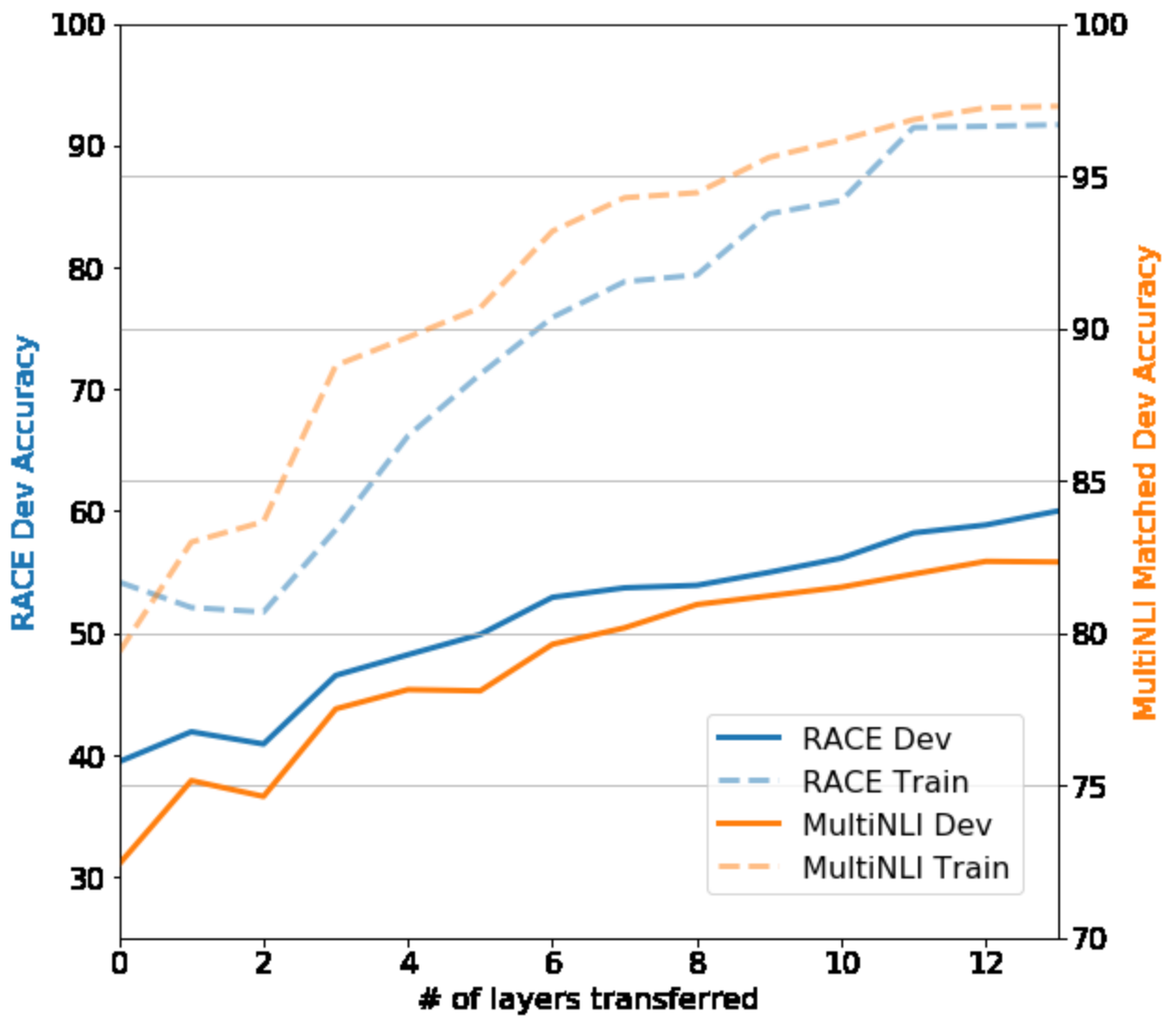
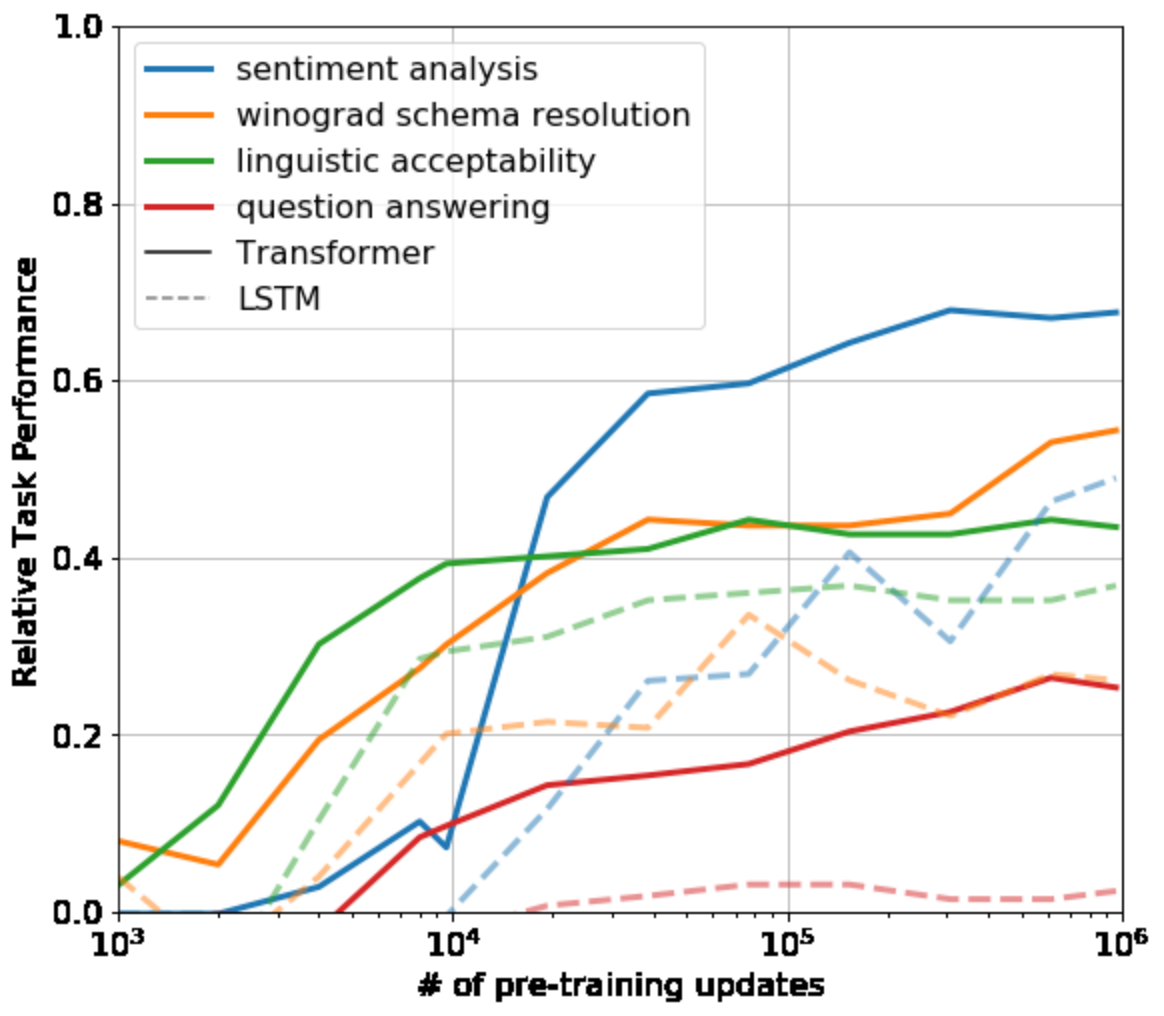
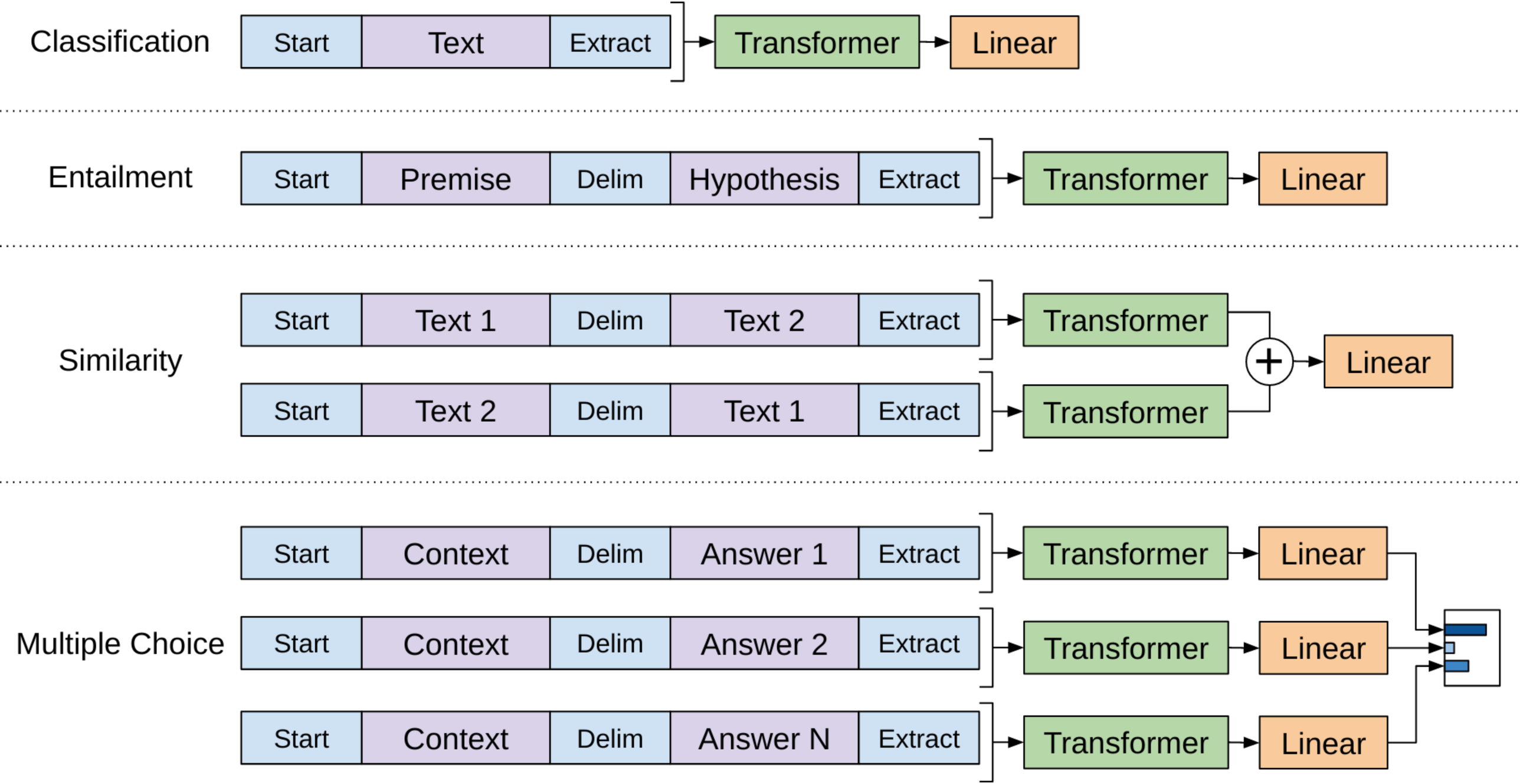
Introduction
the ability to learn effectively from raw text is crucial to alleviating the dependence on supervised learni
Related Work
Semi-supervised learning for natural language has attracted significant interest . The earliest approach
Framework

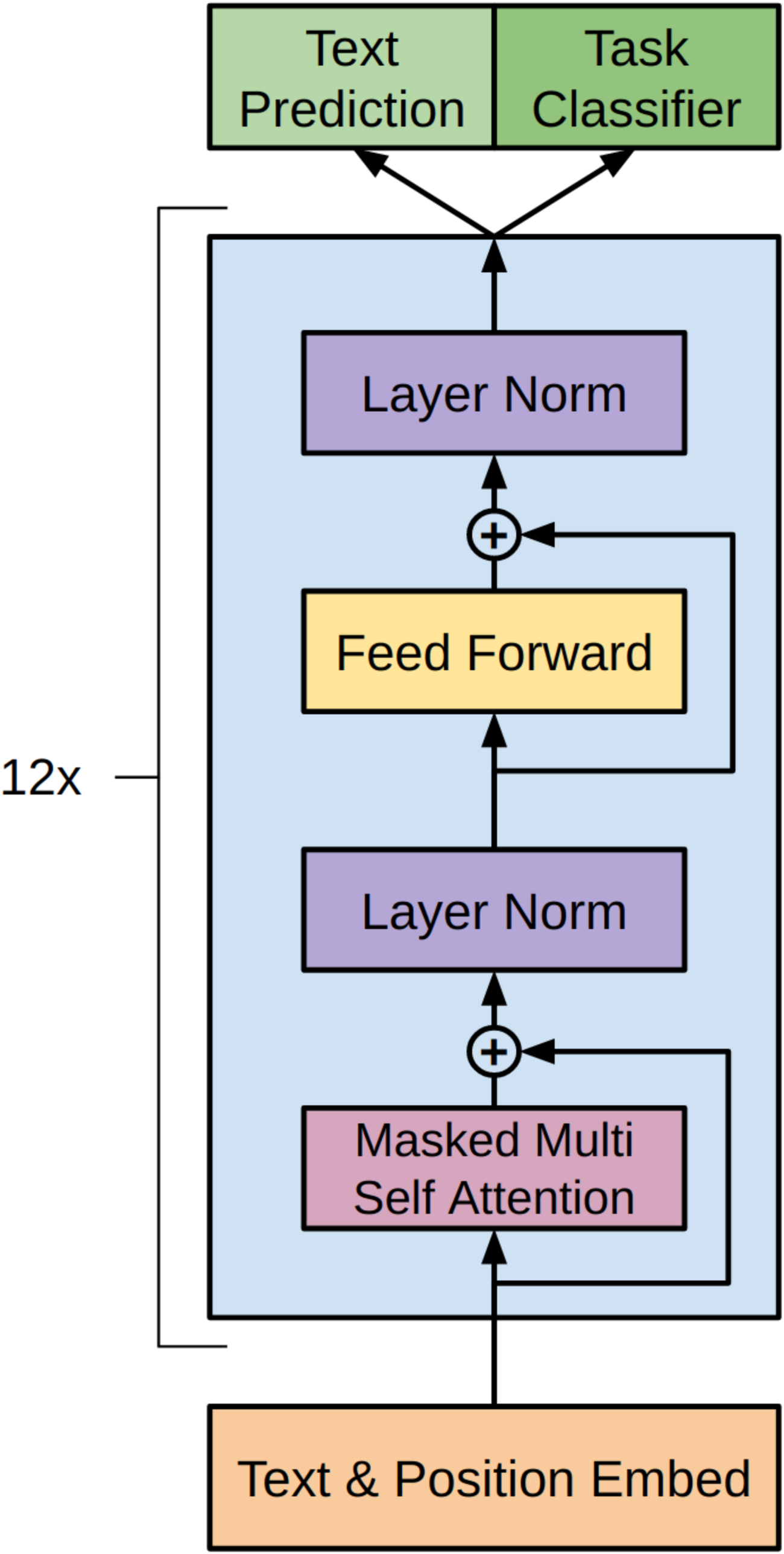
Unsupervised pre-training
Un uses a standard language modeling objective to maximize the following likelihood: L1(U) = i log P(u
Supervised fine-tuning
We perform experiments on a variety of supervised tasks including natural language inference, questio
Task-specific input transformations
previous work proposed learning task specific architectures on top of transferred representations [44] .
Experiments

Setup
Unsupervised pre-training We use the BooksCorpus dataset [71] for training the language model . It co
Analysis
We observed the impact of transferring a variable number of layers from unsupervised pre-training to th
Conclusion
generative pre-training and discriminative fine-tuning introduced a framework for achieving strong natur


RACE Dev Accuracy
MultiNLI Matched Dev Accuracy
# of layers transferred
RACE Dev
RACE Train
MultiNLI Dev
MultiNLI Train
100
90
80
70
60
50
40
30
95
85
75
0
2
4
6
8
10
12


sentiment analysis
winograd schema resolution
linguistic acceptability
question answering
Transformer
LSTM
Relative Task Performance
1.0
0.8
0.6
0.4
0.2
0.0
10^3
10^4
10^5
10^6
# of pre-training updates


Classification
Start
Text
Extract
Transformer
Linear
Entailment
Start
Premise
Delim
Hypothesis
Extract
Transformer
Linear
Similarity
Start
Text 1
Delim
Text 2
Extract
Start
Text 2
Delim
Text 1
Extract
Transformer
Transformer
Linear
Multiple Choice
Start
Context
Delim
Answer 1
Extract
Start
Context
Delim
Answer 2
Extract
Start
Context
Delim
Answer N
Extract
Transformer
Transformer
Transformer
Linear
Linear
Linear


Text Prediction
Task Classifier
Layer Norm
Feed Forward
Layer Norm
Masked Multi Self Attention
12x
Text & Position Embed