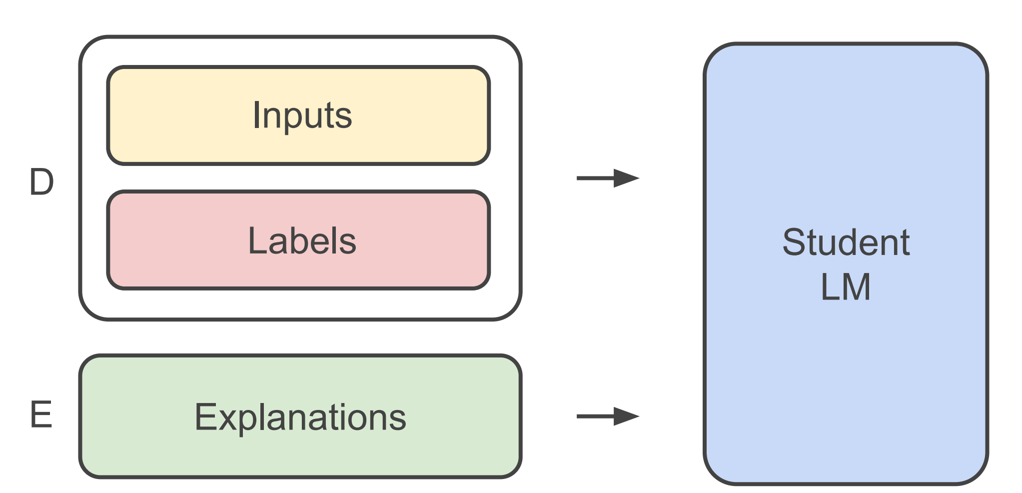


Training Procedure

- Train teacher model F on dataset D, using only L_{task}
- Use F and D to generate explanations E



- Retrain F on (D, E), using $L_{task} + \lambda L_{att}$, where λ is a loss weight hyperparameter



5. Results

Performance on **SST-5[4] sentiment analysis** dataset, using the **BERT-Base[5]** LM and different SalLM variants. Results are averaged over three seeds.

Model (BERT-Base)	Validation Accuracy (%)	Test Accuracy (%)
Vanilla LM	51.07 ± 0.52	53.83 ± 0.42
SalLM	51.77 ± 0.76	54.22 ± 0.19
SalLM (Fine-Tuned)	51.13 ± 0.92	53.27 ± 0.21
SalLM (Iterative)	51.53 ± 0.41	53.45 ± 1.05

6. Next Steps

- Apply SalLM to **other tasks/datasets**
- Try SalLM on **other LM architectures** (e.g., RoBERTa[6])
- Experiment with **non-binarized** explanations
- Investigate **attention head** explanations/regularization
- Adapt SalLM to **semi-supervised** learning settings

7. Acknowledgements

I would like to thank Prof. Xiang Ren, Aaron Chan (and the SalLM team), Cassandra Jeon, Monica Lopez, and Dr. Katie Mills for organizing this unique opportunity for me to work on NLP/AI research at USC.

8. References

[1] Bastings & Filippova. *"The elephant in the interpretability room: Why use attention as explanation when we have saliency methods?."* arXiv 2020.

[2] Vaswani, et al. *"Attention is all you need."* NeurIPS 2017.

[3] Pruthi, et al. *"Evaluating Explanations: How much do explanations from the teacher aid students?."* arXiv 2020.

[4] Socher, et al. *"Recursive deep models for semantic compositionality over a sentiment treebank."* EMNLP 2013.

[5] Devlin, et al. *"Bert: Pre-training of deep bidirectional transformers for language understanding."* NAACL 2019.

[6] Liu, et al. *"Roberta: A robustly optimized bert pretraining approach."* arXiv 2019.