A Deeper Look into the Limitations of Early-Exit Architectures for Single and Multi-Label Classification

Klaudia Bałazy[1,3], Julian McAuley[2], and Jacek Tabor[1]

[1] Jagiellonian University
[2] University of California San Diego
[3] Doctoral School of Exact and Natural Sciences at Jagiellonian University
Corresponding author: klaudia.balazy@doctoral.uj.edu.pl

Abstract. In this study, we explore the limitations of early-exit architectures, which are designed to enhance computational efficiency in neural networks, focusing particularly on both single and multi-label classification tasks within the computer vision domain. We introduce a systematic evaluation framework that not only advances research in this area but also bridges an important gap in understanding how these architectures perform within the complexities of multi-label settings. Our findings reveal a significant challenge: while early-exits improve efficiency without compromising accuracy in single-label tasks, they struggle to offer similar benefits in multi-label classification, necessitating uniquely tailored strategies. Further insights from our ablation suggest that the difficulty in achieving benefits from early-exits in multi-label classification may stem from the varying complexities of processing distinct classes within a single instance. This work lays a solid foundation for future research focused on developing early-exit strategies that effectively handle the complexities of diverse classification contexts.

Keywords: Computational Efficiency · Dynamic Neural Networks · Early-Exit Architecture · Multi-Label Classification.

1 Introduction

Deep neural network models have shown remarkable performance in various fields, including computer vision and natural language processing. The significant improvements in accuracy and predictive capability often come at the cost of increased computational complexity and resource demands [1]. Early-exit strategies [2, 3, 4, 5], which dynamically adjust the computation time based on the complexity of the input, have emerged as a promising approach to mitigate these challenges. Early-exit architectures introduce internal classifiers at different depths within a neural network. These classifiers, also known as "heads", can independently make predictions and stop the inference process early if certain conditions, such as a predefined confidence threshold, are met. By enabling

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97626-1_13

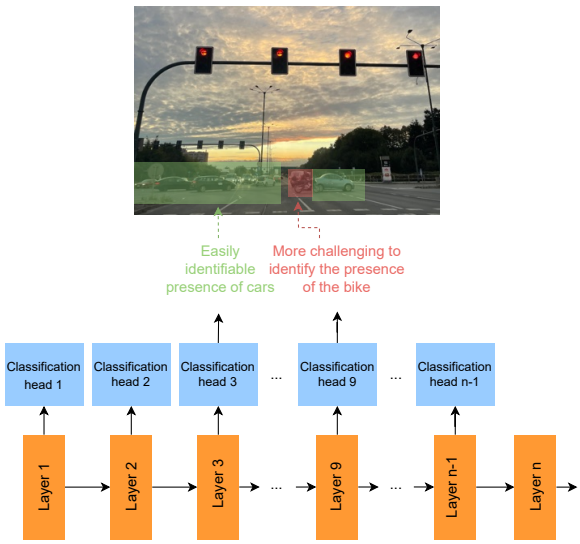

Fig. 1: The underlying complexity of leveraging early-exit models for multi-label classification can be illustrated by this example image with easily identifiable cars and a harder-to-spot bike. Our experiments (see Section 4.3) suggest that while the model may quickly detect easier to notice classes (like cars in this example), identifying the bike could require further processing. This highlights the challenge of finding the optimal exit point when multiple objects with varying recognition difficulty are present (this problem is absent in single-label classification).

models to exit inference early under certain conditions, these strategies can significantly improve computational efficiency.

While early-exit models have been the subject of growing interest, several key aspects remain underexplored. Primarily, the evaluation of these strategies is still a challenging task due to the lack of a universally agreed-upon framework for comparison. Although the Efficient Language Understanding Evaluation (ELUE) [6] has been proposed for natural language processing tasks allowing for evaluation of early-exit methods with different exit thresholds, a broader framework applicable across diverse domains is still lacking. Secondly, the understanding of the limitations and benefits of early-exit strategies needs deeper investigation. Finally, despite considerable research on early-exits for single-label tasks, studies focusing on multi-label tasks are noticeably missing.

In this study, we address these gaps by introducing a systematic framework for the evaluation of early-exit strategies. Our proposed framework offers a structured guideline for future assessments and comparisons within this field. It incorporates a robust evaluation metric that quantifies the performance differences between training each model's head independently and using a collective training

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97626-1_13

approach with a selected early-exit strategy. This approach allows for a deeper understanding of early-exit strategies and can serve as a cornerstone for subsequent research in this field.

For single-label classification, we demonstrate that early-exits are advantageous regardless of the number of output classes. The advantages are particularly pronounced in the initial layers of the network and gradually diminish in later layers. Moreover, our analysis adds nuance to the "big label" problem, a phenomenon identified by Liu et al. [7] in the context of the WOS-46985 dataset, where early-exits struggle with single-label classification tasks involving a large number of output classes. Our investigation reveals that this issue is not universal but dataset-dependent. Specifically, our single-label experiments with multiple output classes did not reproduce the "big label" problem, emphasizing the need for context-aware analyses of early-exit models.

Furthermore, we delve into the analysis for multi-label classification tasks, an area that we have identified as presenting considerable challenges in terms of deriving benefits from early-exit methods. We provide a solid baseline and contribute with insights into this underexplored area. We examine diverse architecture choices and various exiting criteria, including entropy [8, 9, 10], confidence [11, 12, 13], patience [14, 15], learning-to-exit [16], and hybrid approaches [17]. Our investigation reveals that using a confidence or entropy criterion is an effective approach for multi-label classification. To further improve the performance, we explore a learning-based approach for early-exiting that integrates a modified loss function, resulting in enhanced performance at a specific speedup.

We also design an experiment to gain a deeper understanding of the challenges associated with employing early-exit strategies for multi-label classification. Our hypothesis, supported by our findings, indicates that the varying recognition difficulty associated with each class within an instance could be a key challenge when employing early-exits in a multi-label context (see Figure 1). This investigation not only enriches our understanding of the complexities inherent in multi-label classification with early-exit strategies but also points towards intriguing directions for future research.

By offering insights into the limitations of early-exit architectures and shedding light on the complexities of applying these strategies to multi-label classification tasks, our study opens up important directions for future research in this domain. We hope that our work will not only contribute to a better understanding of early-exit models but also inspire further advancements in their design and application.

Our contributions can be summarized as follows:

- We propose a systematic framework for evaluating the benefits of early-exit architectures in deep neural network models.
- We conduct an in-depth examination of various exiting criteria and architecture choices for single-label and multi-label tasks within the computer vision domain, establishing a strong baseline for future research.
- We assess the "big label" problem [7] within the context of early-exit architectures. Our findings reveal that the challenges associated with handling a

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97626-1_13

large number of output classes in single-label classification are influenced by specific dataset characteristics, rather than being inherent limitations of the early-exits.
– We show that applying early-exit models to multi-label classification presents substantial complexities compared to single-label scenarios (see Figure 3). We provide detailed insights into this phenomenon and offer an intuitive understanding of the underlying factors (see Figure 1 and Figure 6).

2 Related work

Early-exit models are types of dynamic neural networks that can accelerate inference by terminating it at an earlier layer. This section provides an overview of the key developments in early-exit methods, and then discusses the approaches for evaluating and analyzing the performance of these methods.

Early-exit methods Adaptive Computation Time (ACT) [2, 3] introduced a trainable halting mechanism for input-adaptive inference. However, training the halting model required extra effort and added more parameters and increased inference costs. To address this issue, BranchyNet [4] utilized the entropy of the prediction probability distribution as a measure of branch classifier confidence to enable early exiting. Shallow-Deep Networks (SDN) [5] employed the softmax scores of branch classifier predictions to counteract the overthinking problem associated with deep neural networks.

Following research in computer vision, there have also been approaches for natural language processing models, as these models are usually large and deep. Some methods rely on predefined confidence thresholds to determine early-exit points. Examples include DeeBERT [8], RightTool [11], FastBERT [9], RomeBERT [10], and SkipBERT [13]. These approaches typically involve training BERT with internal classifiers and using entropy or other metrics to determine when a model's prediction is confident enough to exit early.

Another line of studies recycle the predictions of internal classifiers to improve overall performance and reduce wasted computation. PABEE [14] uses early stopping from model training to jointly train internal classifiers and exits when k consecutive classifiers make the same prediction. LeeBERT [15] encourages consistency among internal classifiers, while Sun et al. [18] introduce diversity loss and voting mechanisms for ensembling. Ensemble-based methods have been shown to improve both efficiency and robustness. Zero Time Waste (ZTW) [12] adds direct connections between internal classifiers and combines previous outputs in an ensemble-like manner, to improve performance.

In contrast to predefined confidence thresholds, some methods learn the early-exit criterion. BERxiT [16] uses a *learning-to-exit* module to predict the correctness of current internal classifiers, while CAT [19] employs a "meta consistency classifier" to determine conformity with the final classifier.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97626-1_13

Evaluation and analysis of early-exits Evaluating and analyzing early-exit models poses unique complexities, and currently, there is no universally agreed-upon method for comparison and evaluation across various contexts. In the context of natural language processing, the Efficient Language Understanding Evaluation (ELUE) [6] has been proposed as a potential standard, which considers multiple speed-accuracy pairs when different thresholds are selected for early-exit methods. However, it is still crucial to agree on the evaluation framework that can facilitate robust evaluation across various applications and domains.

Analyzing the limitations and benefits of early-exit strategies is a complex task, yet it is critical for understanding the conditions under which these models excel or fall short. Liu et al. [7] provide valuable insights into the limitations for single-label classification tasks, uncovering the "big label" issue for the WOS-46985 dataset. The term "big label" problem represents the case in which early-exit models have difficulties accurately processing tasks with a large number of output classes. As a potential solution, they propose a strategy of label reduction to mitigate this issue. Our investigation further deepens the understanding of this issue, demonstrating its dataset-dependent nature and emphasizing the importance of context-aware analyses. While considerable research exists on early-exits for single-label classification tasks, there is a notable gap in the literature regarding multi-label classification tasks. In contrast to existing works, our study comprehensively explores both single- and multi-label classification using early-exit strategies.

3 Evaluation framework of the early-exits effectiveness

In this section, we outline our evaluation framework for assessing the advantages and limitations of using early-exits to speed up inference in neural networks.

Evaluation Setup To evaluate the performance of early-exit strategies, we start with a pre-trained classification model with an unfrozen backbone. We enhance this model by adding additional classification heads, also referred to as internal classifiers, after each layer, aligning with methods used in prior studies [5, 12]. Each of these H heads is trained independently using an appropriate loss function, resulting in H distinct models. These models demonstrate the potential of a static exit strategy, where each head is configured to terminate processing after specific layers. Concurrently, we develop an early-exit model in which all heads are trained collectively. This dynamic model requires a carefully defined exit strategy for each example.

Training Objective In single-label classification tasks, we employ the cross-entropy (CE) loss function coupled with a softmax activation. This setup is chosen for its effectiveness in handling mutually exclusive class predictions typical in single-label scenarios. For multi-label classification, where multiple independent labels may be correct, we use binary cross-entropy (BCE) loss with sigmoid activation at the classification heads. Sigmoid activation allows each label to be treated

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97626-1_13

independently, predicting a label as positive (with a score above a threshold of 0.5) or negative (below this threshold). Alternatively, we also utilize CE loss with softmax activation, requiring the selection of a specific threshold above which a label is considered positive.

Early-exiting strategies We evaluate various commonly used early-exiting strategies: confidence thresholds, entropy thresholds, rank-patience-based strategies, hybrid approaches blending confidence (or entropy) with rank-patience measures, and learning-based (classifier-based) methodologies. For strategies that employ thresholds, we explore a spectrum of potential thresholds, generating a performance-speedup curve in the process. The effectiveness of a strategy is indicated by the area under this curve; the larger the area, the more effective the strategy. In the case of the learning-based approach, we derive a single performance-speedup point.

Confidence-based strategies [5, 11, 12, 13] suggest that the model should exit if a certain confidence threshold is surpassed. This threshold applies to the highest probability label in single-label classification and to all positively classified labels in multi-label case.

Entropy-based approaches [8, 9, 10] operate on the principle that a model should exit if the entropy of its prediction drops below a particular level (low entropy implies a higher degree of certainty in the prediction).

In *patience-based strategies* [14, 15], the decision to exit depends on the consistency of predictions across successive internal classifiers. We adapt this approach for multi-label classification by introducing a rank-based measure, where the exit decision also takes into account the order of predicted classes based on their probability magnitudes. We name this strategy *rank-patience-based approach*, as the model is permitted to exit if the rank of class probabilities remains adequately consistent across consecutive layers. This modification enhances the applicability of patience-based strategies in the context of multi-label classification.

Hybrid approaches [17] combine confidence (or entropy) measures with patience-based measures (rank-patience-based in our multi-label scenario). In this strategy, the model's exit decision is influenced by the confidence or entropy measure, but it also incorporates a rank-patience factor. This factor assesses the stability of the rank (order) of the predictions, allowing the model to process additional layers before deciding to exit, even if the confidence or entropy threshold has been met.

Learning-based approaches Xin et al. [16] introduce an auxiliary classifier with sigmoid activation at each potential exit point. These binary classifiers, trained to predict whether the model should exit at that head, use the corresponding hidden state or the logits from the respective head classifier as input. We employ the Mean Squared Error (MSE) loss function (following previous works [16]), which calculates the difference between the predicted exit probability p_j and the actual binary label y_j (indicating whether an early exit should occur) across all m training instances.

To impose a heavier penalty on incorrect predictions when the model should continue processing, we introduce an additional regularization loss component.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97626-1_13

This component is computed as the MSE of the predicted exit probability p_j in cases where the model should not exit. We also incorporate a weighting factor into the final loss calculation to impose a greater penalty on the earlier heads, as their predictions have the most significant influence on further processing. The weights are determined by $\frac{1}{h_i+1}$, where h_i denotes the head index. Furthermore, we balance the MSE loss and the regularization loss with an additional hyper-parameter α. The final loss $\mathcal{L}$ is formulated as the sum of individual losses $\mathcal{L}_{h_i}$ across all H heads:

$$\mathcal{L} = \sum_{i=1}^{H} \mathcal{L}_{h_i}, \tag{1}$$

where the loss $\mathcal{L}_{h_i}$ for each head h_i is defined as:

$$\mathcal{L}_{h_i} = \frac{1}{m(h_i+1)} \big(\alpha \cdot \sum_{j=1}^{m} (p_j - y_j)^2 + (1-\alpha) \cdot \sum_{j=1}^{m} (p_j \cdot (1-y_j))^2\big). \tag{2}$$

Early-Exit Benefits Evaluation Framework To assess the effectiveness of early-exit architectures, we compare the performance of static classifiers with early-exit models, similarly to the ELUE metric used in NLP [6]. We propose the following systematic approach:

- **Performance evaluation:** We first measure the performance P_{h_j} of each static classifier at head h_j for all n instances in the test set. Simultaneously, we evaluate the performance of the early-exit model $P_{ee}(E)$ at specific average exit layer E. The average exit layer E is calculated as: $E = \frac{\sum_i (e_i)}{n}$, where e_i denotes the exit layer for each instance. The performance $P_{ee}(E)$ represents the aggregated performance across all test instances at their respective exit points.
- **Locating the Nearest Performance Point:** To ensure fair comparisons, we calculate $E_{closest}(h_j)$, which identifies the average exit layer in the early-exit model that most closely matches the computational depth of each static classifier head h_j. We determine the average exit layer that best aligns, in computational terms, with the layer where head h_j is located. This average is computed across all test instances and may vary depending on the thresholds set within our chosen strategy. This alignment ensures that performance comparisons between the early-exit and static models are conducted under equivalent computational conditions:

$$E_{closest}(h_j) = \arg\min_{E} |E - h_j|$$

- **Early-Exits Benefits Quantification:** We then calculate the benefits of early-exit strategies B_{EE_j} for each static head index j by comparing the performance of the early-exit model at the $E_{closest}(h_j)$ average exit layer to the performance of the static classifier at head h_j:

$$B_{EE_j} = P_{ee}(E_{closest}(h_j)) - P_{h_j}$$

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97626-1_13

If the computed value B_{EE_j} is positive, it indicates that the early-exit strategy enhances performance relative to the static model for an equivalent computational effort. This positive benefit suggests that the early-exit model offers a more efficient processing strategy without sacrificing accuracy. A negative B_{EE_j} indicates no benefit from using early-exit models compared to static networks, while a zero value indicates equivalent performance. Our methodology enables us to systematically evaluate and demonstrate the practical advantages of early-exit strategies across various applications.

4 Experiments

In our study, we employ a pre-trained ResNet50 model [20] as the backbone. We extend this model with internal classifiers, which are appended after core bottleneck layers to enable early-exiting. To establish a baseline for comparison, each internal classifier head is fine-tuned independently on a designated dataset, resulting in H distinct static models. These models serve as benchmarks for evaluating the performance at various depths without early-exiting strategies. In parallel, we also develop an integrated early-exit model where all internal classifier heads are collectively optimized. This model is designed to assess the performance and efficiency of dynamic exiting during inference, contrasting directly with the static models that do not incorporate early-exit functionality.

The performance of the models is assessed on specific test sets through F1 score (averaged over all instances) and accuracy score. For static models with independently trained heads, we compute the scores across the entire test set for each head. In contrast, for the early-exit model, we employ diverse strategies, the specifics of which are covered in the subsequent experiment descriptions.

4.1 Early-exits for single-label classification

We start by exploring early-exits for single-label classification. We use the ImageNet [21] dataset with a varying number of output classes. We train models using different learning rates: $\lambda \in \{10^{-3}, 10^{-4}, 10^{-5}\}$. For the early-exit models, we employ a widely-adopted confidence strategy [5, 12] to determine the exit head for specific examples.

Our findings, illustrated in Figure 2, highlight the considerable benefits of using early-exit for single-label. However, these advantages seem to decrease in the final layers for this specific dataset. Moreover, the models with 2-output classes exhibit increased instability, likely attributed to the limited volume of training and test data. Based on these experiments, we draw the conclusion that the incorporation of early-exit architectures for single-label offers significant advantages. The "big label" problem identified for WOS-46985 [7], which refers to poor early-exits performance with many output classes, appears to be dataset-dependent and did not occur in our experiments.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97626-1_13

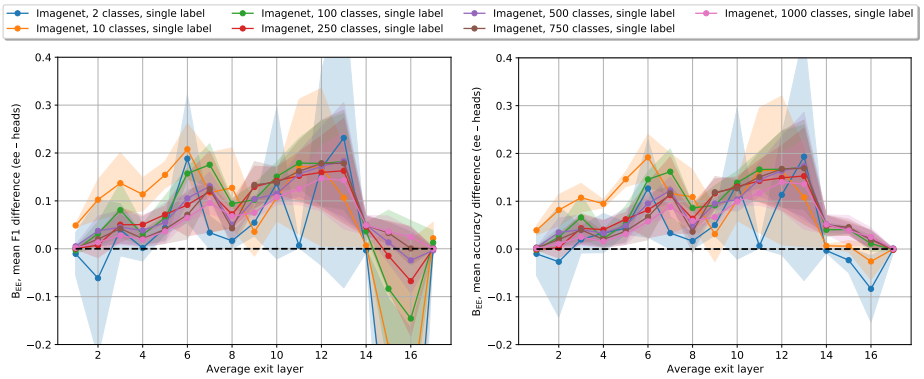

Fig. 2: Benefits of early-exits for single-label ImageNet classification across varying number of output classes. B_{EE} shows the mean F1 and accuracy difference between static exits and a confidence-based early-exit model. The graph shows mean benefits and standard deviations for each exit layer speedup across models trained with different learning rates. Regions above the dashed line indicate favorable performance-speedup trade-offs, highlighting the effectiveness of early-exits for single-label classification.

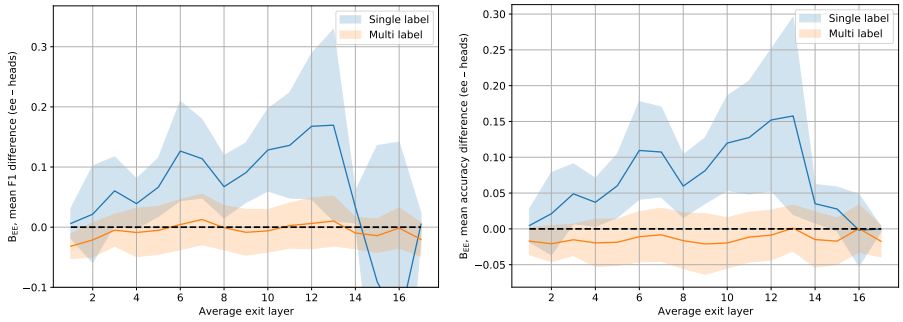

Fig. 3: The disparity in benefits of early-exit architectures for single- and multi-label classification. The graph shows mean benefits B_{EE} and standard deviations from experiments with different hyperparameters. Curves above the dashed line indicate a favorable performance-speedup trade-off. For single-label classification, benefits are predominantly positive, while for multi-label tasks, they are considerably smaller and symmetrically distributed around zero.

4.2 Early-exits for multi-label classification

We next evaluate the effectiveness of early-exit strategies for multi-label classification tasks using three datasets: VOC [22], COCO [23], and a modified version of ImageNet [21]. To adapt the ImageNet dataset for multi-label classification, we randomly select $n = 10$ classes and combine images from m randomly chosen categories ($m \in \{2, 4\}$), creating composite images that belong to multiple classes.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97626-1_13

We train early-exit models using either sigmoid or softmax activation functions at the outputs of the internal classifiers. For loss functions, we employ binary cross-entropy (BCE) for sigmoid activations and cross-entropy (CE) for softmax activations. In the softmax models, we use varying thresholds {0.01, 0.05, 0.1, 0.2, 0.3} during the evaluation phase to determine positive classifications. For our initial experiments, we evaluate early-exit models using two strategies: an entropy-based strategy for softmax-activation models and both confidence and entropy-based strategies for sigmoid-activation models. All models are trained with learning rates $\lambda \in 10^{-3}, 10^{-4}, 10^{-5}$.

Limited benefits for multi-label classification Figure 3 illustrates the mean and the standard deviation of performance outcomes achieved by early-exits with basic exiting strategies, such as confidence and entropy thresholds, for VOC, COCO, and ImageNet test sets. The benefits of implementing early-exit architectures for multi-label classification are significantly less pronounced than those observed for single-label scenarios, with statistical significance levels $p \leq 0.05$. These findings indicate that applying early-exit strategies in a multi-label context poses greater challenges compared to their application in single-label tasks.

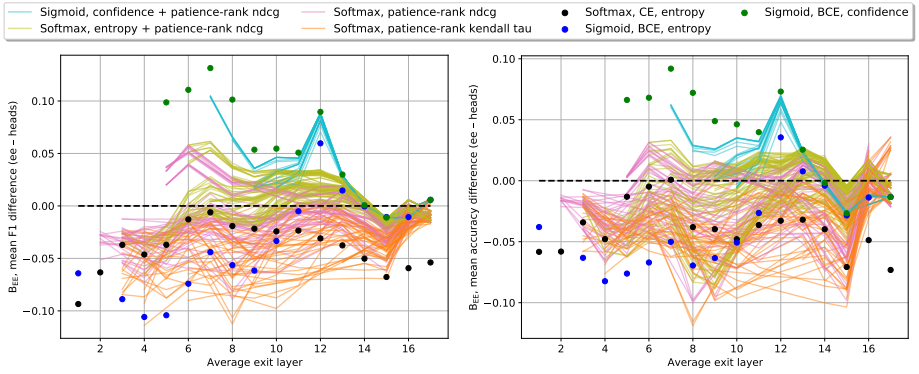

Fig. 4: The benefits B_{EE} of using rank-patience-based early-exiting strategies for multi-label VOC test dataset. Rank-patience-based approaches generally yield less promising results compared to simple confidence-threshold strategies. However, there are notable benefits in employing rank-patience-based approaches to enhance accuracy at higher speedups.

Rank-patience-based early-exiting strategy for multi-label classification Recognizing the performance gap outlined in previous experiments, we decided to further explore the alternate early-exit strategies for multi-label classification. Figure 4 presents the performance outcomes when implementing rank-patience-based strategies and hybrid strategies that combine rank-patience information with confidence or entropy thresholds. We employ two metrics to assess rank agreement: Normalized Discounted Cumulative Gain (nDCG) and Kendall's tau.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97626-1_13

Models are trained using different learning rates ($\lambda \in \{10^{-3}, 10^{-4}, 10^{-5}\}$), patience thresholds ($t \in \{2, 3, 4, 5, 6\}$), and rank-agreement tolerance thresholds (for nDCG scores values below $\{0.001, 0.01, 0.05, 0.1, 0.15\}$ and for Kendall's tau correlation tolerances values above $\{0.7, 0.8, 0.9, 0.95\}$).

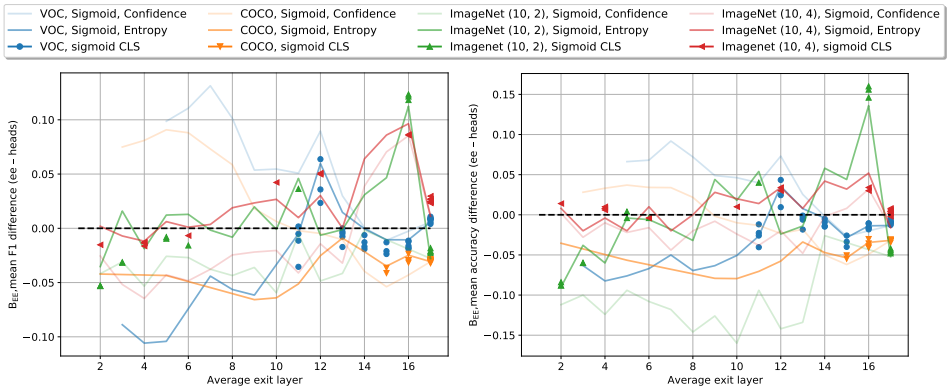

Fig. 5: Benefits B_{EE} of the learning-based early-exit strategy compared to baseline confidence and entropy methods. For each dataset, we show the best results from baseline methods alongside various hyperparameter configurations of the learning-based approach. Each point of the learning-based strategy represents a distinct model setup, illustrating that while there are modest performance improvements at certain speed-ups, these gains vary depending on the dataset.

Figure 4 showcases the results from the VOC dataset, where we tested various learning rates, consecutive head agreements, and prediction order tolerance thresholds. Our analysis reveals that despite rigorous testing, strategies based on patience and predicted probability order did not surpass basic methods such as the confidence-based strategy, with all results being statistically significant ($p \leq 0.05$). However, certain experiments demonstrated modest accuracy improvements at higher average exit layers, likely due to enhanced decision consensus in deeper layers, resulting in more accurate and reliable predictions.

Learning-based early-exiting We further explore the learning-based (classifier-based) early-exiting strategy. We introduce an auxiliary classifier added to each classification head, deciding whether an example should exit at that head. During evaluation, the output value (following sigmoid activation) ranges from 0 to 1, with a value exceeding 0.5 indicating a decision to exit. This single-point decision criterion directly determines the model's performance and specific speedup. While this strategy eliminates extra threshold selection, it also restricts the flexibility of controlling the performance-speedup ratio, offered by other strategies.

For the input to the early-exit decision classifiers, we use the logits from the corresponding classification head outputs (experiments with hidden states did not notably improve performance). We evaluated several common binary

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97626-1_13

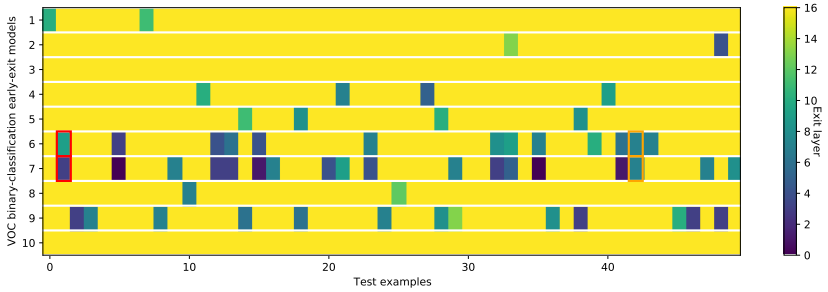

Fig. 6: Illustration of the challenge in early-exit models for multi-label classification. The x-axis shows test examples; the y-axis, ten class-specific binary models. Colors indicate the chosen exit head, highlighting varied recognition difficulty across classes. Rare instances where two different models exit at the same layer are shown in orange, while an example with objects of significantly different recognition difficulty, is highlighted in red.

classification loss functions: Binary Cross-Entropy (BCE), Mean Squared Error (MSE), L1 loss, and Hinge loss, across different hyperparameter configurations. Mean Squared Error loss with regularization, as shown in Equation (1), yielded the best performance. To maintain conciseness, we present only the results from this loss function. Notably, the parameters of the main model remain frozen while training the early-exit decision classifiers, allowing these classifiers to be trained independently.

Figure 5 illustrates the performance of the learning-based early-exiting strategy in comparison to the baseline confidence and entropy strategies across different multi-label datasets: VOC, COCO, and ImageNet. While the learning-based approach achieves performance improvements at certain speedups, the extent of these improvements varies significantly across datasets. This variability highlights the importance of customizing early-exit strategies to fit specific dataset characteristics. Further exploration of learning-based strategies with different architectures and learning schemes may lead to more robust and universally effective solutions for multi-label classification tasks.

4.3 Ablation study

Our exploration of the early-exit limitations in multi-label classification reveals that they provide fewer benefits compared to single-label scenarios. We suppose this complexity arises from the distinct recognition difficulty associated with each class within a single example. In other words, when we have a lot of classes present in one image, the network recognizes the presence of objects from different classes at different processing stages. To validate our hypothesis, we conduct a simple experiment, depicted in Figure 6.

We train ten early-exit models, each dedicated to a unique class from the VOC dataset. Each model consists of binary classifiers heads that determine the

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97626-1_13

presence or absence of its designated class. If the classifier recognizes its class as present, it exits at the current head. We selected test examples that included at least two of the ten classes we focused on. In Figure 6, each point on the x-axis represents a test example, with its color indicating the exit head chosen by the respective model on y-axis. We highlight with orange color when two distinct models, each tasked with a different class, exit at the same layer. The diversity of exit heads across the models illustrates that different classes within an example indeed present varying degrees of recognition difficulty, thereby reinforcing our initial hypothesis.

Our findings suggest that in a multi-label settings, the idea of early-exit must extend beyond the traditional notion of determining an optimal exit point based on the overall prediction confidence. Instead, it should consider the varied recognition difficulty of each class within an example, hence posing a more nuanced problem. Our intuition is that the key characteristic impacting the effectiveness of early-exits could be the similarity of the classes within a dataset. When classes are very similar, the problem resembles multi-label case, making it more challenging for the model to distinguish between them, thus requiring more processing. Conversely, when classes are very distinct, they are easier to recognize, potentially benefiting more from early-exit strategies.

5 Conclusions

We introduced a systematic framework for evaluating early-exit architectures in single- and multi-label classification tasks, demonstrating its effectiveness in computer vision. We found that early-exits reduce computational time in single-label tasks with minimal accuracy loss, while their benefits in multi-label tasks are limited due to the complexity of recognizing multiple classes (see Section 4.3). We revisited the "big label" problem, suggesting that challenges in single-label tasks with many output classes arise more from dataset characteristics than from early-exit limitations. For single-label tasks, widely used confidence-based strategies proved highly effective. In contrast, multi-label classification presents greater challenges, underscoring the need for adaptive exit strategies tailored to class-specific difficulties. This presents a promising direction for future research aimed at improving both performance and efficiency.

Acknowledgements The work of Klaudia Bałazy was carried out within the research project "Bio-inspired artificial neural network" (grant no. POIR.04.04.00-00-14DE/18-00) within the Team-Net program of the Foundation for Polish Science co-financed by the European Union under the European Regional Development Fund. This research was partially funded by the National Science Centre, Poland, grants no. 2023/49/B/ST6/01137 (work by Jacek Tabor). Some experiments were performed on servers purchased with funds from the flagship project entitled “Artificial Intelligence Computing Center Core Facility” from the DigiWorld Priority Research Area within the Excellence Initiative – Research University program at Jagiellonian University in Krakow.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97626-1_13

Bibliography

[1] Canwen Xu and Julian J. McAuley. A survey on dynamic neural networks for natural language processing. In *EACL (Findings)*, pages 2325–2336. Association for Computational Linguistics, 2023.

[2] Alex Graves. Adaptive computation time for recurrent neural networks. *arXiv preprint arXiv:1603.08983*, 2016.

[3] Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Lukasz Kaiser. Universal transformers. In *ICLR*, 2019.

[4] Surat Teerapittayanon, Bradley McDanel, and Hsiang-Tsung Kung. Branchynet: Fast inference via early exiting from deep neural networks. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 2464–2469. IEEE, 2016.

[5] Yigitcan Kaya, Sanghyun Hong, and Tudor Dumitras. Shallow-deep networks: Understanding and mitigating network overthinking. In *International conference on machine learning*, pages 3301–3310. PMLR, 2019.

[6] Xiangyang Liu, Tianxiang Sun, Junliang He, Jiawen Wu, Lingling Wu, Xinyu Zhang, Hao Jiang, Zhao Cao, Xuanjing Huang, and Xipeng Qiu. Towards efficient nlp: A standard evaluation and a strong baseline. *arXiv preprint arXiv:2110.07038*, 2021.

[7] Weijie Liu, Xin Zhao, Zhe Zhao, Qi Ju, Xuefeng Yang, and Wei Lu. An empirical study on adaptive inference for pretrained language model. *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[8] Ji Xin, Raphael Tang, Jaejun Lee, Yaoliang Yu, and Jimmy Lin. Deebert: Dynamic early exiting for accelerating BERT inference. In *ACL*, pages 2246–2251. Association for Computational Linguistics, 2020.

[9] Weijie Liu, Peng Zhou, Zhiruo Wang, Zhe Zhao, Haotang Deng, and Qi Ju. Fastbert: a self-distilling BERT with adaptive inference time. In *ACL*, pages 6035–6044. Association for Computational Linguistics, 2020.

[10] Shijie Geng, Peng Gao, Zuohui Fu, and Yongfeng Zhang. Romebert: Robust training of multi-exit bert. *arXiv preprint arXiv:2101.09755*, 2021.

[11] Roy Schwartz, Gabriel Stanovsky, Swabha Swayamdipta, Jesse Dodge, and Noah A. Smith. The right tool for the job: Matching model and instance complexities. In *ACL*, pages 6640–6651. Association for Computational Linguistics, 2020.

[12] Maciej Wołczyk, Bartosz Wójcik, Klaudia Bałazy, Igor T Podolak, Jacek Tabor, Marek Śmieja, and Tomasz Trzcinski. Zero time waste: recycling predictions in early exit neural networks. *Advances in Neural Information Processing Systems*, 34:2516–2528, 2021.

[13] Jue Wang, Ke Chen, Gang Chen, et al. Skipbert: Efficient inference with shallow layer skipping. In *ACL*, 2022.

[14] Wangchunshu Zhou, Canwen Xu, Tao Ge, Julian McAuley, Ke Xu, and Furu Wei. Bert loses patience: Fast and robust inference with early exit. *Advances in Neural Information Processing Systems*, 33:18330–18341, 2020.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97626-1_13

[15] Wei Zhu. Leebert: Learned early exit for BERT with cross-level optimization. In *ACL-IJCNLP*, pages 2968–2980. Association for Computational Linguistics, 2021.

[16] Ji Xin, Raphael Tang, Yaoliang Yu, and Jimmy Lin. Berxit: Early exiting for BERT with better fine-tuning and extension to regression. In *EACL*, pages 91–104. Association for Computational Linguistics, 2021.

[17] Zhen Zhang, Wei Zhu, Jinfan Zhang, Peng Wang, Rize Jin, and Tae-Sun Chung. PCEE-BERT: Accelerating BERT inference via patient and confident early exiting. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 327–338, Seattle, United States, July 2022. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.findings-naacl.25. URL `https://aclanthology.org/2022.findings-naacl.25`.

[18] Tianxiang Sun, Yunhua Zhou, Xiangyang Liu, Xinyu Zhang, Hao Jiang, Zhao Cao, Xuanjing Huang, and Xipeng Qiu. Early exiting with ensemble internal classifiers. *arXiv preprint arXiv:2105.13792*, 2021.

[19] Tal Schuster, Adam Fisch, Tommi Jaakkola, and Regina Barzilay. Consistent accelerated inference via confident adaptive transformers. *arXiv preprint arXiv:2104.08803*, 2021.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[21] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. https://doi.org/10.1109/CVPR.2009.5206848.

[22] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.

[23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97626-1_13