Shrinking embeddings, not accuracy: Performance-preserving reduction of facial embeddings for complex face verification computations

Philipp Hofer
Johannes Kepler University Linz
Linz, Austria
philipp.hofer@ins.jku.at

Michael Roland
Johannes Kepler University Linz
Linz, Austria
michael.roland@ins.jku.at

Philipp Schwarz
Johannes Kepler University Linz
Linz, Austria
philipp.schwarz@jku.at

René Mayrhofer
Johannes Kepler University Linz
Linz, Austria
rm@ins.jku.at

***Abstract*—Conventional embeddings employed in facial verification systems typically consist of hundreds of floating-point numbers, a widely accepted design paradigm that primarily stems from the swift computation of vector distance metrics for identification and authentication such as the L2 norm. However, the utility of such high-dimensional embeddings can become a potential concern when they are integrated into complex comparative strategies, for example multi-party computations. In this study, we challenge the presumption that larger embedding sizes are always superior and provide a comprehensive analysis of the effects and implications of substantially reducing the dimensions of these embeddings (by a factor of 29). We demonstrate that this dramatic size reduction incurs only a minimal compromise in the quality-performance trade-off. This discovery could lead to enhancements in computation efficiency without sacrificing system performance, potentially opening avenues for more sophisticated and decentral uses of facial verification technology. To enable other researchers to validate and build upon our findings, the Rust code used in this paper has been made publicly accessible and can be found at https://github.com/mobilesec/reduced-embeddings-analysis-icprs.**

***Index Terms*—computational efficiency, embedding reduction, data quantization, decentralized**

I. Introduction

To obviate the need for individualized retraining of biometric systems for every (new) person, modern biometric systems generate embeddings, numerical representations that succinctly capture the unique features of a biometric trait, such as a face. State-of-the-art facial verification algorithms typically employ high-dimensional floating-point values for their embeddings: 4,096 [14], 2,048 [2], 512 [3], [10], [12], and 128 dimensions [15].

These high-dimensional embeddings have proven incredibly useful in facial verification and recognition systems. The use of numerous floating-point numbers optimizes verification accuracy and ensures high computational efficiency, contributing to their broad acceptance as an industry standard.

Despite the unparalleled accuracy of these embeddings in state-of-the-art facial verification systems, there is a growing motivation to reduce their size for three primary advantages: (1) Reduced-size embeddings significantly enhance multi-party computation capabilities. Systems like Funshade [9] efficiently calculate whether the distance between two embeddings is below a threshold without revealing the actual embeddings, ensuring privacy and efficiency. (2) Improved transmission efficiency, especially in environments not reliant on traditional TCP connections. Specifically, embeddings compact enough to fit within a 509-byte Tor cell [6] can be transmitted more swiftly. Furthermore, the necessity for embeddings to be small enough for inclusion in modified Tor introduction packets, as detailed by recent research [8], highlights their importance in scenarios with strict data size constraints. Consequently, smaller embeddings offer significant advantages in data transfer speed and efficiency, particularly beneficial in settings with limited bandwidth or data capacity. (3) Reduced storage requirements, which is especially beneficial for contexts with limited space, such as smart cards. Smaller embeddings allow for more efficient space utilization and increase storage capacity, enhancing device utility and application scope.

Our study investigates how reducing the embedding size affects facial verification system performance, focusing on the trade-offs between efficiency, privacy, accuracy. We aim to provide a detailed understanding of the practical implications of optimizing embedding sizes for better computational efficiency and system performance.

We challenge the common belief that larger embedding sizes always yield better results in facial verification systems by

979-8-3503-7565-7/24/$31.00 ©2024 IEEE

significantly reducing these dimensions.

Our hypothesis suggests that while drastically reducing the embedding size may not proportionally decrease performance, it could enhance computational efficiency. This could allow for more complex comparison functions, such as multi-party computations, potentially improving the decentralization of biometric systems.

In our investigation, we explore two options for embedding reduction: (1) reducing the number of elements within an embedding (dimension reduction) and (2) utilizing smaller data types to represent the elements. Both strategies come with their inherent advantages and potential drawbacks. Dimension reduction may allow for substantial computational savings, but it may also compromise the richness of the data represented. Using smaller data types can similarly reduce computational demand, yet it raises the concern of losing precision and increasing quantization errors.

The following two sections will delve into each of these approaches in detail. We aim to illuminate the consequences and benefits of these strategies and ultimately determine whether the trade-off between efficiency and performance is viable.

II. Related work

The exploration of efficient and compact biometric embeddings is part of the larger field of neural network optimization and model compression. While the specific focus on reducing the size of biometric embeddings is underrepresented in current literature, the extensive research into neural network model minimization offers valuable insights and methodologies that are applicable to this challenge. This chapter provides a focused summary of selected key techniques in model compression, highlighting their relevance and possible applications in shrinking biometric embeddings. The citations included are representative and not exhaustive, aiming to introduce the most significant and pertinent contributions to this area of study.

1) Pruning and Sparsity: One of the primary methods in model compression is pruning, which involves the removal of redundant or non-critical parameters from a neural network. Research by Yang et al. [17] demonstrates a novel approach to enhance neural network efficiency. They introduce a low-cost technique using winners-take-all dropout to regulate dynamic activation sparsity, leading to structured activation sparsity with improved levels. This method, when combined with weight pruning, shows significant runtime speedups with minimal accuracy loss, underscoring the potential of pruning in neural network optimization. Furthermore, Shao et al. [16] propose a dynamic scheme for imposing sparse constraints based on filter weights. Their method demonstrates superior pruning performance, achieving substantial reductions in parameters and computational costs. These studies collectively highlight the significance of pruning and sparsity in enhancing the efficiency of neural networks, a concept that can be transferred to the optimization of biometric embeddings.

2) Quantization: Quantization, another key technique in model compression, involves reducing the precision of the network's parameters. Marinò et al. [13] explore this concept and propose a novel lossless storage format for CNNs leveraging both weight pruning and quantization. Their findings indicate that such compression techniques can drastically reduce the space occupancy of neural networks while maintaining competitive performance levels. This approach is directly applicable to biometric embeddings, as it entails representing data with fewer bits, suggesting that lower precision may be sufficient for maintaining the integrity of biometric data.

A. Datasets

In this study, we utilize two distinct datasets, each with its unique characteristics which are described in this section.

1) LFW: The Labeled Faces in the Wild (LFW) dataset [7] is a well-known, public collection designed for testing face verification technologies under uncontrolled conditions. It contains 13,233 images of 5,749 individuals, showcasing a variety of challenges such as differences in lighting, pose, and expression. Despite this, many images are portrait-like with consistent lighting, making some aspects of the dataset easier for advanced verification systems, which can achieve high accuracy levels. The dataset includes a validation subset of 6,000 image pairs, highlighting its value for developing and benchmarking face verification methods.

2) CPLFW: The Cross-Pose LFW (CPLFW) dataset [18] provides a more challenging environment for testing face verification technologies due to its focus on pose variations and diverse conditions, including lighting and expressions. Featuring over 11,652 images of 3,000 individuals, CPLFW offers a rich diversity of non-ideal scenarios that significantly diverge from the mostly portrait-like images in the LFW dataset. This complexity, especially in pose variation, makes achieving high accuracy more challenging for face verification models. CPLFW also includes a validation subset of 6,000 image pairs to facilitate detailed assessments, paralleling the LFW dataset's structure.

III. Element reduction

Under ideal circumstances, the elements within a biometric embedding would exhibit a balanced distribution, where all elements contribute equally, implying a potential compromise in accuracy should dimensionality reduction occur. This section seeks to understand the impact of reducing these dimensions on model performance.

This study began by evaluating 6,000 test pairs from the LFW dataset using the L2 norm as the distance metric, selected for its widespread use and effectiveness in similar research. Hofer et al.'s suggestion that the choice of distance metric is not crucial supported the decision to use the L2 norm, given its proven efficiency in related empirical studies [5].

To evaluate facial verification models' verification, a threshold is established. Embeddings for face pairs are calculated, and their L2 distance is measured. Pairs are then classified as the same person if the distance is below the threshold, or different individuals if above.

We optimized the threshold to minimize both false positives and negatives by systematically testing every threshold value

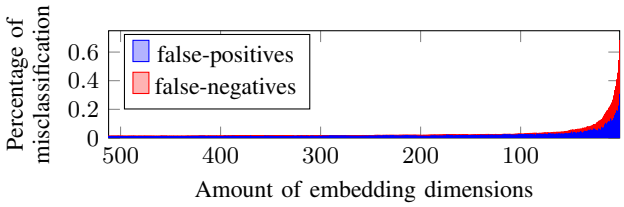

Fig. 1. The error rate on the LFW dataset correlates with embedding dimensionality, rapidly converging to 40/6000 errors. Using 100-dim. embeddings results in slightly more errors (69) than the full 512 dimensions (40).

TABLE I
BRUTE-FORCE SEARCH OF THE BEST ELEMENTS IN THE LFW DATASET.

Iteration	Elements	Percentage of Misclassification
1	16	32.2%
2	16, 31	25.2%
3	16, 25, 31	20.9%
4	16, 25, 29, 31	17.7%
5	14, 16, 28, 29, 31	15.8%
6	14, 15, 16, 28, 29, 31	14.1%
7	1, 3, 16, 17, 24, 26, 31	12.4%

that altered at least one outcome. For example, if the set of distances were 1.0, 1.2, 2.0, we evaluated threshold values as 1.1 (between 1.0 and 1.2), and 1.6 (between 1.2 and 2.0) to ensure comprehensive coverage and precise adjustments.

Tests on three face detection and two verification models showed consistent trends, with the combination of RetinaFace and ArcFace being the most effective. Therefore, in this paper, we will employ this combination, which utilizes an embedding comprised of 512 dimensions of 32-bit floating points. This establishes our baseline: These models achieved 99.3 % accuracy on the LFW dataset, using all 512 dimensions, where accuracy is defined as the ratio of correct predictions (true positives and true negatives) to the total number of predictions.

We examined the accuracy impact of using lower-dimensional embeddings by sequentially removing elements and recalculating the error rate and optimal threshold for each reduced dimension. This process continued until a single-dimensional embedding was reached, illustrating the accuracy trade-offs at each reduction stage. Despite its impracticality, a single-dimensional embedding was included to fully represent the effects of dimensionality variations. The outcome of this iterative process is depicted in Fig. 1.

The findings indicate an excess in embedding dimensions, with a reduction in embeddings not leading to a significant increase in errors initially, suggesting possible data streamlining without major performance loss. Further robustness checks, involving 100 reruns with randomly selected indices on sets with 7, 32, 120, and 200 dimensions, confirmed the initial observation's consistency across different dimensions, underscoring the likelihood that many facial verification systems operate with unnecessary data. These specific dimensions were selected for further investigation due to their intriguing characteristics observed in the raw data presented in Fig. 1. This consistency adds weight to our initial finding: many facial verification systems likely carry more data than necessary.

Some index subsets perform better due to lower error rates, but small differences among all tested combinations suggest that choosing a specific subset may not greatly affect the outcome. Still, steady performance across 100 random indices does not rule out the possibility of an optimal set.

In order to verify the existence of this optimal set, we identify the subset with the lowest error rate, within our experimental framework, requiring evaluation of all combinations. However, enumerating all $\sum_{n=1}^{512} \binom{n}{512}$ combinations is computationally infeasible.

Analysis of data in Fig. 1 shows using only the first 32 indices yields a 96.1 % accuracy, close to the 99.3 % accuracy achieved with all 512 indices. This highlights the effectiveness of our simplified model. Guided by these insights, we resolved to scrutinize every possible combination encapsulated within these initial 32 elements. It presents a viable opportunity to conduct an exhaustive exploration while retaining the potential to yield a satisfactory degree of accuracy.

It is crucial to note that we approached this analysis with a holistic view of all subsets' potential combined performances. The performance of a particular index, for example, index 16 in an initial round, does not necessarily dictate a superior result in subsequent rounds. It is entirely plausible that two separate indices, despite their individual performances not reaching the same peak as index 16, could in conjunction yield a superior outcome.

To account for these variations in possible performance, reliance on results from prior iterations is avoided. Each new round commences with a fresh, comprehensive exploration of all possible subsets. This approach enables the identification of any advantageous combinations that might otherwise be overlooked if merely relying on preceding results.

Our consumer-grade hardware (Intel Core i7-10510U) processes our pipeline at ∼25 iterations per second with a reasonable low power consumption of 15 W. The most effective subset and its corresponding error rate for each iteration are reported in Table I. The seventh iteration took 35 hours, with the eighth and ninth projected to take 116 and 311 hours, respectively. Extrapolating, an exhaustive analysis of all 32-element subsets would take an estimated 5.5 years on this setup.

Our code is yet to be optimized. Dedicated optimization efforts could significantly reduce computation time. A complementary strategy is employing a greedy search instead of a full grid search. A greedy search iteratively adds the best element to an optimal set, reducing the search space to $32 * n$, where n is the amount of elements, and enabling completion within minutes. This non-exhaustive method offers a practical solution with lower computational demand.

To evaluate the effectiveness of greedy search compared to exhaustive brute-force search, we conducted a comparative analysis. Results for the first seven elements suggest that greedy search can effectively substitute for brute-force search, as shown in Fig. 2.

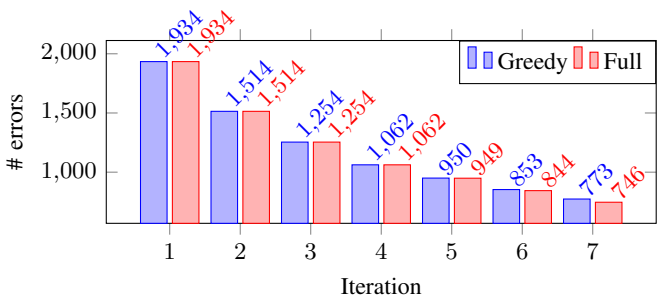

Fig. 2. Error rates obtained from the greedy search are compared with those from the exhaustive brute-force search in the LFW dataset. The first four elements align perfectly; thereafter, the performance begins to exhibit a slight decline. Nevertheless, the similar shape suggests that the greedy search serves as a satisfactory proxy.

Subsequently, in our follow-up experiment, we utilized the top-performing indices identified by the greedy search instead of the initial elements. This modified approach produced promising results, achieving an accuracy rate of 96.1 % with just 32 elements. This performance is notably close to the original 99.3 % accuracy achieved using all 512 elements, illustrating the potential of the greedy search method in reducing computational load while maintaining a high degree of accuracy.

A significant benefit of the greedy search method is its flexibility, as it is not confined to 32 elements and can instead evaluate error rates across all 512 dimensions efficiently. While the absolute values may be lower, the trend observed in this greedy-search setting closely aligns with that seen in Fig. 1. Additionally, it is important to note a slight but noticeable increase in the error rate beyond the 230th index marker. This suggests that the presence of certain elements may indeed be detrimental to the performance of face verification. Such an inference reiterates the notion that a reduction of the embedding size, particularly during the training phase, may even prove beneficial in enhancing accuracy.

The results of the greedy search reinforce our observations from Fig. 1, confirming that a significantly high level of accuracy can be maintained with a reduced subset of elements. To validate the robustness and applicability of our findings, we employed the greedy search method on the more demanding CPLFW dataset, characterized by its array of complexities including unfavorable angles and diverse lighting conditions, as depicted in Fig. 4.

While the graphical representation shares notable resemblance to Fig. 1, an increased optimal error rate, reflective of the greater complexity inherent in the CPLFW dataset, is observed. However, a meticulous examination reveals that the greedy algorithm selects distinct elements for each dataset. Even so, employing a rank-1 approach reveals that the top-performing index from the LFW dataset can still deliver substantial results on the CPLFW, albeit not necessarily as the foremost choice.

To evaluate the cross-dataset applicability of these selected indices, the impact of each on the resulting L2 distance is evaluated. For indices increasing the error in classifying two images of the same individual (indicating a misclassification), the respective index value is increased by the associated distance. Conversely, for indices that err in distinguishing between different individuals (indicating a correct classification), the index value is reduced by the corresponding distance. Thus, a higher value of a particular index reflects its contribution to an increased overall classification error, demonstrating its propensity to introduce “wrongness” into the total distance metric. Finally, to standardize the results and enable a balanced comparison, all index values are normalized to a range between 0 and 1. The distribution of the first 32 indices is depicted in Fig. 5, providing insights into their respective influence on classification accuracy.

Interestingly, and contrary to initial expectations, the indices that minimally contribute to the error differ from those identified by the greedy search. For instance, despite index 16’s superior performance in the greedy search, it is ranked among the least effective in the heatmap. This discrepancy could potentially be explained by a uniform contribution of these indices to the total error, rendering the selection of a specific index less crucial.

To further validate these observations and compare the relative effectiveness of various configurations, we assessed the performance of the greedy search relative to other methods, including the use of initial elements and random selection. The comparative results are detailed in Fig. 3. Moreover, an examination of each index’s individual contribution shows that the choice of the starting index has a negligible impact on the overall error.

IV. Data Quantization

As the field of machine learning evolves, researchers are exploring efficient ways to compress and optimize models. Data quantization is a key technique that reduces data size and complexity by mapping inputs to fewer outputs, improving storage and processing efficiency with little impact on performance. The concept of data quantization is established in machine learning. Liang et al. [11] showed that quantization might not significantly affect system performance. Our study differs by focusing on output quantization, specifically of facial embeddings. This chapter will cover quantization techniques, their impact on facial verification systems, and their implementation to minimize facial embeddings size with minimal performance loss.

The concept of data quantization for optimization is not novel in the machine learning field. Liang et al. [11] demonstrated that quantization inside models might not have a significant impact on the overall system performance. However, it is crucial to note that this work focused on quantizing elements inside the neural networks, whereas our study focuses on the quantization of the output, specifically the facial embeddings.

In this chapter, we will discuss quantization techniques, their effects on facial verification systems, and how they can be implemented to reduce the size of facial embeddings with minimal compromise on performance.

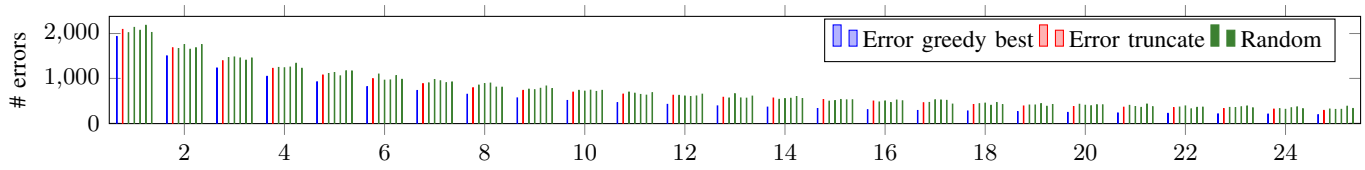

Fig. 3. Comparative analysis of the greedy search against our other configurations (initial and random elements). X-axis: Amount of dimensions used

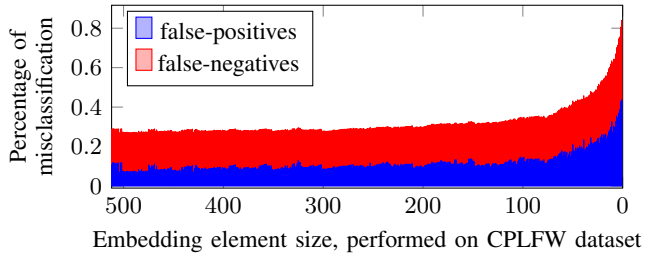

Fig. 4. The shape of the error rate on the CPLFW dataset (shown here) is similar to the error rate of LFW dataset (Fig. 1).

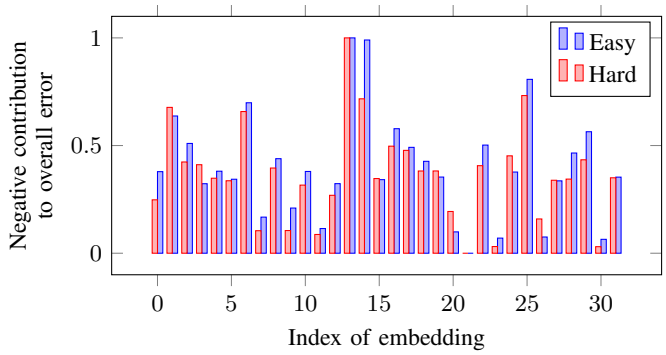

Fig. 5. Cross-dataset index evaluation: Analysis of index contributions to overall classification error in L2 distance metrics. A larger difference between the two bars signifies a higher degree of classification inaccuracies related to a specific index. The blue bar represents a visualization of the LFW dataset, whereas the CPLFW dataset is depicted in the red bar. Notably, the index contributions to the total distance appear to exhibit striking similarities across both datasets.

This study aligns with the element reduction approach outlined in Section III, targeting the optimization of facial embeddings through distinct methods. Element reduction eliminates redundant dimensions, whereas data quantization enhances value representation efficiency. In our approach, we assume uniform quantization for the embeddings, where each level of quantization uniformly represents a segment of the input data range, simplifying the complexity and ensuring more predictable effects on system performance. Section V examines their synergistic potential to efficiently minimize facial embeddings without compromising verification system quality and performance.

Quantization of facial embeddings involves converting the 32-bit floating-point values to alternative data types. This study assesses the impact of such conversions on error rates, emphasizing the importance of precision beyond optimal fixed point range in filter design, as small differences might have significant impact, especially if they are near the threshold

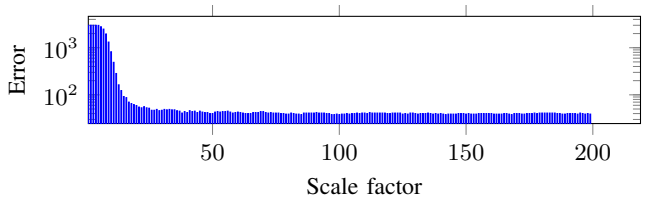

Fig. 6. Visualization of different scale factors. The optimal threshold, which minimizes the combined rate of False Positives (FP) and False Negatives (FN), is dynamically recalculated for each respective scale factor.

range. The goal is to analyze the balance between data compactness and accuracy.

The original 32-bit floating point values fall within a relatively constrained range: approximately -0.2 to $+0.2$. To investigate the contribution of the float mantissa and exponent to the overall information conveyed, we employ calibration. We multiply the original values by a range of factors (between 1 and 200) and subsequently, convert the scaled values into an integer format (which incurs loss of information).

Fig. 6 illustrates the relationship between each scaling factor and the corresponding error rate, providing a detailed overview of the quantization impact.

Analysis of error rates across different scaling factors reveals a plateau effect commencing at a calibration value of 70. Beyond this point, we observe no significant decline in the error rate, as evidenced by a marginal difference in verification accuracy (99.32 % as opposed to the original 99.33 %).

The range of the resulting values with a scale factor of 70 spans from -19 to 21. This suggests the adequacy of a 6-bit signed integer datatype for representing our data post-scaling.

An alternative approach involves adjusting the scaled values with an offset of +19, thus repositioning the range to span from 0 to 40. Consequently, the values can be efficiently represented using an unsigned 6-bit integer datatype.

V. Proposed Pipeline

We suggest a pipeline that operates on only 70 indices determined by the heatmap of each index over the LFW dataset (which can be found in *final_indices.txt* in the accompanying Git repository) and cast the embeddings to an 8-bit integer format (as there is no 6-bit integer data type in most programming languages) with a calibration value of 70.

This modification leads to a reduction of over 29 times the bit requirement (from 16,384 bits to 560 bits), with only a slight decrease in accuracy (99.3 % to 98.6 %), corresponding to a net increase of 44 errors (out of 6,000 comparisons).

When evaluated on the challenging CPLFW dataset, we observe a slightly larger reduction in accuracy (from 85.4 % to

79.87 %), resulting in an increase of 331 errors (out of 5,964 comparisons).

Nonetheless, this approach preserves the computational efficiency, exhibiting a theoretical reduction in computation by a factor of 29.

Given that we are using only 6 out of the 8 available bits, we could apply a more practical approach to leverage the standard 8-bit hardware platforms. By encoding the 70 sets of 6 bits into approximately 53 sets of 8 bits, we optimize the use of existing storage and computational capacity. This adjustment could notably enhance storage efficiency up to a factor of 38 (424 bits / 16,384 bits), while preserving the precision of the results. This strategy enables us to make more efficient use of hardware capabilities without compromising the accuracy of our computations.

An additional advantage of this compact size is its compatibility with an SHA-512 hash function due to similar sizes, presenting potential benefits for specific applications. For instance, cryptographic algorithms that operate with such data can readily use these embeddings without necessitating any modifications.

VI. Discussion: Practical Implications of Compact Embeddings

Beyond the sheer academic fascination and the computational benefits lies the real-world applicability of these compact facial embeddings. Given the current trends towards decentralized and edge computing, the reduction in size becomes even more paramount. Edge devices, such as smartphones or embedded devices, often have limited computational power and storage capacities. By employing compact embeddings, we can deploy facial verification capabilities on these devices without overburdening them.

Furthermore, in the realm of security and data privacy, smaller embeddings imply faster encryption and decryption processes. If facial verification data needs to be transmitted over a network, compact embeddings mean fewer data to send, resulting in quicker transmission times and reduced chances of interception.

VII. Future work

The findings of this research trigger several directions for future inquiries. Our investigation elucidated the ability of facial verification systems to retain considerable accuracy even after a significant reduction in embedding size. However, it also raised intriguing questions that warrant exploration to further enhance the effectiveness and efficiency of facial verification technology.

a) Embedding index importance: One aspect that was not comprehensively addressed in this research is the relative importance of the various indices in embeddings. The pivotal question that surfaces here is: why are certain indices more significant than others? Answering this question would require an intricate examination of the contributing factors that make specific indices of the embeddings more important. Unraveling this has the potential to offer insights into the inner workings of face verification models and could potentially guide strategies for further streamlining the embedding process.

b) Constrained output layer for facial verification models: The findings of our current study hint at another possible line of investigation: improving facial verification models by using a smaller output layer for embeddings. Given our demonstration that a significant level of accuracy can be achieved using considerably less computational performance, it might be plausible to hypothesize that if we constrain the output layer of face verification models to fewer dimensions, we can encode the same amount of information using significantly fewer data. This approach, if successful, could lead to the development of leaner, more efficient facial verification models without compromising on their efficacy.

c) Knowledge distillation: Knowledge distillation is a technique where a smaller, more efficient model is trained to mimic the behavior of a larger, more complex model. Hinton et al. [4] introduced this concept, demonstrating that smaller models could achieve comparable performance to their larger counterparts by learning from them. This approach could be adapted for biometric embeddings, where a smaller embedding could be trained to retain the critical information of a larger, more complex embedding.

d) Data formats: Explore additional data formats, including the increasingly popular bfloat16 [1], to assess their impact on computational efficiency and accuracy.

VIII. Conclusion

We conducted a extensive examination of the effects of embedding reduction on the accuracy of facial verification algorithms, particularly by decreasing the number of elements and altering the data type. Contrary to common beliefs that high-dimensionality embeddings are essential, our findings show that high accuracy levels (over 90 %) are achievable even with a significant reduction in the bit-size of embeddings—approximately 29-fold. This discovery holds substantial promise for use in settings with constrained computational and storage capacities. The reduced size of the embeddings facilitates more efficient comparison, storage, and transmission: Smaller, yet effective, embeddings enable efficient comparison in complex facial verification tasks, which is particularly advantageous for decentralized systems. Storage efficiency is enhanced as these compact embeddings require less space, making them ideal for devices with limited storage capabilities like smart cards. Additionally, the reduced size allows for faster data transmission over networks, which is crucial in environments with limited network capacity or where rapid data transmission is essential.

Our methodology was rigorously tested using the challenging CPLFW dataset. The robustness of our findings amidst these complexities underscores the potential for optimization in face verification models through strategic bit reductions. This opens up new avenues for new developments in creating efficient yet high-performing facial verification systems, particularly for use in decentralized applications.

ACKNOWLEDGMENT

This work has been carried out within the scope of Digidow, the Christian Doppler Laboratory for Private Digital Authentication in the Physical World and has partially been supported by the LIT Secure and Correct Systems Lab. We gratefully acknowledge financial support by the Austrian Federal Ministry of Labour and Economy, the National Foundation for Research, Technology and Development, the Christian Doppler Research Association, 3 Banken IT GmbH, ekey biometric systems GmbH, Kepler Universitätsklinikum GmbH, NXP Semiconductors Austria GmbH & Co KG, Österreichische Staatsdruckerei GmbH, and the State of Upper Austria.

REFERENCES

[1] N. Burgess, J. Milanovic, N. Stephens, K. Monachopoulos, and D. Mansell. Bfloat16 processing for neural networks. In *2019 IEEE 26th Symposium on Computer Arithmetic (ARITH)*, pages 88–91. IEEE, 2019.

[2] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE, 2018.

[3] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019.

[4] G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[5] P. Hofer, P. Schwarz, M. Roland, and R. Mayrhofer. Face to Face with Efficiency: Real-Time Face Recognition Pipelines on Embedded Devices. In *MoMM '23: Proceedings of the 21st International Conference on Advances in Mobile Computing & Multimedia Intelligence*. ACM, Dec. 2023. Accepted for publication.

[6] https://spec.torproject.org/tor-spec/preliminaries.html?highlight=msg-len%20preliminaries#msg-len. Tor specifications: Message lengths, 2024. Accessed May 5, 2024.

[7] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[8] T. Höller. *A Privacy Preserving Networking Approach for Distributed Digital Identity Systems*. PhD thesis, Johannes Kepler University Linz, Institute of Networks and Security, Linz, Austria, Oct. 2022.

[9] A. Ibarrondo, H. Chabanne, and M. Önen. Funshade: Functional secret sharing for two-party secure thresholded distance evaluation. *Cryptology ePrint Archive*, 2022.

[10] M. Kim, A. K. Jain, and X. Liu. Adaface: Quality adaptive margin for face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18750–18759, 2022.

[11] T. Liang, J. Glossner, L. Wang, S. Shi, and X. Zhang. Pruning and quantization for deep neural network acceleration: A survey. *Neurocomputing*, 461:370–403, 2021.

[12] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 212–220, 2017.

[13] G. C. Marinò, A. Petrini, D. Malchiodi, and M. Frasca. Compact representations of convolutional neural networks via weight pruning and quantization. *arXiv preprint arXiv:2108.12704*, 2021.

[14] O. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *BMVC 2015-Proceedings of the British Machine Vision Conference 2015*. British Machine Vision Association, 2015.

[15] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[16] T. Shao and D. Shin. Structured pruning for deep convolutional neural networks via adaptive sparsity regularization. In *2022 IEEE 46th Annual Computers, Software, and Applications Conference (COMPSAC)*, pages 982–987. IEEE, 2022.

[17] Q. Yang, J. Mao, Z. Wang, and L. Hai. Dynamic regularization on activation sparsity for neural network efficiency improvement. *ACM Journal on Emerging Technologies in Computing Systems (JETC)*, 17(4):1–16, 2021.

[18] T. Zheng and W. Deng. Cross-pose lfw: A database for studying cross-pose face recognition in unconstrained environments. *Beijing University of Posts and Telecommunications, Tech. Rep*, 5(7), 2018.