Attribute Injection in the Age of Pretrained Language Models

Reinald Kim Amplayo

@ Naver AI Lab

Online texts are attached with **metadata attributes**

Customer reviews

4.7 out of 5

121 global ratings

5 star 77%
4 star 17%
3 star 2%
2 star 1%
1 star 2%

How are ratings calculated?

Samsung Galaxy Book Ion 15.6 Inch 8 ...

by Samsung

Name: 15 Inch | Pattern Name: Intel i5 | Change

Write a review

Top positive review

All positive reviews ›

The Man Who

Incorrect Spec by Samsung!

Reviewed in the United Kingdom on 12 August 2020

Wonderful craft, smart and neat! The latest and the greatest! Good just to look at and joy to use! However, a big letdown by graphic card: it's not Nvidia MX250 as claimed here by Samsung itself (see 'Customer Questions' below on this page). And it was my question since there were plenty contradictions all over internet. After a quick search, I found out that

Read more

Top critical review

All critical reviews ›

carol

Pretty looks but not functioning, slow and bad after sale ser

Reviewed in the United Kingdom on 7 September 2020

It looks good but is not ready to go on the market. Mine had to go to repair for a hardware issue after 2 months of purchase and is there for at least a second week. Also the battery never charges to the

Read more

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova

We introduce a new language representation model called BERT, which stands for Bidirectional Encoder Representations from Transformers. Unlike recent language representation models, BERT is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task-specific architecture modifications.

BERT is conceptually simple and empirically powerful. It obtains new state-of-the-art results on eleven natural language processing tasks, including pushing the GLUE score to 80.5% (7.7% point absolute improvement), MultiNLI accuracy to 86.7% (4.6% absolute improvement), SQuAD v1.1 question answering Test F1 to 93.2 (1.5 point absolute improvement) and SQuAD v2.0 Test F1 to 83.1 (5.1 point absolute improvement).

Subjects: **Computation and Language (cs.CL)**

Cite as: arXiv:1810.04805 **[cs.CL]**

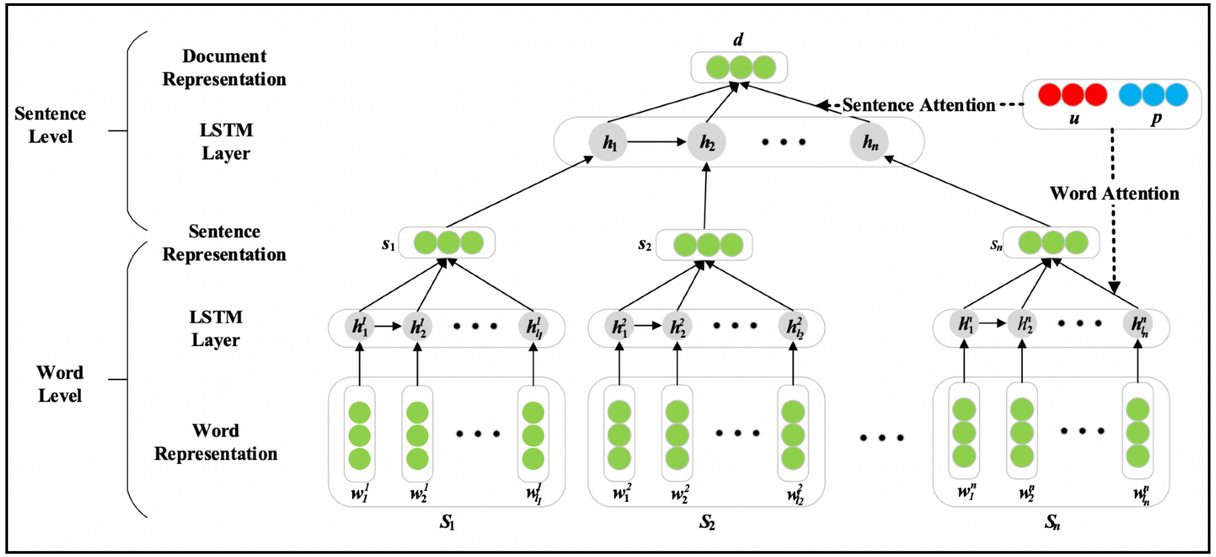
(or arXiv:1810.04805v2 **[cs.CL]** for this version)

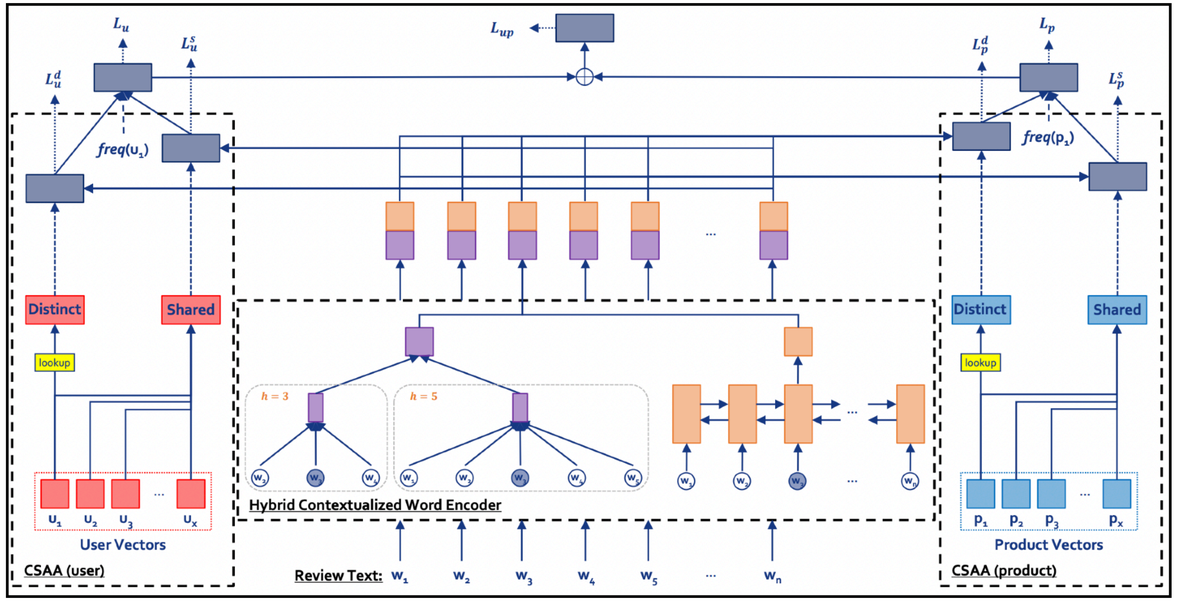
Using these attributes improves NLP models

Model	IMDB	Yelp
HierLSTM	48.7	63.1
+ UPA (Chen et al., 2016)	53.3	65.0
+ CSAA (Amplayo et al., 2018)	52.7	65.4
+ CHIM (Amplayo, 2019)	**56.4 (+5.7)**	**67.8 (+4.7)**

Sentiment Classification Accuracy

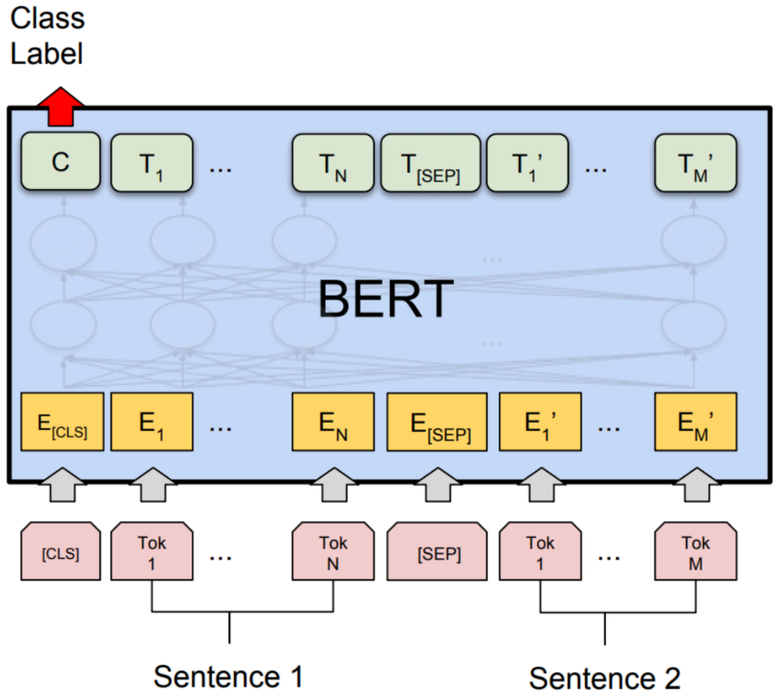
Using these attributes improves NLP models… How?

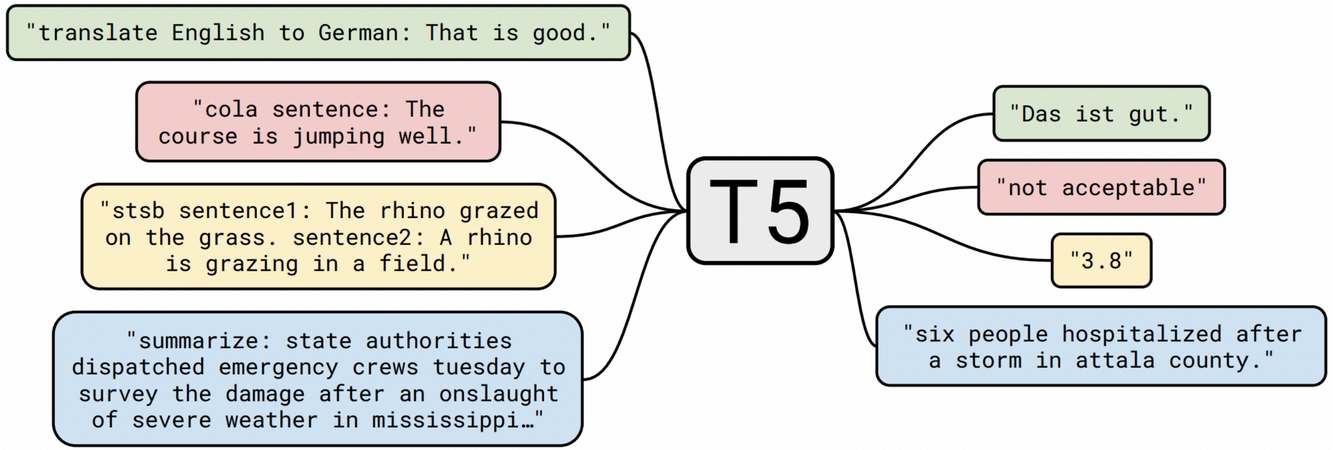




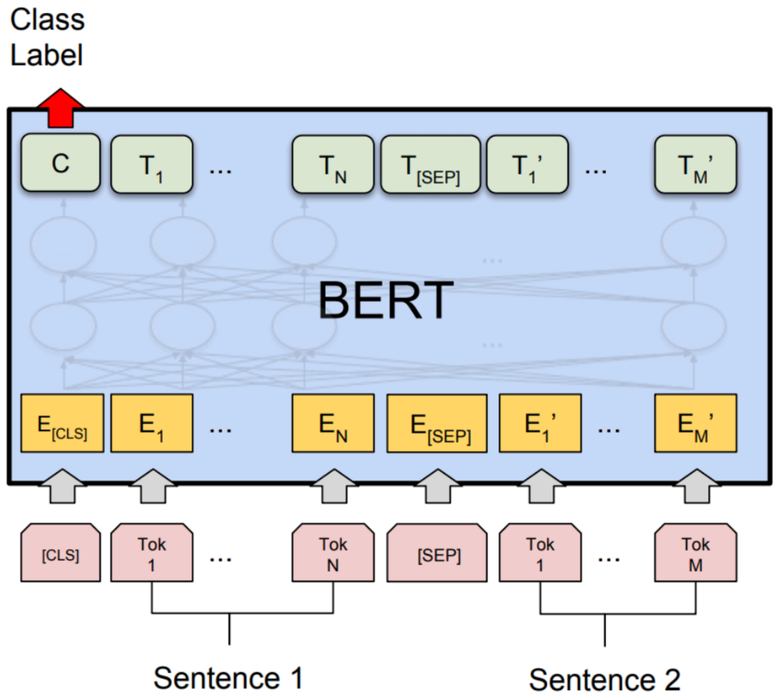
…by modifying the neural architecture

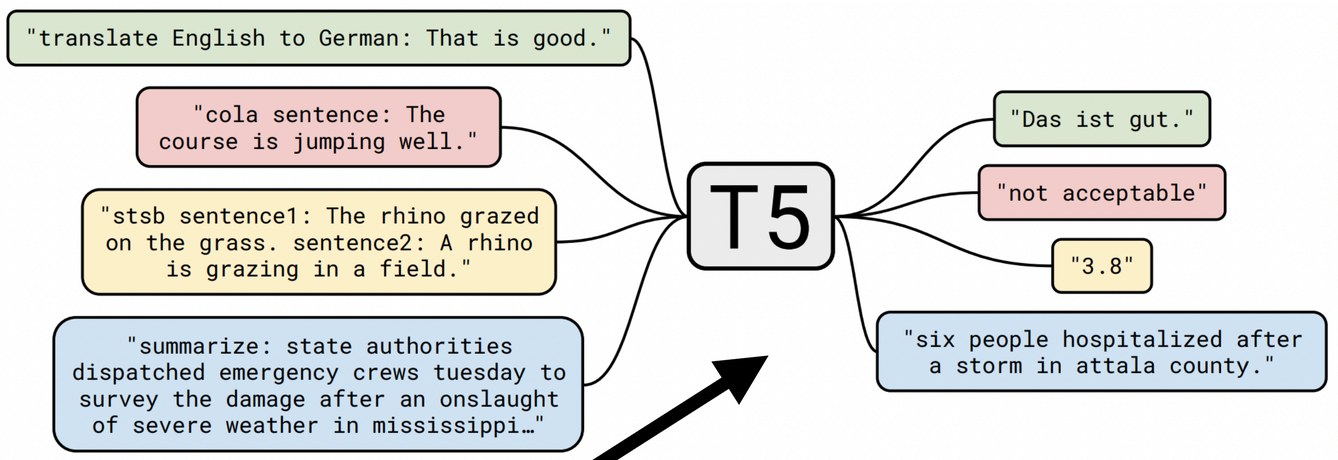
NLP Today: “Age of Pretrained LMs”

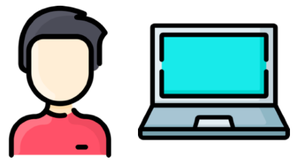




PLMs + Attributes?



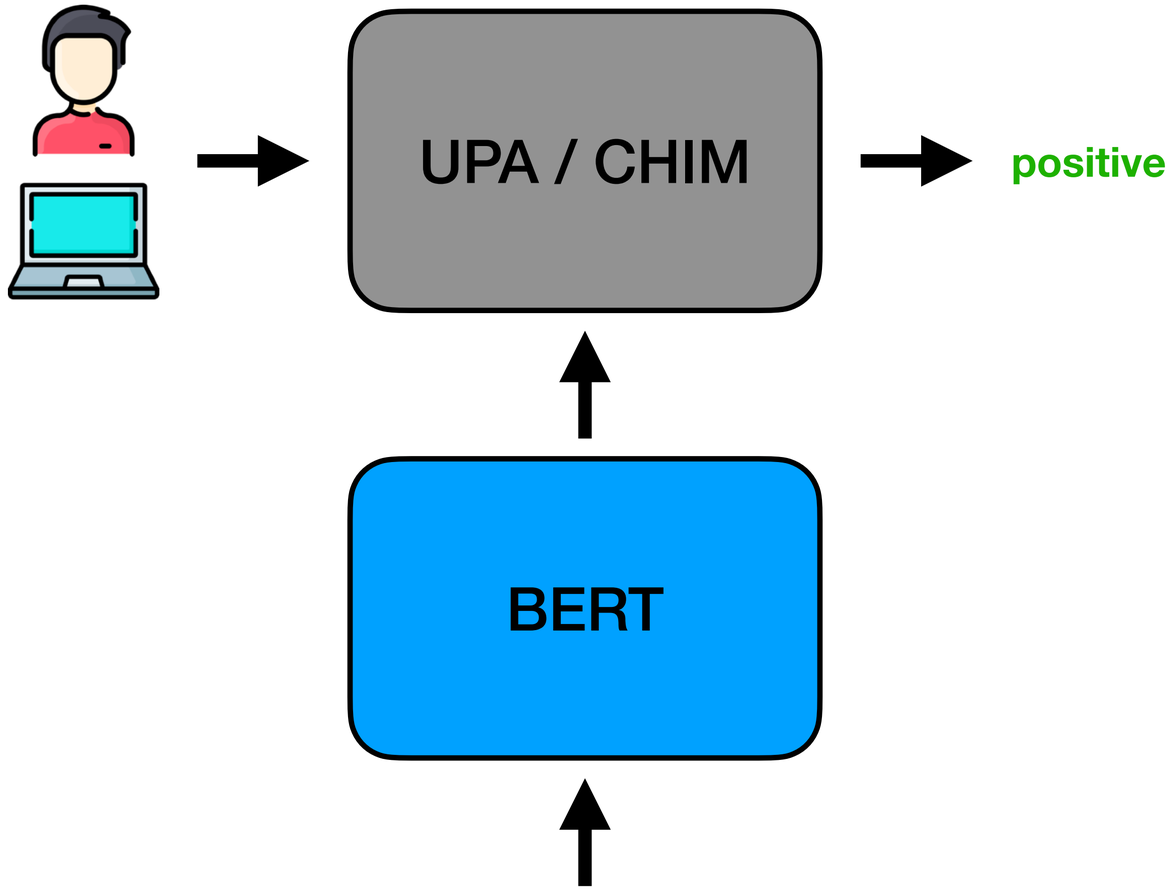




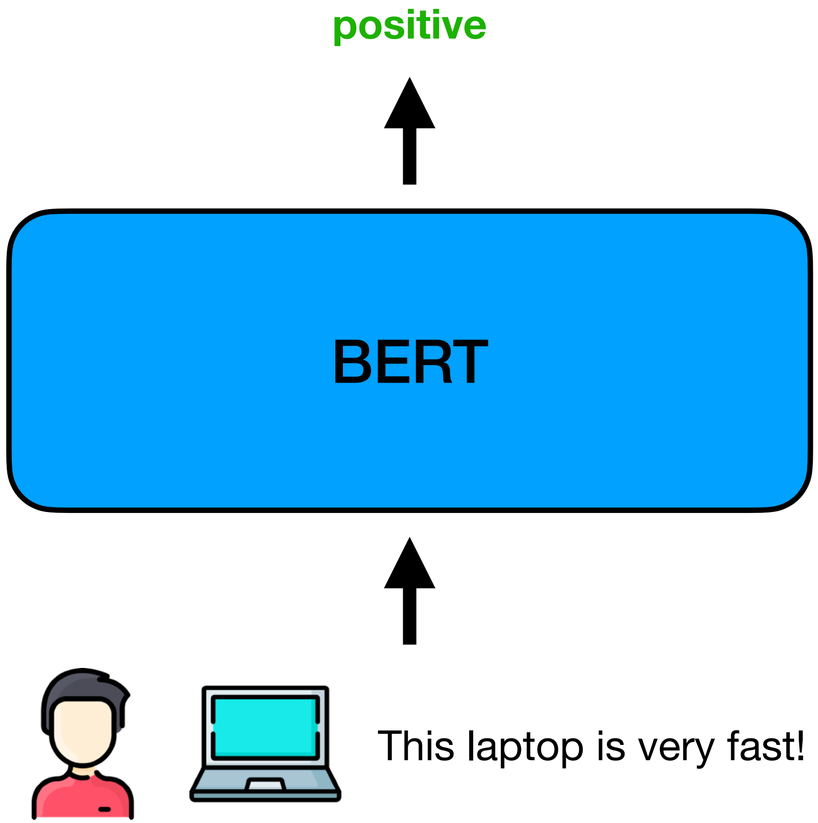
Attributes

We cannot modify the architecture without forgetting!

Baseline 1: Apply outside of BERT



Baseline 2: Use as special tokens

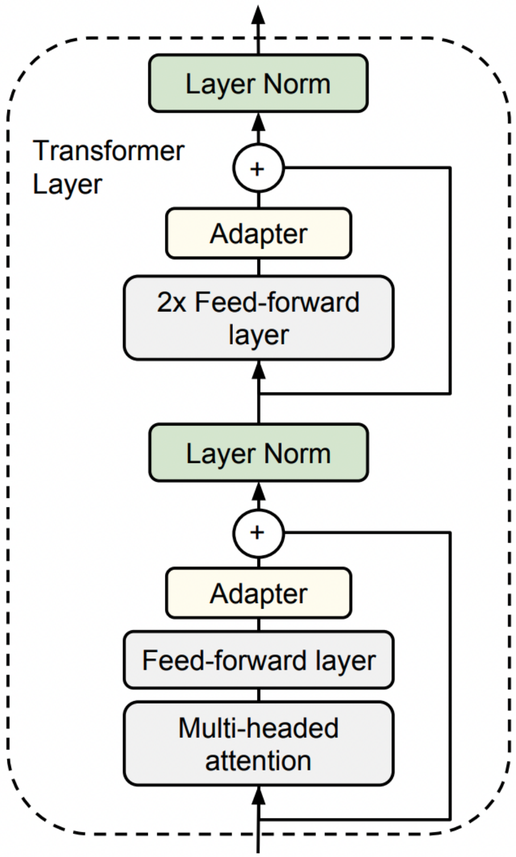


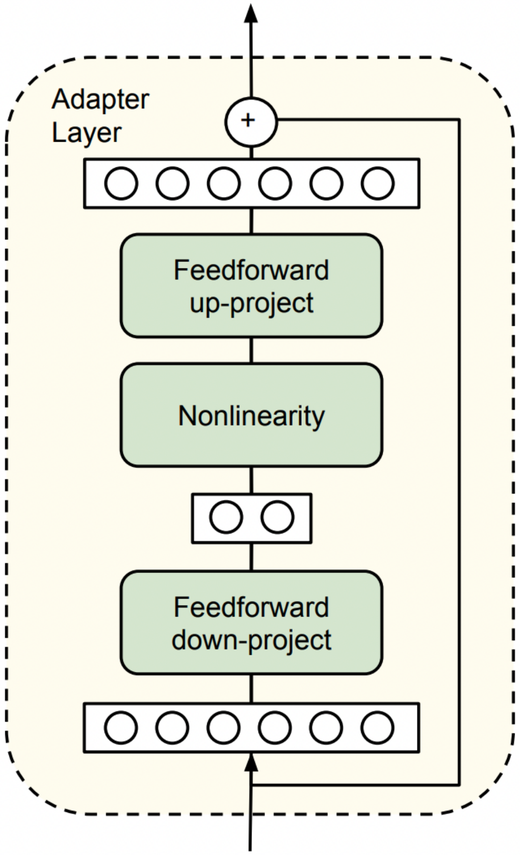
Preliminary Experiments

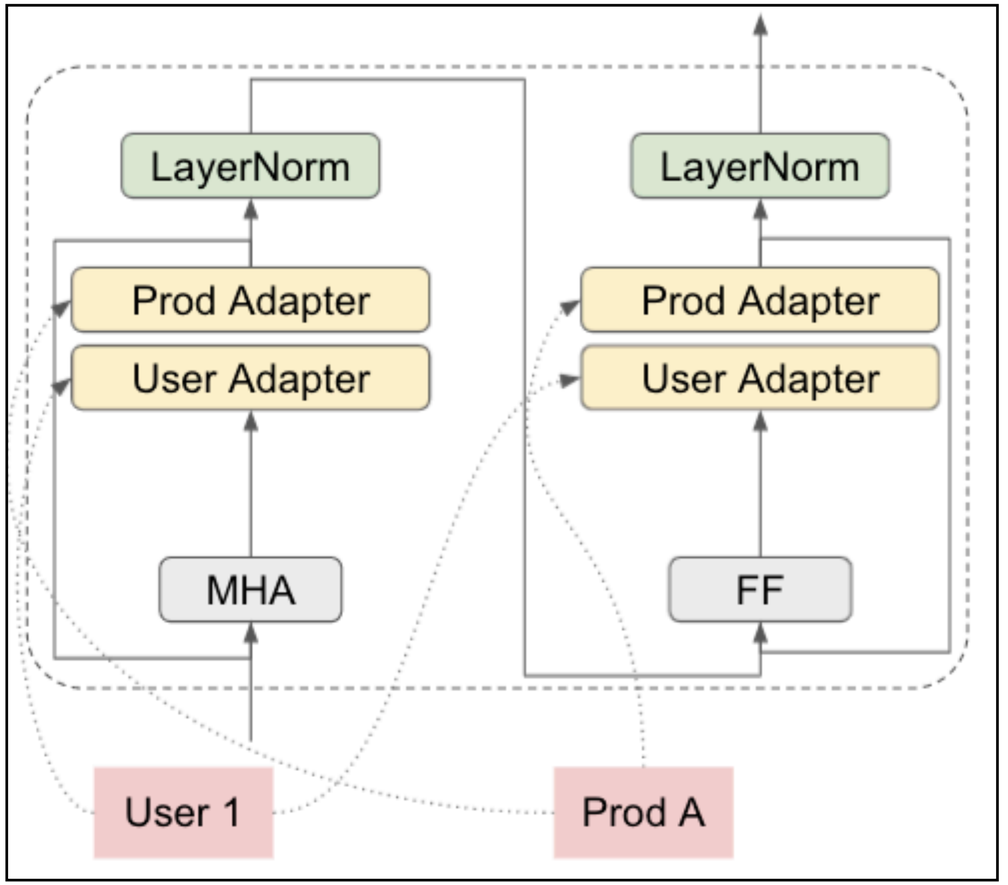
Model	Yelp
HierLSTM	63.1
+ UPA (Chen et al., 2016)	65.0
+ CSAA (Amplayo et al., 2018)	65.4
+ CHIM (Amplayo, 2019)	**67.8 (+4.7)**
BERT-base	66.0
+ UPA	65.7 (-0.3)
+ CHIM	67.1 (+1.1)
+ special tokens	66.5 (+0.5)

Sentiment Classification Accuracy

Idea: Use Adapters







Types of attribute adapters

Normal adapters: $f(x) = W(Vx + b) + c$

Adapters with attributes z:

1. Concat with input z
   $f(x, z) = W(V[x, z] + b) + c$

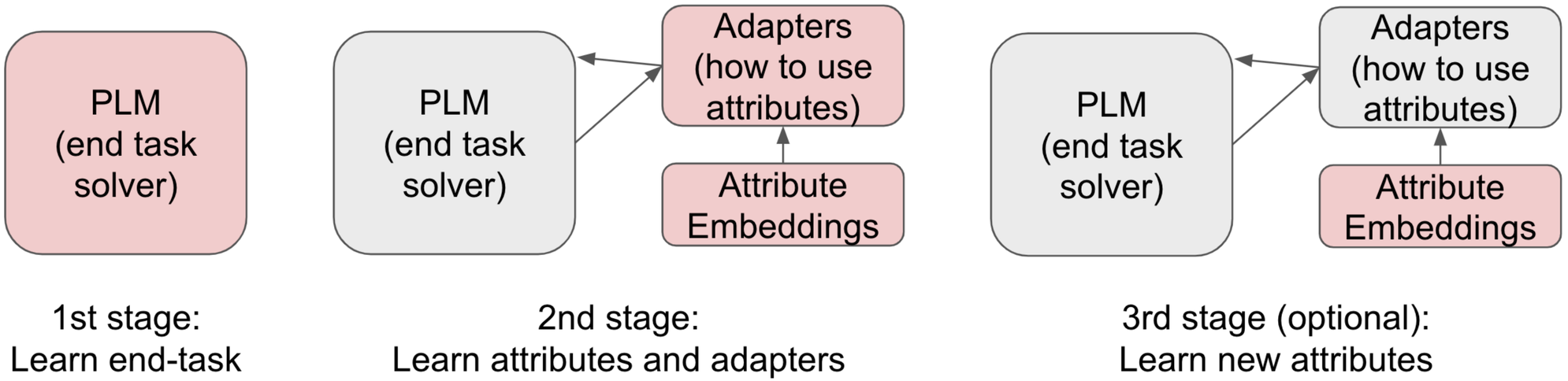
2. Weighted dot-product
   $f(x, z) = W(Vx + xRz + b) + c$

3. As small changes to weight matrix V
   $f(x, z) = W(V'x + b) + c$ $\quad\quad\quad\quad$ $V' = V + \text{lambda} * \tanh(Uz)$

4. As importance weights to weight matrix V (i.e., sparsify V)
   $f(x, z) = W(V'x + b) + c$ $\quad\quad\quad\quad$ $V' = V * \text{sigmoid}(Uz)$

Training the model

1. **Finetune PLM** on the end task *without attributes*
2. Freeze PLM, and **train attribute-specific adapters**
3. When there are new attributes available (i.e., incremental learning), we train **only their representations**



Datasets and Tasks

Text classification

- Sentiment Classification (IMDB, Yelp 2013, Yelp 2014)
- Paper Acceptance Classification (AAPR)
- Message Type Classification (PolMed)

Text generation

- Review Expansion (Amazon)
- Paper Abstract Generation (Arxiv)

Things to consider

- Multi-label attributes (e.g., multiple authors of a paper)
- Cold-start attributes (e.g., new users with zero reviews)
- Large number of attributes (e.g., millions of users in Amazon)