# LANGUAGE MODELS ARE UNSUPERVISED MULTITASK LEARNERS

Alec Radford, Jeffrey Wu, Rewon Child,
David Luan, Dario Amodei, Ilya Sutskever

OpenAI

GPT-2

The scientist named the population, after their distinctive horn, Ovid's Unicorn.

Presented by: Akila Peiris

# We will cover…

- Introduction
- Approach
- Experiments
- Generalization vs Memorization
- Conclusion

# INTRODUCTION

## LANGUAGE MODELING

- Language model (LM)
  - Machine learning model
  - Predict next word of a sentence

- Single task probabilistic framework
  - p(output|input) [1,2]

- General system probabilistic framework
  - p(output|input, task) [3,4,5]


Figure 1: Predictive text suggestion feature on a smart phone


[1] Jelinek, F. and Mercer, R. L. Interpolated estimation of markov source parameters from sparse data. *In Proceedings of the Workshop on Pattern Recognition in Practice, Amsterdam, The Netherlands: North-Holland, May.*, 1980.
[2] Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155, 2003.
[3] Kaiser, L., Gomez, A. N., Shazeer, N., Vaswani, A., Parmar, N., Jones, L., and Uszkoreit, J. One model to learn them all. *arXiv preprint arXiv:1706.05137*, 2017.
[4] Finn, C., Abbeel, P., and Levine, S. Model-agnostic meta- learning for fast adaptation of deep networks. *arXiv preprint arXiv:1703.03400*, 2017.
[5] McCann, B., Keskar, N. S., Xiong, C., and Socher, R. The natural language decathlon: Multitask learning as question answering. *arXiv preprint arXiv:1806.08730*, 2018.

# APPROACH

## DATASET

- Common Crawl
  - Significant data quality issues
  - Best when using a subset similar to the target dataset [6]
  - GPT-2 wanted to avoid making assumptions about the tasks to be performed
- Solution: Human curated web pages
- WebText
  - Emphasizes document quality
  - Outbound links from Reddit
  - Extracted using Dragnet (Peters & Lecocq, 2013) + Newspaper* content extractor
  - 45 million links -> 8 million documents (40 GB text)
  - Sans Wikipedia data

[6] Trinh, T. H. and Le, Q. V. A simple method for commonsense reasoning. *arXiv preprint arXiv:1806.02847*, 2018.
[7] Peters, M. E. and Lecocq, D. Content extraction using diverse fea- ture sets. In *Proceedings of the 22nd International Conference on World Wide Web*, pp. 89–90. ACM, 2013.
[*] https://github.com/codelucas/newspaper

# APPROACH

## INPUT REPRESENTATION

- Byte-level LMs are not competitive with word-level LMs on large scale datasets [8]
- Byte Pair Encoding (BPE) [9]
  - Middle ground between character and word level language modelling
- "Effectively interpolates between word level inputs for frequent symbol sequences and character level inputs for infrequent symbol sequences"
- Benefits:
  - Empirical benefits of word-level LMs
  - Generality of byte-level
- Assigns a probability to any Unicode string
  - Regardless of pre-processing, tokenization, or vocab size

[8] Al-Rfou, R., Choe, D., Constant, N., Guo, M., and Jones, L. Character-level language modeling with deeper self-attention. *arXiv preprint arXiv:1808.04444*, 2018.
[9] Sennrich, R., Haddow, B., and Birch, A. Neural machine trans- lation of rare words with subword units. *arXiv preprint arXiv:1508.07909*, 2015.

# APPROACH

## MODEL

- Transformer [10] based architecture
- Follows OpenAI’s original GPT model [11] except
  - Layer normalization [12] moved to the input of each sub-block
  - Additional layer normalization added after the final self-attention block
- Vocabulary 50,257
- Context increased 512 -> 1024 tokens
- Batch size 512

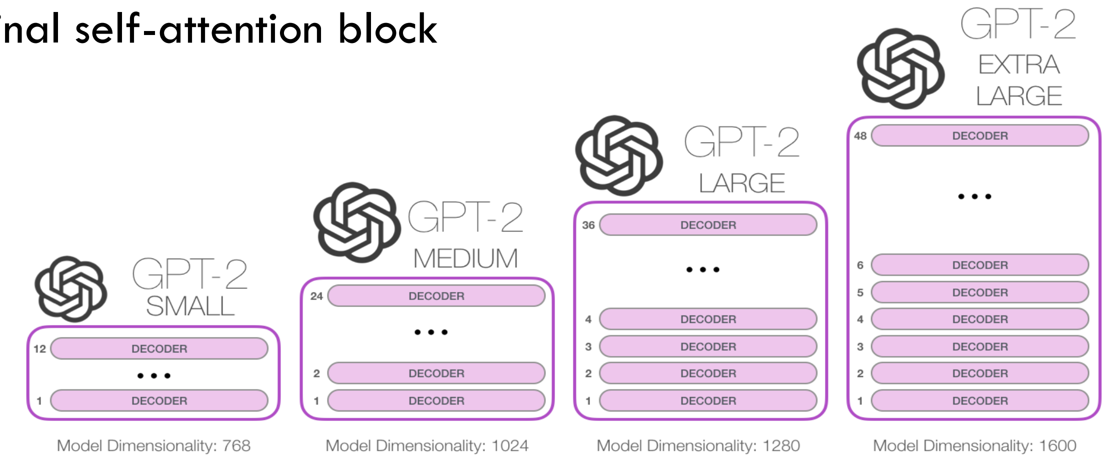


Figure 2: Architecture hyperparameters for the 4 model sizes.

[10] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*, pp. 5998–6008, 2017.
[11] Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. Improving language understanding by generative pre-training. 2018.
[12] Ba, J. L., Kiros, J. R., and Hinton, G. E. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

# EXPERIMENTS

## ZERO-SHOT RESULTS

| | 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 |

Table 1: Zero-shot results on many datasets. No training or fine-tuning was performed for any of these results. PTB and WikiText-2 results are from (Gong et al., 2018) [13]. CBT results are from (Bajgar et al., 2016) [14]. LAMBADA accuracy result is from (Hoang et al., 2018) [15] and LAMBADA perplexity result is from (Grave et al., 2016) [16]. Other results are from (Dai et al., 2019) [17].


[13] Gong, C., He, D., Tan, X., Qin, T., Wang, L., and Liu, T.-Y. Frage: frequency-agnostic word representation. In *Advances in Neural Information Processing Systems*, pp. 1341–1352, 2018.

[14] Bajgar, O., Kadlec, R., and Kleindienst, J. Embracing data abun- dance: Booktest dataset for reading comprehension. *arXiv preprint arXiv:1610.00956*, 2016.

[15] Hoang, L., Wiseman, S., and Rush, A. M. Entity tracking im- proves cloze-style reading comprehension. *arXiv preprint arXiv:1810.02891*, 2018.

[16] Grave, E., Joulin, A., and Usunier, N. Improving neural language models with a continuous cache. *arXiv preprint arXiv:1612.04426*, 2016.

[17] Dai, Z., Yang, Z., Yang, Y., Cohen, W. W., Carbonell, J., Le, Q. V., and Salakhutdinov, R. Transformer-xl: Attentive lan- guage models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*, 2019.

# EXPERIMENTS

## CHILDREN'S BOOK TEST

- Hill et al., in 2015 [18] created this test to examine the performance of LMs on different categories of words:
  - named entities, nouns, verbs, and prepositions
- Reports accuracy on automatically constructed cloze test
  - The task: predict which of 10 possible choices for an omitted word is correct.
- Same approach introduced in original paper
  - Compute probability of each choice and the rest of the sentence conditioned on this choice according to the LM, and predict the one with the highest probability

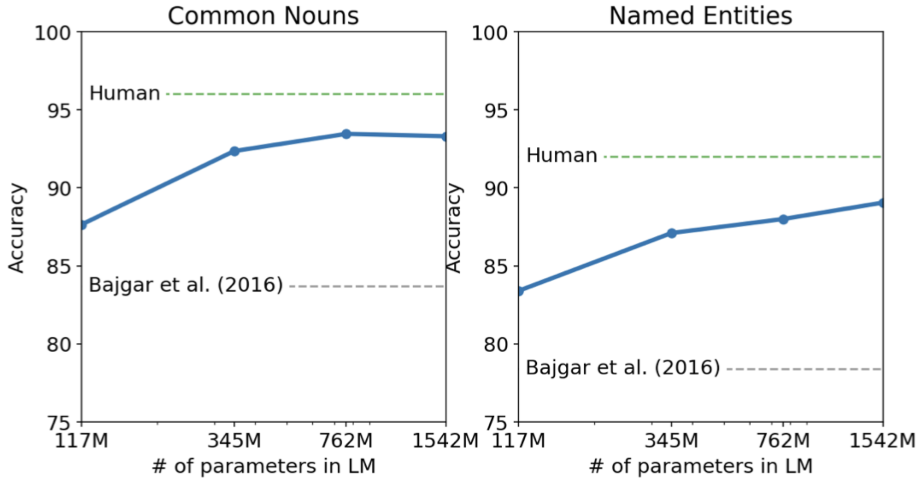


Figure 3: Performance on the Children's Book Test as a function of model capacity.


[18] Hill, F., Bordes, A., Chopra, S., and Weston, J. The goldilocks principle: Reading children's books with explicit memory rep- resentations. *arXiv preprint arXiv:1511.02301*, 2015.

# EXPERIMENTS

## LAMBADA

- LAnguage Modeling Broadened to Account for Discourse Aspects
- Predict the final word of sentences
  - Requires at least 50 tokens of context for a human to successfully predict
- Perplexity score increased from 99.8 [19] to 8.6
- Accuracy increased from 19% [20] to 52.66%
- GPT-2's errors
  - most predictions are valid continuations of the sentence
  - but not valid final words
- Adding a stop-word filter increases accuracy by 4%

[19] Grave, E., Joulin, A., and Usunier, N. Improving neural language models with a continuous cache. *arXiv preprint arXiv:1612.04426*, 2016.
[20] Dehghani, M., Gouws, S., Vinyals, O., Uszkoreit, J., and Kaiser, Ł. Universal transformers. *arXiv preprint arXiv:1807.03819*, 2018.

# EXPERIMENTS

## WINOGRAD SCHEMA CHALLENGE

- Levesque et al., 2012 [21]
  - Measure capability of a system to perform common sense reasoning
  - Done by measuring its ability to resolve ambiguities in text.
- Improves state of the art accuracy by 7%

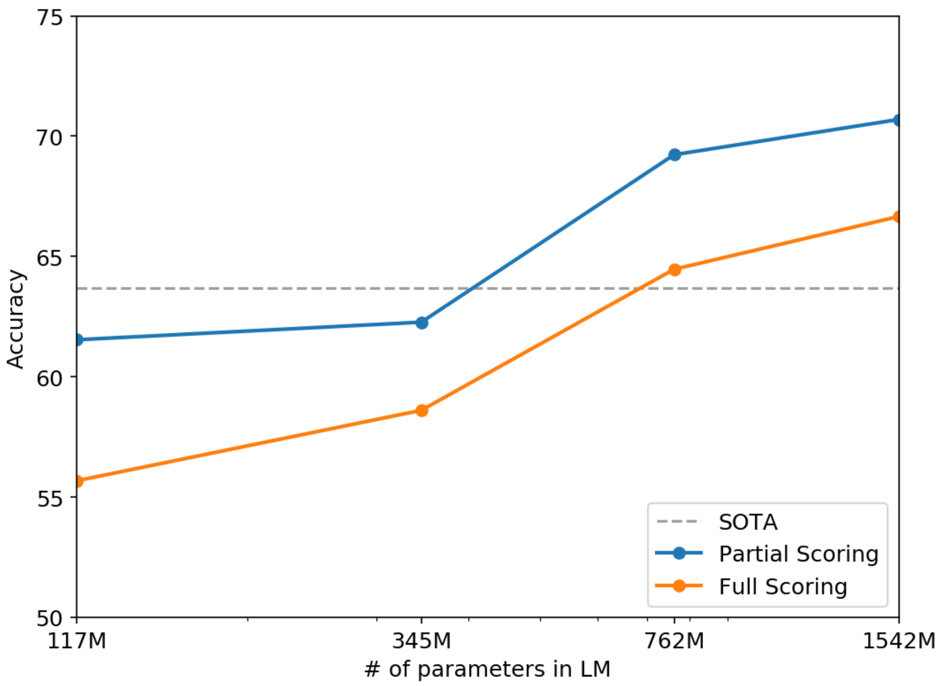


Figure 4: Performance on the Winograd Schema Challenge as a function of model capacity.


[21] Levesque, H., Davis, E., and Morgenstern, L. The winograd schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*, 2012.

# EXPERIMENTS

## READING COMPREHENSION

- Conversation Question Answering dataset (CoQA) Reddy et al. (2018)
  - documents from 7 different domains paired with natural language dialogues of questions and answerers about the document.
- GPT-2 performs well enough for a system without any supervised training
  - But it often uses simple retrieval based heuristics

[22] Reddy, S., Chen, D., and Manning, C. D. Coqa: A conversational question answering challenge. *arXiv preprint arXiv:1808.07042*, 2018.

# EXPERIMENTS

## SUMMARIZATION

- CNN and Daily Mail dataset (Nallapati et al., 2016)
- Steps
  - Add the text TL;DR: after the article
  - Generate 100 tokens with Top-k random sampling (Fan et al., 2018) with k = 2
- Tested with ROUGE 1,2,L metrics
  - Performance similar to classic neural baselines

| | R-1 | R-2 | R-L | R-AVG |
|---|---|---|---|---|
| Bottom-Up Sum | **41.22** | **18.68** | **38.34** | **32.75** |
| Lede-3 | 40.38 | 17.66 | 36.62 | 31.55 |
| Seq2Seq + Attn | 31.33 | 11.81 | 28.83 | 23.99 |
| GPT-2 `TL;DR:` | 29.34 | 8.27 | 26.58 | 21.40 |
| Random-3 | 28.78 | 8.63 | 25.52 | 20.98 |
| GPT-2 no hint | 21.58 | 4.03 | 19.47 | 15.03 |

Table 2: Summarization performance as measured by ROUGE F1 metrics on the CNN and Daily Mail dataset. Bottom-Up Sum is the SOTA model from (Gehrmann et al., 2018)

[23] Nallapati, R., Zhou, B., Gulcehre, C., Xiang, B., et al. Abstrac- tive text summarization using sequence-to-sequence rnns and beyond. *arXiv preprint arXiv:1602.06023*, 2016.

# EXPERIMENTS

## TRANSLATION

- Sample format
  - English sentence = French sentence
- Sample using prompt
  - English sentence =
- Greedy decoding – sample first sentence
- WMT-14 English-French test set – 5 BLEU
  - Worse than word-by-word substitution [24]
- WMT-14 French-English test set – 11.5 BLEU
  - Outperforms several unsupervised machine translation baselines (2017)
  - But lacking compared to state of the art model by [25] which scored 33.5
- WebText only contains 10MB of data in the French language
  - 500x smaller than the monolingual French corpus

[24] Conneau, A., Lample, G., Ranzato, M., Denoyer, L., and Je´gou, H. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*, 2017b.
[25] Artetxe, M., Labaka, G., and Agirre, E. An effective ap- proach to unsupervised machine translation. *arXiv preprint arXiv:1902.01313*, 2019.

# EXPERIMENTS

## QUESTION ANSWERING

- Natural Questions dataset [26]
- Sample format
  - Question = Answer
- Answers 4.1% of questions correctly when evaluated by the exact match metric
- Has an accuracy of 63.1% on the 1% of questions it is most confident in
- Open domain question answering systems can answer 30-50% [27]

[26] Kwiatkowski, T., Palomaki, J., Rhinehart, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Kelcey, M., Devlin, J., et al. Natural questions: a benchmark for question answering research. 2019.
[27] Alberti, C., Lee, K., and Collins, M. A bert baseline for the natural questions. *arXiv preprint arXiv:1901.08634*, 2019.

# GENERALIZATION VS MEMORIZATION

- Important to analyze how much test data also shows up in the training data.
  - Results in an over-reporting of the generalization performance
- Bloom filters
  - Calculate percentage of 8-grams from that dataset that are also found in the WebText training set

| | PTB | WikiText-2 | enwik8 | text8 | Wikitext-103 | 1BW |
|---|---|---|---|---|---|---|
| Dataset train | **2.67%** | 0.66% | **7.50%** | 2.34% | **9.09%** | **13.19%** |
| WebText train | 0.88% | **1.63%** | 6.31% | **3.94%** | 2.42% | 3.75% |

Table 3: Percentage of test set 8 grams overlapping with training sets.

  - Overlap between WebText and specific evaluation datasets provides a small consistent benefit
  - For most datasets there is no significantly larger overlaps than those already existing between standard training and test sets

# CONCLUSIONS

- Performance of GPT-2 is competitive with supervised baselines in a zero-shot setting on reading comprehension
- Does not perform well with tasks such as summarization, translation, question answering, etc.
- Studied zero-shot performance of WebText LMs on many canonical NLP tasks

# LINKS

- OpenAI blog
  - https://openai.com/blog/better-language-models/
  - 1.5 billion version: https://openai.com/blog/gpt-2-1-5b-release/
- Git hub
  - https://github.com/openai/gpt-2
- Huggingface
  - https://huggingface.co/gpt2
  - Transformers doc: https://huggingface.co/docs/transformers/model_doc/gpt2

# REFERENCES


[1] Jelinek, F. and Mercer, R. L. Interpolated estimation of markov source parameters from sparse data. *In Proceedings of the Workshop on Pattern Recognition in Practice, Amsterdam, The Netherlands: North-Holland, May.*, 1980.

[2] Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155, 2003.

[3] Kaiser, L., Gomez, A. N., Shazeer, N., Vaswani, A., Parmar, N., Jones, L., and Uszkoreit, J. One model to learn them all. *arXiv preprint arXiv:1706.05137*, 2017.

[4] Finn, C., Abbeel, P., and Levine, S. Model-agnostic meta- learning for fast adaptation of deep networks. *arXiv preprint arXiv:1703.03400*, 2017.

[5] McCann, B., Keskar, N. S., Xiong, C., and Socher, R. The natural language decathlon: Multitask learning as question answering. *arXiv preprint arXiv:1806.08730*, 2018.

[6] Trinh, T. H. and Le, Q. V. A simple method for commonsense reasoning. *arXiv preprint arXiv:1806.02847*, 2018.

[7] Peters, M. E. and Lecocq, D. Content extraction using diverse fea- ture sets. In *Proceedings of the 22nd International Conference on World Wide Web*, pp. 89–90. ACM, 2013.

[8] Al-Rfou, R., Choe, D., Constant, N., Guo, M., and Jones, L. Character-level language modeling with deeper self-attention. *arXiv preprint arXiv:1808.04444*, 2018.

[9] Sennrich, R., Haddow, B., and Birch, A. Neural machine trans- lation of rare words with subword units. *arXiv preprint arXiv:1508.07909*, 2015.

[10] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*, pp. 5998–6008, 2017.

[11] Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. Improving language understanding by generative pre-training. 2018.

[12] Ba, J. L., Kiros, J. R., and Hinton, G. E. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[13] Gong, C., He, D., Tan, X., Qin, T., Wang, L., and Liu, T.-Y. Frage: frequency-agnostic word representation. In *Advances in Neural Information Processing Systems*, pp. 1341–1352, 2018.

# REFERENCES


[14] Bajgar, O., Kadlec, R., and Kleindienst, J. Embracing data abun- dance: Booktest dataset for reading comprehension. *arXiv preprint arXiv:1610.00956*, 2016.

[15] Hoang, L., Wiseman, S., and Rush, A. M. Entity tracking im- proves cloze-style reading comprehension. *arXiv preprint arXiv:1810.02891*, 2018.

[16] Grave, E., Joulin, A., and Usunier, N. Improving neural language models with a continuous cache. *arXiv preprint arXiv:1612.04426*, 2016.

[17] Dai, Z., Yang, Z., Yang, Y., Cohen, W. W., Carbonell, J., Le, Q. V., and Salakhutdinov, R. Transformer-xl: Attentive lan- guage models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*, 2019.

[18] Hill, F., Bordes, A., Chopra, S., and Weston, J. The goldilocks principle: Reading children's books with explicit memory rep- resentations. *arXiv preprint arXiv:1511.02301*, 2015.

[19] Grave, E., Joulin, A., and Usunier, N. Improving neural language models with a continuous cache. *arXiv preprint arXiv:1612.04426*, 2016.

[20] Dehghani, M., Gouws, S., Vinyals, O., Uszkoreit, J., and Kaiser, Ł. Universal transformers. *arXiv preprint arXiv:1807.03819*, 2018.

[21] Levesque, H., Davis, E., and Morgenstern, L. The winograd schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*, 2012.

[22] Reddy, S., Chen, D., and Manning, C. D. Coqa: A conversational question answering challenge. *arXiv preprint arXiv:1808.07042*, 2018.

[23] Nallapati, R., Zhou, B., Gulcehre, C., Xiang, B., et al. Abstrac- tive text summarization using sequence-to-sequence rnns and beyond. *arXiv preprint arXiv:1602.06023*, 2016.

[24] Conneau, A., Lample, G., Ranzato, M., Denoyer, L., and Je´gou, H. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*, 2017b.

[25] Artetxe, M., Labaka, G., and Agirre, E. An effective ap- proach to unsupervised machine translation. *arXiv preprint arXiv:1902.01313*, 2019.

# REFERENCES

[27] Alberti, C., Lee, K., and Collins, M. A bert baseline for the natural questions. *arXiv preprint arXiv:1901.08634*, 2019.