TinyFSL: Tiny Machine Learning for Filipino Sign Language

Loben Klien A. Tipan[1], **Alyanna Mari Abalos**[2], **Alyana Erin Bondoc**[3], **Justin Jarrett To**[4], **Joanna Pauline Rivera**[5], **Ann Franchesca Laguna**, and **Edward Tighe**
De La Salle University
2401 Taft Avenue, Manila, Philippines 0922
[1] loben_klien_tipan@dlsu.edu.ph, [2] alyanna_mari_abalos@dlsu.edu.ph,
[3] alyana_erin_bondoc@dlsu.edu.ph, [4] justin_jarrett_to@dlsu.edu.ph,
[5] joanna.rivera@dlsu.edu.ph

Abstract

A Sign Language Translation (SLT) model is one example of a large-scale model, resulting from the use of video dataset and deep learning models. For practical use of the Deaf community, SLT models are meant to be eventually deployed on mobile devices, for instance. However, large-scale models entail high resource requirements from mobile devices with limited capacity. Tiny Machine Learning (TinyML) is a rapidly emerging field that can condense large-scale models for deployment on low-resource devices. By leveraging TinyML techniques, this research refines an adapted 2D Convolutional Neural Networks (CNN) and Transformer Neural Networks (TNN) model by Camgoz et al. (2020). The teacher model is trained on the 2D CNN and TNN model using the Filipino Sign Language - Non-manual Signals (FSL-NMS) dataset by Rivera and Ong (2018b). Through knowledge distillation, the student model achieved 45% higher BLEU-4 score compared to the teacher model, and a 5.4 compression ratio. These results highlight the potential of knowledge distillation techniques on compressing and improving SLT models. This work paves the way for the development of more accessible communication tools for the Filipino Deaf community and non-signers.

1 Introduction

Sign Language Recognition (SLR) often requires multi-modal large-scale models that converts videos into words, phrases, or sentences. Continuous SLR (CSLR) is a further improvement of SLR which interprets multiple sign language gestures without delineation between gestures. These CSLR models, however, typically require a vast amount of data to train to achieve high accuracy. This makes most CSLR models more complex and larger compared to isolated SLR models (Zhou et al., 2022). Another model that aims to develop a more robust approach for translating sign language videos to text that learns the grammar and morphology of the sign language is a Sign Language Translation (SLT) model. However, state-of-the-art SLT models are trained on large datasets using deep learning techniques, also often resulting to larger models.

The main goal of SLT models is to help the Deaf community, thus it should be deployed eventually to be used. A few FSL-related applications are interpreters that are mostly used for learning FSL (e.g. Senyas by (Alberto et al., 2022), and 3D animation of Aesop's Fable by (Cueto et al., 2020)). These are manually translated FSL signs to text, and vice versa, that may benefit from automatic translation systems.

As several studies have shown isolated Filipino CSLR models performing with over 90% accuracy (shown in Section 2.1), and SLT studies reaching a BLEU-4 of over 20 as demonstrated by Camgoz et al. (2020), it is about time to also consider the possibilities of deployment to reach the intended users. However, all of these studies produced large models which entail high resource requirements on mobile devices with limited capacity. This makes deep learning applications difficult to deploy on mobile devices (Wang et al., 2018).

TinyML is a growing sub-field of machine learning that is dedicated to run Artificial Intelligence (AI) algorithms on devices with limited resources, without needing heavy computation or internet connectivity. It minimizes dependability and latency issues. Additionally, it provides enhanced privacy by reducing the need to send personal data to the cloud (Kallimani et al., 2023).

Several TinyML applications include detection of eating habits (Nyamukuru and Odame, 2020), and detection of medical face mask (Mohan et al., 2021). In addition, TinyML has already been explored in various fields such as audio analysis (e.g. audio wake words (Zhang et al., 2017)), image recognition (e.g. visual wake words (Chowdhery et al., 2019), gesture recognition (Amir et al.,

2017)), psychological/behavioral metrics (e.g. activity detection (Hassan et al., 2018)), and industry telemetry (e.g. anomaly detection (Koizumi et al., 2019)) (Dutta and Bharali, 2021).

This work utilizes TinyML techniques to condense a large-scale model for Filipino Sign Language (FSL) to a lightweight and efficient model that can potentially be deployed on a variety of devices, including smartphones, wearable devices, and even embedded systems.

FSL is a mode of communication by the Deaf community in the Philippines. According to Newall et al. (2020), approximately 15% of Filipinos suffer from moderate to severe hearing impairment. By developing innovative TinyML-powered FSL tools, there is an opportunity to enhance communication avenues for the Filipino Deaf community. This is important for promoting inclusivity, as well as enabling their fuller engagement in a variety of social activities.

The rest of this paper is organized as follows. Section 2 enumerates works related to TinyML and FSL. Section 3 describes the characteristics and preparation of the Filipino Sign Language - Non-manual Signals (FSL-NMS) dataset (Rivera and Ong, 2018b). Section 4 discusses the methodology used in applying TinyML in Filipino SLT, wherein a 2D Convolutional Neural Networks (CNN) and Transformer Neural Networks (TNN) model by Camgoz et al. (2020) is adapted and trained on the FSL-NMS dataset for the teacher model. Section 5 reports the results and analysis of the teacher model and the student model using the BLEU scores and ROUGE metric. Lastly, the conclusions and recommendations are presented in Section 6.

2 Related Work

2.1 Filipino Sign Language Recognition

In recent years, there has been growing interest in developing deep learning-based approaches for FSL recognition. Deep learning models have the potential to learn the complex patterns in FSL signs and phrases, and to achieve high accuracy on image-based recognition tasks.

In the study of Cabalfin et al. (2012), they used Manifold Projection Learning model where signs are predicted based on the computation and comparison of Dynamic Time Warping (DTW), and Longest Common Sub-sequence Similarity Matching (LCSSM). Their dataset consists of 72 isolated Filipino signs. Their highest recognition rates using DTW are 89% on 10 signs and 40% on all 72 signs. Using LCSSM, their highest recognition rates are 93% on 10 signs 31% on 72 signs.

As machine learning techniques become more prominent, the study by Ramos et al. (2019) focused on using Support Vector Machine (SVM) and K-Nearest Neighbor (KNN) which are both classification techniques. They used Histogram of Oriented Gradients (HOG) for feature extraction on 26 isolated gestures of FSL alphabets, achieving 94.49% accuracy.

A paper by Montefalcon et al. (2021) takes this further by applying a deep learning-based approach for Filipino Sign Language (FSL) recognition using CNN architecture, specifically ResNet-50, which extracts features from static images of FSL number signs ranging from (0-9). The model achieves a validation accuracy of 86.7% when the epoch value equals 15. Similarly, their subsequent study (Montefalcon et al., 2023) proposes a continuous SLR model for FSL recognition using Long Short Term Memory (LSTM) model. MediaPipe Holistic is used to extract features from video files of 15 Filipino phrases performed by three FSL signers. The LSTM model achieves an accuracy of 94% on the test set, outperforming their previous ResNet model with an accuracy of 87%. They have indicated on their analysis that facial components affect the performances, marking it an important set of features for recognition.

A similar study by Tupal et al. (2022) utilizes MediaPipe Holistic and LSTM. Their FSL recognition models applied InceptionV3, LSTM, and Gated Recurrent Units (GRU). When trained on a dataset comprising 20 foundational FSL words with at least 20 samples each, the model, leveraging the GRU achieved the highest accuracy of 86.74%.

A different study that focuses on facial expressions in FSL is conducted by (Rivera and Ong, 2018a), wherein 3D Animation Units (AU) extracted using Microsft Kinect are used as features. SVM is also used, achieving 87.14% as the highest accuracy. However, it was emphasized that hand signs must be recognized together with the facial expressions for the model to understand the context.

Overall, all the mentioned studies make significant contributions to the field of FSL recognition. As can be seen, different approaches yield different performances. However, despite using the same approach, performances can still differ as the amount, structure, and quality of the dataset differs.

Nonetheless, although it is still hampered by the challenge of limited available datasets, the burgeoning field of FSL recognition has made remarkable progress with the aid of deep learning approaches.

2.2 TinyML

Despite the increasing studies on SLR for FSL, there is still minimal studies focusing on the possibility of deployment on, for instance, mobile devices for practical use. Deep learning approaches in FSL may have promising results, but, despite using small datasets, it results to large models that are difficult to port to mobile or wearable devices.

Recent advancements in TinyML have focused on the development and optimization of machine learning models for deployment on resource-constrained devices. These techniques aim to reduce model size, power consumption, and computational requirements while maintaining acceptable levels of accuracy.

Model pruning has emerged as a pivotal technique in TinyML, addressing the challenge of deploying neural networks on devices with stringent memory constraints.

Han et al. (2015) demonstrated that by systematically removing weights with minimal impact on the output, the size of neural networks could be significantly reduced without a substantial loss in accuracy. Complementing this, quantization has been recognized for shrinking the model's memory footprint further.

Gupta et al. (2015) showcased that converting weights and activations from floating-point to lower-precision formats not only reduces the size but also accelerates inference, making it a vital technique for TinyML applications.

Knowledge distillation is another technique that has gained traction in TinyML. Hinton et al. (2015) introduced the concept of training a smaller "student" model to emulate the behavior of a larger "teacher" model. This process effectively compresses the knowledge of a complex network into a more compact and efficient form, making it suitable for deployment on low-power devices.

The automatic discovery of efficient architectures through Network Architecture Search (NAS) has also been a recent research focus. Zoph et al. (2018) explored the use of NAS to find models that are not only accurate but also computationally efficient for TinyML. This approach leverages the power of machine learning itself to design architectures that are tailored for performance on resource-constrained devices.

The convergence of machine learning and embedded systems is at the heart of TinyML. Warden and Situnayake (2019) emphasized that the goal of TinyML is to enable the deployment of AI in environments where traditional models would be impractical. By leveraging techniques like model pruning, quantization, knowledge distillation, and NAS, TinyML seeks to make AI ubiquitous, extending its reach to the most resource-constrained environments.

3 FSL-NMS Dataset Preparation

The dataset used in this study is the Filipino Sign Language - Non-manual Signals (FSL-NMS) dataset by Rivera and Ong (2018b). It is originally used for studying the different types of facial expressions in FSL.

The dataset contains a total of 50 sentences, featuring a broader array of signs and more specific emotions, including common phrases such as 'thank you' and 'good morning'. Among these are expressions like 'I am proud of you!' and 'Our team won!', as well as more complex sentiments like 'I am heartbroken' and situation-specific statements such as 'I saw a ghost.' Additionally, the dataset included various questions and time-specific greetings, enhancing its diversity and applicability in different contexts.

The dataset incorporated five videos, each featuring a different signer who sequentially signed the 50 sentences. The group of signers included three females and two males, providing a variety of signing styles and body languages. This diversity is crucial in enriching the dataset's value. These videos are then carefully edited and trimmed to ensure each sign is clearly presented, with each sign tailored to showcase a specific sign for about five seconds, totaling to 250 videos.

3.1 Data Annotation

The model used in this study (to be discussed further in Section 4.1) requires gloss translations (e.g. you how), in addition to the sentence translations (e.g. How are you?). Gloss translations are literal translations of each sign as it appears to its equivalent word or phrase, while sentence translations follow the English grammar. Since the dataset is created for the study of facial expressions, it initially did not include glosses, necessitating the annotation of glosses for the 50 sentences. Some

Words/Sentence	# of Sentences	
	Original	Augmented
1	0	0
2	15	111
3	115	777
4	85	518
5	30	222

Table 1: Distribution of Sentence Lengths in the FSL-NMS Dataset before and after Augmentation

entries were unintentionally skipped, while some have similar glosses that are only differentiated by facial expressions. This reduced the dataset to a total 44 sentences with unique gloss annotations. Refer to Appendix A for the complete list of gloss annotations.

The FSL-NMS dataset consists of a total of 245 samples, with the distribution of sentence lengths (n-grams) shown in Table 1.

3.2 Dataset Augmentation

As the dataset is particularly small for training a CNN-based SLT model, data augmentation is used to increase the diversity and volume of training data. This is crucial in enhancing the robustness of the model against various visual and environmental conditions. The FSL-NMS dataset, originally consisting of 245 samples, is expanded through mirroring, shifting and padding, adding noise, adding minimal motion to mimic jitters, and color adjustment, such as converting the videos to greyscale. The augmentation helped simulate a wider range of signing scenarios, thereby preparing the model to perform reliably in diverse settings.

After augmentation, the FSL-NMS dataset consists of a total of 1628 samples, distributed as shown in Table 1.

3.3 Sentence Distribution

The distribution of the sentences across the train, development, and test sets follows the 70-15-15 ratio, respectively. This structured allocation extends to each individual sign translation, ensuring that the counts of each sign are proportionately split according to these percentages across the different sets. This approach ensures a balanced representation of each sign in every subset, which is crucial for preventing model bias towards over-represented signs in any particular set.

4 TinyFSL Model

This study adapted a transformer-based architecture for an end-to-end training of a combination of CSLR and SLT model using the FSL-NMS dataset (Rivera and Ong, 2018b). Due to the complexity of the adapted model, knowledge distillation is applied to compress it to a lightweight and efficient model. Knowledge distillation is a technique where a smaller and more computationally efficient model (the 'student') is trained to approximate the performance of a larger, more complex model (the 'teacher') by learning from the teacher's outputs (Hinton et al., 2015). The basic architecture of knowledge distillation is illustrated in Figure 1.

4.1 Teacher Model Training

A 2D Convolutional Neural Networks (CNN) and Transformer Neural Networks (TNN) model by Camgoz et al. (2020) is adapted in this study. It is a transformer-based architecture that combines CSLR and SLT, and allows training in an end-to-end manner. To produce the teacher model, it is trained using the augmented FSL-NMS dataset. It has two parts: sign to gloss recognition, and gloss to text translation.

In the sign to gloss recognition, Squeezenet (Iandola et al., 2016) is first used to embed video frames. Second, these spatial embeddings are positionally encoded and then fed to the self-attention layer of the sign to gloss recognition part to learn the contextual relationship between frames. Lastly, the output of the self-attention layer is passed through a feed forward layer that produces the spatio-temporal representations.

In the gloss to text translation, a linear layer is first used embed the words of the target sentence. Second, these word embeddings are positionally encoded, and then fed to a masked self-attention layer of the gloss to text translation part to extract contextual information. The self-attention layer is similar to the one utilized in the sign to gloss recognition part, but it is masked to ensure that context was only modeled between previous words. Third, the extracted representations are combined with the spatio-temporal representations previously learned from the sign to gloss recognition. It is then given to the encoder and decoder module that learns the mapping between the video frames and the output text. Lastly, the output of the encoder and decoder module is passed through a feed forward layer that learns to generate one word at a time

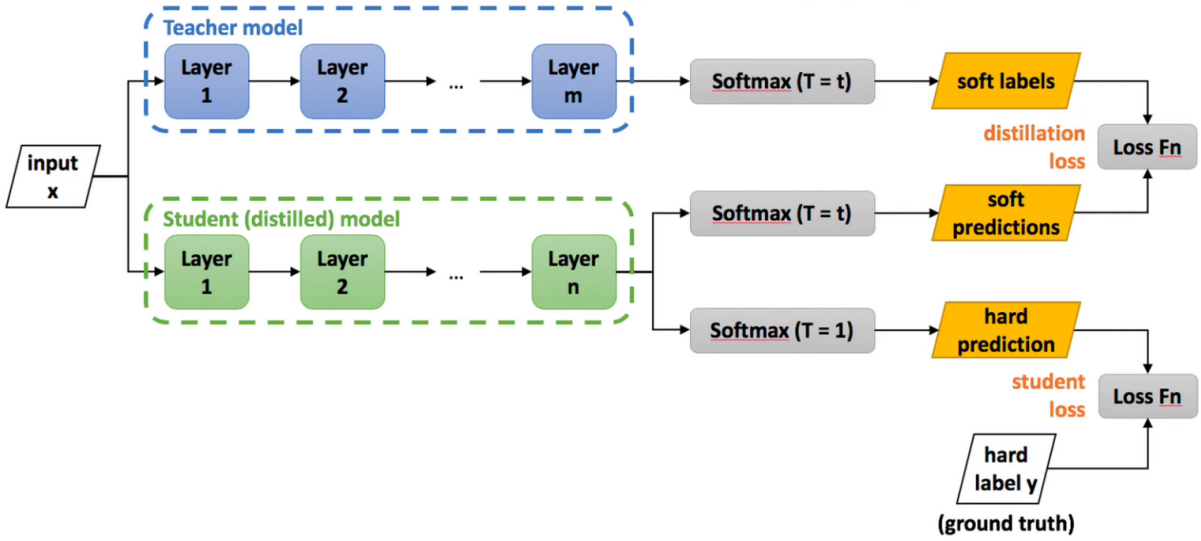

Figure 1: Knowledge Distillation Architecture (Sachdeva, 2023)

until it produces an <EOS> token which signifies the end of a sentence.

For this study, a dimension size of 512 and Xavier initialization (Glorot and Bengio, 2010) is used for both the spatial and the word embeddings. Three transformer layers with 8 heads each are used for the encoder and decoder, while the feed forward layer has a size of 2048. This teacher model is trained using the augmented FSL-NMS dataset (Rivera and Ong, 2018b) and has served as the foundation for distilling knowledge to the student model.

The teacher model, with its greater capacity, is initially trained on a given task, producing "soft targets", which are the output probabilities that contained nuanced information about the inter-class relationships learned by the model. A key aspect of the soft targets generation process is temperature scaling, which is introduced via a temperature parameter T in the softmax function. This parameter controls the "softness" of the probability distribution over classes. A higher temperature can lead to a softer distribution, which is crucial for aiding the student model's learning from the teacher's outputs (Hinton et al., 2015).

Utilizing the trained teacher model, soft targets are generated by processing the dataset through the teacher model and applying temperature scaling to the softmax function as shown in Equation 1, where q_i is the softened probability for class i, z_i is the logit for class i, and T is the temperature parameter.

$$q_i = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)} \quad (1)$$

These softened probabilities provided a richer signal than hard labels alone, allowing the student model to learn more effectively.

4.2 Student Model Design and Training

Designing the student model involved determining the appropriate architecture that balanced performance with computational efficiency. Inspired by the success of TinyBERT (Jiao et al., 2020), where the student model contained 4 and 6 layers compared to the teacher model's 12 layers, the study proposed starting with a student model with approximately 30% of the teacher model's layers. This served as a starting point, and the architecture could be adjusted iteratively based on empirical performance.

With that, the student model's architecture is adjusted from the teacher model's 3 layers to the student model's 2 layers. Additionally, the student model's embeddings and hidden sizes are reduced from 512 to 256, and the feed-forward size is reduced from 2048 to 1024, all aimed at simplifying the student model while maintaining performance. The difference between the student and teacher model is summarized in Table. 2.

The student model is trained on the same dataset. It is trained not only on the hard targets (the actual labels) but also to mimic the soft targets produced by the teacher model. This is achieved through a

	Teacher	Student
Layers	3	2
Embedding Size	512	256
Feedforward Size	2048	1024

Table 2: Number of Parameters for the student and teacher model

loss function that combined the traditional loss (i.e. cross-entropy with the hard targets) with a distillation loss that measured the discrepancy between the soft targets of the teacher and student models (Hinton et al., 2015). The traditional cross-entropy loss is shown in Equation 2, where y_i is the true label and p_i is the predicted probability for class i.

$$L_{CE} = -\sum_i y_i \log(p_i) \tag{2}$$

The distillation loss is often computed using the Kullback-Leibler divergence between the softened outputs of the teacher and student models, which is defined as shown in Equation 3, where q_i^T and q_i^S are the softened probabilities for class i from the teacher and student models, respectively.

$$L_{KD} = \sum_i q_i^T \log\left(\frac{q_i^T}{q_i^S}\right) \tag{3}$$

The overall loss function is a weighted sum of the traditional loss and the distillation loss, shown in Equation 4, where α is a hyperparameter that balanced the two loss components, and T^2 is a scaling factor for the distillation loss.

$$L = \alpha L_{CE} + (1-\alpha)T^2 L_{KD} \tag{4}$$

4.3 Hyperparameter Tuning

Optimizing the student model's performance hinged on the careful tuning of hyperparameters. Key parameters such as temperature, hard label weight, and the loss weight for different types of knowledge are crucial in refining the distillation process (Lu et al., 2022).

Grid search is initially applied with the original dataset (i.e. no data augmentation performed yet) to find the optimal combination of temperature (T) until the highest performance is achieved on the validation set. The initial values of the temperature range from 1.5 to 3.0 with an interval of 0.5, while the alpha is set to 0.5. The search was not started from $T = 1$ anymore as it indicates no temperature scaling at all. The initial results indicated the top three T for further analysis are 1.5, 2.5, and 3. These values are then used for training on the augmented dataset.

After the temperature yielding the highest performance is determined, grid search is applied with the augmented dataset to find the optimal alpha (α). The values of α range from 0.3 to 0.7 with an interval of 0.2. This method provided a practical yet effective means of hyperparameter tuning.

The student model is iteratively trained with different values of T and α, then its performance is evaluated on the validation set. The combination of T and α that yielded the highest performance is then selected for the final student model.

4.4 Evaluation and Iteration

The teacher and student model's performances are evaluated using separate test sets. The Bilingual Evaluation Understudy (BLEU) metric is used for evaluation to measure the quality of machine-translated output. The ROUGE metric is also employed to measure the recall of the student model.

The results of the teacher and student model are then compared to analyze the impact of the proposed knowledge distillation methods. If the performance, measured by both BLEU and ROUGE, did not meet the desired criteria, iteration on the previous steps and refinement of the student model's architecture and hyperparameters are re-conducted.

5 Results and Discussion

Significant adjustments are made to adapt the student model for efficiency. This included reducing the number of layers, embedding size and hidden layer sizes of the student model compared to the teacher model. These modifications aim to create a model with reduced capacity, optimizing it for efficiency while striving to maintain performance levels of the translation. The performance of the translation model is measured by using BLEU and ROUGE, while model compression is measured by using compression ratio.

5.1 BLEU and ROUGE Performances

As shown in Table 3, the combination of the hyperparameters $T = 3, \alpha = 0.5$ yielded the highest BLEU and ROUGE scores among the top three combinations from the hyperparameter tuning that is initially performed on the original dataset. In line with this, further experiments are conducted with the nearby hyperparameters, $T = 3, \alpha = 0.3$

Model	T	α	Set	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE
Teacher	**-**	**-**	**DEV**	**21.69**	**17.5**	**10.88**	**11.38**	**22.98**
			TEST	**22.89**	**18.44**	**11.19**	**10.22**	**24.46**
Student	1.5	0.5	DEV	22.37	17.84	11.91	10.3	22.62
			TEST	22.79	18.3	12.1	9.81	23.53
Student	2.5	0.5	DEV	22.87	18.49	12.58	10.9	23.59
			TEST	23.9	19.25	12.59	9.66	25.24
Student	**3**	**0.5**	**DEV**	**23.29**	**18.72**	**12.93**	**11.67**	**23.2**
			TEST	**25.61**	**21.25**	**16.03**	**14.84**	**25.69**
Student	3	0.7	DEV	21.19	17.13	11.07	12	22.68
			TEST	22.9	18.75	12.33	13.05	24.98
Student	3	0.3	DEV	22.15	16.76	11.45	9.69	23.13
			TEST	22.68	17.2	11.91	10.51	23.44
Student	3.5	0.5	DEV	21.4	17.23	10.61	11.69	23.08
			TEST	22.43	18.15	10.68	10.65	24.59

Table 3: BLEU and ROUGE Scores of the Student and Teach Models on the developement set (DEV) and test set (TEST) using different T and α. BLEU-n scores reflect the model's precision in matching n-grams to reference translations from single words (BLEU-1) to four-word phrases (BLEU-4). ROUGE assesses recall, showing how well the model captures the reference's n-grams.

and $T = 3, \alpha = 0.7$ focusing on α, as well as increasing the temperature to $T = 3.5$ with $\alpha = 0.5$ to explore potential improvements. These explorations aimed to determine if a slight adjustment in α or an increase in T would enhance model performance even further.

The combination of $T = 3, \alpha = 0.5$ performed better in terms of BLEU scores across the different combinations tested and compared to their respective teacher models' results. The utilization of temperature T in the softmax function for knowledge distillation is pivotal in the experiments. A higher T led to a softer probability distribution, crucial for effective knowledge transfer from the teacher model to the student model. This is particularly evident in the improvements in BLEU scores with $T = 3$, demonstrating that a softer distribution can enhance learning in more complex configurations.

The cumulative BLEU score provides a single, comprehensive measure of translation quality. As shown in Table 3, the student model exhibits higher cumulative BLEU scores across all n-gram levels compared to the teacher model, indicating a more robust performance. Its BLEU-1 (BLEU for 1-gram) score of **25.61** in the TEST set suggests a more effective word matching, while its BLEU-4 (BLEU for 4-gram) score of **14.84** shows stronger performance in generating accurate four-word sequences compared to the teacher model.

Both the teacher model and the student model demonstrate strong performance with more frequent and shorter n-grams, particularly 1-grams and 2-grams. The teacher model is able to correctly predict the following words across different sentences: 'i', 'am', 'you', 'are', 'so', 'slow', 'not', 'fine', 'shocked', 'my', 'worried', 'nervous', and 'tired'. The student model is able to correctly predict the same set of words except 'my', but with the addition of the following words: 'old', 'proud', 'of', '12', 'years'. Majority of these words have higher frequency across different sentences. Sentences with a combination of these words also has higher accuracy than sentences that are composed of words that do not frequently appear in the dataset. This explains why its accuracy diminishes as the n-gram length increases, indicating a need for further training and exposure to a broader variety of sequences. Incorporating more diverse and complex n-grams into the training dataset could improve the model's robustness and accuracy across different n-gram lengths.

The better performance of the student models compared to the teacher model, as observed in the experiments, can be traced back to several factors integral to the distillation process itself. First, knowledge distillation efficiently transfers "soft target" from the teacher to the student model, not only reducing over-fitting, but also acts as a form of regularization, optimizing error learning from the teacher model and preventing the student from becoming too confident prematurely. Second, the student model inherits robust features from the teacher,

facilitating a more streamlined learning process. Third, the student models often show enhanced adaptability to specific tasks or datasets, thanks to tailored adjustments like the softmax temperature, focusing learning on task-relevant aspects of the data. These collective advantages contribute to the distilled models' improved performance in terms of accuracy, robustness, and efficiency, underlining the value of knowledge distillation in resource-constrained environments.

5.2 Compression Ratio

While the translation performance of the student model showed favorable results compared to the teacher model, it is also important to measure the compression ratio. This can show if the model size is reduced, while maintaining performance.

Results revealed a significant reduction in the model size from the original teacher model to the compressed version, the student model. The file size of the teacher model is 320.98 MB, It represents a baseline for performance but is impractical for deployment in memory-limited environments. In contrast, the student model is compressed to 59.33 MB. This indicates a **5.4** compression ratio, indicating effective compression without compromising the model's utility.

This drastic reduction showcases the potential of advanced model compression techniques, such as quantization and pruning, which are essential for deploying deep learning models on mobile and embedded devices.

6 Conclusions and Recommendations

This research marks a significant breakthrough in Filipino Sign Language (FSL) recognition and translation, employing Tiny Machine Learning (TinyML) to refine and enhance a sophisticated model that integrates 2D Convolutional Neural Networks (CNN) and Transformer Neural Networks (TNN) trained on an FSL dataset of sentences. The potential of TinyML techniques, specifically knowledge distillation, in compressing and improving a large-scale model is shown in the comparison of the teacher and student model performances in terms of BLUE and ROUGE scores, and compression ratio.

The student model achieved a BLEU-4 score of 14.84 and a ROUGE score of 24.46, which is 45% and 5% higher than the teacher model respectively. Although the highest BLEU-4 score of the original 2D CNN and TNN model by Camgoz et al. (2020) adapted in this study is 21.59, the performance of our model is still promising given the use of a relatively small dataset. The augmented FSL-NMS dataset (Rivera and Ong, 2018b) used by our model comprises of 1628 samples which are composed of 2 to 5 words each, while the Phoenix14-T dataset used by Camgoz et al. (2020) comprises of 8257 samples which are composed of 1 to 52 words each. As mentioned in Section 2.1, use of larger datasets can possibly lead to better performances in translation. For an SLT task, the model would benefit more from longer and continuous sentences, as it can learn the context and morphology of the language.

Aside from improved performances in translation, its capability in condensing a large-scale model is evident as the student model has reached a 5.4 compression ratio, with respect to the teacher model. As there are other TinyML techniques as enumerated in Section 2.2, there are still a lot of room for improvements. This study opens the opportunities for future enhancements and deployments of SLT models on mobile, and wearable devices. Looking ahead, this lays a solid foundation for future technological enhancements and deeper integration of the Deaf community into the societal fabric, underscoring the profound societal benefits of inclusive technology.

Acknowledgments

This research is funded by DOST-PCIEERD Project No. 1211355 in cooperation with DLSU-RGMO (Project No. 24N 2TAY22-3TAY23.) and DLSU-Science Foundation, Philippines.

The authors would like to extend their appreciation to Juls Andrada and Joi Villareal for the dataset annotation, and the De La Salle University-College of Computer Studies for providing the tools and facilities.

References

Arra Alberto, Hanna Mangampo, Macario Lou Presto, and Tita Herradura. 2022. Senyas: A 3d animated filipino sign language interpreter using speech recognition.

Arnon Amir, Brian Taba, David Berg, Timothy Melano, Jeffrey McKinstry, Carmelo Di Nolfo, Tapan Nayak, Alexander Andreopoulos, Guillaume Garreau, Marcela Mendoza, et al. 2017. A low power, fully event-based gesture recognition system. In *Pro-*

ceedings of the IEEE conference on computer vision and pattern recognition, pages 7243–7252.

Ed Peter Cabalfin, Liza B. Martinez, Rowena Cristina L. Guevara, and Prospero C. Naval. 2012. Filipino sign language recognition using manifold projection learning. In *TENCON 2012 IEEE Region 10 Conference*, pages 1–5.

Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020. Sign language transformers: Joint end-to-end sign language recognition and translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10023–10033.

Aakanksha Chowdhery, Pete Warden, Jonathon Shlens, Andrew Howard, and Rocky Rhodes. 2019. Visual wake words dataset. *arXiv preprint arXiv:1906.05721*.

Mark Cueto, Winnie He, Rei Untiveros, Josh Zuñiga, and Joanna Pauline Rivera. 2020. Translating an Aesop's fable to Filipino Sign Language through 3D animation. In *Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives*, pages 39–44, Marseille, France. European Language Resources Association (ELRA).

Dr. Lachit Dutta and Swapna Bharali. 2021. Tinyml meets iot: A comprehensive survey. *Internet of Things*, 16:100461.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 249–256, Chia Laguna Resort, Sardinia, Italy. PMLR.

Suyog Gupta, Ankur Agrawal, Kailash Gopalakrishnan, and Pritish Narayanan. 2015. Deep learning with limited numerical precision. In *International conference on machine learning*, pages 1737–1746. PMLR.

Song Han, Jeff Pool, John Tran, and William Dally. 2015. Learning both weights and connections for efficient neural network. *Advances in neural information processing systems*, 28.

Mohammed Mehedi Hassan, Md Zia Uddin, Amr Mohamed, and Ahmad Almogren. 2018. A robust human activity recognition system using smartphone sensors and deep learning. *Future Generation Computer Systems*, 81:307–313.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer. 2016. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and< 0.5 mb model size. *arXiv preprint arXiv:1602.07360*.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. TinyBERT: Distilling BERT for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4163–4174, Online. Association for Computational Linguistics.

Rakhee Kallimani, Krishna Pai, Prasoon Raghuwanshi, Sridhar Iyer, and Onel LA López. 2023. Tinyml: Tools, applications, challenges, and future research directions. *arXiv preprint arXiv:2303.13569*.

Yuma Koizumi, Shoichiro Saito, Hisashi Uematsu, Noboru Harada, and Keisuke Imoto. 2019. Toyadmos: A dataset of miniature-machine operating sounds for anomalous sound detection. In *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 313–317. IEEE.

Chengqiang Lu, Jianwei Zhang, Yunfei Chu, Zhengyu Chen, Jingren Zhou, Fei Wu, Haiqing Chen, and Hongxia Yang. 2022. Knowledge distillation of transformer-based language models revisited. *arXiv preprint arXiv:2206.14366*.

Puranjay Mohan, Aditya Jyoti Paul, and Abhay Chirania. 2021. A tiny cnn architecture for medical face mask detection for resource-constrained endpoints. In *Innovations in Electrical and Electronic Engineering: Proceedings of ICEEE 2021*, pages 657–670. Springer.

Myron Darrel Montefalcon, Jay Padilla, and Ramon Rodriguez. 2023. *Filipino Sign Language Recognition Using Long Short-Term Memory and Residual Network Architecture*, pages 489–497.

Myron Darrel Montefalcon, Jay Rhald Padilla, and Ramon Llabanes Rodriguez. 2021. Filipino sign language recognition using deep learning. ICSET 2021, page 219–225, New York, NY, USA. Association for Computing Machinery.

John Newall, Norberto Martinez, DeWet Swanepoel, and Catherine McMahon. 2020. A national survey of hearing loss in the philippines. *Asia Pacific Journal of Public Health*, 32:101053952093708.

Maria T. Nyamukuru and Kofi M. Odame. 2020. Tiny eats: Eating detection on a microcontroller. In *2020 IEEE Second Workshop on Machine Learning on Edge in Sensor Systems (SenSys-ML)*, pages 19–23.

A. L. A. Ramos, G. D. M. Dalhag, M. L. D. Daygon, J. Omar, K. D. La Cruz, A. A. Macaranas, and K. L. J. Almodovar. 2019. Alphabet hand gesture recognition using histogram of oriented gradients, support vector machine and k-nearest neighbor algorithm. *International Research Journal of Computer Science (IRJCS)*, 6:200–205.

Joanna Pauline Rivera and Clement Ong. 2018a. Facial expression recognition in filipino sign language: Classification using 3d animation units. In *Proceedings of the 18th Philippine Computing Science Congress (PCSC)*.

Joanna Pauline Rivera and Clement Ong. 2018b. Recognizing non-manual signals in Filipino Sign Language. In *Proceedings of the LREC2018 8th Workshop on the Representation and Processing of Sign Languages: Involving the Language Community*, pages 177–184, Miyazaki, Japan. European Language Resources Association (ELRA).

K. Sachdeva. 2023. [knowledge distillation] distilling the knowledge in a neural network. *Medium*. Retrieved on November 20, 2023 from https://towardsdatascience.com/paper-summary-distilling-the-knowledge-in-a-neural-network-dc8efd9813cc.

Isaiah Tupal, Melvin Cabatuan, and Michael Manguerra. 2022. Recognizing filipino sign language with inceptionv3, lstm, and gru. In *2022 IEEE 14th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM)*, pages 1–5.

Ji Wang, Bokai Cao, Philip Yu, Lichao Sun, Weidong Bao, and Xiaomin Zhu. 2018. Deep learning towards mobile applications. In *2018 IEEE 38th International Conference on Distributed Computing Systems (ICDCS)*, pages 1385–1393.

Pete Warden and Daniel Situnayake. 2019. *Tinyml: Machine learning with tensorflow lite on arduino and ultra-low-power microcontrollers*. O'Reilly Media.

Yundong Zhang, Naveen Suda, Liangzhen Lai, and Vikas Chandra. 2017. Hello edge: Keyword spotting on microcontrollers. *arXiv preprint arXiv:1711.07128*.

Zhenxing Zhou, Vincent WL Tam, and Edmund Y Lam. 2022. A portable sign language collection and translation platform with smart watches using a blstm-based multi-feature framework. *Micromachines*, 13(2):333.

Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V Le. 2018. Learning transferable architectures for scalable image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8697–8710.

A Gloss Translations of the Sentences from FSL-NMS

The complete list of sentences and its corresponding gloss translations from the FSL-NMS dataset is shown in Table 4. Gloss translations are literal translations of the word or phrase as they appear when signed, separated by a space. This do not follow the English grammar yet.

Sentences	**Glosses**
John likes Mary.	john likes mary
You are sick.	you sick
Is it new year?	new fireworks
How are you?	you how
How old are you?	age you how much
You are sick!	you sick
I am fine.	fine
I am 12 years old.	old 12 year
Does john like Mary?	john like mary
Happy new year!	happy new fireworks
Good morning!	good morning
Good noon!	good noon
My head is not painful.	headache not
I do not like you.	not like you
I am not tired.	not tired
You are not slow.	you not slow
This is not hard.	not hard
My head is painful.	headache
I like you.	like you
I am tired.	tired
You are slow.	slow
This is hard.	hard
My head is very painful.	headache very
I like you very much.	like you very much
I am so tired.	much tired
You are so slow!	you much slow / much slow
This is very hard.	much hard
I hate you!	hate
You are disgusting!	disgusting
I am scared.	scared
I am nervous.	nervous
I am worried.	worry
I am shocked!	shocked
I saw a ghost.	ghost
Thank you.	thank you
The trip is exciting.	trip exciting/joyful
The show is amazing.	show amazing
I am proud of you!	proud you
Our team won!	class/group win
I am sorry.	sorry
My dog died.	dog die
I am alone.	alone
I am heartbroken.	heartache
I failed the exam.	fail exam

Table 4: Gloss Translations of Sentences from FSL-NMS dataset