Solid-State Electronics 199 (2023) 108500

Contents lists available at ScienceDirect

Solid State Electronics

journal homepage: www.elsevier.com/locate/sse

Hierarchical Mixture-of-Experts approach for neural compact modeling of MOSFETs[☆],[☆☆]

Chanwoo Park[*], Premkumar Vincent, Soogine Chong, Junghwan Park, Ye Sle Cha, Hyunbo Cho

Research & Development center, Alsemy Inc., 34, Seolleung-ro 90-gil, Gangnam-gu, Seoul, 06193, South Korea

ARTICLE INFO

Keywords:
Neural compact model
MOSFET
Mixture-of-Experts
Artificial neural network

ABSTRACT

With scaling, physics-based analytical MOSFET compact models are becoming more complex. Parameter extraction based on measured or simulated data consumes a significant time in the compact model generation process. To tackle this problem, ANN-based approaches have shown promising performance improvements in terms of accuracy and speed. However, most previous studies used a multilayer perceptron (MLP) architecture which commonly requires a large number of parameters and train data to guarantee accuracy. In this article, we present a Mixture-of-Experts approach to neural compact modeling. It is 78.43% more parameter-efficient and achieves higher accuracy using fewer data when compared to a conventional neural compact modeling approach. It also uses 43.8% less time to train, thus, demonstrating its computational efficiency.

1. Introduction

Compact modeling acts as a bridge between device fabrication and circuit design. It has two main goals: computational efficiency and accuracy. Conventionally, to achieve such contradicting goals, analytical approximations and empirical fitting parameters have been used. Threshold-voltage-based models had issues in solving harmonic distortion analysis which took a considerable time for surface-potential-based models to solve [1].

Since the conventional compact models are technology dependent, it could take years to develop new models for upcoming devices. Hence, there is an immediate need for fast and efficient compact model generation. The use of artificial neural network (ANN)-based compact modeling, or neural compact modeling, to increase accuracy and shorten the model generation period has been studied in the literature.

It has been shown that the performance of ANN-based compact modeling can be improved through appropriate data preprocessing and conversion function in [2,3]. Other works tried to solve the nonphysical behavior of simple MLP-based neural compact models by incorporating device physics into network architecture and loss function [4,5]. Despite the promising results, building ANN simply by stacking fully-connected layer to achieve reasonable accuracy can increase the number of parameters, memory usage, and computational cost. Requirements such as large data set and longer training time make them unattractive for circuit simulation applications (e.g. SPICE).

In this paper, we propose Mixture-of-Experts (MoE) to offset the aforementioned limitations by sub-categorizing the problem into different operation regimes and solving them using dedicated experts. Compared to the baseline ANN, our MoE approach showed 78.43% better parameter efficiency, 56.69% less training data, 79.97% reduction in the number of multiply-accumulate (MAC) operations, and 43.8% reduction in training time to achieve the same target mean-squared-error (MSE) of 0.0025. We found that the MoE approach is robust, easily expandable, and efficient for neural compact modeling applications.

2. Mixture-of-experts for neural compact modeling

The Mixture-of-Experts is based on the *divide and conquer* principle, which was first introduced in [6]. The MoE works on the idea that the whole input space can be partitioned into smaller regimes of distinct characteristics. The MoE systems have been applied in various research fields, such as speech recognition [7] and aerodynamic design [8], and their performance has been verified.

The main idea of MoE can also be effective for device modeling if the input regimes of a device have meaningful distinctions as in the

[☆] This work (Grants No. S3031427) was supported by Business for Startup growth and technological development (TIPS Program) funded by Korea Ministry of SMEs and Startups in 2022. This work was also supported by WISET-2022-487 in Korea.

[☆☆] The review of this paper was arranged by Francisco Gamiz.

[*] Corresponding author.
E-mail address: chanwoo.park@alsemy.com (C. Park).

https://doi.org/10.1016/j.sse.2022.108500

Available online 4 November 2022

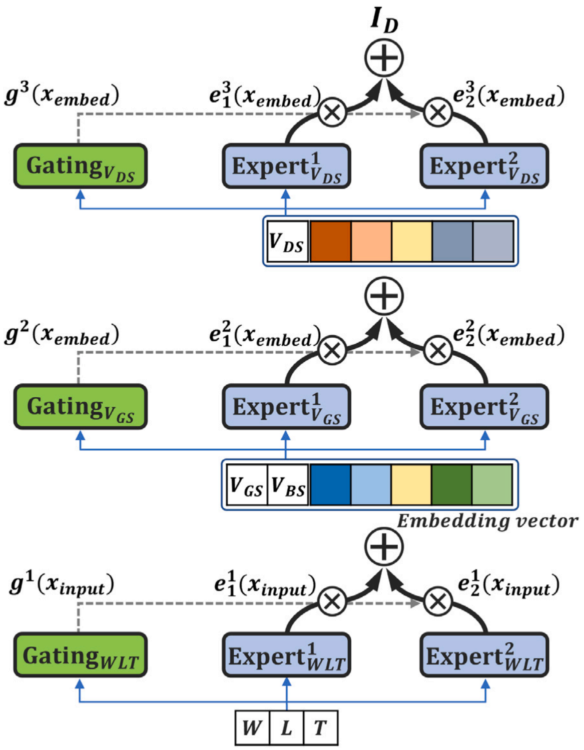


Fig. 1. Mixture-of-Experts based neural compact modeling. The gating network categorizes the input vector and assigns weights to the output of each expert network. The weighted sum of the outputs is passed on to the next level.

case of MOSFETs. For accurate MOSFET modeling, each of the distinguishable regions and their sub-characteristics, including gate-induced drain leakage, short-channel effect, needs to be precisely modeled. If all of the distinct MOSFET characteristics are simultaneously learned via a single large ANN, the training objectives of the sub-regions may conflict with each other, and the performance of the whole ANN-based model can be limited. We address this conflicting sub-tasks optimization by automatically partitioning the device characteristics and assigning specialized local models for each partition.

The entire network consists of three-level mixture-of-experts as shown in Fig. 1. At each level, the outputs of the expert networks are provided with importance weights by their gating network. The gating network learns automatically to assign larger weights to the more appropriate experts based on the input to improve the performance of the entire network. The experts gradually learn to focus more on their own specialized regions. In our experiments, all experts and gating networks are simple feed-forward neural networks, where the output of gating networks is softmax, assigning mixing weights to each expert. We denote $g^k(x)$ and $e_i^k(x)$ as the output of the gating network and the output of the ith expert network at the kth level for a given input vector x, respectively. The final embedding vector $f^k(x)$ at each level is the linearly weighted sum of each expert's output, expressed as follows:

$$f^k(x) = \sum_{i=1}^{N} g^k(x) e_i^k(x) \tag{1}$$

where N is the number of experts at the same level and $\sum_i g^k(x) = 1$. In the hierarchical MoE, the information is processed sequentially from the bottom level, and the additional input vector, which is necessary to assign importance weights to the experts, is fed at each level.

The final embedding vector (weighted sum of each experts' output) is a fixed dimensional embedding vector which contains information on all the modeled MOSFET characteristics. This is passed on to the next gating network where different experts take control for a different input regime.

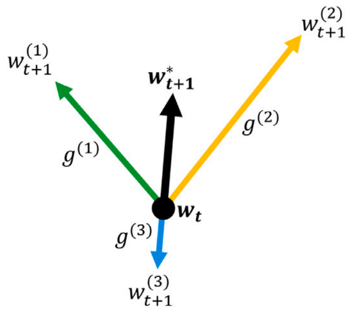


Fig. 2. Illustration of conflicting gradients. The final weight w_{t+1}^* is the average of the other three weights. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

2.1. Analytical formulation

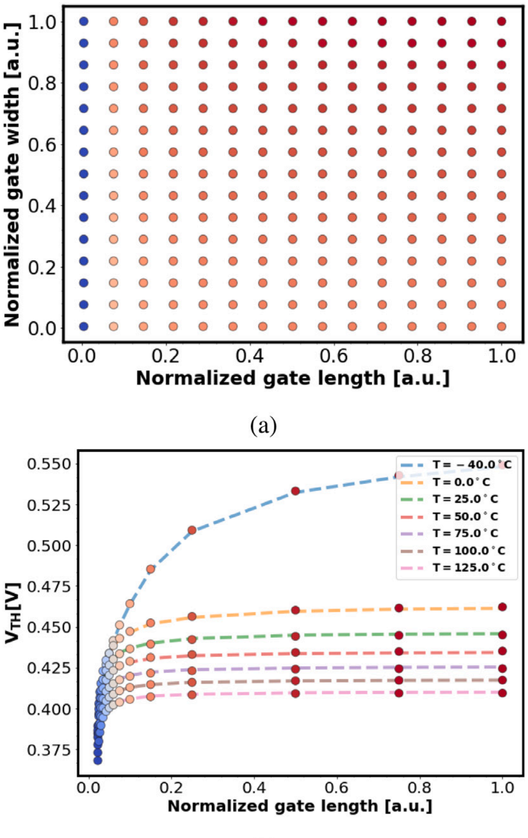
As mentioned above, the input regime of a device consists of very distinctive sub-regions. The training objectives of these unique sub-regions may conflict with each other if trained simultaneously and can lead to a phenomenon called *conflicting gradients*.

$$w_{t+1}^* = w_t + \frac{1}{k} \sum_{i=1}^{k} g^{(i)} \tag{2}$$

We denoted w_t as the parameters of the network at the current training iteration, $w_{t+1}^{(i)}$ as the optimal update for each sub-region, and k is the number of different sub-regions in a mini-batch. We want to find out the direction of the next parameter update by measuring the gradients of the loss function with respect to the weights using samples in a mini-batch. As shown in Fig. 2, let us assume we have three unique sub-regions in one training mini-batch: the optimal gradients for each sub-region can be substantially different. For example, data points from the saturation region of a short channel device would tend to update the whole parameters in the green direction, while another group of data points from the linear region of another narrow width device would prefer to update the network in the yellow direction, and the last set of data points from the cut-off region would direct the weight updates in the blue direction. However, in gradient-based optimization, the update direction for the current iteration is determined by simply averaging the gradients for the samples in a mini-batch, which inevitably leads to sub-optimal update for all sub-regions and eventually degrades the performance of neural compact modeling. This phenomenon is very common in other ML tasks. For examples, in the long tailed problems, gradient from the major class will dominate the gradients from the minor class and the final model will have poor performance on the minor classes. We observe this conflicting sub-task problem can particularly more matter in device modeling because of its unique sub-region nature. Thus, we proposed a new architecture, mixture of experts, using device domain knowledge as an inductive bias and validated it experimentally. Our MoE approach can be more parameter and sample efficient compared to the baseline MLP.

3. Experiments and results

We performed experiments by while changing the structure of the baseline MLP and the proposed MoE method in various ways in order to figure out how the number of parameters and network architecture affect the accuracy, the amount of data required for training, and the training time. The dataset was generated from the SPICE simulation of 45 nm PTM HP model card [9]. In our discussion, $e_1^k(x)$ is denoted by blue, and $e_2^k(x)$ by red. Intermediate shades represent a mixed contribution from both experts.



(a)

(b)

Fig. 3. Representation of categorized regions: (a) W vs L and (b) V_{TH} vs L for short and long gate length regimes.

3.1. *WLT gating*

At the *WLT* gating network, device characteristics were subcategorized according to gate width (W), gate length (L), and temperature (T) by the experts. $e_1^1(x)$ expertized in short-channel effects, while $e_2^1(x)$ was responsible for long-channel lengths as seen from Fig. 3(a). Fig. 3(b) shows threshold voltage (V_{TH}) vs varying gate length, where the decreased short-channel V_{TH} was modeled primarily by $e_1^1(x)$, while the long-channel V_{TH} was modeled by $e_2^1(x)$. The reverse narrow-width effect was negligible, and thus a separate expert was not required. Our MoE approach is robust enough to accommodate additional experts to model more non-linear MOSFET characteristics if required.

3.2. *V_{GS} Gating*

The embedding vector from the *WLT* gate was combined with V_{GS} and V_{BS} to create the input vector for the V_{GS} gating network. $e_1^2(x)$ was designed to take V_{GS} as an input to capture the approximately linear or quadratic dependence of I_D on V_{GS} in the ON-state, while $e_2^2(x)$ was designed to take $\exp(V_{GS})$ as the input to model the exponential dependence of I_D on V_{GS} in the sub-threshold region (Fig. 4). The transition region exhibited a smooth continuous change in prioritizing the expert based on the operation region.

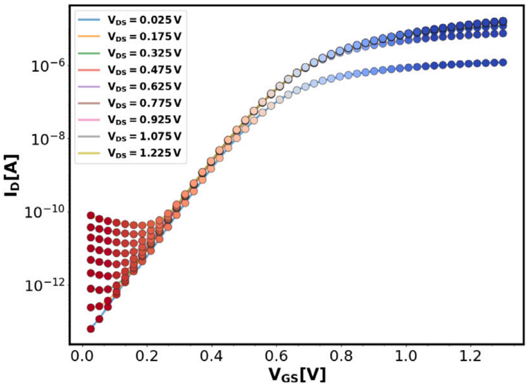


Fig. 4. Categorized $I_D - V_{GS}$ regions: $I_D \propto \exp(V_{GS})$ in OFF region, and $I_D \propto V_{GS}$ in ON region.

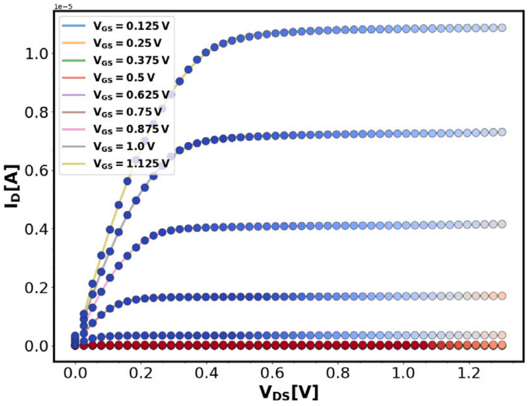


Fig. 5. Categorized $I_D - V_{DS}$ regions: $I_D \propto \exp(V_{DS})$ in the cut-off region, $I_D \propto V_{DS}$ in the linear region, and I_D is weakly dependent on V_{DS} in the saturation region.

Table 1
Conventional ANN vs MoE approach.

	Conventional	MoE	Gain [%]
# of parameters	2049	442	78.43
# of data	47800	20700	56.69
Multiply-accumulate	1792	359	79.97
Training time [s]	66.2	37.2	43.80

*The above data were calculated at MSE = 0.0025.

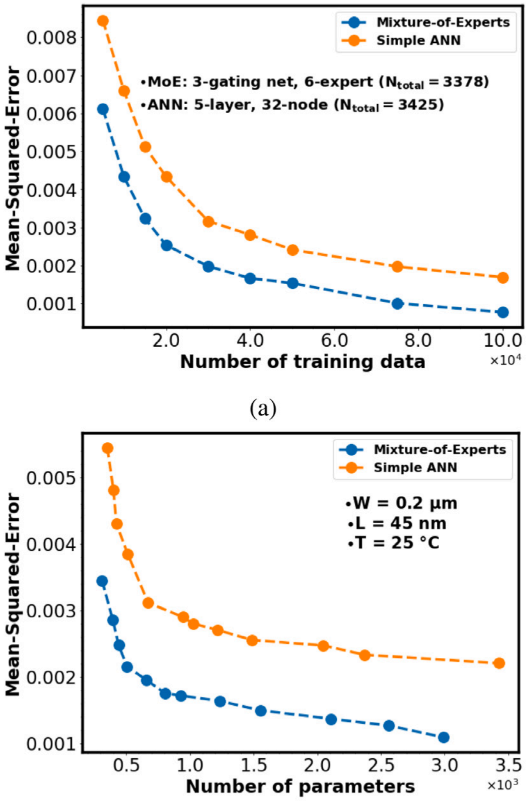
3.3. *V_{DS} Gating*

The final I_D was predicted by the gating network and the experts which were in control of V_{DS} regions, (*i.e.*, cut-off, linear, or saturation region). While I_D in the cut-off and the saturation regions do not heavily depend on V_{DS}, the secondary effects introduce a V_{DS} dependency. While the leakage current in the cut-off region shows a $\exp(V_{DS})$ dependency, channel length modulation, drain-induced barrier lowering, and substrate current induced body effect in the saturation region depend on V_{DS}. Hence, the gating network assigned $e_1^3(x)$ to model the cut-off region and $e_2^3(x)$ to model the linear region, while the saturation region was modeled by both the experts (Fig. 5).

In our method, the choice of gating decisions was continuous rather than discrete. Any changes of physical phenomena, (*i.e.* the linear/saturation regions in the transistor), were not strictly partitioned, in fact, they underwent a continuous transition. Therefore, if two experts respectively controlled the linear and the saturation regions, we designed the gating network to be able to continuously mix the two experts rather than discretely select one of them.

4. Performance comparison

Fig. 6 shows that, with a similar number of parameters (N_{total}), the MoE consistently achieved higher accuracy compared to the baseline ANN with a wide range of training data size. In Table 1, to obtain the target MSE of 0.0025, the baseline ANN required about 2049



(a)

(b)

Fig. 6. Comparison of test accuracy as a function of (a) the number of training data with the similar model capacity, (b) the number of parameters on the same W, L, *T* dataset.

parameters, while our MoE approach required only 442 parameters. This contributed to a 78.43% improvement in parameter efficiency. We also observed that the MoE approach required 56.69% less training data than a baseline ANN. Since the experimental data is scarce in most cases, this is an important distinction. Our approach also showed a 79.97% decrease in the number of MAC operations and 43.8% improvement over the baseline approach in terms of training time.

5. Conclusion

In this paper, we propose a novel MoE structure for neural compact modeling which utilizes the fact that MOSFETs have distinct characteristics for each input region. A parameter-efficient neural compact model is demonstrated using the MoE structure where we have lightweight experts specialized in each region rather than a single large neural network that has to learn the entire input region. The gating network automatically determines which particular input region each expert will be in charge of in the end-to-end training process. We show that the proposed MoE architecture is 78.43% more parameter-efficient and achieves higher accuracy using fewer training data while also being less computationally intensive.

CRediT authorship contribution statement

Chanwoo Park: Conceptualization, Methodology, Software. **Premkumar Vincent:** Data curation, Writing – original draft.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Chanwoo Park has patent pending to Alsemy Inc.

Data availability

Data will be made available on request.

References

[1] McAndrew C. Compact modeling: Principles, techniques, and applications. In: Statistical modeling using backward propagation of variance. Springer New York; 2010.

[2] Zhang L, Chan M. Artificial neural network design for compact modeling of generic transistors. J Comput Electron 2017;16(3):825–32.

[3] Wang J, Kim Y-H, Ryu J, Jeong C, Choi W, Kim D. Artificial neural network-based compact modeling methodology for advanced transistors. IEEE Trans Electron Devices 2021;68(3):1318–25.

[4] Li M, İrsoy O, Cardie C, Xing HG. Physics-inspired neural networks for efficient device compact modeling. IEEE J Explor Solid-State Comput Devices Circuits 2016;2:44–9.

[5] Kim Y, Myung S, Ryu J, Jeong C, Kim DS. Physics-augmented neural compact model for emerging device technologies. In: 2020 International conference on simulation of semiconductor processes and devices. SISPAD, IEEE; 2020, p. 257–60.

[6] Jacobs RA, Jordan MI, Nowlan SJ, Hinton GE. Adaptive mixtures of local experts. Neural Comput 1991;3(1):79–87.

[7] Jain A, Singh VP, Rath SP. A multi-accent acoustic model using mixture of experts for speech recognition. In: Interspeech. 2019, p. 779–83.

[8] Bartoli N, Lefebvre T, Dubreuil S, Olivanti R, Priem R, Bons N, et al. Adaptive modeling strategy for constrained global optimization with application to aerodynamic wing design. Aerosp Sci Technol 2019;90:85–102.

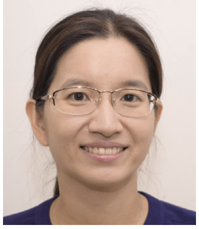
[9] Nanoscale Integration and Modeling Group, ASU. PTM high performance 45nm MOSFET. 2008, http://ptm.asu.edu/modelcard/HP/45nm_HP.pm.

Chanwoo Park received his B.S. and M.S. in Electrical Engineering from Seoul National University in 2009 and 2012, respectively. He joined Alsemy Inc. in 2020, as a Chief AI Officer. His current research interests include neural compact modeling, physics-based neural network, meta learning, and Bayesian inference.

Premkumar Vincent received his Ph.D. from Kyungpook National University, South Korea (2020). He completed his B.E. in Electronics and Communication Engineering in 2014. His research interests include TCAD simulations of MOSFET, TFTs, and solar cells. He is currently a Research Engineer with Alsemy Inc., South Korea.

Soogine Chong received the B.S. degree in Electrical Engineering from Seoul National University in 2003, and the M.S. and Ph.D. degrees in Electrical Engineering from Stanford University, Stanford, CA, U.S.A. in 2006 and 2012, respectively. She is currently a Chief Scientist with Alsemy Inc., South Korea.

Junghwan Park received the B.S. degree in Space Science from Kyunghee University, South Korea in 2021. He currently works as a Research Engineer in Alsemy Inc., South Korea.



Ye Sle Cha received her Ph.D. in Mathematics from Stony Brook University, U.S.A. in 2013. She had completed her B.S. in Mathematics at KAIST, South Korea in 2007. She currently works as a Research Scientist in Alsemy Inc., South Korea. Her research interests include scientific machine learning, meta learning, and neural compact modeling.

Hyunbo Cho is founder and CEO of Alsemy Inc. Previously, he received the B.S. degree in Electrical Engineering from Seoul National University in 2005, and the M.S. degree in Electrical Engineering from Stanford University, Stanford, CA, U.S.A. in 2007.