Scalable Multi-Teacher Gated Knowledge Distillation for Accurate & Efficient Leukemia Classification

Tahsen Islam Sajon
Dept. of CSE, Rajshahi University of Engineering & Technology, Bangladesh
Email:sajon.tahsen@gmail.com | Code:GitHub

ABSTRACT: Recent advancements in deep learning have driven significant progress in medical imaging, notably in automated leukemia classification. However, deploying state-of-the-art models in real-world applications can be prohibitive due to their computational and memory overhead. In this context, this study introduces a multi-teacher knowledge distillation framework, complemented by a learned gating mechanism. This gate is designed to dynamically allocate weights to the teacher models during the distillation process, promoting accurate predictions in the student model. Swin Transformer and ConvNext models serve as the chosen teachers, with the intention to harness both the robust representational capabilities of transformers and the locality-preserving attributes of convolutions. Employing this strategy to guide a lightweight MobileNetV2 student model, an evident improvement in performance was achieved over conventional training techniques.

1 Introduction

The classification of Acute Lymphoblastic Leukemia (ALL) through deep learning has been extensively studied, with models like Zakir et al.'s attention-based CNN using VGG16 and ECA to emphasize the semantic features of ALL cells [7], and Niranjana et al.'s specialized ALLNET registering 95.54% accuracy on the C-NMC-2019 dataset [6]. Efforts towards lightweight architectures have also been noted; Krzysztof et al. combined MobileNetV2 with traditional classifiers like DT and RF to obtain a 97.4% accuracy score on the ALL-IDB dataset [5]. Similarly, Faro et al. integrated neural networks with classifiers like KNN and SVM, aiming at computational efficiency [1]. In the domain of knowledge distillation, Genovese et al. proposed DL4ALL, a multi-task model utilizing various cross-dataset transfer learning techniques, with the 'greedy' variant showing the most promise [2].

Despite these strides, discernible gaps remain in the KD literature for leukemia classification, opening avenues for further research. Considering this, the objectives for this exploration are as follows:

- **Robust KD Framework:** I propose a knowledge distillation approach that:
  - Moves beyond the limitations inherent in single-teacher models.
  - Incorporates a fluid, trainable gating mechanism to make the most of the knowledge from diverse teachers.
- **Versatile Model Design:** Both teacher models, with convolutional and transformer-centric feature extractors, outpace the majority of current literature, underscoring the efficacy of the model architecture refined through hyperparameter tuning and optimization.

2 Materials and Methods

2.1 Data Collection & Preprocessing

From the C-NMC 2019 dataset [3], I sourced 10,661 images, comprising 7,272 blast cells and 3,389 healthy cells. These were divided into training, validation, and test sets in a 70:15:15 ratio. Augmentation methods used were random flips, rotation, and conditional zoom and crop.

2.2 Model Architecture

The ConvNext and Swin Transformer models [4], given their established efficiency, were employed as teacher models. The crafted model architecture, depicted in Fig. 1, incorporates Batch Normal-

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

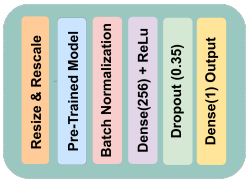

Figure 1: Model Architecure

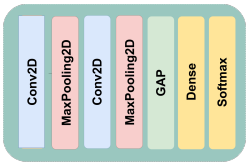

Figure 2: Gating Network

Table 1: Key Configurations and Corresponding Values for Experiments

Configuration	Teacher Models	KD Student Model
Image Size	224x224	224x224
Batch Size	64	64
Training Epoch	150	650
Optimizer	RAdam + Lookahead	Adam
Kernel Regularizer	L2 (0.016)	L2 (0.016)
Activity Regularizer	L1 (0.006)	L1 (0.006)
Bias Regularizer	L1 (0.006)	L1 (0.006)

ization and Regularization layers to address the dataset's class imbalance and mitigate overfitting. Additionally, the gating network, illustrated in Fig. 2, is purposefully lightweight, ensuring seamless integration without overshadowing the compact MobileNetV2 student model.

2.3 Training Configuration

In addition to network design, the training process significantly influences performance. Table 1 lists key parameters used.

2.4 Knowledge Distillation

The core algorithm of the proposed KD approach is presented in Algorithm 1.

Algorithm 1 Pseudocode Describing the Proposed Multi Teacher KD Learning Procedure

procedure TRAIN_STEP($data$)
  $x, y \leftarrow data$
  $student\_predictions \leftarrow student(x, training = True)$
  $student\_loss \leftarrow student\_loss\_function(y, student\_predictions)$
  for each $teacher$ in $teachers$ **do**
    $teacher\_predictions \leftarrow teacher(x, training = False)$
  end for
  $gating\_weights \leftarrow gating\_network(x, training = True)$
  $weighted\_teacher\_predictions = 0$
  for $i, teacher\_prediction$ in $teacher\_predictions$ **do**
    $weighted\_teacher\_predictions$ += $gating\_weights[:, i : i + 1] \times teacher\_prediction$
  end for
  $distillation\_loss \leftarrow distillation\_loss\_function(weighted\_teacher\_predictions, student\_predictions)$
  // Alpha Factor to Weight the Student and Distillation Loss
  $total\_loss \leftarrow \alpha \times distillation\_loss + (1 - \alpha) \times student\_loss$
  // Gradient Calculation and Weight Update of Student & Gating Model
  $trainable\_vars \leftarrow student.trainable\_variables + gating\_network.trainable\_variables$
  $gradients \leftarrow gradient(total\_loss, trainable\_vars)$
  $optimizer.apply\_gradients(gradients, trainable\_vars)$
end procedure

3 Preliminary Results

Experiments reveal improved classification performance in the student model over its standalone counterpart, as shown in Table 2. Both the teacher and student models demonstrate results comparable to, or even surpassing, current state-of-the-art and mobile-focused studies.

Table 2: Preliminary Results (Provided Precision, Recall, and F1-Score are Weighted Average values).

Model	Parameters	Accuracy	Precision	Recall	F1-Score
Swin Base	87.35M	0.985	0.98	0.98	0.98
ConvNext Base	87.35M	0.982	0.98	0.98	0.98
MobileNetV2	2.59M	0.948	0.95	0.95	0.95
Distilled MobileNetV2	**2.59M**	**0.961**	**0.96**	**0.96**	**0.96**

4 Future Scope

While promising in its progress, this work views itself as part of an evolving narrative with several forward-looking paths. The gating mechanism currently processes raw images. However, incorporating the confidence of teacher predictions—using their proximity to ground truth labels—may offer a nuanced approach to optimizing its decisions.

As it stands, both the teacher and student models handle 224x224 images. There's potential for teachers to process larger resolutions to enrich feature capture, while the student retains its smaller 224x224 input size. The dynamic resizing from the high-resolution source can ensure detailed extraction without sacrificing efficiency.

Modern training strategies beckon; sophisticated learning rate schedules and advanced augmentation methods such as CutMix, RandAugment are on the horizon. I am keen to implement Grad-CAM visualizations to shed light on the distillation process, adding interpretability.

Perhaps the most exciting prospect is the scalability of the proposed architecture. With just two teacher models now, it is designed to accommodate an expanded array. The unique application of such a framework to ALL classification remains relatively uncharted, highlighting the possibility of expansive inter-domain knowledge transfer. Ultimately, the goal is to leverage insights from these varied models to elevate leukemia classification, potentially in a format that is both lightweight and widely accessible.

References

[1] Yúri Faro Dantas de Sant'Anna, José Elwyslan Maurício de Oliveira, and Daniel Oliveira Dantas. Interpretable lightweight ensemble classification of normal versus leukemic cells. *Computers*, 11(8):125, 2022.

[2] Angelo Genovese, Vincenzo Piuri, Konstantinos N Plataniotis, and Fabio Scotti. Dl4all: Multi-task cross-dataset transfer learning for acute lymphoblastic leukemia detection. *IEEE Access*, 2023.

[3] Anubha Gupta and Ritu Gupta. Isbi 2019 c-nmc challenge: Classification in cancer cell imaging. *Select Proceedings*, 2019.

[4] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022.

[5] Krzysztof Pałczyński, Sandra Śmigiel, Marta Gackowska, Damian Ledziński, Sławomir Bujnowski, and Zbigniew Lutowski. Iot application of transfer learning in hybrid artificial intelligence systems for acute lymphoblastic leukemia classification. *Sensors*, 21(23):8025, 2021.

[6] Niranjana Sampathila, Krishnaraj Chadaga, Neelankit Goswami, Rajagopala P Chadaga, Mayur Pandya, Srikanth Prabhu, Muralidhar G Bairy, Swathi S Katta, Devadas Bhat, and Sudhakara P Upadya. Customized deep learning classifier for detection of acute lymphoblastic leukemia using blood smear images. In *Healthcare*, volume 10, page 1812. MDPI, 2022.

[7] Muhammad Zakir Ullah, Yuanjie Zheng, Jingqi Song, Sehrish Aslam, Chenxi Xu, Gogo Dauda Kiazolu, and Liping Wang. An attention-based convolutional neural network for acute lymphoblastic leukemia classification. *Applied Sciences*, 11(22):10662, 2021.