Scaling Laws for Transfer

•••

By Danny Hernandez, Jared Kaplan, Tom Henighan, Sam McCandlish

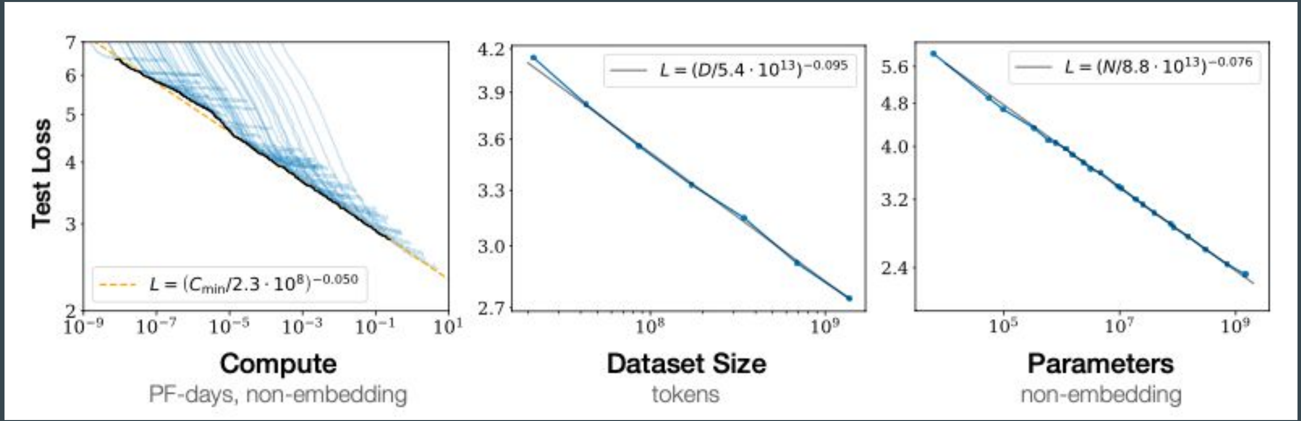
Presented by Balaji Balasubramanian and Eshwanth Baskaran

Index

1. What are scaling laws?
2. What is transfer learning?
3. What are scaling laws for transfer?
4. Experiment 1: Transfer from English to Python programming language
5. Experiment 2: Transfer from English to other human languages
6. Conclusion

What are scaling laws?

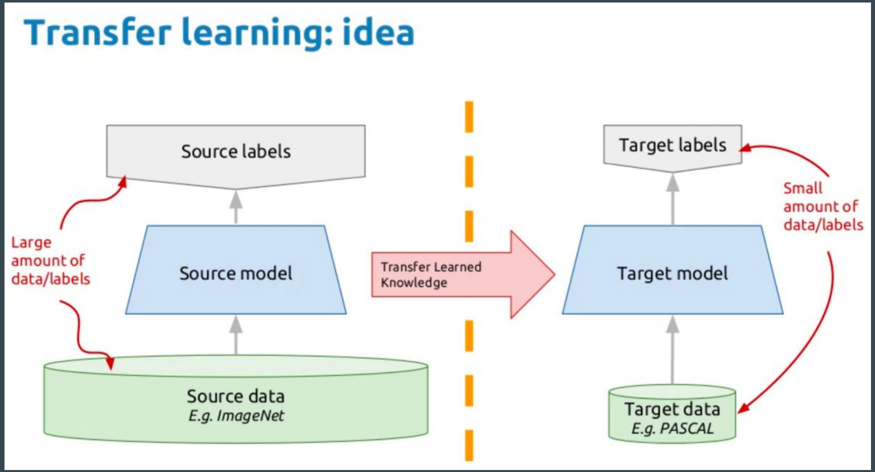
- The Language Model's performance increases smoothly as we increase the amount of compute used for training, dataset size and the model size.
- In each of the three graphs below, one of the three metrics is fixed and the others can be freely varied to obtain the best performance.



Kaplan et al. - Scaling Laws for Neural Language Models

What is transfer learning?

- Transfer the knowledge gained in one task to use it in another task.
- In deep learning, use a pre-trained network and apply it to a custom task.
- Useful in cases where the labelled data is very less.



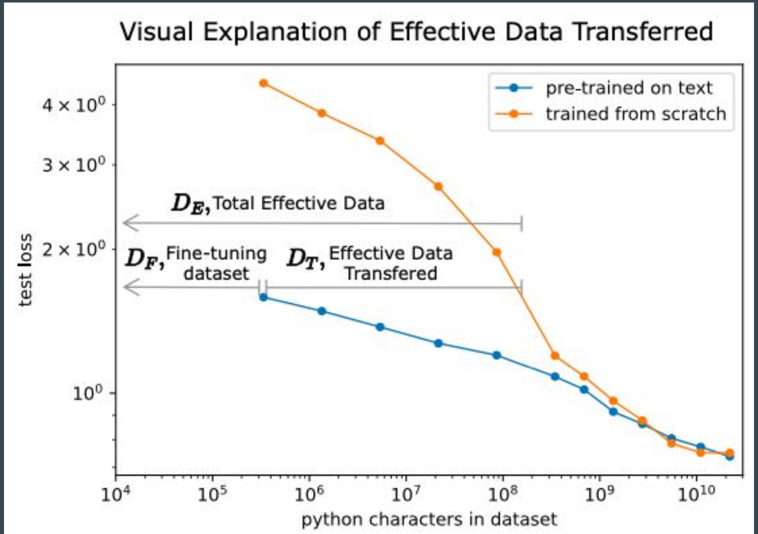
https://www.pinterest.com/pin/424745808604824736/

What are scaling laws for transfer?

- Most of the work in scaling assumes the availability of infinite amount of data.
- But most real world problems have limited availability of data.
- This paper focuses on benefits of transfer learning for low data regime and scaling law equations for quantifying the benefits.

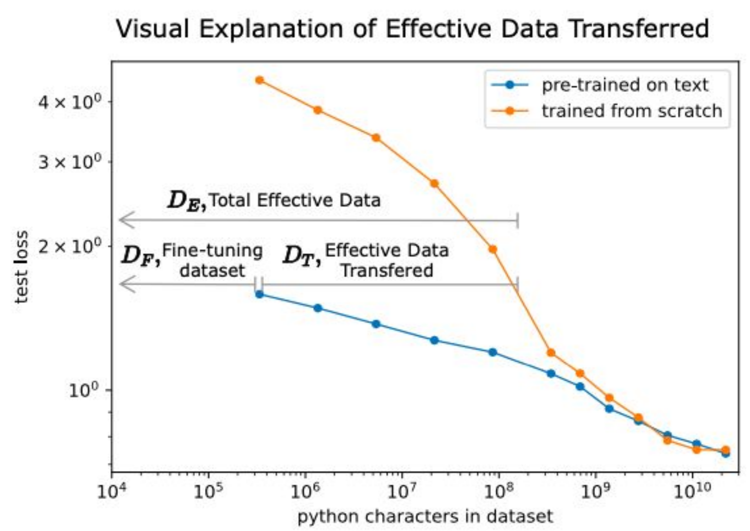
Experiment 1: Transfer from English to Python language

- Explores transfer learning for a model pretrained on English text and finetuned on python code.
- This was compared to the model that was trained from scratch on Python.



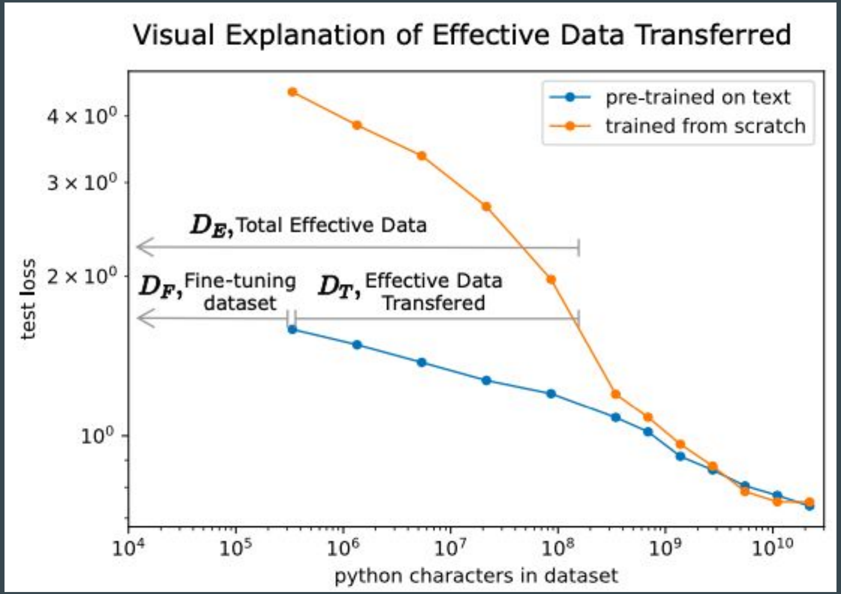
Effective Data Transfer

- Effective Data Transfer is useful for measuring the importance of pretraining
- This graph shows the performance of a 40M parameter model
- Blue- model pretrained on text and finetuned on python.
- Orange- model trained from scratch on python.



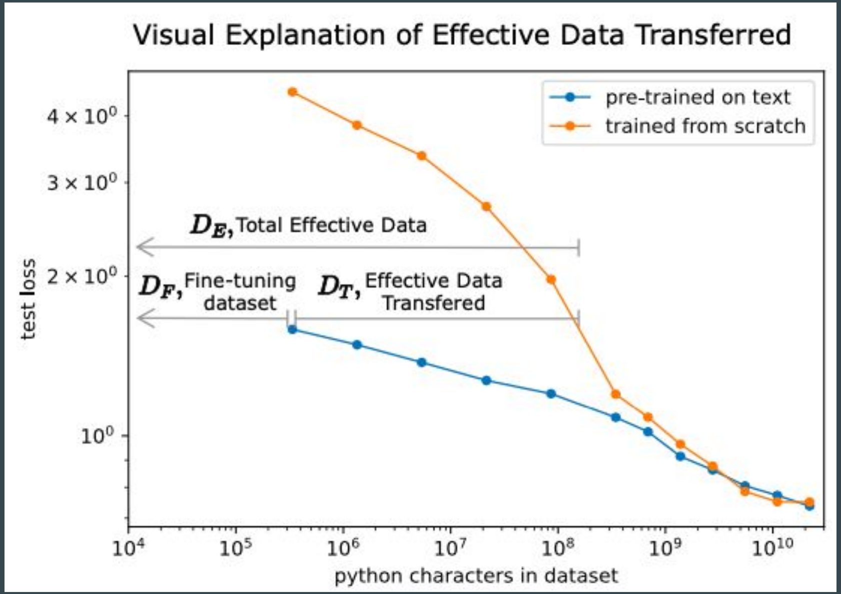
Effective Data Transfer

- D_F - Finetuning Dataset size (in characters)
- D_E - Total effective data the model trained from scratch requires to reach the same performance.
- D_T (Effective Data Transferred) = D_E-D_F



Effective Data Transfer

- For a 40M parameter Transformer model finetuned on 3e5 characters, D_E is approximately 1000x of D_F
- Effectiveness of transfer learning is inversely proportional to the size of the finetuning dataset (D_F)



Datasets

Text data was created using a mix of WebText 2, Common Crawl, English Wikipedia and publicly available Internet books

Webtext2

- Created using the text from outbound links of Reddit posts with atleast 3 karmas.
- Karma is heuristic to know whether people found the post relevant and interesting.
- Consists if 20.3M documents with 1.62 x 1010 words.

Common Crawl

- Common Crawl produces 20TB of web scraped data every month.
- C4 version of Common Crawl used which is a filtered version of Common Crawl used in T5.

Python code dataset consisted of 22 billion character dataset collected from github.

Kaplan et al- Scaling laws for neural language models
Raffel et al- Exploring the limits of transfer learning with a unified text-to-text transformer

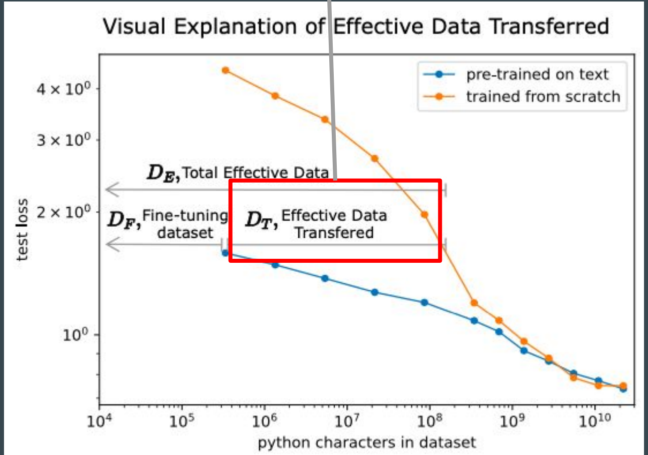
Experiment 1: Training setup

- Trained series of **transformer** language models on 3 different dataset categories:
  - Train from-scratch on python code
  - Pre-train on natural language, then fine-tune on python code
  - Pre-train on natural language and non-python code, then fine-tune on python
- Model and data sizes span 4 order of magnitudes i.e. 10^6 to 10^{10} and 10^5 to 10^9.
- Test loss on a held-out python dataset

Experiment 1: Observations

1. The effective data transferred is well-described by a power-law in the low-data regime.

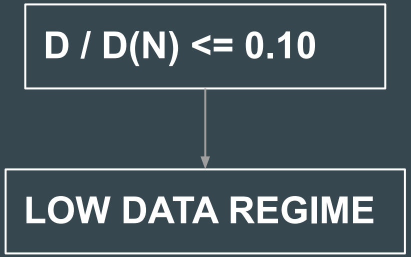
$$D_T = \text{effective data transferred} = k(D_F)^{\alpha}(N)^{\beta}$$



- D_F - Fine-tuning dataset size
- N - Number of non-embedding parameters of model
- α - Scaling law exponent
- β - Scaling law exponent
- k - Constant

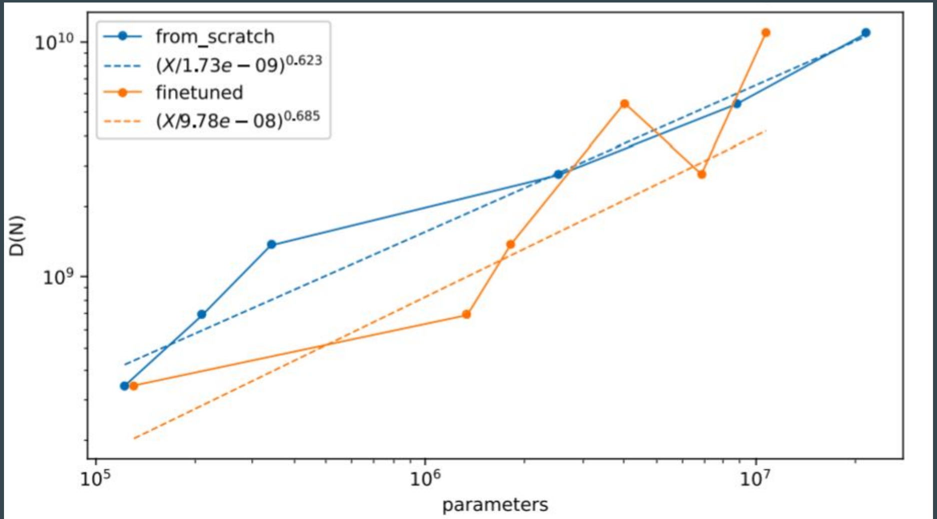
What exactly is a low-data regime?

Notation: **D(N)** - Amount of data it takes to reach 99% of the performance that infinite python data would yield for a given model size.



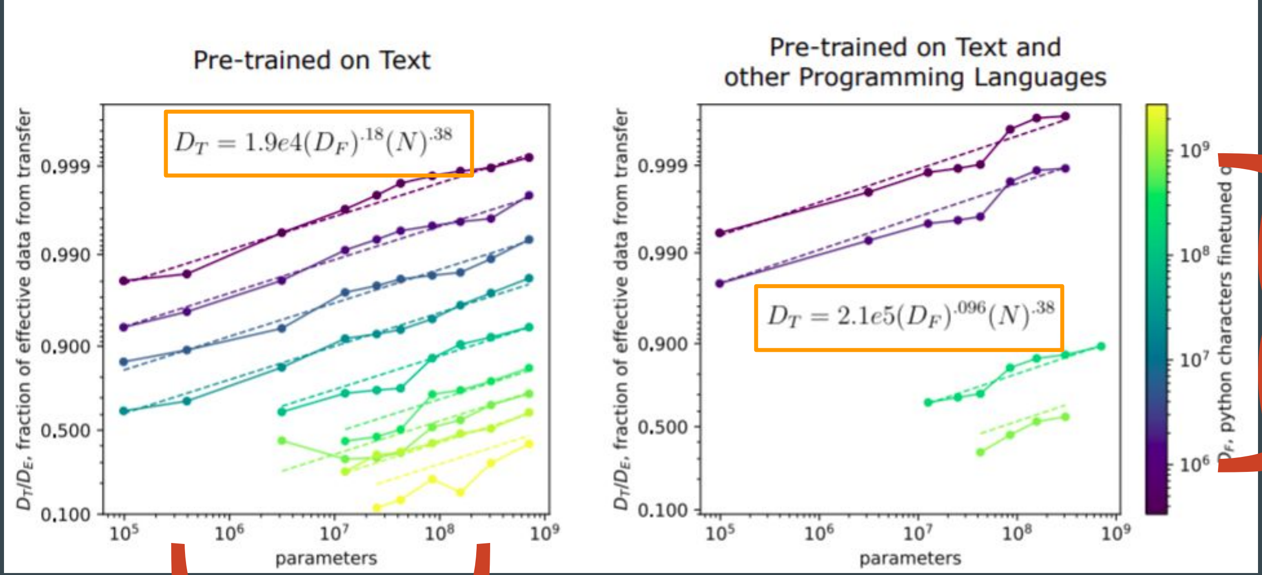
D / D(N) <= 0.10

LOW DATA REGIME



Experiment 1: Observations

1. The effective data transferred is well-described by a power-law in the low-data regime.



good fit for over 3 orders of magnitude in fine-tuning dataset size.

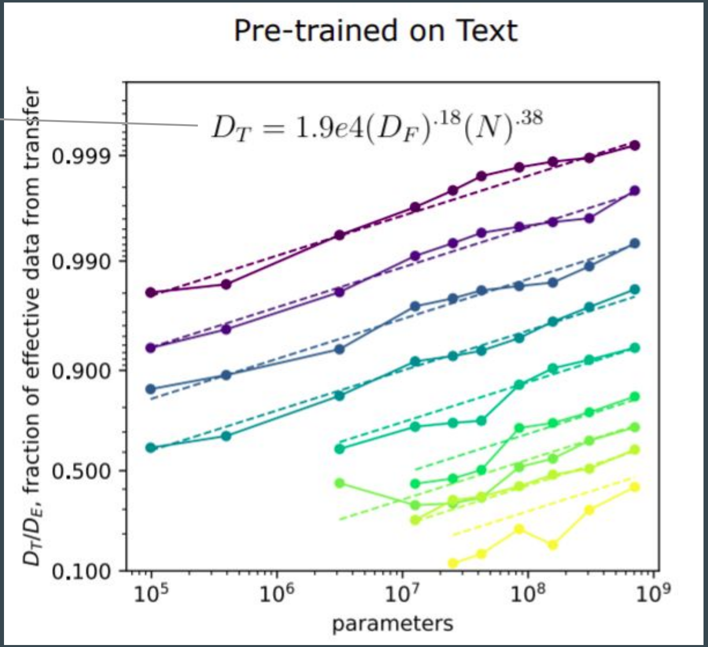
good fit for over 4 orders of magnitude in model size.

Experiment 1: Observations

1. The effective data transferred is well-described by a power-law in the low-data regime.

$\beta \approx 2\alpha$

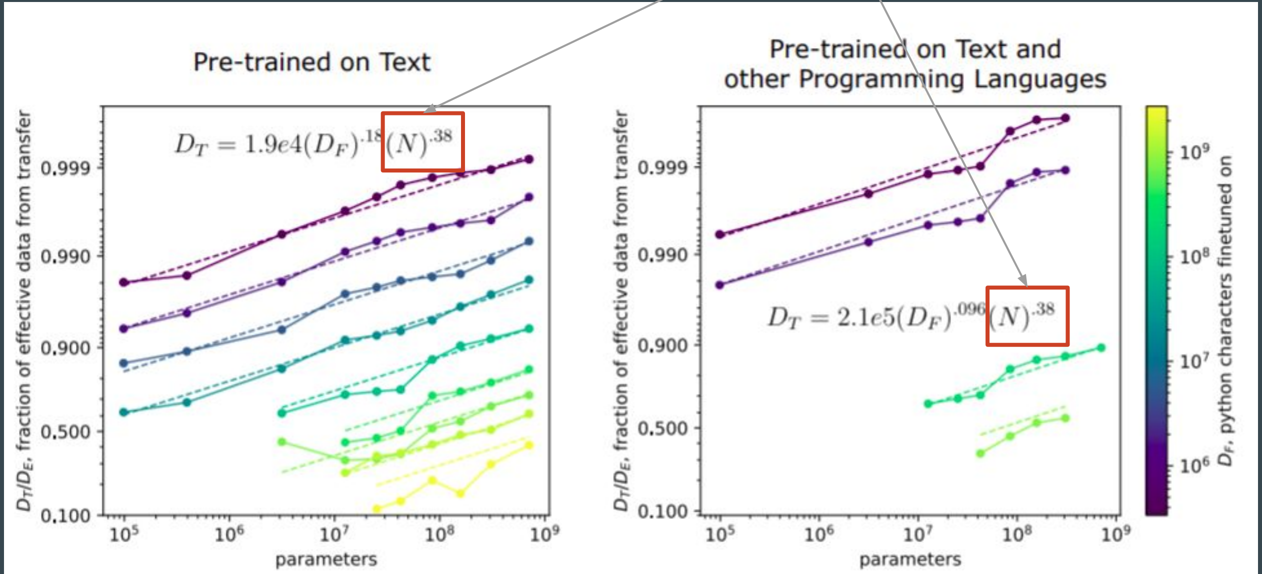
10x increase in model size, N, equivalent to 100x increase in fine-tuning dataset size, D_F.



Experiment 1: Observations

2. Identical scaling with model size, the exponent β = 0.38

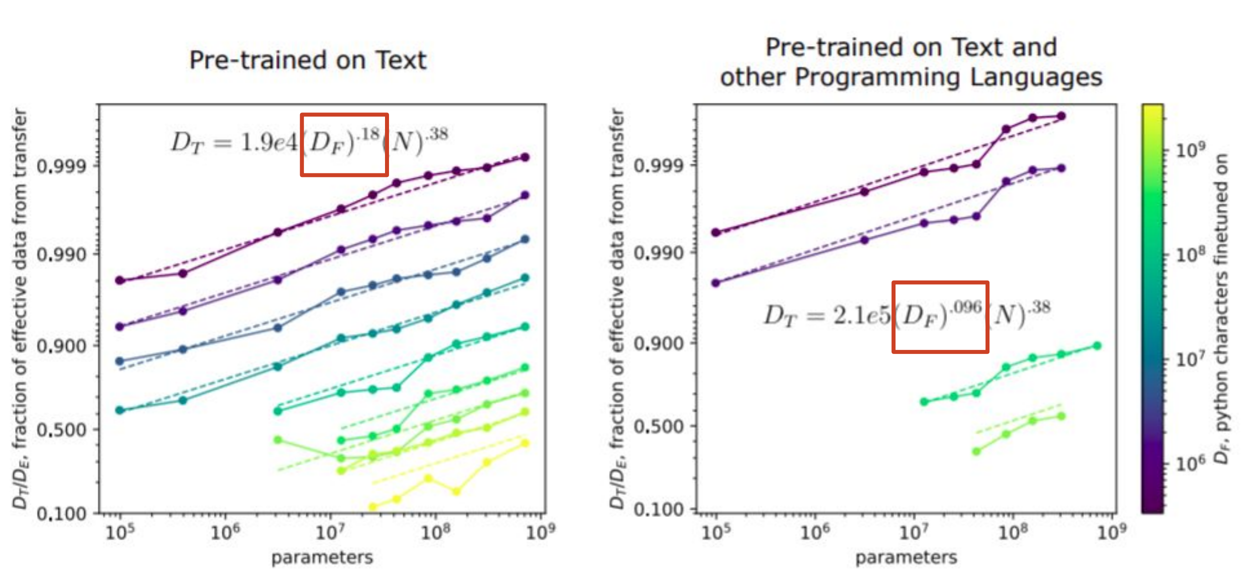
Appears to depend only on the target distribution



Experiment 1: Observations

3. Exponent α - Measure of the proximity of two distributions

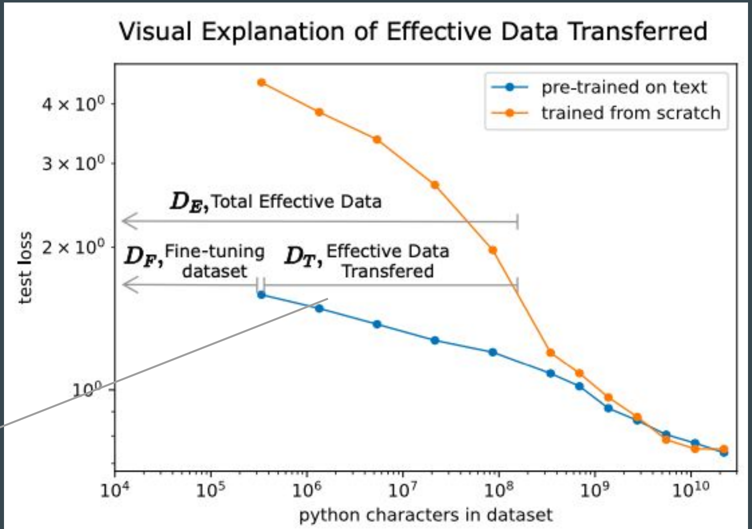
Smaller α ⇒ closer proximity



Experiment 1: Observations

4. Pre-training multiplies the fine-tuning dataset in the low-data regime.

$$\text{effective data multiplier} = \frac{D_F + D_T}{D_F} \approx \frac{D_T}{D_F} = \frac{k(N)^{\beta}}{(D_F)^{1-\alpha}}$$

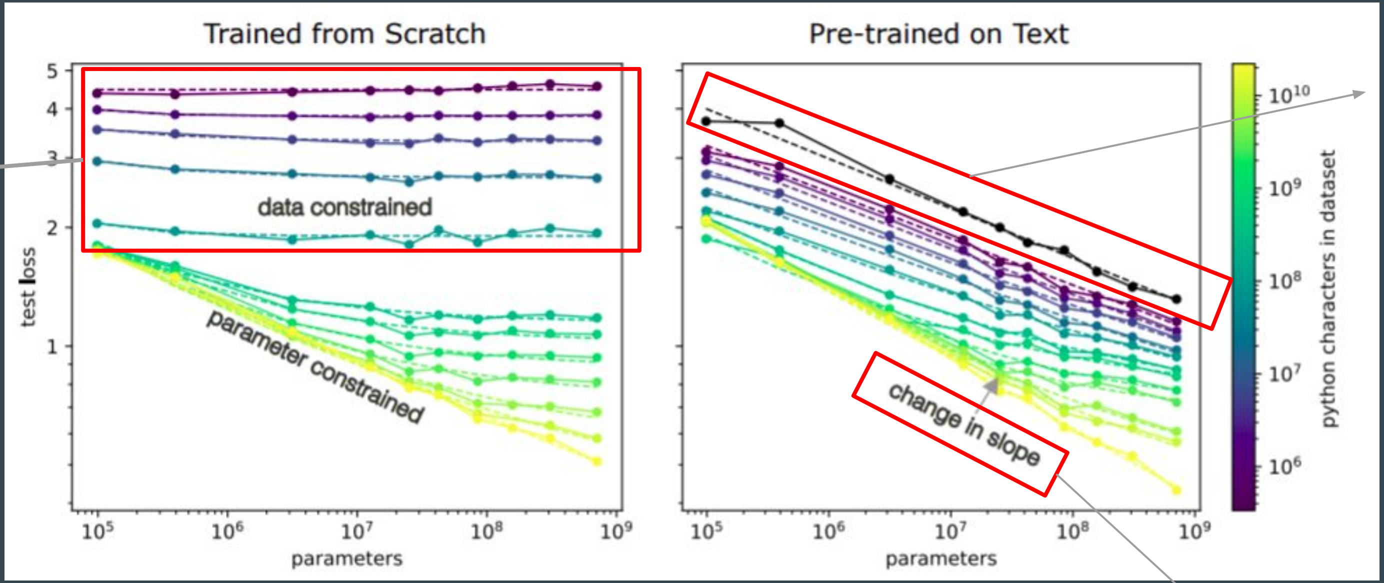


Visual Explanation of Effective Data Transferred

D_T is approximately 1000x bigger than D_F

Experiment 1: Observations

5. When data limits performance, the pre-trained models have a better scaling law.



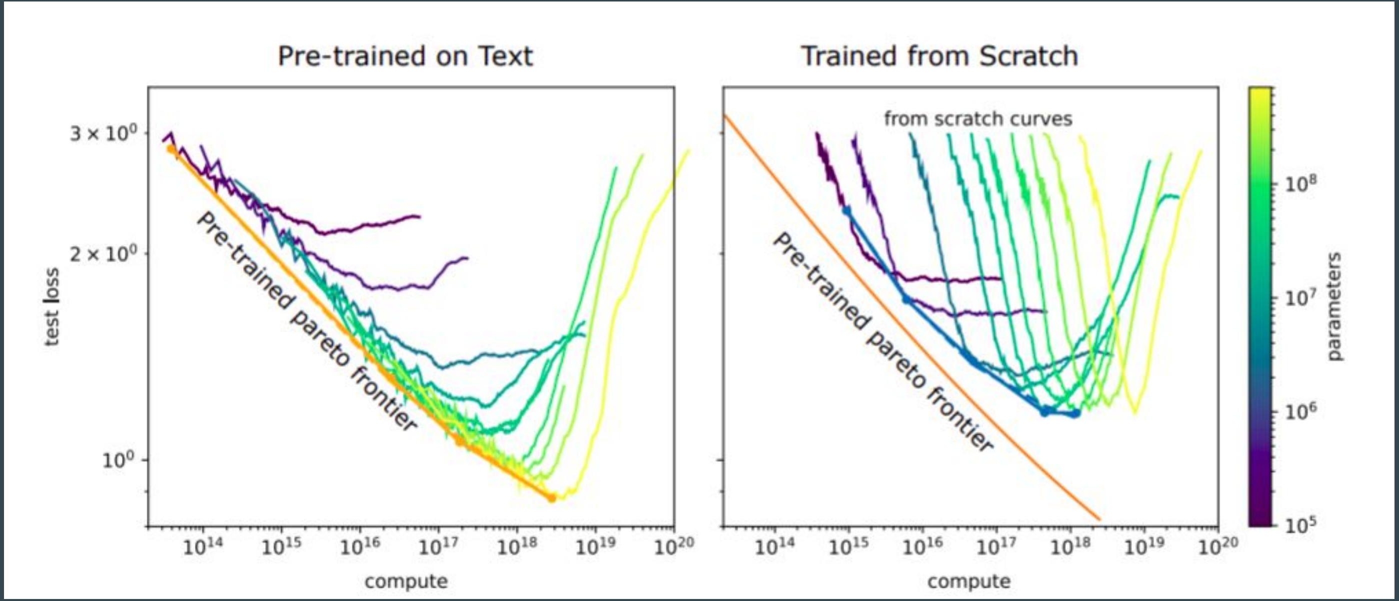
Limited data and increasing model size

Zero-shot performance

Steeper line with large dataset

Experiment 1: Observations

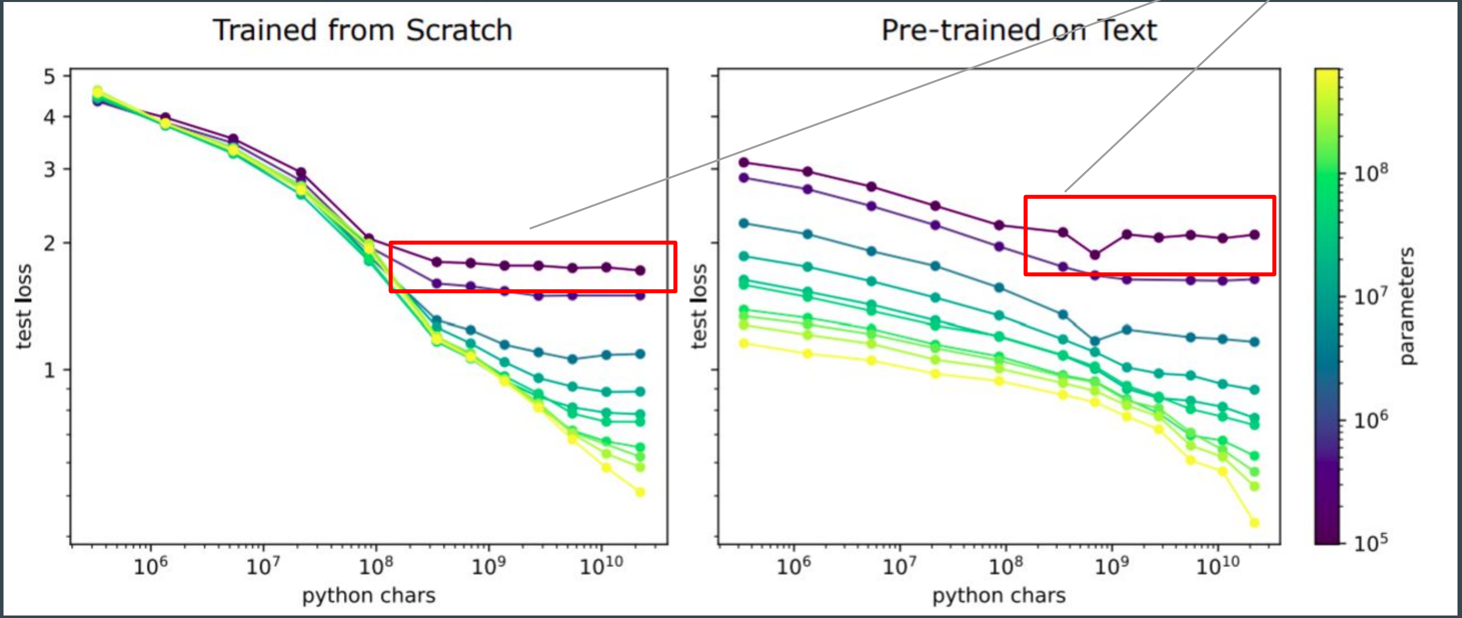
6. Fine-tuned models are more compute efficient in the low data regime (ignoring pre-training).



CAN PRE-TRAINING HARM PERFORMANCE?

OSSIFICATION

Pretraining harms the model performance in high data regime

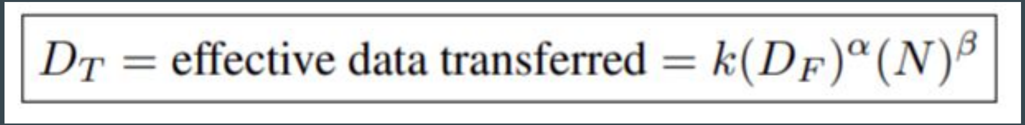


Why does ossification occur?

For a 1M parameter model trained in high data regime (>1e8) ⇒ model trained from scratch performs better

Experiment 1: Discussion

1. Potential applications of these scaling laws



$$\boxed{D_T = \text{effective data transferred} = k(D_F)^{\alpha}(N)^{\beta}}$$

- Whether to gather more data for fine-tuning?
  - Collecting more data is expensive
  - Use power law ⇒ cheap experiment to check

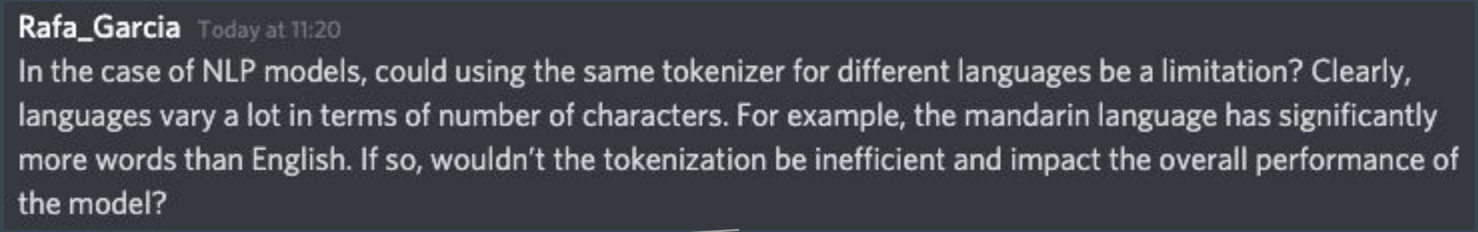
Experiment 1: Discussion

2. Distance between pretrained and fine-tuned distribution

- Dissimilarity of English and Python is representative of transfer between distant distributions.
  - There is English within python code.
- **Closer distributions** like *English:French* and **farther distributions** like *English:Math -* Interest for the future.

Experiment 1: Limitations

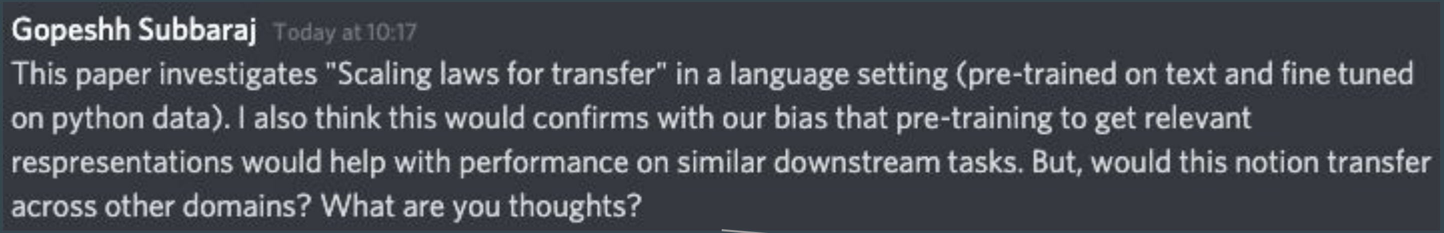
1. Models weren't tuned for fine-tuning -> leveraged hyperparameters from "Scaling laws for Neural Language Models".
2. Unclear if the power law fit would be observed for a **broad set of distribution pairs**.
3. Transfer between distributions was measured in an unsupervised setting.
   - Unclear as to what degree the findings would generalize to a supervised or RL setup.
4. Performance was only measured on transformers.



Rafa_Garcia Today at 11:20

In the case of NLP models, could using the same tokenizer for different languages be a limitation? Clearly, languages vary a lot in terms of number of characters. For example, the mandarin language has significantly more words than English. If so, wouldn't the tokenization be inefficient and impact the overall performance of the model?

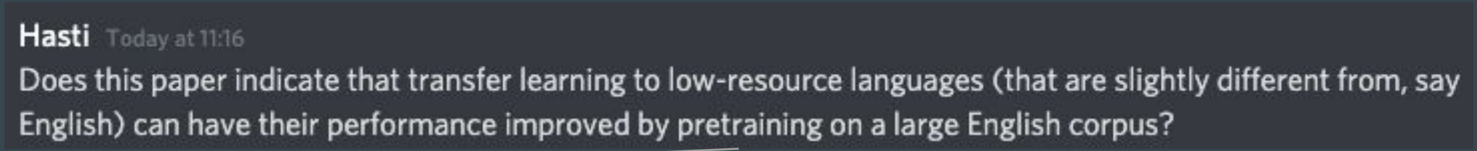
Yes! Its a limitation



Gopeshh Subbaraj Today at 10:17

This paper investigates "Scaling laws for transfer" in a language setting (pre-trained on text and fine tuned on python data). I also think this would confirms with our bias that pre-training to get relevant respresentations would help with performance on similar downstream tasks. But, would this notion transfer across other domains? What are you thoughts?

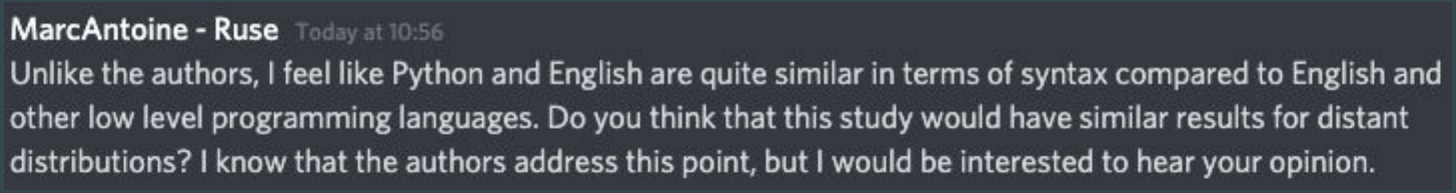
It should



Hasti Today at 11:16

Does this paper indicate that transfer learning to low-resource languages (that are slightly different from, say English) can have their performance improved by pretraining on a large English corpus?

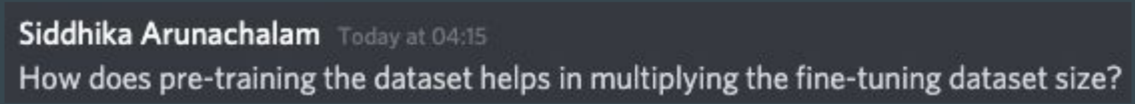
Potentially, yes!



MarcAntoine - Ruse Today at 10:56

Unlike the authors, I feel like Python and English are quite similar in terms of syntax compared to English and other low level programming languages. Do you think that this study would have similar results for distant distributions? I know that the authors address this point, but I would be interested to hear your opinion.

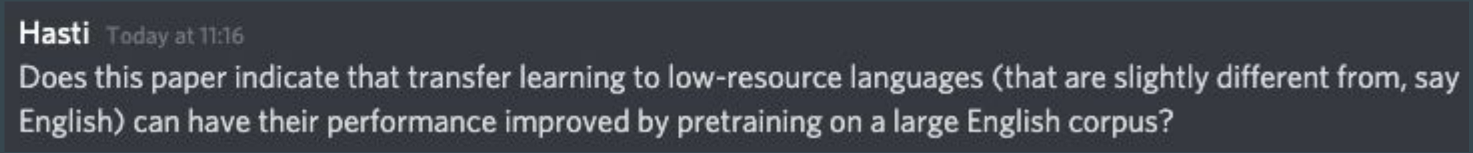
English is used to communication of information between humans and Python is used to transfer instructions to a computer. So, their purpose is different.



Siddhika Arunachalam Today at 04:15

How does pre-training the dataset helps in multiplying the fine-tuning dataset size?

You can get a good performance using a small finetuning dataset which is equivalent to training from scratch using a large dataset. Intuitively speaking, the target dataset has been multiplied.



Hasti Today at 11:16

Does this paper indicate that transfer learning to low-resource languages (that are slightly different from, say English) can have their performance improved by pretraining on a large English corpus?

It will be discussed in more detail soon, but the short answer is that the amount of effective transfer depends on linguistic similarity between the two languages (English and Chinese vs English and German).

Experiment 2: Transfer from English to other human languages

- Another experiment which explores the scaling laws for transfer.
- Explores and discusses scaling laws discovered **while fine-tuning across different languages** with pre-trained English language models.

Christina Kim- https://christina.kim/2021/04/11/scaling-laws-for-language-transfer-learning/

Experiment 2: Training setup

Dataset

OpenWebText2 was used to pretrain the model which is the open sourced version of WebText created by EleutherAI.

For German and Spanish, OSCAR (Open Super-large Crawled Aggregated coRpus) was used which is a multilingual corpus collected from Common Crawl.

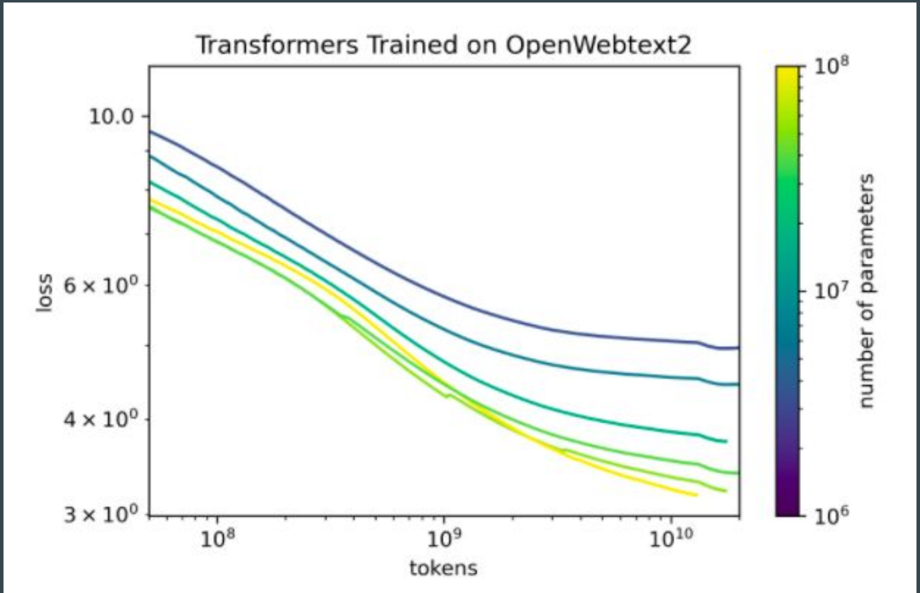
For Chinese, Community QA was used (details not given).

Model

- Transformer with non embedding parameters ranging from 3.3M to 124M.
- Used GPT-2's Byte level Byte Pair Encoding

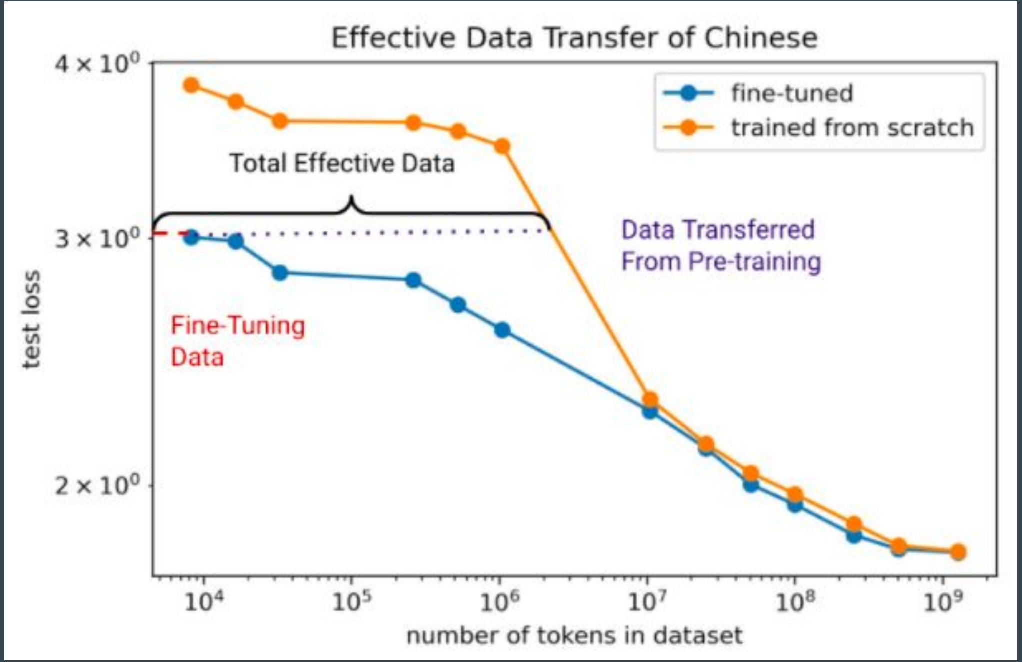
Experiment 2: Pretraining

- The loss for the transformer trained on OpenWebText2 decrease as the number of tokens and model parameters increase
- The decrease is not linear which is because the larger models were undertrained and hyperparameters were not tuned properly.



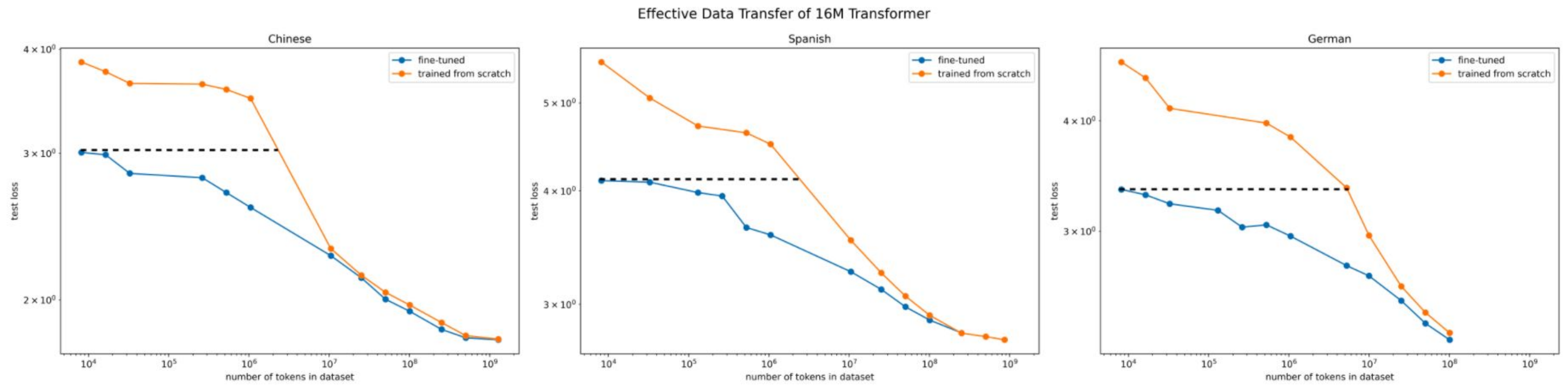
Experiment 2: Effective data transfer

- Similar to original scaling laws paper, D_T (Effective Data Transferred) = D_E-D_F
- The curves are not smooth/linear because larger models are undertrained.



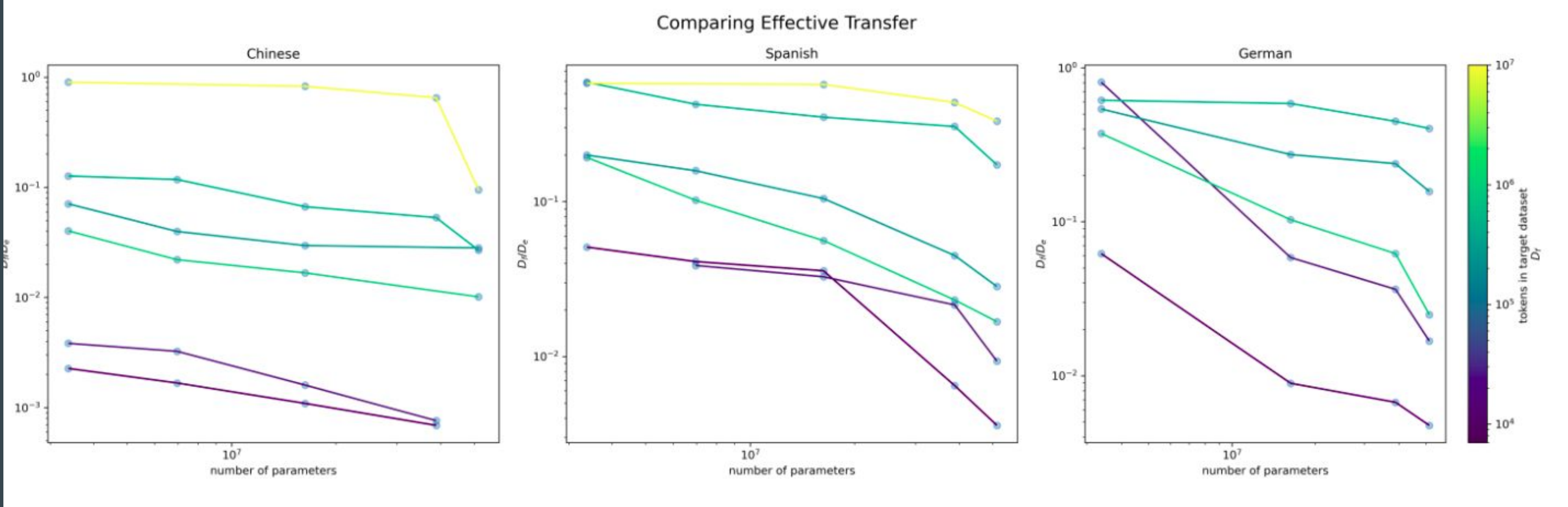
Experiment 2: Results

- Effective data transfer from English to Chinese, Spanish and German.
- Lowest Effective transfer for Chinese. Pretraining helps for one magnitude less(10^7 vs 10^8).
- Highest Effective Transfer for German because of high linguistic similarity (Germanic Languages).
- Less effective transfer for Spanish (Romance Language).



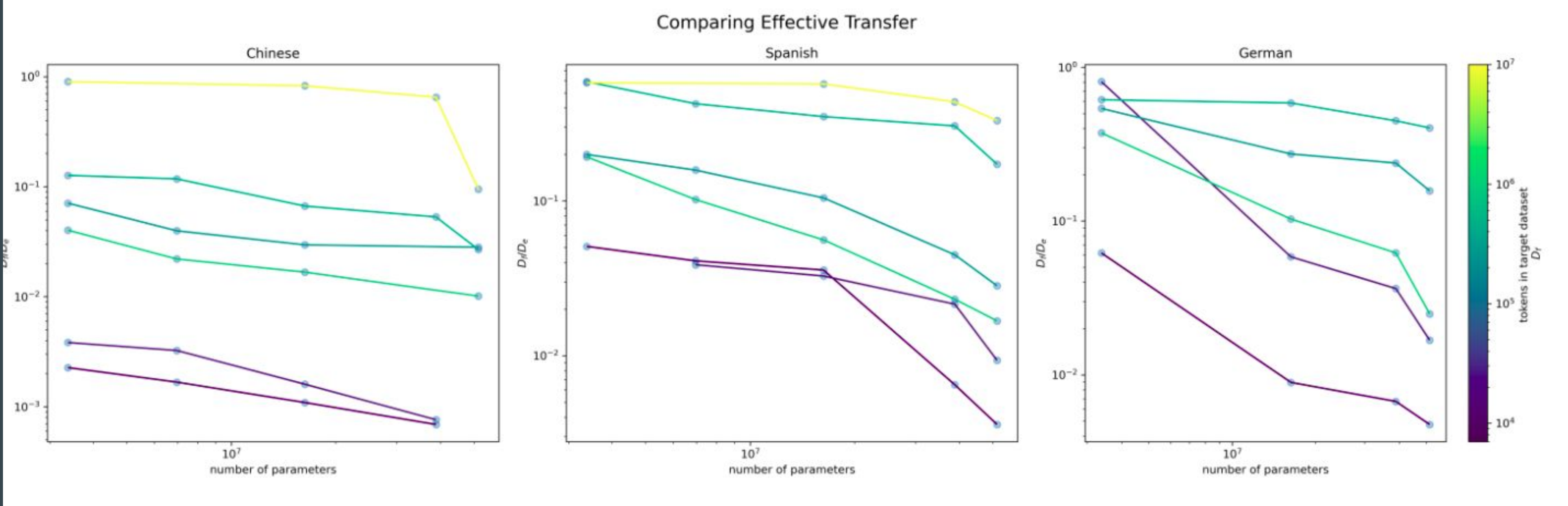
Experiment 2: Results

- Small d_f / d_e is better.
- d_f/d_e increases as finetuning dataset size increases.



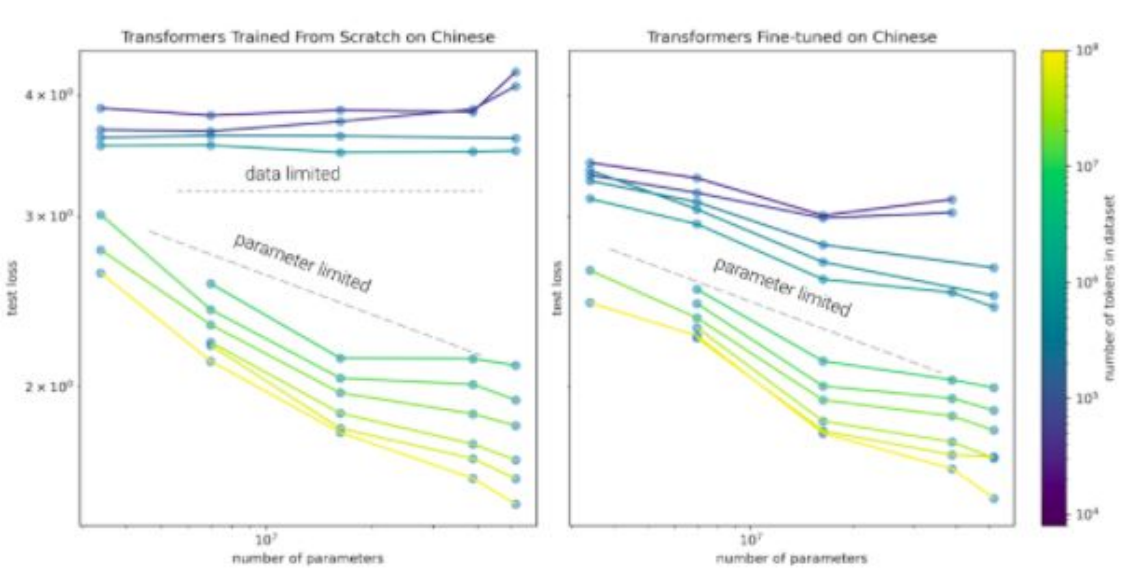
Experiment 2: Results

- d_f/d_e decreases as model size increases
- Slopes of the curve for german is higher than other language due to higher effective transfer.
- Some of the curves for spanish and german are inconsistent because the training is not done properly.



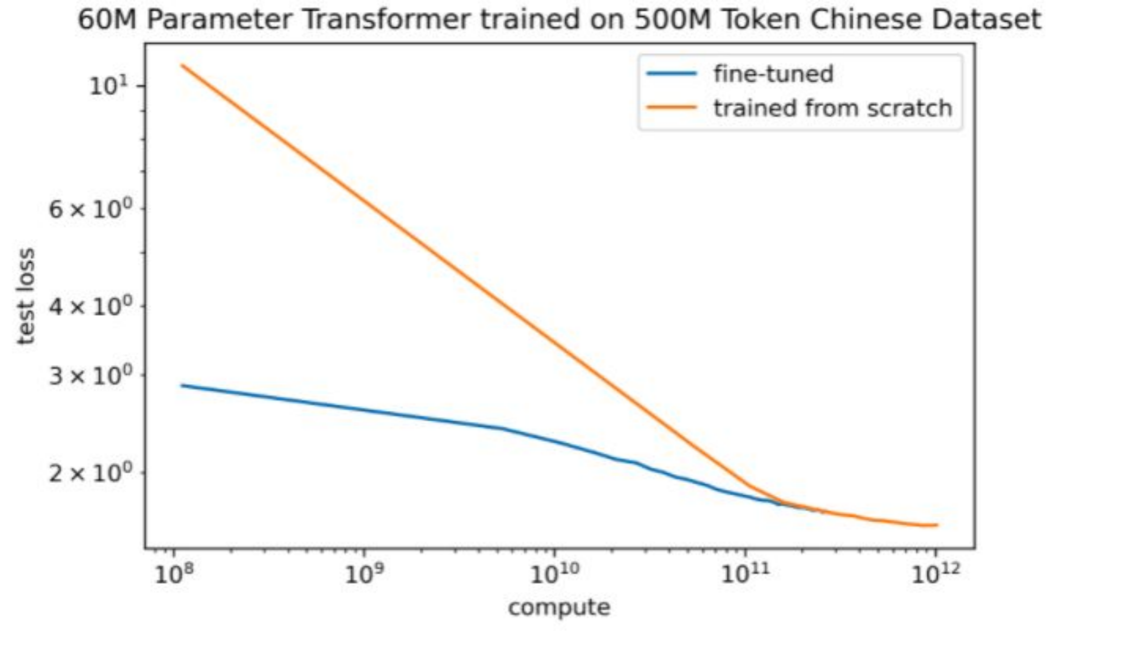
Experiment 2: Results

- Similar to the results of the previous experiment.
- Model trained from scratch on low data regime is data limited, not the case for finetuned model.
- In high data regime, both curves are similar.



Experiment 2: Results

Finetuned model is more compute efficient in low compute regime.



Experiment 2: Limitations and Future Works

Limitations

- Same tokenizer used for all languages. Most problematic for Chinese.
- Larger models are undertrained.
- Better hyperparameter tuning can be performed.

Future Works

- Finetuning the models for low resource languages.
- Finding the scaling law equations and comparing those for different languages.
- Finding a good ratio for finetuning and pretraining for different languages for a given budget.

Conclusion

- Scaling Laws for transfer have been studied for transfer from English Language to Python code and English Language to Different Languages.
- Linguistic Similarity had a huge impact in language transfer. Hence, it is necessary to make sure that the source and target datasets have some amount of similarity.
- It will be interesting to see results for transfer in different domains like image and video.