AttendNets: Tiny Deep Image Recognition Neural Networks for the Edge via Visual Attention Condensers

Alexander Wong[1,2,3,*], **Mahmoud Famouri**[3], **and Mohammad Javad Shafiee**[1,2,3]
[1] Vision and Image Processing Research Group, University of Waterloo, Waterloo, ON, Canada
[2] Waterloo Artificial Intelligence Institute, University of Waterloo, Waterloo, ON, Canada
[3] DarwinAI Corp., Waterloo, ON, Canada
[*] a28wong@uwaterloo.ca

Abstract

While significant advances in deep learning has resulted in state-of-the-art performance across a large number of complex visual perception tasks, the widespread deployment of deep neural networks for TinyML applications involving on-device, low-power image recognition remains a big challenge given the complexity of deep neural networks. In this study, we introduce **AttendNets**, low-precision, highly compact deep neural networks tailored for on-device image recognition. AttendNets possess deep self-attention architectures based on **visual attention condensers**, which extends on the recently introduced stand-alone attention condensers to improve spatial-channel selective attention. Furthermore, AttendNets have unique machine-designed macroarchitecture and microarchitecture designs achieved via a machine-driven design exploration strategy. Experimental results on ImageNet_{50} benchmark dataset for the task of on-device image recognition showed that AttendNets have significantly lower architectural and computational complexity when compared to several deep neural networks in research literature designed for efficiency while achieving highest accuracies (with the smallest AttendNet achieving $\sim$7.2% higher accuracy, while requiring $\sim 3\times$ fewer multiply-add operations, $\sim 4.17\times$ fewer parameters, and $\sim 16.7\times$ lower weight memory requirements than MobileNet-V1). Based on these promising results, AttendNets illustrate the effectiveness of visual attention condensers as building blocks for enabling various on-device visual perception tasks for TinyML applications.

1 Introduction

Deep learning [11] has resulted in significant breakthroughs in the area of computer vision, with state-of-the-art performance in a wide range of visual perception tasks such as image recognition [6, 9], object detection [4], and segmentation [1, 7]. Despite these breakthroughs, the widespread deployment of deep neural networks for tiny machine learning (TinyML) applications involving on-device visual perception on low-cost, low-power devices remains a major challenge given the increasing complexities of deep neural networks. Motivated by the tremendous potential of deep learning empowering TinyML applications and inspired to tackle the aforementioned complexity challenge, there has been significant effort in recent years on the creation of highly efficient deep neural networks for edge scenarios. These efforts in efficient deep learning have yielded a number of effective strategies, and can be typically grouped into two main categories: i) model compression [10, 5], and ii) efficient architecture design [8, 13, 12, 6]. In the realm of efficient architecture design, a number of architecture design patterns have been introduced leveraging bottlenecks [6, 13], factorized convolutions [8, 13], pointwise group convolutions and channel shuffling [12]. One particular area that has not been well explored and is ripe for innovation is to leverage the concept of self-attention [14, 9],

6th Workshop on Energy Efficient Machine Learning and Cognitive Computing (EMC2 2020), Online (San Jose, California).

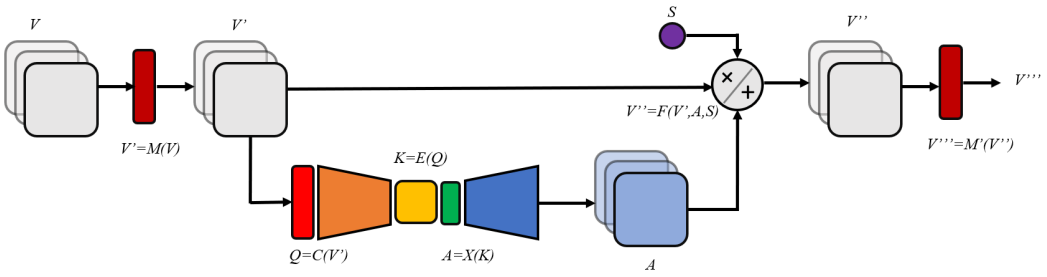

Figure 1: A visual attention condenser (VAC) is a self-attention mechanism consisting of a down-mixing layer, $M(V)$, condensation layer $C(V')$, an embedding structure $E(Q)$, an expansion layer $X(K)$, a selective attention mechanism $F(V', A, S)$, and an up-mixing layer $M'(V'')$.

seen as one of the recent big breakthroughs in deep learning, for the purpose of building highly efficient deep neural network architectures.

In this study, we introduce **AttendNets**, low-precision, highly compact deep neural networks tailored for on-device image recognition. AttendNets possess deep self-attention architectures based on **visual attention condensers**, which extends on the recently introduced stand-alone attention condensers to improve spatial-channel selective attention. AttendNets have unique macroarchitecture and microarchitecture designs achieved via a machine-driven design exploration strategy, making them particularly tailored for TinyML applications on edge devices.

2 Method

2.1 Visual Attention Condensers

The first concept we leverage to construct the proposed AttendNets is the concept of visual attention condensers. The concept of self-attention in deep learning has led to significant advances in recent years [14, 9], particularly with the advent of Transformers [14] that has reshaped the landscape of machine learning for natural language processing. It can be said that much of research on self-attention in deep learning has focused on improving accuracy, and this has had a heavy influence over the design of self-attention mechanisms. Motivated to explore the design of self-attention mechanisms in the direction efficiency instead of accuracy, Wong et al. [15] introduce the concept of attention condensers as a stand-alone building block for deep neural networks geared around condensed self-attention. They were able to demonstrate the efficacy of attention condensers on the task of limited-vocabulary speech recognition, achieving state-of-the-art in network efficiency.

Inspired by the promise of attention condensers, we extend upon the attention condenser design to further improve their efficiency and effectiveness for tackling visual perception tasks such as image recognition. We take inspiration from the observation that deep neural network architectures for tackling complex visual perception tasks often have very high channel dimensionality. As such, while the existing attention condenser design can still achieve significant reductions on network complexity under such scenarios, we hypothesize that further complexity reductions can be gained through better handling of the high channel dimensionality when learning the condensed embedding of joint spatial-channel activation relationships. As such, we introduce an extended visual attention condenser design where we introduce a pair of learned channel mixing layers that further reduces spatial-channel embedding dimensionality while preserving selective attention performance.

An overview of the proposed visual attention condenser (VAC) is shown in Figure 1. More specifically, a visual attention condenser is a self-attention mechanism consisting of a down-mixing layer, $M(V)$, condensation layer $C(V')$, an embedding structure $E(Q)$, an expansion layer $X(K)$, a selective attention mechanism $F(V', A, S)$, and an up-mixing layer $M'(V'')$. The down-mixing layer $V' = M(V)$ learns and produces a projection of the input activations V to a reduced channel dimensionality to obtain V'. The condensation layer (i.e., $Q = C(V')$) condenses V' for reduced dimensionality to Q with an emphasis on strong activation proximity to better promote relevant region of interest despite the condensed nature of the spatial-channel representation. An embedding structure (i.e., $K = E(Q)$) then learns and produces a condensed embedding K from Q characterizing joint spatial-channel activation relationships. An expansion layer (i.e., $A = X(K)$) then projects the condensed embedding K to an increased dimensionality to produce self-attention values A emphasizing regions of interest in the same domain as V'. The output V'' is a product of V', self-attention values A, and scale S via selective attention (i.e., $V'' = F(V', A, S)$). Finally, the up-mixing layer $M'(V'')$ learns and prodcues a projection of V'' to a higher channel dimensionality for final output V''' that

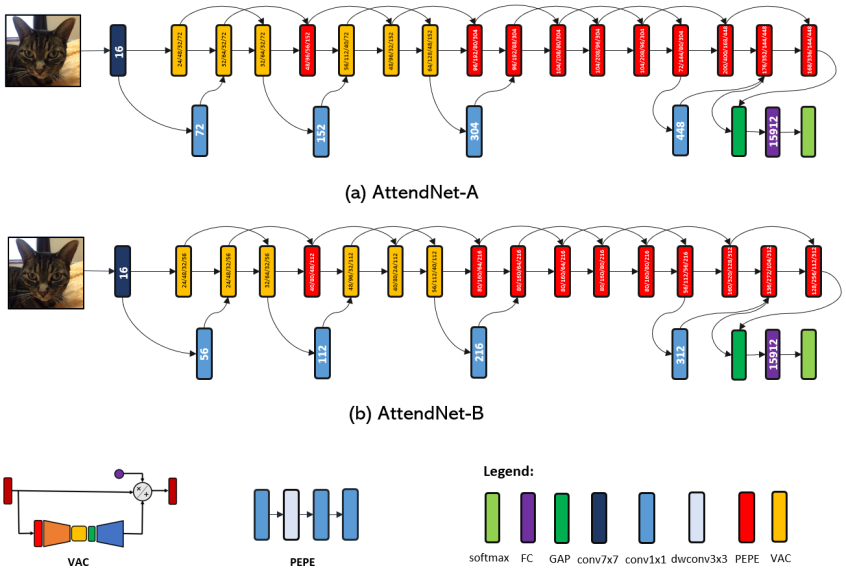

Figure 2: AttendNet architectures for image recognition. The number in each module represents the number of channels. The number in each fully-connected layer represents the number of synapses.

has the same channel dimensionality as the input activation V. Overall, through the introduction of the pair of learned mixing layers into the attention condenser design, a better balance between joint spatial-channel embedding dimensionality and selective attention performance can be achieved for building highly efficient deep neural networks for tackling visual perception problems on the edge.

2.2 Machine-driven Design Exploration

The second concept we leverage to construct the proposed AttendNets is the concept of machine-driven design exploration. In this study, we leverage a generative synthesis [17] design exploration approach to automate the process of generating the macroarchitecture and microarchitecture designs of the final AttendNet network architectures such that they are tailored specifically for the purpose of on-device image recognition in computational and memory constrained scenarios such as on low-cost, low-power edge devices, with an optimal balance between image recognition accuracy and network efficiency. As the goal for the proposed AttendNets is to achieve a strong balance between accuracy and network efficiency for the task of on-device image recognition, we imposed two key constraints during the machine-driven design exploration process: i) the top-1 validation accuracy is greater than or equal to 71% on the ImageNet$_{50}$ edge vision benchmark dataset introduced by Fang et al. [3] for evaluating performance of deep neural networks for on-device vision applications, and ii) 8-bit weight precision. First, a top-1 validation accuracy constraint of 71% validation accuracy was chosen to make AttendNets comparable in accuracy to a state-of-the-art efficient deep neural network proposed in [16] for on-device image recognition. Second, an 8-bit weight precision constraint was chosen to account for the memory constraints of low-cost edge devices.

Taking advantage of the fact that the generative synthesis process is iterative and produces a number of successive generators [17], we leverage two of the constructed generators at different stages to automatically generate two compact deep image recognition networks (AttendNet-A and AttendNet-B) with different tradeoffs between image recognition accuracy and network efficiency. Finally, to realize the concept of visual attention condensers in a way that enables the learning of condensed embeddings characterizing joint spatial-channel activation relationships in an efficient yet effective manner, we leveraged max pooling, a lightweight two-layer neural network (grouped then pointwise convolution), unpooling, and pointwise convolution for the condensation layer $C(V')$, the embedding structure $E(Q)$, the expansion layer $X(K)$, and the mixing layers $M(V)$ and $M'(V'')$, respectively, within a visual attention condenser.

3 AttendNet Architecture Designs

Figure 2 shows the architecture designs of the two AttendNets, produced via machine-driven design exploration that incorporates visual attention condensers in its design considerations. A number of interesting observations can be made about the AttendNet architecture designs. First, the AttendNet architecture designs are comprised of a mix of consecutive stand-alone visual attention condensers performing consecutive visual selective attention and projection-expansion-projection-expansion

Table 1: Top-1 accuracy, number of parameters, and number of multiply-add operations of AttendNets in comparison to four efficient networks (MobileNet-V1 [8], MobileNet-V2 [13], AttoNet-A [16], AttoNet-B [16]). Best results are in **bold**. Results for AttendNets based on 8-bit low precision weights, while results for other tested networks based on 32-bit full precision weights

Model	Top-1 Accuracy	Params	Mult-Adds
MobileNet-V1	64.5%	3260K	567.5M
MobileNet-V2	68.7%	2290K	299.7M
AttoNet-A	73.0%	2970K	424.8M
AttoNet-B	71.1%	1870K	277.5M
AttendNet-A	**73.2%**	1386K	276.8M
AttendNet-B	71.7%	**782K**	**191.3M**

(PEPE) modules for efficient feature representation. The PEPE module was discovered by the machine-driven exploration strategy and comprises of a projection layer that reduces dimensionality via pointwise convolution, an expansion layer that increases dimensionality efficiently via depthwise convolution, a projection layer that reduces dimensionality again via pointwise convolution, and finally an expansion layer that increases dimensionality again via pointwise convolution.

Second, it can be observed that there is a heavy use of visual attention condensers early on within the AttendNet architecture by the machine-driven design exploration strategy, while relying on PEPE modules later in the network architecture. This interesting design choice by the machine-driven design exploration strategy may be a result of selective attention being more important earlier on low-level to medium-level visual abstraction for image recognition to enable better focus on irrelevant regions of interest critical to decision-making within a complex scene.

Third and finally, it can be observed that the AttendNet network architectures exhibits high architectural diversity both at the macroarchitecture level and microarchitecture level. For example, at the macroarchitecture level, there is a heterogeneous mix of visual attention condensers, PEPE modules, spatial and pointwise convolutions, and fully-connected layers. At the microarchitecture level, the visual attention condensers and PEPE modules have a diversity of microarchitecture designs as seen by the differences in channel configurations. This level of architectural diversity is a result of the machine-driven design exploration process, which has the benefit of determining the best architecture design at a fine grained level to achieve a strong balance of network efficiency and accuracy for the specific task at hand. Based on these three interesting observations, it can be seen the proposed AttendNet network architectures is highly tailored for on-device image recognition for the edge, and also shows the merits of leveraging both visual attention condensers and machine-driven design exploration for achieving such highly efficient, high-performance deep neural networks.

4 Results and Discussion

In this study, we evaluate the efficacy of the proposed low-precision AttendNets on the task of image recognition to empirically study the balance between accuracy and network efficiency. More specifically, we leverage ImageNet_{50}, a benchmark dataset that was introduced by Fang et al. [3] for evaluating performance of deep neural networks for on-device vision applications on the edge derived from the popular ImageNet [2] dataset. To quantify accuracy and network efficiency, we computed the following performance metrics: i) top-1 accuracy, ii) the number of parameters (to quantify architectural complexity), and iii) the number of multiply-add operations (to quantify computational complexity). For comparative purposes, the same performance metrics were also evaluated on MobileNet-V1 [8], MobileNet-V2 [13], AttoNet-A [16], and AttoNet-B [16]), four highly efficient deep image recognition networks that were all designed for on-device image recognition purposes.

Table 1 shows the top-1 accuracy, the number of parameters, and the number of multiply-add operations of the AttendNets alongside the four other tested efficient deep image recognition networks. It can be clearly observed that the proposed AttendNets achieved a significantly better balance between accuracies and architectural and computational complexity when compared to the other tested efficient deep neural networks. In terms of lowest architectural and computational complexity, AttendNet-B achieved significantly higher accuracy compared to MobileNet-V1 ($\sim$**7.2%** higher) but requires $\sim$**4.17**$\times$ fewer parameters, $\sim$**16.7**$\times$ lower weight memory requirements, and $\sim$**3**$\times$ fewer multiply-add operations than MobileNet-V1. Compared to similarly-accurate state-of-the-art AttoNet-B, AttendNet-B achieved $\sim$**0.6%** higher accuracy but requires $\sim$**2.4**$\times$ fewer parameters, $\sim$**9.6**$\times$ lower weight memory requirements, and $\sim$**1.45**$\times$ fewer multiply-add operations. In terms of the highest top-1 accuracy, AttendNet-A achieved significantly higher accuracy compared to MobileNet-V1

and MobileNet-V2 (∼**8.7%** higher and ∼**4.5%** higher, respectively) but requires ∼**2.35**× fewer parameters, ∼**9.4**× lower weight memory requirements, and ∼**2.1**× fewer multiply-add operations than MobileNet-V1 and ∼**1.65**× fewer parameters, ∼**6.6**× lower weight memory requirements, and ∼**1.1**× fewer multiply-add operations than MobileNet-V2. These quantitative performance results illustrate the efficacy of leveraging both visual attention condensers and machine-driven design exploration to creating highly-efficient deep neural network architectures tailored for on-device image recognition that striking a strong balance between accuracy and network complexity.

Given the promising results, future work involves exploring the effectiveness of AttendNets on downstream tasks such as object detection and segmentation to empower a wider variety of vision-related TinyML applications ranging from autonomous vehicles to wearable assistive technologies to remote sensing. We also aim to explore different design choices for the individual components of the visual attention condenser (e.g., mixing layers, embedding structure, condensation layer, expansion layer) and their impact on accuracy and efficiency. Finally, the exploration of self-attention architectures based on visual attention condensers and their adversarial robustness is a worthwhile endeavor, particularly given recent focuses on robustness and dependability of deep learning.

References

[1] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE TPAMI*, 40(4):834–848, 2018.

[2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. 2009.

[3] B. Fang, X. Zeng, and M. Zhang. Nestdnn: Resource-aware multi-tenant on-device deep learning for continuous mobile vision. *MobiCom*, 2018.

[4] R. Girshick. Fast r-cnn. In *ICCV*, pages 1440–1448, 2015.

[5] S. Han, H. Mao, and W. J. Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149*, 2015.

[6] K. He et al. Deep residual learning for image recognition. *arXiv:1512.03385*, 2015.

[7] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[8] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[9] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu. Squeeze-and-excitation networks, 2017.

[10] B. Jacob et al. Quantization and training of neural networks for efficient integer-arithmetic-only inference. *arXiv:1712.05877*, 2017.

[11] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *nature*, 521(7553):436, 2015.

[12] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *ECCV*, pages 116–131, 2018.

[13] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *CVPR*, pages 4510–4520, 2018.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need, 2017.

[15] A. Wong, M. Famouri, M. Pavlova, and S. Surana. Tinyspeech: Attention condensers for deep speech recognition neural networks on edge devices. *arXiv preprint: 2008.04245*, 2020.

[16] A. Wong, Z. Q. Lin, and B. Chwyl. Attonets: Compact and efficient deep neural networks for the edge via human-machine collaborative design. 2019.

[17] A. Wong, M. J. Shafiee, B. Chwyl, and F. Li. Ferminets: Learning generative machines to generate efficient neural networks via generative synthesis. *NIPS Workshops*, 2018.