Supplementary Material

Experimental Setup

In our experiments, given K resource categories, we simulate the resource capacity of each category by enforcing an even spacing principle if possible such that $l_k = l_0 + k\Delta L$ for $k \in \{1, 2, \dots K-1\}$, where l_0 is the location of the first exit, $\Delta L = \lfloor \frac{N-l_0}{K-1} \rfloor$ and $l_K = N$ is the last exit, i.e. the full model. We perform experiments with ResNet56 [11] on CIFAR-10 and with DenseNet121 [14] on CIFAR-100. We use the default ResNet settings for 56-layer architecture and insert two evenly spaced early exits at the 18th and 36th layers (K=3). For DenseNet, we follow the default settings for 121-layer configuration and insert three early exit layers at the 12th, 36th and 84th layers due to transition layers (K=4). We train these models using Adam optimizer [15] for 150 epochs (first 20 epochs without early exits) and a batch size of 128, with the initial learning rate of 0.1 (decays by 0.1 at 50th and 100th epochs). On the ImageNet dataset, we use MSDNet [13] with 35 layers, 4 scales and 32 initial hidden dimensions. We insert four evenly spaced early exits at the 7th, 14th, 21th and 28th layers (K=5). Each early exit classifier consists of three 3x3 convolutional layers with ReLU activations. We set $\alpha_{KL} = 0.01$ and activate it after completing 75% of the training. For HRNet, we inject exits after the 2nd and 3rd stages with structures as in [21]. For BERT, we insert three evenly spaced early exits at the 3rd, 6th and 9th layers (K=4). Each early exit classifier consists of a fully-connected layer. We finetune the models for 20 epochs using gradient descent with a learning rate of 3e-5 and batch size of 16.

For our approach, based on validation performance, we set $D_h = 0.5D$ and $D_h = 2D$ for image/text classification experiments respectively. For image segmentation, we first downsample predictions by four with bilinear sampling and then operate on the mean of pixel-level computations. We optimize the weights using Adam optimizer with the learning rate of 3e-5 on validation data and set $\alpha_{cost} = 10$. We observe that our algorithm satisfies the budget constraint with this setting and it is also robust with respect to selections within the range of $\alpha_{cost} \in [1, 100]$. In all experiments, we stop the optimization if the loss does not decrease for 50 consecutive epochs on the validation set. Inference measurements for CIFAR experiments are carried out on a machine with an 8-core 2.9GHz CPU, other experiments on a machine with RTX3060 GPU, and repeated ten times. The extra inference time caused by the exit score computations is also included in the reported latency measurements, and the cost is much smaller compared to the cost of the forward pass of the model as shown in Table 3.

Effect of Self-Distillation

We use the same model trained with self-distillation while comparing our scheduling policy with other early exit methodologies in all reported results. To analyze the improvements obtained with self-distillation, we also report the results on CIFAR datasets without applying self-distillation during training in Table 4. Compared to the results provided in Table 1, we observe up to 1% accuracy decrease in EENet and BranchyNet, and up to 1.6% accuracy decrease in MSDNet when self-distillation during training is disabled.

Dataset	Budget	BranchyNet	MSDNet	EENet
CIFAR-10	3.50 ms	93.55	93.60	93.69
	3.00 ms	92.11	92.39	92.58
	2.50 ms	86.98	88.10	88.51
CIFAR-100	7.50 ms	73.90	73.88	74.05
	6.75 ms	70.99	70.96	71.54
	6.00 ms	67.09	67.15	68.56

Table 4. Accuracy (%) values under various budget settings on CIFAR datasets without self-distillation during multi-exit model training.

Ablation Study of Design Components

We also analyze the effect of different components in EENet on performance by in-depth investigation of the results for SST-2. Figure 6 provides the plot of average inference time vs. accuracy for two additional variants of EENet and compares with MSDNet and BranchyNet. The first variant of our approach shows the results of our system without optimizing the exit scoring, and instead, directly using maximum prediction scores. The second variant shows

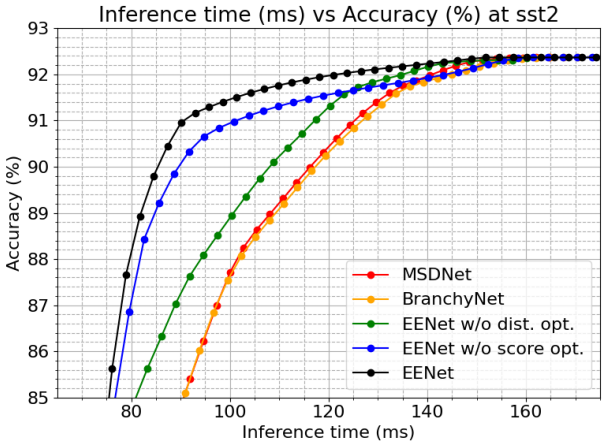


Figure 6. Average latency (ms) vs Accuracy (%) results at SST-2 for BranchyNet, MSDNet, and EENet variations (without distribution/scoring optimization).

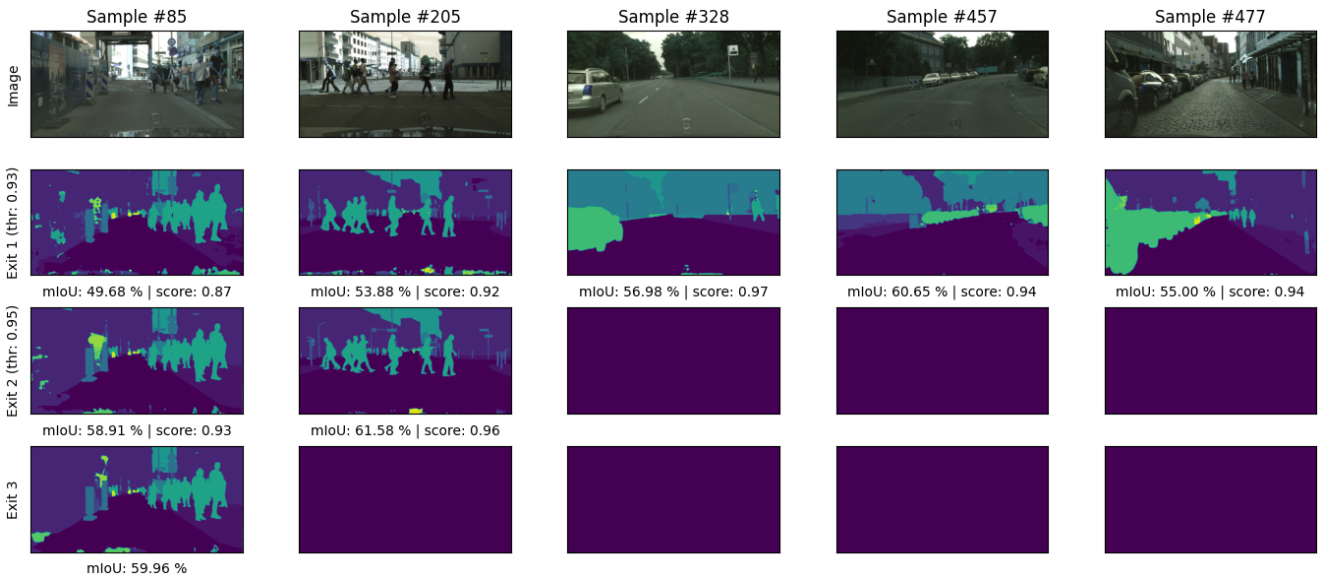


Figure 7. Visual comparison of randomly selected image segmentation examples from Cityscapes dataset for 50 ms/sample budget. Empty images indicate that the sample has exited previously as a result of having a higher exit score than the computed exit threshold.

the results of EENet without optimizing exit distributions through our budget-constrained learning, and instead, directly using geometric distribution. We observe that optimization of both exit scoring functions and distributions to obtain thresholds contributes to the superior performance of EENet.

Visual Analysis of Early Exit Behavior

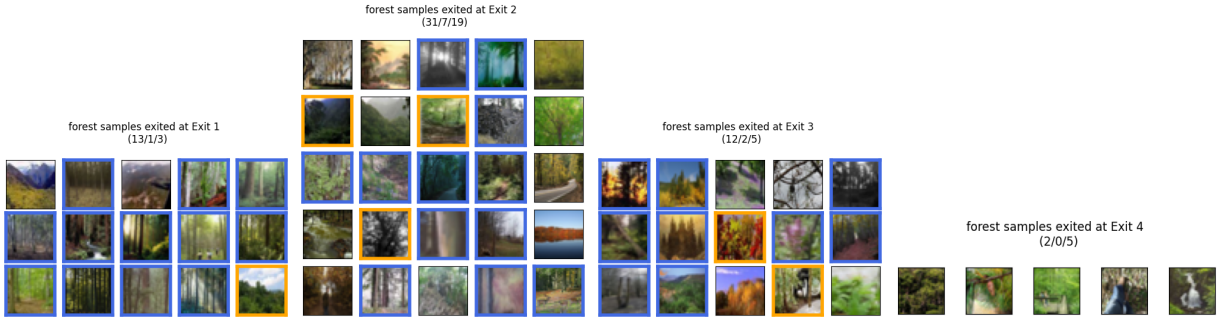
We analyze the early exiting behavior of EENet on different tasks by qualitatively investigating the samples. On image segmentation, we observe that frames with less objects tend to exit earlier as shown in Figure 7. We also illustrate test samples exited at each exit for four different classes from CIFAR-100 data in Figure 8. It is visually clear that the easier samples exit earlier to utilize the provided more efficiently. We observe that EENet utilizes the second exit in this particular scenario very efficiently by assigning easy samples and obtaining higher accuracy with significantly lower average latency. We conduct a similar analysis for the experiment set on AgNews test data in Figure 9. Similarly, EENet utilizes the second exit efficiently by assigning easy samples and obtaining higher accuracy with significantly lower latency.

Adapting to Dynamic Budget Settings

In practice, the test dataset may contain significantly easier/harder or out-of-distribution samples. Data distribution can also change over time. However, none of the existing adaptive inference algorithms strictly meets the latency budget over test data since the optimization is performed over the training or validation dataset. We also optimize the scheduling policy over the validation dataset however, our solution provides lightweight easy-to-optimize schedulers without requiring any changes on the full model itself. Therefore, a simple yet effective approach of switching between a few schedulers (optimized for different budget values) in an online manner is possible in our framework and can be utilized if necessary. For instance, we consider the scenario for CIFAR experiments, where during the test, we can switch between schedulers trained for three different budget values as provided in Table 1. For the 3.0 ms/sample budget setting on CIFAR-10, if the test samples are easier/harder than expected and the realized latency per sample is getting lower/higher than the provided budget, we can switch to the scheduler optimized for the budget of 3.50/2.50 ms per sample. To this end, we compute the remaining budget per sample after each inference operation and switch to the scheduler trained under the closest budget setting. We provide the results in Table 5.

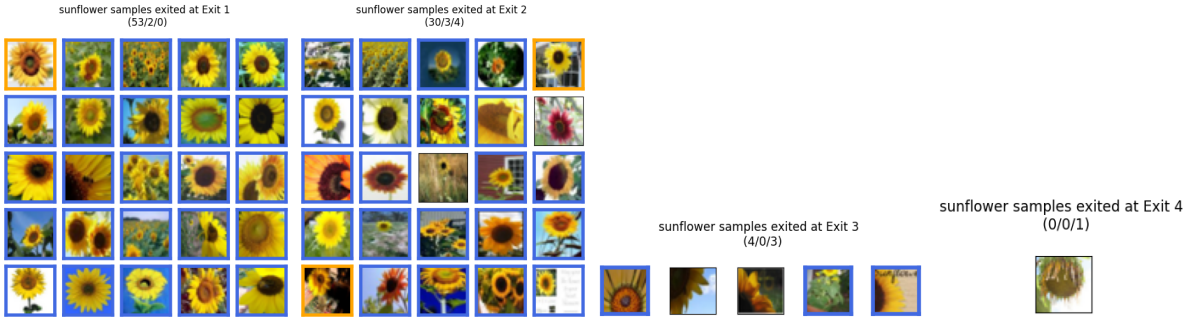
Dataset	Budget	Method	Latency	Accuracy (%)
CIFAR-10	3.00 ms	BranchyNet	2.87 ms	92.57
		EENet	2.85 ms	92.90
		EENet w/ online switch	2.98 ms	92.92
CIFAR-100	6.75 ms	BranchyNet	6.55 ms	71.65
		EENet	6.61 ms	72.12
		EENet w/ online switch	6.74 ms	72.11

Table 5. Test accuracy values on CIFAR datasets and the realized latency/sample values during the test for BranchyNet, EENet, and EENet with online switching during inference.

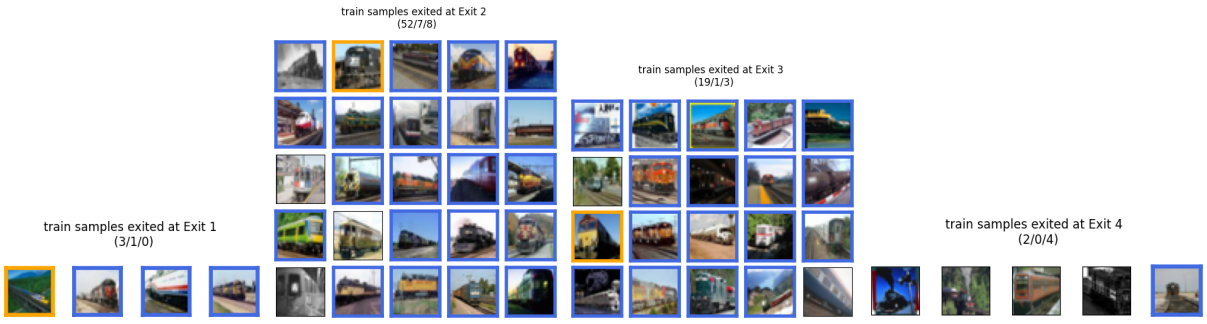


(a) Exit 1 - forest (b) Exit 2 - forest (c) Exit 3 - forest (d) Exit 4 - forest

(e) Exit 1 - mountain (f) Exit 2 - mountain (g) Exit 3 - mountain (h) Exit 4 - mountain

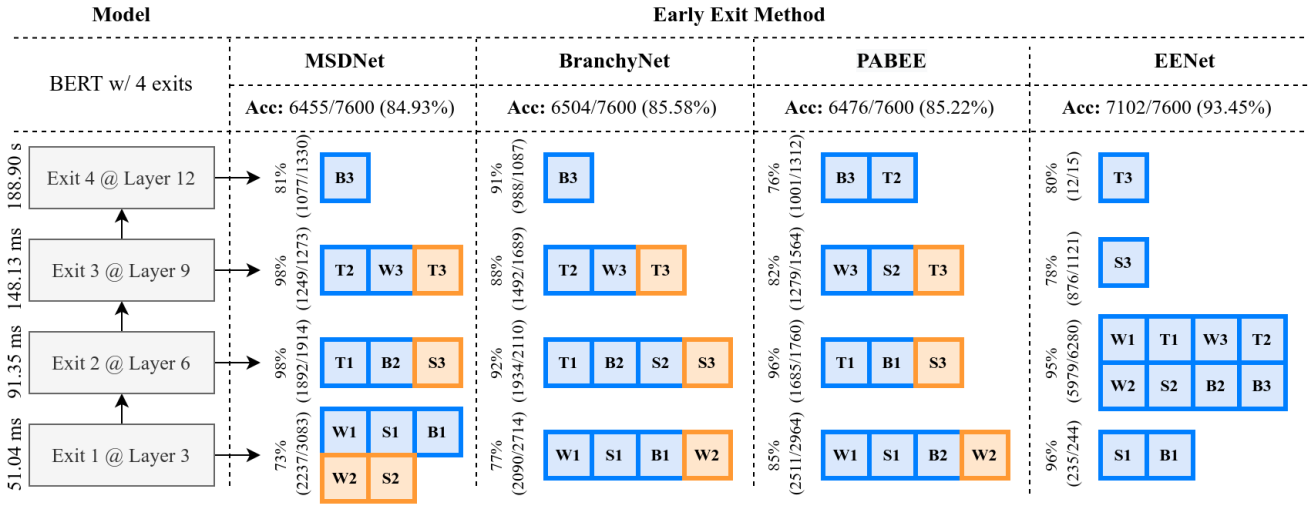


(i) Exit 1 - sunflower (j) Exit 2 - sunflower (k) Exit 3 - sunflower (l) Exit 4 - sunflower



(m) Exit 1 - train (n) Exit 2 - train (o) Exit 3 - train (p) Exit 4 - train

Figure 8. Samples from CIFAR-100 test set for forest/mountain/sunflower/train classes. Each subfigure illustrates the samples at the corresponding exit of DenseNet121 with four exits under the average inference budget of 6.25 milliseconds/sample. Blue borders indicate correct predictions. Orange borders indicate incorrectly predicted samples and if it were exiting at the last exit, it would be correctly predicted. No borders indicate the samples incorrectly predicted by EENet at that exit and also the last exit. The values on the titles indicate the number of samples from these categories (correctly predicted / incorrectly predicted at that exit / also incorrectly predicted at the last exit).



World-1	Venezuelans voted resoundingly to keep firebrand populist Hugo Chavez as their president in a victory that drew noisy reactions yesterday from...
World-2	Students at the Mount Sinai School of Medicine learn that diet and culture shape health in East Harlem.
World-3	Fewer Americans lined up to claim first-time jobless benefits last week but analysts said the modest decline said very little about the current state...
Sports-1	World 100 meters champion Torri Edwards will miss the Athens Olympics after her appeal against a two-year drugs ban was dismissed on Tuesday...
Sports-2	One day after placing a waiver claim on troubled cornerback Derek Ross, the Saints did an about-face and released the former Ohio State...
Sports-3	Tim Henman confirmed he was in good health, despite being diagnosed with a magnesium deficiency, after a straight-sets win over Antony Dupuis in the...

Business-1	With the much-ballyhooed initial public offering of Google behind them and oil chugging to a new record high, investors took a step back today.
Business-2	The price of oil charged to a new high above $47 a barrel yesterday amid nagging concerns about instability in Iraq, the uncertain fate of Russian...
Business-3	Samsung Electronics, the world #39;s second largest computer chip manufacturer, yesterday said that it would invest Won25,000bn (\$24bn)...
Tech-1	Dell Inc. (DELL.O), the world's\largest PC maker, could announce an expanded selection of its\consumer electronics line in the next several...
Tech-2	New antispam technology standards are on the way that promise to hit spammers where it hurts the most--their wallets. At issue is the ability to...
Tech-3	An experiment using two orbiting satellites has proved that as the Earth turns it drags space and time around itself, like a spinning top in treacle.

Figure 9. Visual comparison of the early exit approaches on AgNews test data with BERT (4 exits) for the average latency budget of 100 milliseconds. We illustrate the randomly selected twelve samples from four classes and the exit location that they were assigned. Images with green/red borders are predicted correctly/incorrectly at the corresponding exit. We also report the number of correct predictions and exited samples at each exit. Our approach does not make costly assignments to the last two exits and uses the second exit more effectively.