Scaling Cross-lingual Transfer via Continual Pre-training

Anonymous ACL submission

Abstract

Large Language models(LLMs) have developed rapidly in recent years, and are remarked as a brand new milestone in the information age. However, powerful LLMs in vogue are predominantly mainstream language speakers, especially in English, and the desire for native LLMs remains strong. Inspired by these demands, we examine the scaling of multilingual models, focusing on the interplay between language-specific computational requirements and universal scaling laws. Our findings demonstrate that continual pre-training of an other-language model on an English base effectively maintains English proficiency while improving other-language performance, challenging traditional notions of cross-lingual transfer, which is commonly equated with fine-tuning. We propose a strategic approach for efficient multilingual training, emphasizing the balance between computational resource allocation and avoiding catastrophic forgetting. Our work helps to understand language-independent model scaling behaviors and transform "outsiders" into "locals" with basic capacities mostly preserved.

1 Introduction

In recent years, Large Language Models(LLMs) based on the GPT (Radford et al., 2018, 2019; Brown et al., 2020; Achiam et al., 2023) architecture, known as decoder-only architecture, have gained ground and gradually become essential infrastructure in our daily lives in the information age. They hold great potential for realizing the grand vision of Artificial General Intelligence (AGI). During the dramatic evolution of the modern Natural Language Process(NLP), a growing number of real-world tasks can be surrogated by artificial intelligence, which brings considerable convenience. Hence, in order to enhance productivity and fully utilize existing resources, the demand for LLMs is becoming increasingly intense. However, according to Ethnologue, despite over 7000 living languages worldwide, most speech and language technologies are concentrated on only a tiny subset of them (ACL, 2021; Balachandran, 2023; Ojo et al., 2023), particularly those based on predominantly mainstream language, such as English.

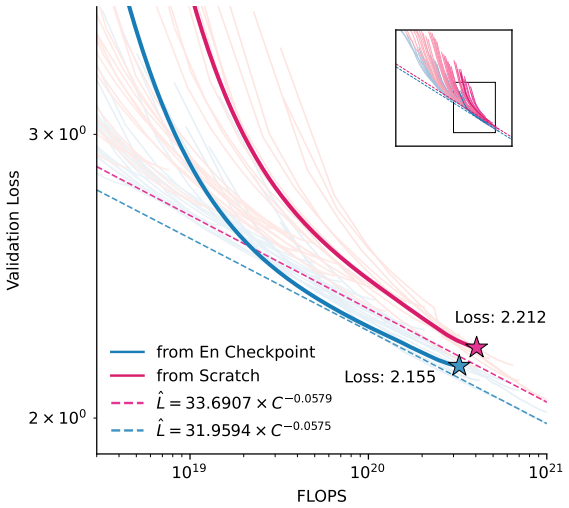

Figure 1: Scaling Behaviour Comparison between Pre-training and Continual Pre-training on a New Language. Models with 2B parameters are highlighted.

To achieve satisfaction with localized LLMs for non-dominant languages, related studies have been conducted from different perspectives and in diverse fields. There are two main approaches: before and after pre-training. Cross-lingual Transfer following pre-training tends to conform to an Encoder-Decoder architecture and translation tasks, as parallel corpus and shared embeddings are commonly used. However these measures were viewed lack of effectiveness during our experiments, and conventional transfer learning via fine-tuning appears sluggish due to inadequacy of slight data for transfer from various languages instead of analogous downstream tasks. Consequently, we extend it to continual pre-training. As for approaches before pre-training, previous works fed carefully processed mixing data from numerous target languages into

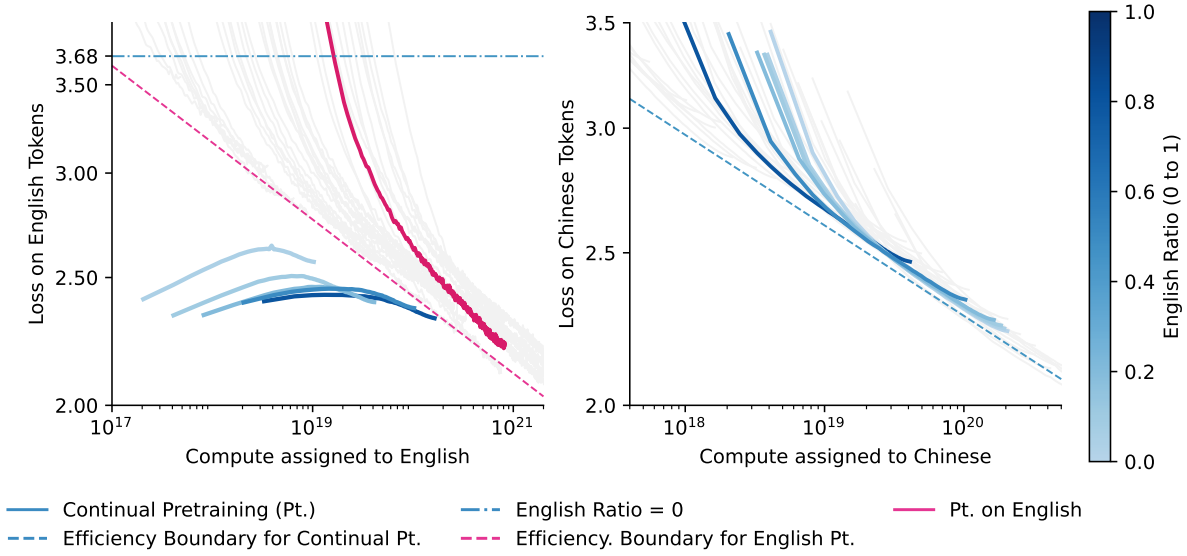

Figure 2: Scaling Behaviour of Continual Pre-training when Mixing Original Distributions at Different Ratios. Loss in different languages is obtained by evaluating the same 1.4B models on two validation sets of English and Chinese during continual pre-training.

LLMs for universal multilingual language models (Workshop et al., 2022), or provided data with elaborated repetitions to LLMs following certain Scaling laws for monolingual native-speaker models trained from scratch (Muennighoff et al., 2023). While a universal multilingual language model seems to be an "almighty formula", performance regarding lesser-represented languages fails to meet practical expectations for natives, and it's too early for "universalization" nowadays. The latter work is similar to our work in some ways. Nevertheless, we take into account the limitations of previous scaling law studies and current powerful English LLMs together and hold the opinion that since the pre-trained models are far from optimal convergence, we can naturally attempt to transform high-performance models already in place to native speakers without from the outset. Thus, we concentrate on figuring out the scaling law of continual pre-training in the field of cross-lingual transfer and ensure training settings comply with it. Building upon this, we demonstrate empirically that better-performing local models can be derived from mature mainstream models possessing a more rapid convergence rate throughout training, in terms of basic capacities such as math and logic abilities. Additionally, our method proves beneficial in scenarios with limited source data as a result of the power law trend, which means we can anticipate an expected loss curve to reasonably adjust computational resource allocation and strategies according to the total budget. To avoid the dilemma of catastrophic forgetting, akin to multi-task learning, we also mix a small proportion of data from the origin language into the continual pre-training process and identify the optimal balance between preserving proficiency in the original language and expanding capabilities in other languages.

In summary, the main contributions of our work are as follows:

- We demonstrate the scaling law that exists in cross-lingual transfer continual pre-training.
- We figure out the form of the power law formulation of the scaling law and the hyperparameters involved, as well as the explanation relevant to the impacting factors.
- We provide effective strategies to support the continual pre-training within the process of cross-lingual transfer.

Above all, our work paves the way for ensuring open source LLMs are more "open" to broader regions and countries, considering not the most mainstream but equally invaluable languages all over the world. That's the point why we are committed to transferring under the cross-lingual scenery.

Model	E	A	B	α	β	γ
Chinchilla (MassiveText)	1.69	406.4	410.7	0.34	0.28	-
Ours (from Scratch)	1.55	420.0	719.5	0.40	0.30	-
Ours (Continual Pre-training)	1.55	420.0	433.3	0.40	0.20	0.08

Table 1: Comparison of parameter fitting results among Chinchilla (Hoffmann et al., 2022), our model trained from scratch, and our model with continual pre-training.

2 Method

2.1 Scaling Law for Pre-training from Scratch

We conducted continual pre-training in other languages based on a basic English model, using a large-scale training of Chinese models as an example to investigate the scaling law of this continuous training process. By comparing with training from scratch under the same settings and data, we explored the optimal computational allocation and returns for both approaches and discussed their differences.

First, let's review the Parametric Fit method used by Hoffmann et al. (2022), where they derived and fit a formula for the loss. They decompose the loss $L(N, D)$ into:

$$\begin{aligned} L(N, D) &\triangleq L(\hat{f}_{N,D}) \\ &= L(f^*) + \Big(L(\hat{f}_N) - L(f^*)\Big) \\ &\quad + \Big(L(\bar{f}_{N,D}) - L(\hat{f}_N)\Big) \\ &= E + \frac{A}{N^\alpha} + \frac{B}{D^\beta} \end{aligned} \tag{1}$$

Here, N represents the parameters, D represents the training tokens. E, A, B, α, β are learned variables.

In the Equation 1, f^* represents the optimal Bayesian classifier, $\hat{f}_N$ denotes the optimal transformer model under the constraint of parameters N, $\bar{f}_{N,D}$ represents the outcome obtained through gradient descent under the constraints of parameters N and training tokens D in the experiments. The loss includes three parts:the Bayes risk, which is the smallest possible loss for predicting the next token based on the full distribution P, also known as the "entropy of natural text", a term related to how well the function approximates based on the hypothesis space size, and a stochastic approximation term.

Accordingly, the formula they proposed for the optimal allocation of computational resources C with respect to N (model size) and D (dataset size) is:

$$\begin{aligned} N_{\text{opt}}(C) &= G\left(\frac{C}{6}\right)^a \\ D_{\text{opt}}(C) &= G^{-1}\left(\frac{C}{6}\right)^b \end{aligned} \tag{2}$$

$$\text{where} \quad G = \left(\frac{\alpha A}{\beta B}\right)^{\frac{1}{\alpha+\beta}},$$

$$\text{and} \quad a = \frac{\beta}{\alpha+\beta}, \quad b = \frac{\alpha}{\alpha+\beta}$$

2.2 Scaling Law for Continual Pre-training

Our goal is to modify the Equation 1 to fit the scenario of continual pre-training. Based on the meaning of the decomposition formula, we think while keeping the dataset and training process consistent, the first two terms, which respectively represent the natural language entropy and the function estimation capability under the hypothesis space size, should essentially remain consistent with those of pre-training from scratch.

Building on the premise of keeping the first two terms fixed, our study of continual pre-training involves separately controlling the variables D (dataset size) and N (model size) to fit the final stochastic approximation term. We discovered that both variables exhibit a power-law relationship. Therefore, inspired by the Equation 1 and also drawing inspiration from the concept of effective data transferred $= k(D_F)^\alpha(N)^\beta$ as proposed in the study by Hernandez et al. (2021), we believe the final error term in this scenario can be represented as $\frac{B'}{(D)^{\beta\prime}(N)^\gamma}$.

Consequently, the loss formula for continual pre-training is:

$$\hat{L}(N, D) \equiv E + \frac{A}{N^\alpha} + \frac{B'}{(D)^{\beta\prime}(N)^\gamma} \tag{3}$$

Here, E, A, α are the same parameters inherited from Equation 1, which applies to training from scratch under the same dataset conditions.B, β', γ are newly learned variables.

Model	Coeff. a where $N_{\text{opt}} \propto C^a$	Coeff. b where $D_{\text{opt}} \propto C^b$
Chinchilla (MassiveText)	0.46	0.54
Ours (from Scratch)	0.429	0.571
Ours (Continual Pre-training)	0.385	0.615

Table 2: Comparison of optimization coefficients among Chinchilla (Hoffmann et al., 2022), our model trained from scratch, and our model with continual pre-training.

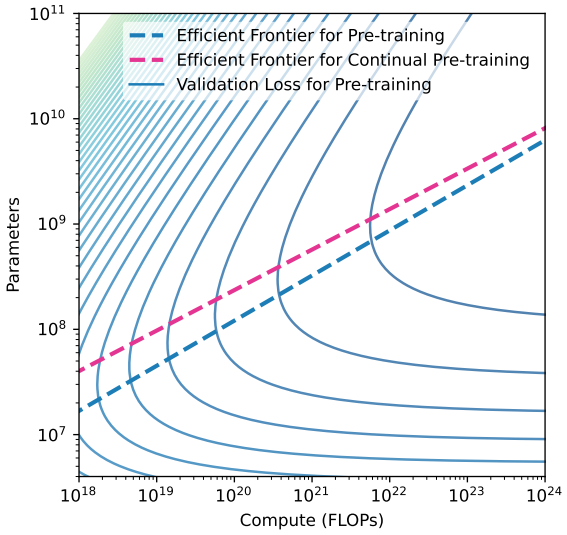

Figure 3: Predicted Compute-optimal Efficient frontier on Iso-loss Contour.

Similarly, based on the constraint $C = 6ND$ (Kaplan et al., 2020), the minimum value of the modified formula can be found by minimizing the loss function, which also conforms to the power-law relationship with respect to C:

$$N_{\text{opt}}(C) = G\left(\frac{C}{6}\right)^{a}$$
$$D_{\text{opt}}(C) = G^{-1}\left(\frac{C}{6}\right)^{b} \tag{4}$$

$$\text{where} \quad G = \left(\frac{\alpha A}{(\beta' - \gamma)B'}\right)^{\frac{1}{\alpha+\beta'-\gamma}},$$
$$\text{and} \quad a = \frac{\beta'}{\alpha+\beta'-\gamma}, \quad b = \frac{\alpha-\gamma}{\alpha+\beta'-\gamma}$$

2.3 Parametric Fit

Following Hoffmann et al. (2022), we also minimize the Huber loss (Huber, 1992) between the predicted and observed log loss, with δ set to 10^{-3}.

First, we need to optimize the following problem using data from training from scratch:

$$\min_{a,b,e,\alpha,\beta} \sum_{\text{Run } i} \text{Huber}_\delta\left(\text{LSE}(a - \alpha \log N_i, b - \beta \log D_i, e) - \log L_i\right) \tag{5}$$

Next, based on the value of a, α, e obtained from the results, we substitute it and optimize the following continual pre-training problem:

$$\min_{b',\beta',\gamma} \sum_{\text{Run } i} \text{Huber}_\delta\left(\text{LSE}(a - \alpha \log N_i, b' - \beta' \log D_i - \gamma \log N_i, e) - \log L_i\right) \tag{6}$$

where LSE is the log-sum-exp operator. and we can set $A, B, E, B' = exp(a), exp(b), exp(e), exp(b')$.

During the fitting process, we utilized the Optuna library for hyperparameter search iterations and employed the L-BFGS (Nocedal, 1980) algorithm for optimal local search, resulting in the identification of the best hyperparameters. The final fitted parameter results are shown in Table 1.

Based on the parameter fitting results, we can calculate the optimization allocation coefficients, as shown in Table 2. The efficient frontier results for both training from scratch and continual pre-training are shown in Figure 3.

3 Experiments

3.1 Pre-training Setup

We utilize a decode-only transformer language model based on the LLaMA2 architecture(Touvron et al., 2023b) for our experiments, encompassing a variety of model sizes. This range includes 40 distinct model types, with capacities varying from 4 million to 5 billion parameters. Adhering to the principles outlined in (Hoffmann et al., 2022) and aligning with the typical configurations of open-source language models, we set the token count to be 20 times the model size for base English models. Our approach employs cosine learning rate schedules that feature a decay factor of 10× the initial learning rate.

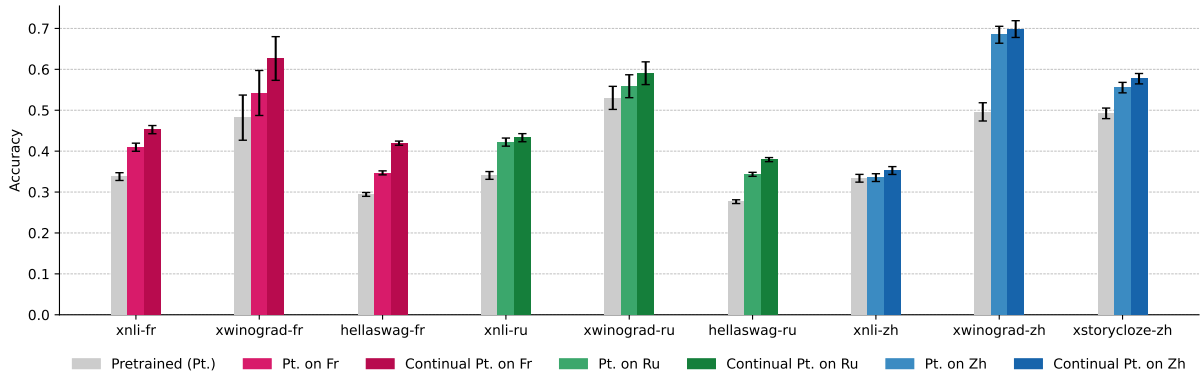

Figure 4: Benchmark Performance Comparison of Pre-trained and Continually Pre-trained Language Models. Continually pre-trained models of different languages are continuations of the same checkpoint (colored in gray)

For the continued training of the model, we adopted the re-warmup technique according to (Gupta et al., 2023), which involves repeating the same warmup process during continual pre-training, and conducted experiments with 5 $\times$ and 20 $\times$ the length of training by observing changes in the loss. This approach was adopted to achieve more precise optimization calculations. Following previous experiences and the LLaMA settings, we set our learning rate at $2 * 10^{-4}$). We also adopted different batch sizes for models of varying sizes: for models smaller than 1 billion parameters, we used a batch size of 512; for those between 1 and 2.5 billion parameters, we used 1024; and for models larger than 2.5 billion parameters, we used a batch size of 2048.

Our English training data primarily originates from the Redpajama dataset(Computer, 2023), while the Chinese training data was sourced privately and has undergone filtering and deduplication processes. For other languages, the data is mainly derived from the mC4 dataset(Raffel et al., 2019). We randomly partition this data into training and validation sets. Our training principle is to ensure that when using larger datasets, they encompass the smaller ones, thus maintaining a comprehensive and inclusive approach to data coverage across different languages.

3.1.1 Evaluation Benchmarks

We evaluate the cross-lingual language model on some widely adopted multi-lingual benchmarks. Specifically, we use splits of French, Russian and Chinese from XNLI (Conneau et al., 2018), Winograde (Sakaguchi et al., 2019), Multi-lingual Hellaswag (Dac Lai et al., 2023) and XStorycloze (Lin et al., 2021) to evaluate language understanding and commonsense reasoning ability of continual pre-trained model for above three languages. To analyze the impact of mixing English at the different ratios for training, we additionally evaluate the above models on XCopa (Ponti et al., 2020) and PiQA (Bisk et al., 2019).

3.2 Results

We perform extensive parametric fitting over training across various parameters and training settings to qualitatively model the scaling properties of cross-lingual pre-training. To avoid biases caused by language contamination in the validation set, we also evaluated models on several widely used benchmarks for language modeling.

3.2.1 Comparative Analysis on Scaling Behaviours

We optimize models of different sizes on the Chinese data splits from both scratch and existing English Checkpoints. For all different runs, we do parametric fits for checkpoints with the lowest validation loss, adhering to the assumption of a power-law relationship between the validation loss and raw compute to train the model, total training data or model parameters.

Scaling by Parameters and Data We compare our estimated parameters with previous works in Table 1. It can be seen that in continual pre-training, we have $B = 433$ and $r = 0.08$, which indicates the cross-lingual transfer effect scales with model parameters.

Scaling by Compute The training curves for continual pre-training are shown in Figure 1, where the fittes relationship between validation loss and compute is visualized. It's worth noting that not

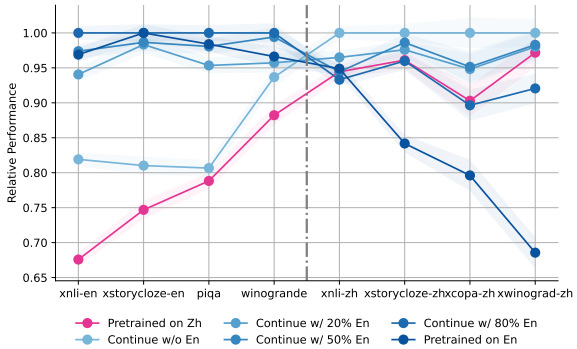

Figure 5: Model performance on English and Chinese benchmarks at different English data mix ratios with 1.4B parameters trained. Relative Performance refers to accuracy relative to the highest accuracy achieved across different training settings with 1.4B parameters.

all checkpoints contribute to the predicted scaling relationship as the lowest loss value of some checkpoints is higher than models of other parameters at the same compute budget and therefore is not counted.

It can be seen that pre-training language models from existing checkpoints yield lower final loss consistently across different parameter numbers. Results of parametric fit indicate that the advantages of lower loss are generally uniform over each unit of compute devoted, which can be supported by the significantly lower scaling factor of the computing term (33.6907 to 31.9594) and nearly unchanged exponent (-0.0579 to -0.0575).

3.2.2 Downstream Performance of Pre-trained checkpoints

We further evaluate the impact of cross-lingual continual pre-training on several widely used benchmarks. On 1.4B parameters, we continually trained language models of language French (Fr.), Russian (Ru.), and Chinese (Zh) from the same English checkpoint. We then evaluate these models to compare them to their equivalent model trained from scratch as well as the English checkpoints. The results are shown in Figure 5.

We evaluate models of different languages on their corresponding language split of multi-lingual benchmarks to provide a fair comparison. The results showed that models of all three different languages gain improved language ability to various extents compared to the original pre-trained model. It can also be seen that continual pre-training universally helps with benchmark performance under different languages and different scenarios.

Interestingly, French models benefit most from continual pre-training, which can be explained from the perspective of language similarity. French share lots of common words and grammar structures with English, hence the cross-lingual transferring is more obvious than Russian and Chinese, which on the other hand share less similarities with English.

3.2.3 Compute-optimal Scaling for Cross-lingual Transfering

Estimated Parameters under FLOPS budget According to the theoretical model established in Section 2, we can solve the optimal trade-off between model parameters and data to be trained under a compute-constraint scenario, which is a practical problem under modern large-scale training. More specifically, according to Equation 2, when only a certain amount of compute power (FLOPS) is allowed, we fit optimal training data and model parameters for continual pre-training to be:

$$\hat{N}_{opt}(C) = 0.324C^{0.429}, \hat{D}_{opt}(C) = 0.514C^{0.571} \tag{7}$$

Comparatively, for pre-training from scratch:

$$N_{opt}(C) = 4.79C^{0.385}, D_{opt}(C) = 0.035C^{0.615} \tag{8}$$

Visualization for the efficient frontier of model parameter N regarding computes over the Iso-loss contour is shown in Figure 3. We find that the optimal parameters for continual pre-training are slightly deviated from pre-training from scratch, favoring fewer computes for the same model sizes. This aligns with the nature of cross-lingual transfer learning, where the model of continuation is "premature" due to prior knowledge acquired in the base language.

Optimal English Mixing Ratio We also investigate methods to prevent catastrophic forgetting of original distributions during continual pre-training in another language. At 1.4B parameters, we continually train several models with mixed training corpus by mixing various ratios of training data from the base pre-training into the continual pre-training processes.

We visualize the training curves of English-mixing models in Figure 2. Notably, for validation

loss in English and Chinese, the compute is calculated as the computer devoted to tokens of the corresponding language.

Figure 2 reveals that the mixed data of the original language and the continually trained language behave differently in terms of scaling behavior. As shown in the right of Figure 2, mixing different ratios of original data only affects the early stage of training. Models converge to the same validation loss when the same amount of computing is involved, despite they are mixed with various ratios of original data, ranging from 1% up to 80%. This suggested that a higher level of mixing original data is welcomed as mixing does not hinder scaling property while preserving the model's performance on original distribution.

The left of Figure 2 compares the relationship between compute and validation loss on the original distribution throughout continual pre-training, which can be viewed as the "Scaling law of forgetting". Interestingly, the scaling behavior depicts a power-law relationship resembling the one during pre-training from scratch. Validation losses of models at different English mixing ratios increase at the early stage of training and then decline, maintaining lower than the original pre-trained run. It suggests that a large amount of original knowledge is preserved throughout the continual training, even at a very low English mixing ratio (1%).

To further analyze impacts brought by mixing original data in continual pre-training, we evaluate the model performance on English and Chinese benchmarks at different English data mix ratios in Figure 5. It shows that while pre-training purely on one language yields sub-optimal results on the other language, any non-skewed English ratio can effectively preserve performance on the original distributions. In practice, we find that around 30% of original data could be capable of keeping the validation loss lower than the start of continual pre-training.

4 Related Work

4.1 Scaling Law

Scaling Law suggests the predictable influences that salient factors (e.g., dataset specifications, number of parameters, batch size, etc.) exert on model performance through scaling behavior across multiple orders of magnitude.

Building upon prior research before the extensive usage of attention mechanisms (Hestness et al., 2017, 2019), the modern perspective on Scaling Law has identified power-law scaling of test loss in relation to training data size as well as model size (Rosenfeld et al., 2019). The study of Scaling Law gradually deepened, becoming more theoretical and empirical, which included learning curvey theory (Hutter, 2021), explanation of Scaling Law (Bahri et al., 2021), and Elucidation from the dimension of the data manifold (Sharma and Kaplan, 2020). Until 2021, our understanding of scaling laws has become clearer. In particular, Kaplan et al. (2020) directed attention toward larger-scale models, providing the general form with variables taken into account to the maximum extent. The apparent general applicability of the scaling law theory has been successively verified through numerous experimental results (Henighan et al., 2020). However, the generality previously identified may not align with practical applications. For instance, Hoffmann et al. (2022) revealed the sub-optimality of models adhering to the recommended ratio of model size to the dataset. And the debate sparked further consideration towards modifying space, more experimentally validated hyperparameter choices have emerged concurrently, such as the possible influence intrinsic to model architecture (Tay et al., 2022; Frantar et al., 2023), value of repeated tokens (Hernandez et al., 2022; Muennighoff et al., 2023)(contrast to utilize once in previous works) double descent & the delayed phenomena during scaling behavior (Caballero et al., 2022), scaling law for forgetting situation when exerting PEFT rather than learning (Kalajdzievski, 2024), Incorporation the inference cost into the formula (Sardana and Frankle, 2023) and expolate to other modals (Alabdulmohsin et al., 2023; Aghajanyan et al., 2023) or specific domains (Rang et al., 2023; Zhang et al., 2023a; Wu et al., 2024) .

4.2 Transfer Learning

Transfer Learning is dedicated to enhancing generalization during the transition from one data distribution to another similar distribution, provided with an appropriate volume of domain-specific data. This facilitates downstream tasks in adapting to such transitions and achieving optimal performance, sidestepping the need to relearn knowledge from the ground up (Pan and Yang, 2009; Zhuang et al., 2020).

While conventional approaches involved fine-tuning models on a relatively small scale, yielding

satisfactory results (Brown et al., 2020), current research indicates that additional pre-training, specifically Continual Pre-training, involving adapting a Language Model (LM) to the target domain using a corresponding corpus, can significantly enhance end-task performance (Gururangan et al., 2020; Xu et al., 2019; Sun et al., 2020). This is crucial not only due to the necessity for generation pattern models to learn but also because the substantial knowledge contained in the original corpus requires digestion (Ke et al., 2023, 2022). Additionally, continual pre-training mitigates catastrophic forgetting (Kirkpatrick et al., 2017; Cossu et al., 2022). Positioned as an intermediate stage between the base pre-trained model and the fine-tuning operation (Zhang et al., 2023b), Continual Pre-training is analogous to equipping base models with incremental knowledge oriented towards the target domain. Regarding Scaling Law, Hernandez et al. (2021) empirically studied scaling laws for the effective transfer of data when fine-tuning instead of pre-training. This study involved data following natural language distribution and Python code distribution, concluding the validation of predicting cross-entropy loss during scaling behavior in terms of parameters, data, and compute.

4.3 Multi-lingual Language Model

During the era of the encoder-decoder architecture (Vaswani et al., 2017), substantial efforts were undertaken, primarily focusing on optimization in various aspects, including data blending (Patil et al., 2023; Srivastava and Singh, 2021; Choudhury et al., 2019), parallel corpus (Kimera et al., 2023; Dabre et al., 2021; Azunre et al., 2021; Rabinovich et al., 2018), and shared embeddings (Liu et al., 2020; Takase and Kobayashi, 2020; Jones et al., 2021; Wu and Monz, 2023). With the emergence of the GPT (Radford et al., 2018, 2019; Brown et al., 2020; Achiam et al., 2023) accompanied by In-Context Learning (ICL) (Brown et al., 2020; Dong et al., 2022), the utilization of fine-tuning and ICL played a crucial role (Tanwar et al., 2023) in terms of multilingual language models. The currently popular universal multilingual language models, particularly BLOOM (Workshop et al., 2022), exemplify a highly representative category. These models achieve many-to-one transfer by undergoing pre-training on hundreds of languages followed by fine-tuning on the target language, a methodology referred to as multilingual learning (Lai et al., 2023). However, they underscore the persistent demand for native speakers. In response to this, cross-lingual transfer learning (Ostendorff and Rehm, 2023), which emphasizes one-to-one transfer, has addressed this limitation, facilitating a more effective transition from potent English language models to models in indigenous tongues. Notable LLMs, including LLaMA (Touvron et al., 2023a,b), Falcon (Penedo et al., 2023), and OPT (Zhang et al., 2022), have demonstrated remarkable capabilities across numerous tasks. Therefore, cross-lingual transfer holds the promise of being a blueprint for the momentum of rapid development in LLMs shared across all languages.

Previous works discussing scaling laws tended to focus on data utilization efficiency (Muennighoff et al., 2023) or were limited to considerations of cross-task transfer (Hernandez et al., 2021). However, the impact on the efficiency of cross-lingual transfer lacks thorough investigation in that variations in both expression and knowledge between natural language and code are noteworthy and disparate languages only differ in expression, with unchanged underlying knowledge. Simultaneously, though Scaling Laws for forgetting captured attention (Kalajdzievski, 2024), the LoRA (Hu et al., 2021) method used is lightweight compared to ours. Thus, the focal point of our work is to elucidate the Scaling Law in the field of cross-lingual transfer utilizing continual pre-training for competitive performance.

5 Conclusion

In this paper, we presents a comprehensive study on the scaling law of cross-lingual transfer through the lens of continual pre-training. We successfully demonstrate that continual pre-training not only preserves the model's original language capabilities but also significantly improves performance across various languages. The study uncovers that the strategic mixing of original training data and optimizing computational resources based on language-specific demands are crucial for enhancing cross-lingual transfer. Our work lays the groundwork for future explorations into effective multilingual model training strategies, highlighting the need for tailored approaches that consider the unique characteristics of each language while leveraging the benefits of large-scale language models.

Limitations

Language contamination In this study, we utilized publicly accessible datasets for pre-training. Although the Chinese dataset and mC4 dataset attempt to clean and create language-specific training splits, they cannot entirely prevent the contamination of English at a more granular level. This is particularly challenging due to the inherent nature of many languages, such as French, which often incorporate English words. To estimate the computational effort for different languages, we counted the number of samples processed in each language training split. This approach may be imprecise if the dataset contains a large amount of text in other languages. This issue highlights the need for future research to conduct a more in-depth analysis of the impact of language contamination in multilingual pre-training.

Hyper-parameter sensitivity In the training of models across various scales, we selected hyper-parameters based on experience and trial and error. Our preliminary results showed that deviating from optimal hyper-parameters can significantly harm model optimization and disrupt the scaling laws. To maintain consistency, we selected a constant learning rate, optimizer, learning rate scheduler, and batch size that matched the scale of the model for different experiments. This approach is in line with the conclusions of previous studies. Future research should explore the finding of optimal hyper-parameters from the perspective of language-specific scaling laws, which could lead to more effective pre-training configurations.

Limited Scale Due to computational limitations, many experiments were not covered, especially in cases where the training data was excessive or the model size was too large. This limitation means that our findings may have limited reference value for larger scaling-up scenarios.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

ACL. 2021. ACL 2022 Theme Track: "Language Diversity: From Low-Resource to Endangered Languages". https://www.2022.aclweb.org/post/acl-2022-theme-track-language\-diversity-from-low-resource-to\-endangered-languages.

Armen Aghajanyan, Lili Yu, Alexis Conneau, Wei-Ning Hsu, Karen Hambardzumyan, Susan Zhang, Stephen Roller, Naman Goyal, Omer Levy, and Luke Zettlemoyer. 2023. Scaling laws for generative mixed-modal language models. *arXiv preprint arXiv:2301.03728*.

Ibrahim Alabdulmohsin, Xiaohua Zhai, Alexander Kolesnikov, and Lucas Beyer. 2023. Getting vit in shape: Scaling laws for compute-optimal model design. *arXiv preprint arXiv:2305.13035*.

Paul Azunre, Salomey Osei, Salomey Addo, Lawrence Asamoah Adu-Gyamfi, Stephen Moore, Bernard Adabankah, Bernard Opoku, Clara Asare-Nyarko, Samuel Nyarko, Cynthia Amoaba, et al. 2021. English-twi parallel corpus for machine translation. *arXiv preprint arXiv:2103.15625*.

Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee, and Utkarsh Sharma. 2021. Explaining neural scaling laws. *arXiv preprint arXiv:2102.06701*.

Abhinand Balachandran. 2023. Tamil-llama: A new tamil language model based on llama 2. *arXiv preprint arXiv:2311.05845*.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2019. Piqa: Reasoning about physical commonsense in natural language. *ArXiv*, abs/1911.11641.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Ethan Caballero, Kshitij Gupta, Irina Rish, and David Krueger. 2022. Broken neural scaling laws. *arXiv preprint arXiv:2210.14891*.

Monojit Choudhury, Anirudh Srinivasan, and Sandipan Dandapat. 2019. Processing and understanding mixed language data. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): Tutorial Abstracts*.

Together Computer. 2023. Redpajama: an open dataset for training large language models.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Andrea Cossu, Tinne Tuytelaars, Antonio Carta, Lucia Passaro, Vincenzo Lomonaco, and Davide Bacciu. 2022. Continual pre-training mitigates forgetting in language and vision. *arXiv preprint arXiv:2205.09357*.

Raj Dabre, Aizhan Imankulova, Masahiro Kaneko, and Abhisek Chakrabarty. 2021. Simultaneous multi-pivot neural machine translation. *arXiv preprint arXiv:2104.07410*.

Viet Dac Lai, Chien Van Nguyen, Nghia Trung Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan A Rossi, and Thien Huu Nguyen. 2023. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. *arXiv e-prints*, pages arXiv–2307.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey for in-context learning. *arXiv preprint arXiv:2301.00234*.

Elias Frantar, Carlos Riquelme, Neil Houlsby, Dan Alistarh, and Utku Evci. 2023. Scaling laws for sparsely-connected foundation models. *arXiv preprint arXiv:2309.08520*.

Kshitij Gupta, Benjamin Thérien, Adam Ibrahim, Mats L Richter, Quentin Anthony, Eugene Belilovsky, Irina Rish, and Timothée Lesort. 2023. Continual pre-training of large language models: How to (re) warm your model? *arXiv preprint arXiv:2308.04014*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*.

Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B Brown, Prafulla Dhariwal, Scott Gray, et al. 2020. Scaling laws for autoregressive generative modeling. *arXiv preprint arXiv:2010.14701*.

Danny Hernandez, Tom Brown, Tom Conerly, Nova DasSarma, Dawn Drain, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Tom Henighan, Tristan Hume, et al. 2022. Scaling laws and interpretability of learning from repeated data. *arXiv preprint arXiv:2205.10487*.

Danny Hernandez, Jared Kaplan, Tom Henighan, and Sam McCandlish. 2021. Scaling laws for transfer. *arXiv preprint arXiv:2102.01293*.

Joel Hestness, Newsha Ardalani, and Gregory Diamos. 2019. Beyond human-level accuracy: Computational challenges in deep learning. In *Proceedings of the 24th Symposium on Principles and Practice of Parallel Programming*, pages 1–14.

Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. 2017. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Peter J Huber. 1992. Robust estimation of a location parameter. In *Breakthroughs in statistics: Methodology and distribution*, pages 492–518. Springer.

Marcus Hutter. 2021. Learning curve theory. *arXiv preprint arXiv:2102.04074*.

Alex Jones, William Yang Wang, and Kyle Mahowald. 2021. A massively multilingual analysis of cross-linguality in shared embedding space. *arXiv preprint arXiv:2109.06324*.

Damjan Kalajdzievski. 2024. Scaling laws for forgetting when fine-tuning large language models. *arXiv preprint arXiv:2401.05605*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Zixuan Ke, Yijia Shao, Haowei Lin, Tatsuya Konishi, Gyuhak Kim, and Bing Liu. 2022. Continual pre-training of language models. In *The Eleventh International Conference on Learning Representations*.

Zixuan Ke, Yijia Shao, Haowei Lin, Hu Xu, Lei Shu, and Bing Liu. 2023. Adapting a language model while preserving its general knowledge. *arXiv preprint arXiv:2301.08986*.

Richard Kimera, Daniela N Rim, and Heeyoul Choi. 2023. Building a parallel corpus and training translation models between luganda and english. *arXiv preprint arXiv:2301.02773*.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

Viet Dac Lai, Nghia Trung Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023. Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning. *arXiv preprint arXiv:2304.05613*.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona T. Diab, Ves Stoyanov, and Xian Li. 2021. Few-shot learning with multilingual generative language models. In *Conference on Empirical Methods in Natural Language Processing*.

Jinyang Liu, Yujia Zhai, and Zizhong Chen. 2020. Normalization of input-output shared embeddings in text generation models. *arXiv preprint arXiv:2001.07885*.

Niklas Muennighoff, Alexander M Rush, Boaz Barak, Teven Le Scao, Aleksandra Piktus, Nouamane Tazi, Sampo Pyysalo, Thomas Wolf, and Colin Raffel. 2023. Scaling data-constrained language models. *arXiv preprint arXiv:2305.16264*.

Jorge Nocedal. 1980. Updating quasi-newton matrices with limited storage. *Mathematics of computation*, 35(151):773–782.

Jessica Ojo, Kelechi Ogueji, Pontus Stenetorp, and David I Adelani. 2023. How good are large language models on african languages? *arXiv preprint arXiv:2311.07978*.

Malte Ostendorff and Georg Rehm. 2023. Efficient language model training through cross-lingual and progressive transfer learning. *arXiv preprint arXiv:2301.09626*.

Sinno Jialin Pan and Qiang Yang. 2009. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.

Aryan Patil, Varad Patwardhan, Abhishek Phaltankar, Gauri Takawane, and Raviraj Joshi. 2023. Comparative study of pre-trained bert models for code-mixed hindi-english data. In *2023 IEEE 8th International Conference for Convergence in Technology (I2CT)*, pages 1–7. IEEE.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.

E. Ponti, Goran Glavavs, Olga Majewska, Qianchu Liu, Ivan Vulic, and Anna Korhonen. 2020. Xcopa: A multilingual dataset for causal commonsense reasoning. In *Conference on Empirical Methods in Natural Language Processing*.

Ella Rabinovich, Shuly Wintner, and Ofek Luis Lewinsohn. 2018. A parallel corpus of translationese. In *Computational Linguistics and Intelligent Text Processing: 17th International Conference, CICLing 2016, Konya, Turkey, April 3–9, 2016, Revised Selected Papers, Part II 17*, pages 140–155. Springer.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv e-prints*.

Miao Rang, Zhenni Bi, Chuanjian Liu, Yunhe Wang, and Kai Han. 2023. Large ocr model: An empirical study of scaling law for ocr. *arXiv preprint arXiv:2401.00028*.

Jonathan S Rosenfeld, Amir Rosenfeld, Yonatan Belinkov, and Nir Shavit. 2019. A constructive prediction of the generalization error across scales. *arXiv preprint arXiv:1909.12673*.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Winogrande. *Communications of the ACM*, 64:99 – 106.

Nikhil Sardana and Jonathan Frankle. 2023. Beyond chinchilla-optimal: Accounting for inference in language model scaling laws. *arXiv preprint arXiv:2401.00448*.

Utkarsh Sharma and Jared Kaplan. 2020. A neural scaling law from the dimension of the data manifold. *arXiv preprint arXiv:2004.10802*.

Vivek Srivastava and Mayank Singh. 2021. Challenges and considerations with code-mixed nlp for multilingual societies. *arXiv preprint arXiv:2106.07823*.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2020. Ernie 2.0: A continual pre-training framework for language understanding. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8968–8975.

Sho Takase and Sosuke Kobayashi. 2020. All word embeddings from one embedding. *Advances in Neural Information Processing Systems*, 33:3775–3785.

Eshaan Tanwar, Manish Borthakur, Subhabrata Dutta, and Tanmoy Chakraborty. 2023. Multilingual llms are better cross-lingual in-context learners with alignment. *arXiv preprint arXiv:2305.05940*.

Yi Tay, Mostafa Dehghani, Samira Abnar, Hyung Won Chung, William Fedus, Jinfeng Rao, Sharan Narang, Vinh Q Tran, Dani Yogatama, and Donald Metzler. 2022. Scaling laws vs model architectures: How does inductive bias influence scaling? *arXiv preprint arXiv:2207.10551*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Di Wu and Christof Monz. 2023. Beyond shared vocabulary: Increasing representational word similarities across languages for multilingual machine translation. *arXiv preprint arXiv:2305.14189*.

Minghao Wu, Thuy-Trang Vu, Lizhen Qu, George Foster, and Gholamreza Haffari. 2024. Adapting large language models for document-level machine translation. *arXiv preprint arXiv:2401.06468*.

Hu Xu, Bing Liu, Lei Shu, and Philip S Yu. 2019. Bert post-training for review reading comprehension and aspect-based sentiment analysis. *arXiv preprint arXiv:1904.02232*.

Gaowei Zhang, Yupeng Hou, Hongyu Lu, Yu Chen, Wayne Xin Zhao, and Ji-Rong Wen. 2023a. Scaling law of large sequential recommendation models. *arXiv preprint arXiv:2311.11351*.

Haode Zhang, Haowen Liang, Liming Zhan, Xiao-Ming Wu, and Albert Lam. 2023b. Revisit few-shot intent classification with plms: Direct fine-tuning vs. continual pre-training. *arXiv preprint arXiv:2306.05278*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. 2020. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76.