Efficient LLM Fine-tuning for Cantonese Text Translation

AIST4010 Foundation of Applied Deep Learning
Report

Hon Kwan Shun Quinson

Abstract

The large population of Cantonese speakers and the release of various technology products surrounding Cantonese shows the strong demand for accurate Cantonese services and translations. At the same time, many advancements are being made in the machine translation field that allow chat-based language models with exceptional capabilities to be customised for specific use cases without a large amount of computing power. This project explores the capabilities of such models on English-Cantonese and Mandarin-Cantonese translation tasks. The project utilizes publicly available corpora and datasets, along with new tools and techniques in LLM training to improve translation performance on the existing Yi-6B model. Two models were trained through pretraining and supervised fine-tuning, both having notable changes in performance on various translation pairs. This paper will go into detail about the model training process, experiments and results of training Cantonese translation LLMs.

I Introduction

Cantonese is a version of Chinese spoken by over 80 million people around the world [1]. Cantonese has become increasingly popular with native speakers online, and technology companies are beginning to provide Cantonese support in their products, such as Galaxy AI [2] and Cantonese typing on Apple devices [3]. It is essential that this large group of users have full access to features provided by technology companies, especially with the use of AI in their products.

On the other hand, the Transformer architecture has revolutionized natural language processing with its immediate application in machine translation (MT) [4], and later becoming the basis of large language models (LLMs) including GPT-4 [5] and LLaMA-2 [6], with applications in human-computer interactions, code writing and translation. Moreover, in recent years, it has become increasingly viable to fit LLMs to particular use cases, with methods such as Low-Rank Adaptation [7] significantly reducing the resources need for fine-tuning an LLM.

To make a step towards building AI tools and assistants tailored for Cantonese-speaking communities, this project will focus on fine-tuning LLMs for English-Cantonese and Mandarin-Cantonese translations. It is hoped that this project can show the feasibility of training and running LLMs with limited computing and Cantonese language resources, opening up the possibilities for future local AI applications.

II Related Work

A. Neural Machine Translation (NMT)

When researchers began using neural networks for machine translation, recurrent neural networks and convolutional neural networks were often used for processing languages. The models proposed in [8] use a convolutional network encoder to obtain a representation of the input sequence, and a basic recurrent neural network acts as a decoder to obtain the translated output. This paper forms the basis of NMT and introduces the encoder-decoder architecture still commonly used today in MT.

[9] proposes an additional weighting on the hidden states of the model before decoding, which would later be called the attention mechanism. The idea of attention in this paper is a core concept used in the development of the Transformers architecture widely used in MT.

B. Transformers in NMT

The Transformer architecture in [4] revolutionized machine translation by using self-attention layers to extract patterns and features in words, setting a new standard for neural machine translation.

Since then, sequence-to-sequence Transformers have been popular in machine translation tasks. The current state-of-the-art sequence-to-sequence model for multilingual machine translation is No Language Left Behind made by Meta [10], which can perform translations in over 200 languages, including English, German, French, Spanish, Mandarin Chinese and Cantonese. This model shows the immense potential of the Transformer architecture in modelling human language and performing MT. In addition to their model, Meta also created the FLORES dataset available for evaluating translation performance on language models.

C. Decoder-LLMs in NMT

Apart from full sequence-to-sequence transformers, causal language models that only use decoder Transformer blocks have gained significant interest, especially with the release of GPT-4 [5] and LLaMA-2 [6] models. The rise of GPT-4 brings public attention to natural language processing and related topics such as MT, while LLaMA-2 enables the open-source community to develop high-quality LLMs for many use cases including MT.

As LLMs continue to gain popularity, the availability of these models towards users of different backgrounds becomes essential. Out of these models, BLOOM [11] has one of the widest ranges of covered languages at 46 natural languages and 13 programming languages. In a subsequent study on the model's machine translation abilities, it is noted BLOOM performs poorly in zero-shot situations but performs well in few-shot scenarios [12]. This indicates the need of good prompting and guidance in using LLMs to perform specific tasks such as MT.

[13] finds that decoder LLMs tend to perform worse then sequence-to-sequence Transformers in machine translation, though they exhibit unique properties such as performing correct translations when languages are not specified, and being able to translate zero-resource languages decently. This shows the possibility of adopting LLMs in more flexible translation scenarios, especially when combined with other tasks in LLM.

D. Training LLMs on Translation Tasks

Parameter-Efficient Fine-Tuning (PEFT) by Hugging Face [14] is a library that enables efficient adaptation of pretrained models to new applications by only tuning a small number of extra parameters using Low-Rank Adaptation (LoRA) [7]. This tool has opened up the possibilities for training custom models on specific use cases with significantly reduced hardware requirements.

[15] describes a two-step fine-tuning approach which LLMs can learn to perform machine translation. The LLM first learns from a large amount of monolingual data in the target language, then learns the translation task with a small set of parallel corpora. Research such as [16] describe methods which LLMs should be prompted to achieve better results. These resources are helpful for defining a suitable training approach for this project.

E. Cantonese Machine Translation

There have been some attempts to study the effectiveness of Transformers in Cantonese translation. [17] focuses on translating between Cantonese and Mandarin using RNN and Transformer architectures. [18] uses a pre-trained BART model for machine translation between Cantonese and English, further training it on monolingual and parallel data from English and Cantonese. Cantonese datasets and LLM models trained in Cantonese can also be found on Hugging Face [19]. These previous attempts at Cantonese MT provide significant groundwork for this project, especially in finding sources with clean Cantonese data.

Among resources that have been found, there is no existing research that focuses on fine-tuning chat LLMs on Cantonese translation, which is a niche that this project targets.

III Data

Three different types of data are used for different stages of the project: monolingual corpora, parallel corpora and evaluation datasets.

A. Monolingual Corpora

Monolingual corpora are used for pre-training the LLM on Cantonese. The majority of content in this stage comes from Cantonese Wikipedia, containing over 600 000 lines of Cantonese text, with a total of 25 million Cantonese characters. The contents of Cantonese Wikipedia are retrieved from a Wikimedia dump [20], and then processed with WikiExtractor [21] to obtain the raw text. Apart from Cantonese Wikipedia, openrice-senti is a dataset that contains over 7 000 restaurant reviews from the OpenRice website in Hong Kong [22]. The reviews have a total of over 2 million characters, with the majority of text written in Cantonese.

There are various sources of monolingual data that are considered but not used in the project due to time constraints. This includes social media websites such as LIHKG, Instagram, Facebook, and YouTube.

B. Parallel Corpora

Parallel corpora are used for supervised fine-tuning (SFT) on the translation task. The ABC Cantonese Parallel Corpus provides 14 000 parallel sentences for English-Cantonese translations [18]. Moreover, Kaifangcidian [23, 17] includes 15 000 Mandarin-Cantonese parallel sentences pairs within its website. Prompt templates are added to these parallel sentence pairs to form conversations that the LLM can be trained on.

Apart from pre-built parallel corpora, an attempt was made to extract video transcripts from the TED-Ed website [24] to construct parallel corpora. However, due to difficulties in aligning video scripts across languages, this approach was not adopted in supervised fine-tuning.

C. Evaluation Datasets

The FLORES+ dataset [10] has 2009 examples in the publicly available dev and devtest splits, which are used to test the performance of translation examples. These examples are extracted from Wikinews, Wikijunior and Wikivoyages, and are available in English, Beijing Mandarin, Taiwan Mandarin and Cantonese. Special precautions were made by the FLORES+ team to protect the data from web crawlers, which ensures the fairness of the dataset as a testing benchmark.

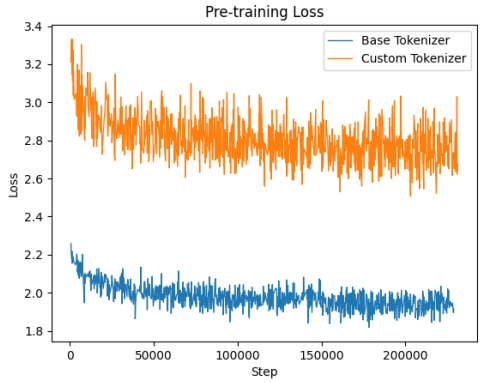

Fig. 1: Pre-training Loss using base model tokenizer and custom tokenizer

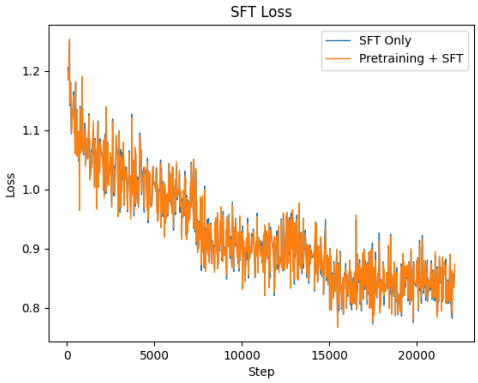

Fig. 2: Supervised Fine-Tuning Loss using models with and without pre-training

Model	Base model	SFT only	Pretraining + SFT
Eng-Can	0.80	0.78	0.79
Can-Eng	0.94	0.93	0.94
Man-Can	0.82	0.83	0.83
Can-Man	0.86	0.86	0.86
TW-Can	0.83	0.84	0.84
Can-TW	0.85	0.85	0.85

TABLE I: BERTscore metrics for Yi-6B base model, SFT model and Pretrained + SFT model on FLORES+ dataset

IV Approach

In this project, an LLM with English and Simplified Chinese capabilities will be fine-tuned to perform Cantonese translation, leveraging its existing English and Mandarin ability to improve translations. Following results in [25], the Yi-6B model [26] is chosen for its lower perplexity and decent performance in the Cantonese language, which is hoped to lead to better performance of the fine-tuned model.

Two new fine-tuned models are trained in this project. One model goes through the full training process described in [15]. The model is first trained on monolingual Cantonese training data to learn Cantonese. At each training step, the model learns on a few short Cantonese paragraphs at a time, with each paragraph limited to 500 characters to reduce memory usage. The model is trained for 1.5 epochs, which took about 18 hours on an NVIDIA GeForce RTX 4090 GPU.

Then, the model is trained on the parallel corpora through SFT, where the training examples are embedded into conversation templates to be learned by the model. The model is trained for 1 epoch to specialize in the machine translation task, which took 1.2 hours on the same GPU. A separate model is trained purely on SFT to analyze the effect of pre-training on translation tasks.

The model is fine-tuned with the Hugging Face Transformers library [27], as it contains automated Trainers that simplify the training process. Parameter-Efficient Fine-Tuning (PEFT) [14] is used in both stages of training to minimize computing requirements, as full-weight fine-tuning is impractical on consumer hardware. LoRA [7] is applied on key, query and value matrices in attention layers with rank 32 and 0.05 dropout. The default AdamW optimizer [28] is used, with the learning rate starting at 3e-4 that decays linearly to 0 during training.

After fine-tuning the model, the model is compared with the original model to compare their translation performance using the FLORES+ dataset. SacreBLEU [29] and BERTscore [30] are used to evaluate results with the BLEU and BERTscore metrics. Six translation directions are tested, which include translating from English, Beijing Mandarin and Taiwan Mandarin to Cantonese, and translating from Cantonese to the three other languages.

Initially, there were plans to modify the pre-trained tokenizer by adding new Cantonese characters from the monolingual corpus. However, there were difficulties in creating a new tokenizer that would be compatible with the existing base model tokenizer, so the model would not need to relearn everything in a new tokenization scheme. The resulting pre-trained model did not perform adequately, as its training loss was significantly higher than pretraining with the original tokenizer, and initial inference tests showed ill-formed tokens and characters being used in the output. Thus, no further testing was done on this model.

V Results

A. Metrics

Out of the 6 translation directions tested in Fig. 3, The pre-trained model showed the best performance when translating from Beijing Mandarin and Taiwan Mandarin to Cantonese, increasing BLEU by 12.73 and 6.05 points respectively over the base model. The model trained on SFT only performed the best when translating from Cantonese to Taiwan Mandarin, increasing BLEU by 5.43 points over the base model. The base model outperforms the trained models in the English-Cantonese translation pair by up to 4 BLEU points, while showing a dominant

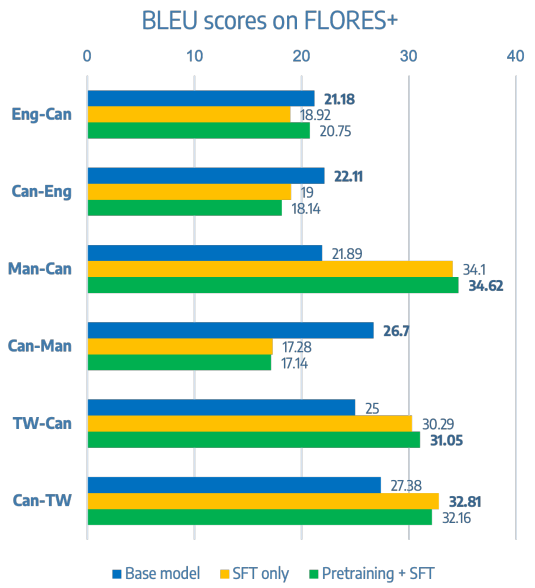

Fig. 3: BLEU scores for Yi-6B base model, SFT model and Pretrained + SFT model on FLORES+ dataset

advantage on translating from Cantonese to Beijing Mandarin by over 9 BLEU points

Comparing results between the SFT only model and the pre-trained model, the pre-trained model achieved a higher BLEU on 3 out of 6 occassions. Moreover, the pre-trained model is always within 2 BLEU points of the SFT model, despite having significantly more training time.

From Table I, All models performed similarly in BERTscore F1 scores, with differences of up to 0.02 between models. This indicates that the translation of semantic meanings in each model are comparable to each other.

System	Translate the given Mandarin words into Cantonese.
User	上午 8 點 46 分時，整座城市陷入一片寂靜，此刻正是第一架噴射機撞擊目標的確切時間。
Human	朝早嘅 8 點 46 分，全城一片寂靜，標誌著第一架飛機襲擊目標嘅時刻。
Base model	早上八點四十六分，整座城市陷入一片寂靜，這正是第一架噴射機撞擊目標的確切時間。
SFT Only	上午八點四十六分，全城靜晒，嗰陣時正係第一架客機撞擊目標嘅確實時間。
Pretrained + SFT	八點四十六分，成個城市靜晒，嗰一刻正係第一架噴射機撞擊目標嘅確實時間。

TABLE II: A translation example from Taiwan Mandarin to Cantonese

B. Manual Evaluation

When translating from other languages into Cantonese, the trained models perform relatively well compared to the base model. As seen in Table II, the SFT and pre-trained

System	Translate the given Cantonese words into English.
User	學生亦都可以喺全日任何時候向導師提交問題，並且可以預期得到幾快嘅回應，而唔係等到下次面授課堂。
Human	Students can also submit questions to instructors at any time of day and expect reasonably quick responses, rather than waiting until the next face-to-face meeting.
Base model	Students can also submit questions to their tutors at any time during the day and can expect a quick response, rather than waiting until the next face-to-face class.
SFT	Students can also submit questions at any time during the day, and can expect a quick response, and not wait until the next class.
Pretrained + SFT	Students can submit questions to their teachers at any time of the day, and can expect to get a quick response, rather than waiting till the next face-to-face class.

TABLE III: A translation example from Cantonese to English

models can utilize word patterns in Cantonese such as ” 靜晒” and ” 成個” better than the base model, sometimes better than the human translations. However, some details in the original sentence may be lost when trained models are translating sentences. In this case, the pre-trained model fails to capture that the sentence is referring to a time in the morning, and omits that information in the translated sentence.

When translating from Cantonese to other languages, the trained models tend to perform slightly worse than the base model and human translations. In Table III, all LLMs failed to capture the idea of ”reasonably quick” in the original sentence. Moreover, the SFT and pre-trained models suffer from minor grammatical inaccuracies, which reduce the conciseness of the sentence but do not affect its meaning.

VI Discussion

From calculating evaluation metrics and analyzing translation results, it is apparent that conducting supervised fine-tuning on a base model has a clear impact on how the LLM performs in translation. In particular, the SFT model dramatically improved the capabilities of the LLM in Taiwan Mandarin-Cantonese translation, as well as translating from Beijing Mandarin to Cantonese.

Despite the improvements with SFT, the model did not seem to improve greatly from additional pre-training. This could be due to several reasons, such as inappropriate grouping of sentences in pre-training, the lack of training epochs, and insufficient variety in Cantonese data. Apart from pre-training issues, another reason why the model may not have learned much from pre-training could be because the base model is trained on CommonCrawl [26],

which includes Wikipedia pages. Therefore, pre-training the model on Cantonese Wikipedia again may not have a significant effect on improving model performance.

Another issue with both SFT and pre-trained models is that their language performance in English and Beijing Mandarin is reduced compared to the base model. In English, the problem may be exacerbated due to the imbalance between English and Chinese data, since the Kaifangcidian dataset consists of only Chinese data, so the English data only consists about 1/4-th of the total parallel dataset. The disadvantage in Cantonese-Beijing Mandarin translation may be attributed to Beijing Mandarin being written in Simplified Chinese in the FLORES+ dataset, and may be fixed with prompting to encourage translation in Simplified Chinese.

Meanwhile, it is noted that the base model occasionally fails to align with the translation task. In some cases, the base model outputs more words than needed to explain what it is trying to do, rather than outputting the translation directly. It also occasionally does not make translations into Cantonese, and instead outputs the original sentence, or in worse cases, output the sentence in languages that are not specified in any of the training or testing data, such as Japanese, Hindi or Thai. This indicates that the base LLM model may have been underperforming since it could not understand the prompt given by the user. Therefore, more work is needed in prompting LLMs to produce desirable outputs, which could include few-shot learning [31] or clearer task specifications in prompts.

In the future, further research may be done on training LLMs on translation tasks. As this project has not been able to improve translations significantly through pre-training, more work can be done to improve the pre-training process, whether that is through a larger quantity or variety of data, more training time, etc. Moreover, research such as [17] has shown that character-based tokenisation on new Cantonese characters can lead to better performance in the translation task, which can be tested on chat LLMs. Research similar to [16] can also be considered to improve the zero-shot performance of LLM translations, as these translations could be combined with additional prompts to perform various tasks. Finally, exploring the use of hyperparameters such as LoRA matrix ranks, learning rate and batch size can also be helpful in training optimal models.

Beyond LLM translations to and from Cantonese, it is hoped that improvements in Cantonese language abilities in LLMs could lead to the development of fully Cantonese chat LLMs that can generalize well on many tasks in Cantonese, especially those related to Cantonese language use. Moreover, integrating Cantonese LLMs into real-life applications is especially important for building products that fully suit the needs of Cantonese users, and more work needs to be done to make this a reality.

VII Conclusion

The project has fine-tuned two models on the English-Cantonese and Mandarin Cantonese translation task, where one model was pre-trained and fine-tuned through supervised fine-tuning, and the other model was trained on supervised fine-tuning only. The results show that the models trained in this project mostly improved translations in the Mandarin-Cantonese translation pair, but failed to improve on the English-Cantonese translation pair. However, the fluency of both models in Cantonese improved considerably after training. Further research is need to ensure that models can achieve a performance boost in more translation pairs through better training techniques. It is hoped that this project can contribute to the development of Cantonese LLMs and its applications in real-life scenarios.

References

[1] D. M. Eberhard, G. F. Simons *et al.*, Eds., *Ethnologue: Languages of the World*, 27th ed. Dallas, Texas: SIL International, 2024. [Online]. Available: http://www.ethnologue.com

[2] Pickle Rick, "Samsung Galaxy AI to Add Cantonese This Spring, 13 New Languages to be Added This Year." [Online]. Available: https://unwire.hk/2024/04/11/samsung-galaxy-ai-3/ai/

[3] Apple, "Chinese and Cantonese Input Method User Guide," Palo Alto, California. [Online]. Available: https://support.apple.com/en-ca/guide/chinese-input-method/welcome/mac

[4] A. Vaswani, N. Shazeer *et al.*, "Attention is All you Need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg *et al.*, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[5] J. Achiam, S. Adler *et al.*, "GPT-4 Technical Report," 2024.

[6] H. Touvron, L. Martin *et al.*, "Llama 2: Open Foundation and Fine-Tuned Chat Models," 2023.

[7] E. J. Hu, Y. Shen *et al.*, "LoRA: Low-Rank Adaptation of Large Language Models," in *International Conference on Learning Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=nZeVKeeFYf9

[8] N. Kalchbrenner and P. Blunsom, "Recurrent Continuous Translation Models," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. ACL, 2013, pp. 1700–1709. [Online]. Available: https://aclanthology.org/D13-1176/

[9] D. Bahdanau, K. Cho, and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: http://arxiv.org/abs/1409.0473

[10] NLLB Team, M. R. Costa-jussà *et al.*, "No Language Left Behind: Scaling Human-Centered Machine Translation," 2022.

[11] T. L. Scao, A. Fan *et al.*, "BLOOM: A 176B-Parameter Open-Access Multilingual Language Model," 2023.

[12] R. Bawden and F. Yvon, "Investigating the Translation Performance of a Large Multilingual Language Model: the Case of BLOOM," 2023.

[13] W. Zhu, H. Liu *et al.*, "Multilingual Machine Translation with Large Language Models: Empirical Results and Analysis," 2023.

[14] S. Mangrulkar, S. Gugger *et al.*, "PEFT: State-of-the-art Parameter-Efficient Fine-Tuning methods," https://github.com/huggingface/peft, 2022.

[15] H. Xu, Y. J. Kim *et al.*, "A Paradigm Shift in Machine Translation: Boosting Translation Performance of Large Language Models," 2024.

[16] B. Zhang, B. Haddow, and A. Birch, "Prompting Large Language Model for Machine Translation: A Case Study," in *Proceedings of the 40th International Conference on Machine Learning*, ser. ICML'23. JMLR.org, 2023.

[17] M. Dare, V. F. Diaz *et al.*, "Unsupervised Mandarin-Cantonese Machine Translation," *arXiv preprint arXiv:2301.03971*, 2023.

[18] A. Mikazuki, "TransCan: An English-to-Cantonese Machine Translation Model," https://github.com/ayaka14732/TransCan, 2023.

[19] "Cantonese Llama 2 7b v1," https://huggingface.co/indiejoseph/cantonese-llama-2-7b-oasst-v1, 2023.

[20] "zh_yuewiki dump progress on 20240301," 2024. [Online]. Available: https://dumps.wikimedia.org/zh_yuewiki/20240301/

[21] attardi, "WikiExtractor." [Online]. Available: https://github.com/attardi/wikiextractor/

[22] toastynews, "openrice-senti." [Online]. Available: https://github.com/toastynews/openrice-senti

[23] "Kaifangcidian." [Online]. Available: https://kaifangcidian.com/han/yue/

[24] "TED-Ed." [Online]. Available: https://ed.ted.com/

[25] J. Cheng, "Fine-tuning Cantonese Large Language Models Part 1 of 3: Choosing a Pre-Trained Model." [Online]. Available: https://hon9kon9ize.com/posts/2023-12-18-llm-finetuning1

[26] A. Young, B. Chen *et al.*, "Yi: Open Foundation Models by 01.AI," 2024.

[27] T. Wolf, L. Debut *et al.*, "Huggingface's transformers: State-of-the-art natural language processing," *CoRR*, vol. abs/1910.03771, 2019. [Online]. Available: http://arxiv.org/abs/1910.03771

[28] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=Bkg6RiCqY7

[29] mjpost, "sacreBLEU." [Online]. Available: https://github.com/mjpost/sacrebleu

[30] T. Zhang, V. Kishore *et al.*, "BERTScore: Evaluating Text Generation with BERT," in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. [Online]. Available: https://openreview.net/forum?id=SkeHuCVFDr

[31] T. Brown, B. Mann *et al.*, "Language Models are Few-Shot Learners," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato *et al.*, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 1877–1901. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf