Language Models are Unsupervised Multitask Learners

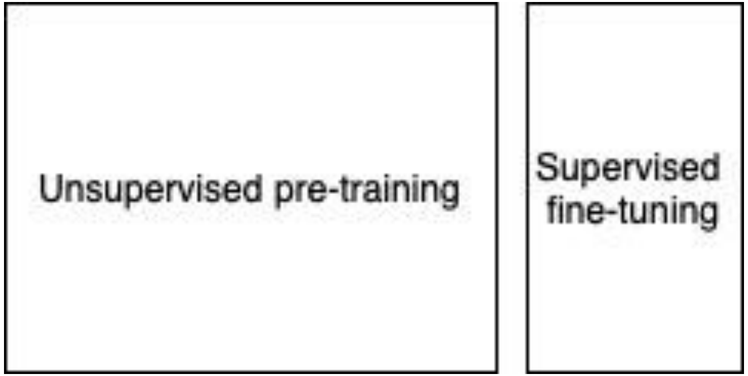
Improving Language Understanding by Generative Pre-Training

Sentemova Olga

Plan

- GPT-1
  - Setting up the framework
  - Task-specific transformation
  - Experiments
  - Results
- GPT-2
  - Approach
  - Training dataset
  - Input transformation and representation
  - Model
  - Experiments
  - Determine and reducing effect from the overlap from the test and train sets

GPT-1: Framework



Unsupervised pre-training

Supervised fine-tuning

$$L_1(\mathcal{U}) = \sum_i \log P(u_i | u_{i-k}, \ldots, u_{i-1}; \Theta)$$

$$P(y|x^1, \ldots, x^m) = \texttt{softmax}(h_l^m W_y).$$

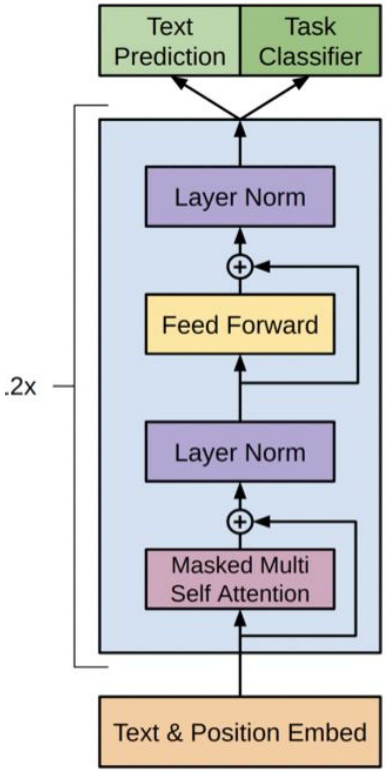
$$L_2(\mathcal{C}) = \sum_{(x,y)} \log P(y|x^1, \ldots, x^m).$$

$$h_0 = UW_e + W_p$$

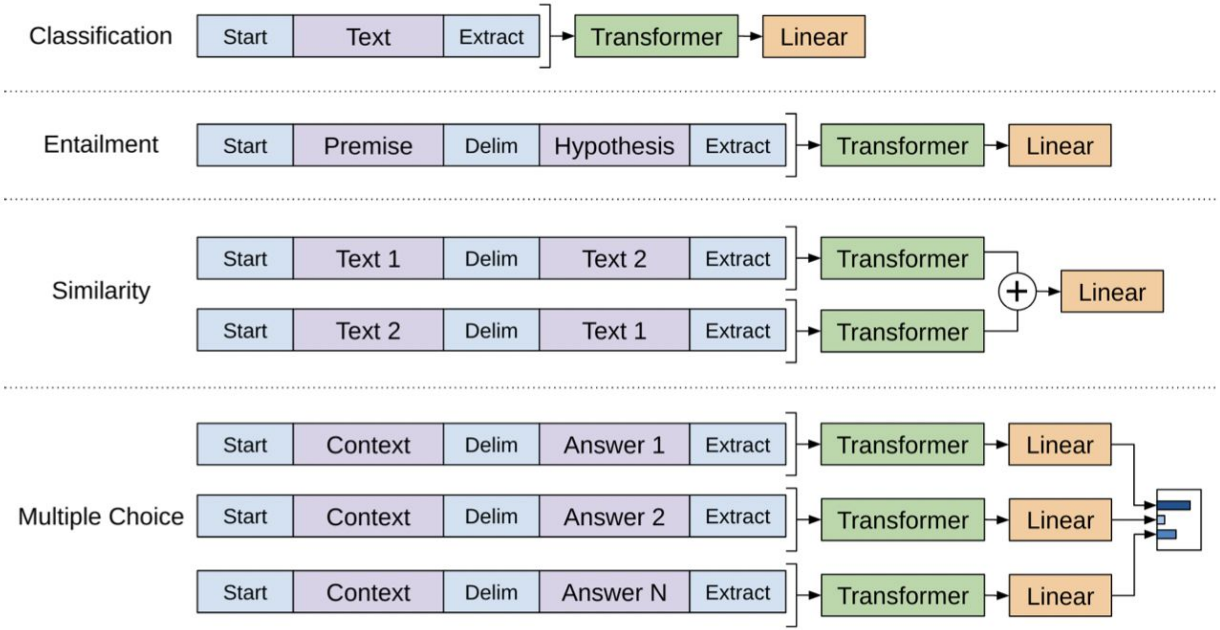
$$h_l = \texttt{transformer\_block}(h_{l-1}) \forall i \in [1, n]$$

$$P(u) = \texttt{softmax}(h_n W_e^T)$$

Transformer architecture



Fine tuning input transformations



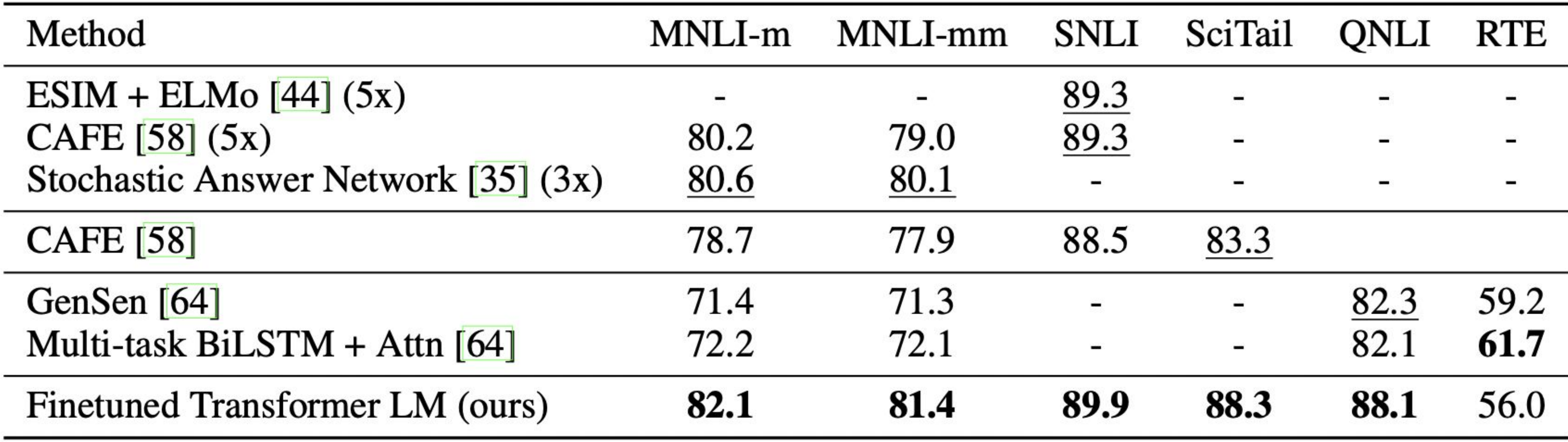
Experiments: Unsupervised pre-training

- Dataset
  - BooksCorpus
- Model parameters:
  - Byte Pair Encoding (40000)
  - 12 layer
  - 768 dimensional states
  - 12 attention heads
  - Positional-wise 3072 dimensional inner states
  - Adam optimisation with max learning rate 2.5*e-4
  - 100 epochs
  - mini batches 64 randomly chosen samples
  - GELU
  - ftfy for cleaning
  - spaCy to tokenize

Experiments: fine tuning

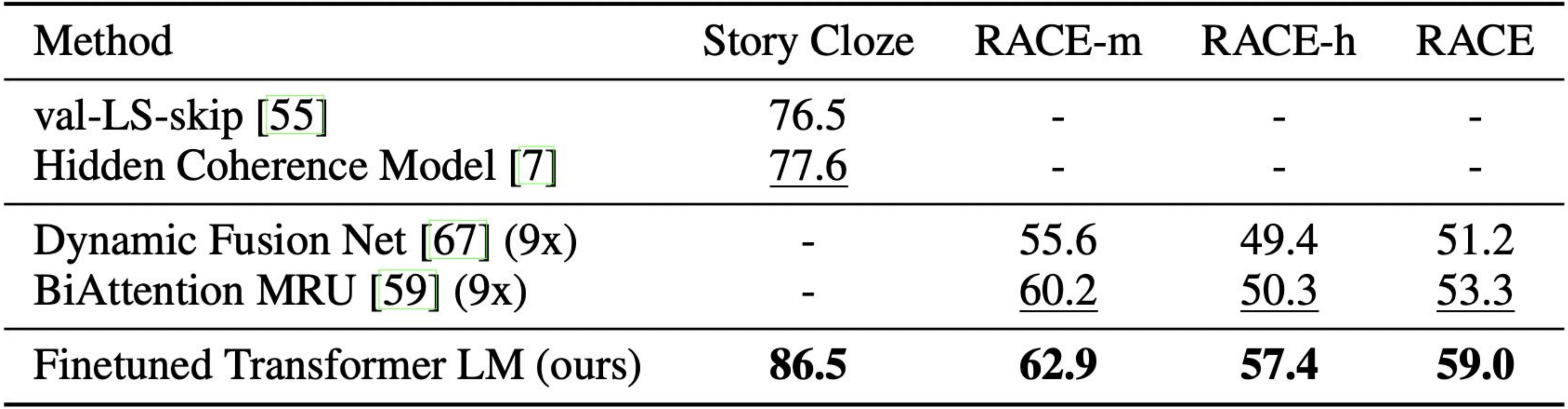
- Model parameters:
  - Hyperparameters are the same as for unsupervised learning
  - dropout 0.1
  - 3 epochs
  - learning rate 6.25e-5

Comparison to another methods: NLI



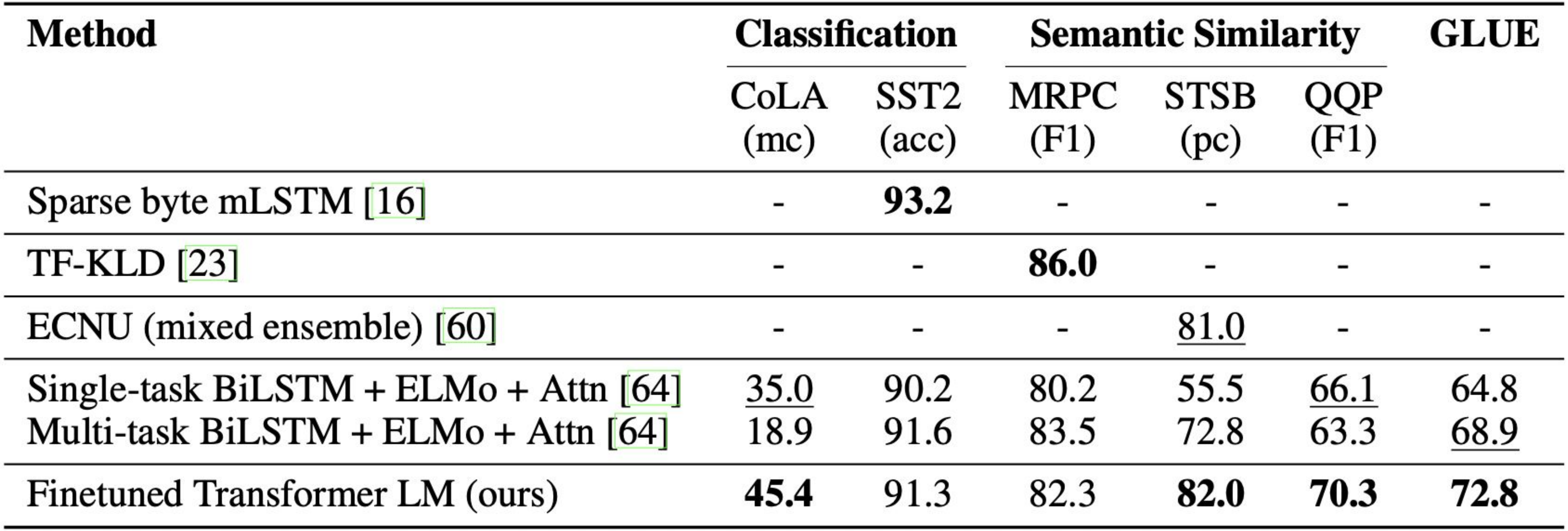
Method	MNLI-m	MNLI-mm	SNLI	SciTail	QNLI	RTE
ESIM + ELMo [44] (5x)	-	-	89.3	-	-	-
CAFE [58] (5x)	80.2	79.0	89.3	-	-	-
Stochastic Answer Network [35] (3x)	80.6	80.1	-	-	-	-
CAFE [58]	78.7	77.9	88.5	83.3		
GenSen [64]	71.4	71.3	-	-	82.3	59.2
Multi-task BiLSTM + Attn [64]	72.2	72.1	-	-	82.1	**61.7**
Finetuned Transformer LM (ours)	**82.1**	**81.4**	**89.9**	**88.3**	**88.1**	56.0

Comparison to another methods: questions answering



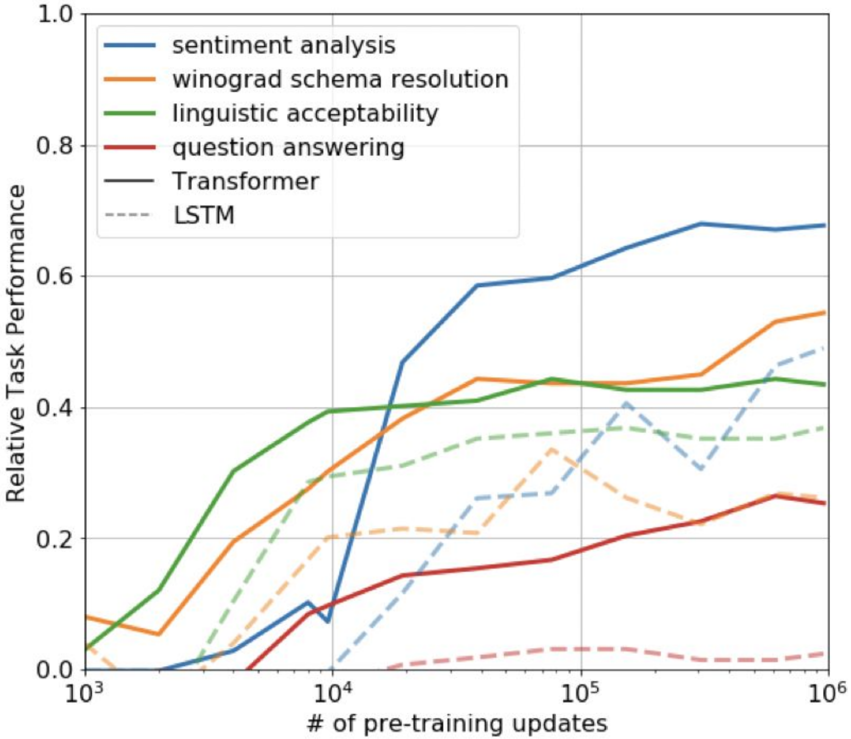
Method	Story Cloze	RACE-m	RACE-h	RACE
val-LS-skip [55]	76.5	-	-	-
Hidden Coherence Model [7]	77.6	-	-	-
Dynamic Fusion Net [67] (9x)	-	55.6	49.4	51.2
BiAttention MRU [59] (9x)	-	60.2	50.3	53.3
Finetuned Transformer LM (ours)	**86.5**	**62.9**	**57.4**	**59.0**

Comparison to another methods: Classification

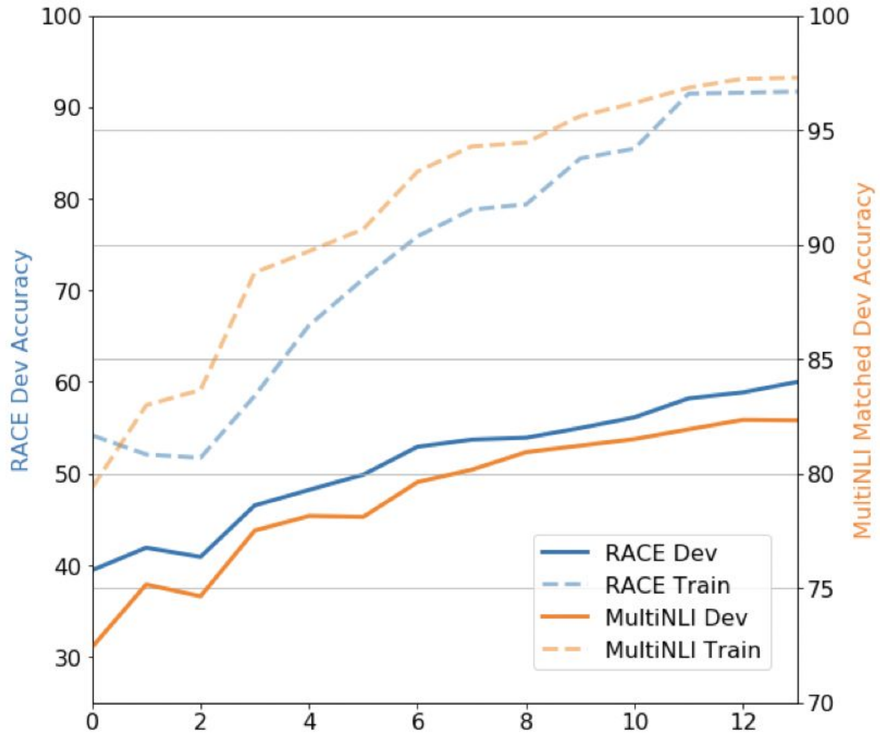


Method	Classification		Semantic Similarity			GLUE
	CoLA (mc)	SST2 (acc)	MRPC (F1)	STSB (pc)	QQP (F1)	
Sparse byte mLSTM [16]	-	**93.2**	-	-	-	-
TF-KLD [23]	-	-	**86.0**	-	-	-
ECNU (mixed ensemble) [60]	-	-	-	81.0	-	-
Single-task BiLSTM + ELMo + Attn [64]	35.0	90.2	80.2	55.5	66.1	64.8
Multi-task BiLSTM + ELMo + Attn [64]	18.9	91.6	83.5	72.8	63.3	68.9
Finetuned Transformer LM (ours)	**45.4**	91.3	82.3	**82.0**	**70.3**	**72.8**

Analysis: pre-train stage



Analysis: fine tuning



GPT-2 vs GPT-1: Task conditioning

$$p(output|input)$$

before GPT-2

$$p(output|input, task).$$

GPT-2

GPT-2 vs GPT-1: Zero Shot Learning and Zero Short Task Transfer

GPT-1:

- Model understands the task based on the input format

GPT-2:

- Model understands the task based on the task text -> it can solve unknown tasks

Dataset

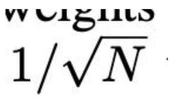
- Common crawl
  - Huge dataset
  - Has quality issues
- -> WebText (Reddit)
  - 8 mln documents

Input representation

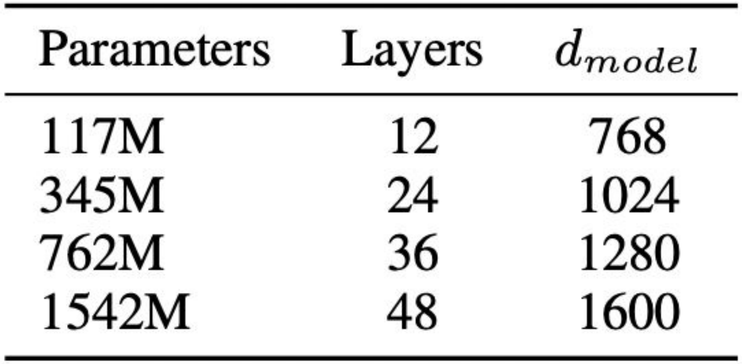
- BTE

Model specification

- Architecture is the same as GPT-1
- Residual layers were scaled with coef


weights

$1/\sqrt{N}$.

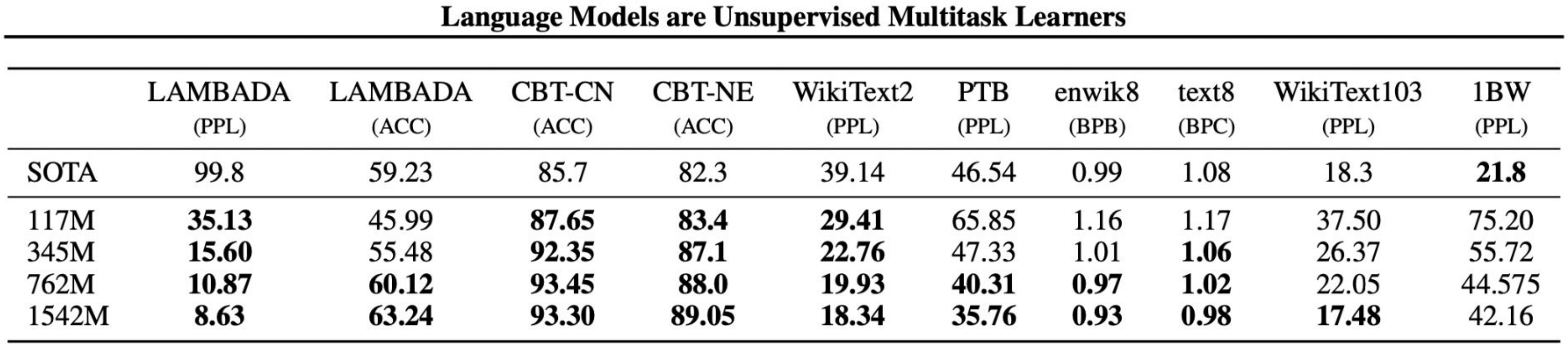


Parameters	Layers	d_{model}
117M	12	768
345M	24	1024
762M	36	1280
1542M	48	1600

Experiments

- Language modeling
- Children’s book test
- LAMBADA
- Winograd schema challenge
- The conversation comprehension
- TLDR
- Translation
- Question answering

Experiments results



Language Models are Unsupervised Multitask Learners

	LAMBADA (PPL)	LAMBADA (ACC)	CBT-CN (ACC)	CBT-NE (ACC)	WikiText2 (PPL)	PTB (PPL)	enwik8 (BPB)	text8 (BPC)	WikiText103 (PPL)	1BW (PPL)
SOTA	99.8	59.23	85.7	82.3	39.14	46.54	0.99	1.08	18.3	**21.8**
117M	**35.13**	45.99	**87.65**	**83.4**	**29.41**	65.85	1.16	1.17	37.50	75.20
345M	**15.60**	55.48	**92.35**	**87.1**	**22.76**	47.33	1.01	**1.06**	26.37	55.72
762M	**10.87**	**60.12**	**93.45**	**88.0**	**19.93**	**40.31**	**0.97**	**1.02**	22.05	44.575
1542M	**8.63**	**63.24**	**93.30**	**89.05**	**18.34**	**35.76**	**0.93**	**0.98**	**17.48**	42.16

What about overlap between test and train?

- Bloom filters (8-grams)
- Result: mostly everything is fine

Links

- GPT-1
- GPT-2

Questions

- What classes of tasks can be solved with GPT-1. Name 3.
- What architecture is used for unsupervised pre-training GPT-1?
- What is Byte Pair Encoding?
- What are the main differences between GPT-1 и GPT-2?