Adaptative Inference Cost With Convolutional Neural Mixture Models
Adrià Ruiz, Jakob Verbeek

To cite this version:

HAL Id: hal-02267564

https://hal.archives-ouvertes.fr/hal-02267564
Submitted on 19 Aug 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Adaptative Inference Cost With Convolutional Neural Mixture Models

Adria Ruiz Jakob Verbeek
Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJJK, 38000 Grenoble, France
firstname.lastname@inria.fr

Abstract

Despite the outstanding performance of convolutional neural networks (CNNs) for many vision tasks, the required computational cost during inference is problematic when resources are limited. In this context, we propose Convolutional Neural Mixture Models (CNMMs), a probabilistic model embedding a large number of CNNs that can be jointly trained and evaluated in an efficient manner. Within the proposed framework, we present different mechanisms to prune subsets of CNNs from the mixture, allowing to easily adapt the computational cost required for inference. Image classification and semantic segmentation experiments show that our method achieves excellent accuracy-compute trade-offs. Moreover, unlike most of previous approaches, a single CNMM provides a large range of operating points along this trade-off, without any re-training.

1. Introduction

Convolutional neural networks (CNNs) form the basis of many state-of-the-art computer vision models. Despite their outstanding performance, the computational cost of inference in these CNN-based models is typically very high. This holds back applications on mobile platforms, such as autonomous vehicles, drones, or phones, where computational resources are limited, concurrent data-streams need to be processed, and low-latency prediction is critical.

To accelerate CNNs we can reduce their complexity before training, e.g. by decreasing the number of filters or network layers. This solution, however, may lead to suboptimal results given that over-parametrization plays a critical role in the optimization of deep networks [7, 9]. Fortunately, other studies have found a complementary phenomena: given a trained CNN, a large number of its filters are redundant and do not have a significant impact on the final prediction [26]. Motivated by these two findings, much research has focused on accelerating CNNs using network pruning [11, 19, 22, 35, 38, 41, 47, 56]. Pruning can be applied at multiple levels, e.g. by removing independent filters [35, 41], groups of them [11, 19], or entire layers [56].

Despite the encouraging results of these methods, their ability to provide a wide range of operating points along the trade-off between accuracy and computation is limited. The reason is that these approaches typically require to train a separate model for each specific pruning level.

In this paper, we propose Convolutional Neural Mixture Models (CNMMs), which provide a novel perspective on network pruning. A CNMM defines a distribution over a large number of CNNs. The mixture is naturally pruned by removing networks with low probabilities, see Figure 1. Despite the appealing simplicity of this approach, it presents several challenges. First, learning a large ensemble of CNNs may require a prohibitive amount of computation. Second, even if many networks in the mixture are pruned, their independent evaluation during inference is likely to be less efficient than computing the output of a single large model.

In order to ensure tractability, we design a parameter-sharing scheme between different CNNs. This enables us to (i) jointly train all the networks, and (ii) efficiently compute an approximation of the mixture output without independently evaluating all the networks.

Image classification and semantic segmentation experiments show that CNMMs achieve an excellent trade-off between prediction accuracy and computational cost. Unlike most previous network pruning approaches, a single
CNMM model achieves a wide range of operating points along this trade-off without any re-training.

2. Related work

Neural network ensembles. Learning ensembles of neural networks is a long-standing research topic. Seminal works explored different strategies to combine the outputs of different networks to obtain more accurate predictions [30, 46, 61]. Recently, the success of deep models has renewed interest in ensemble methods.

For this purpose, many approaches have been explored. For instance, [31, 62] used bagging [3] and boosting [49] to train multiple networks. Other works have considered to learn diverse models by employing different parameter initializations [34], or re-training a subset of layers [60]. While these strategies are effective to learn diverse networks, their main limitation is the required training cost. In practice, training a deep model can take multiple days, and therefore large ensembles may have a prohibitive cost. To reduce the training time, it has been suggested [18, 39] to train a single network and to use parameters from multiple iterations of the optimization process to define the ensemble. Despite the efficiency of this method during training, this approach does not reduce inference cost, since multiple networks must be evaluated independently at test time.

An alternative strategy to allow efficient training and inference is to use implicit ensembles [11, 21, 32, 47]. By relying on sampling, these methods allow to jointly train all the individual components in the ensemble and perform approximate inference during testing. Bayesian neural networks (BNNs) fall in this paradigm and use a distribution over parameters, rather than a single point estimate [11, 28, 40, 47]. A sample from the parameter distribution can be considered as an individual network. Other works have implemented the notion of implicit ensembles by using dropout [52] mechanisms. Dropping neurons can be regarded as sampling over a large ensemble of different networks [2]. Moreover, scaling outputs during testing according to the dropout probability can be understood as an approximated inference mechanism. Motivated by this idea, different works have applied dropout over individual weights [10], network activations [50], or connections in multi-branch architectures [12, 32]. Interestingly, it has been observed that ResNets [13] behave like an ensemble of models, where some residual connections can be removed without significantly reducing prediction accuracy [55]. This idea was used by ResNets with stochastic depth [21], where different dropout probabilities are assigned to the residual connections.

Our proposed Convolutional Neural Mixture Model is an implicit ensemble defining a mixture distribution over an exponential number of CNNs. This allows to use the learned probabilities to prune the model by removing non-relevant networks. Using a mixture of CNNs for model pruning is a novel approach, which contrasts to previous methods employing ensembles for other purposes such as boosting performance [8, 34], improving learning dynamics [21], or uncertainty estimation [24, 31].

Efficient inference in deep networks. A number of strategies have been developed to reduce the inference time of CNNs, including the design of efficient convolutional operators [16, 23, 58], knowledge distillation [4, 15], neural architecture search [14, 63], weight compression [43, 54], and quantization [25, 36]. Network pruning has emerged as one of the most effective frameworks for this purpose [11, 33, 35, 56]. Pruning methods aim to remove weights which do not have a significant impact on the network output. Among these methods, we can differentiate between two main strategies: online and offline pruning. In offline pruning, a network is first optimized for a given task using standard training. Subsequently, non-relevant weights are identified using different heuristics including their norm [35], similarity to other weights [51], or second order derivatives [33]. The main advantage of this strategy is that it can be applied to any pre-trained network. However, these approaches require a costly process involving several prune/retrain cycles in order to recover the original network performance. Online approaches, on the other hand, perform pruning during network training. For example, sparsity inducing regularization can be used over individual weights [38, 40], groups of them [11, 19, 47], or over the connections in multi-branch architectures [1, 56]. These methods typically have a hyper-parameter, to be set before training, determining the trade-off between the final performance and the pruning ratio.

In contrast to previous approaches, we prune entire CNNs by removing the networks with the smallest probabilities in the mixture. This approach offers two main advantages. First, it does not require to define a hyper-parameter before training to determine the balance between the potential compression and the final performance. Second, the number of removed networks can be controlled after optimization. Therefore, a learned CNMM can be deployed at multiple operating points to trade-off computation and prediction accuracy. For example, across different devices with varying computational resources, or on the same device with different computational constraints depending on the processor load of other processes. The recently proposed Slimmable Neural Networks [57] have also focused on adapting the accuracy-efficiency trade-off at run time. This is achieved by embedding a small set of CNNs with varying widths into a single model. Different from this approach, our CNMMs embed a large number of networks with different depths, which allows for a finer granularity to control the computational cost during pruning.
Figure 2. (Left) Illustration of how a large collection of CNNs is represented in a CNMM. Each network is uniquely identified by a non-decreasing sequence $s_{0:T}$, containing numbers from 0 to $T$. Consecutive entries in the sequence determine the functions $f_{s_t-1}^i$ applied to compute the CNN output. In this manner, sequences with common sub-sequences share functions and their parameters in the corresponding networks. (Right) Illustration of the distribution $p(s_{0:T})$ that defines the mixing weights over the CNN models, here $T = 4$. Each $p(s_{t-1}|s_t)$ is a Bernoulli distribution on whether $s_{t-1}$ equals $s_t$ or $t-1$. This defines a binary tree generating all the valid sequences.

### 3. Convolutional Neural Mixture Models

Without loss of generality, we consider a CNN as a function $\mathcal{F}(\mathbf{H}_0) = \mathbf{H}_T$ mapping an RGB image $\mathbf{H}_0 \in \mathbb{R}^{W_0 \times H_0 \times 3}$ to a tensor $\mathbf{H}_T \in \mathbb{R}^{W \times H \times C}$. In particular, we assume that $\mathcal{F}$ is defined as a sequence of $T$ operations:

$$ \mathcal{F}(\mathbf{H}_0) = f_{T-1}^T(\cdots f_1^T(\mathbf{H}_0))) $$

where $\mathbf{H}_t = f_{t-1}^T(\mathbf{H}_{t-1})$ is computed from the previous feature map $\mathbf{H}_{t-1}$. We assume that the functions $f_{t-1}^T$ can be either the identity function, or a standard CNN block composed of different operations such as batch-normalization, convolution, activation functions, or spatial pooling. In this manner, the effective depth of the network, i.e. the number of non-identity layers $f_{t-1}^T$, is at most $T$.

The output tensor $\mathbf{H}_T$ of the CNN is used to make predictions for a specific task. For example, in image classification, a linear classifier over $\mathbf{H}_T$ can be used in order to estimate the class probabilities for the entire image. For semantic segmentation the same linear classifier is used for each spatial position in $\mathbf{H}_T$.

Given these definitions, a convolutional neural mixture model (CNMM) defines a distribution over output $\mathbf{H}_T$ as:

$$ p(\mathbf{H}_T|\mathbf{H}_0) = \sum_{\mathcal{F} \in \mathcal{F}} p(\mathcal{F})p(\mathbf{H}_T|\mathbf{H}_0, \mathcal{F}) $$

where $\mathcal{F} = \{\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_K\}$ is a finite set of CNNs, $p(\mathbf{H}_T|\mathbf{H}_0, \mathcal{F}_k)$ is a delta function centered on the output $\mathcal{F}_k(\mathbf{H}_0)$ of each network, and $p(\mathcal{F})$ defines the mixing weights over the CNNs in $\mathcal{F}$.

#### 3.1. Modelling a distribution over CNNs

We now define mixtures that contain a number of CNNs that is exponential in the maximum depth $T$, in a way that allows us to manipulate these mixtures in a tractable manner.

Each component in the mixture is a chain-structured CNN uniquely characterised by a sequence $s_{0:T}$ of length $T + 1$, where the sequences are constrained to be non-decreasing of integers from 0 to $T$, i.e. with $s_0 = 0$, $s_T = T$ and $s_{t+1} \geq s_t$. This sequence determines the set of functions that are used in Eq. (1). In particular, given a sequence $s_{0:T}$, the output of the corresponding network is computed as:

$$ \mathcal{F}(\mathbf{H}_0) = f_{s_T}^{s_T}(\cdots f_{s_0}^1(\mathbf{H}_0))) $$

For $i < j$ the function $f_i^j$ is a convolutional block as described above with its own parameters, while the functions $f_i^1$ are identity functions that leave the input unchanged.

By, imposing $s_{t-1} \in \{t-1, s_t\}$, there is a one-to-one mapping between sequences $s_{0:T}$ and the corresponding CNNs. If multiple networks use the same function $f_i^j$, these networks share their parameters on this function, which ensures that the total number of parameters of the mixture does not grow exponentially, although there are exponentially many mixture components. For instance, for $T = 4$, the mixture will be composed of eight different networks illustrated in Figure 2 (Left). From the illustration it is easy to see that, in general, the mixture contains $2^T - 1$ components with shared parameters.

In order to define the probabilities $p(\mathcal{F})$ for each network in the mixture, we define a distribution over sequences $s_{0:T}$ as a reversed Markov chain:

$$ p(s_{0:T}) = p(s_T) \prod_{t=1}^{T} p(s_{t-1}|s_t) $$

To ensure that sequences have positive probability if and only if they are valid, i.e. satisfy the constraints defined

\[\text{In particular, this constraint ensures that, e.g., the network } f_1^1(\mathbf{H}_0) \text{ is uniquely encoded by the sequence '01444', ruling out the alternative sequences '01144' and '01114'. See Figure 2 (Left).} \]
above, we set \( p(s_T = T) = p(s_0 = 0|s_1) = 1 \) and define:

\[
p(s_{t-1}|s_t) = \begin{cases} 
\pi_{t-1}^s & \text{if } s_{t-1} = (t-1), \\
1 - \pi_{t-1}^s & \text{if } s_{t-1} = s_t, \\
0 & \text{otherwise.}
\end{cases}
\]

As illustrated in Figure 2 (Right), these constraints generate a binary tree generating valid non-decreasing sequences. The conditional probabilities \( p(s_{t-1}|s_t) \) are modelled by a Bernoulli distribution with probability \( \pi_{t-1}^s \), indicating whether the previous number in the sequence is \( s_t \) or \( t-1 \).

### 3.2. Sampling outputs from CNMMs

The graphical model defined in Figure 3 shows that we can sample from the output distribution \( p(\text{H}_T|\text{H}_0) \) in Eq. (2) by first generating a sequence from \( p(s_{0:T}) \) and then evaluating the associated network with Eq. (3). In the following, we formulate an alternative strategy to sample from the model. This formulation offers two advantages. (i) It is amenable to continuous relaxation, which facilitates learning. (ii) It suggests an iterative algorithm to compute feature map expectations, which can be used instead of sampling for efficient inference.

The conditional \( p(\text{H}_t|s_t = l, \text{H}_0) \) gives the distribution over \( \text{H}_t \) across the networks with \( s_t = l \). For example, \( p(\text{H}_2|s_2 = 4, \text{H}_0) \) consists of two weighted delta peaks, located at \( f_0^4(\text{H}_0) \) and \( f_0^2(\text{H}_0) \), respectively. See Figure 2 (Left). These conditional distributions can be expressed as the forwards recurrence:

\[
p(\text{H}_t|s_t, \text{H}_0) = \sum_{\text{H}_{t-1},s_{t-1}} p(\text{H}_t|\text{H}_{t-1},s_t,s_{t-1}) p(s_{t-1}|s_t) \times p(\text{H}_{t-1}|s_{t-1}, \text{H}_0),
\]

where \( p(\text{H}_t|\text{H}_{t-1}, s_t, s_{t-1}) \) is a delta function centered on \( f_{s_{t-1}}^s(\text{H}_{t-1}) \). Therefore, unbiased samples \( \tilde{\text{h}}_t^s \) from \( p(\text{H}_t|s_t, \text{H}_0) \) can be obtained through sample propagation. Recall from Eq. (5) that, given \( s_t \), there are only two possible values of \( s_{t-1} \) that remain, namely \( s_t \) and \( t-1 \). As a consequence, the sum over \( s_{t-1} \) in Eq. (6) only consists of two terms. Given this observation, samples \( \tilde{\text{h}}_t^s \sim p(\text{H}_t|s_t, \text{H}_0) \) can be obtained from samples

\[
\tilde{\text{h}}_{t-1}^{s_t} \sim p(\text{H}_{t-1}|s_{t-1}, \text{H}_0),
\]

where for a given value of \( s_t \) we sample \( s_{t-1} \) from \( p(s_{t-1}|s_t) \) to compute a binary indicator \( \pi_{t-1}^s = [s_{t-1} = t-1] \), which signals whether the resulting \( \tilde{\text{h}}_{t-1}^{s_t} \) is equal to \( \tilde{\text{h}}_{t-1}^{s_t} \) or \( f_{t-1}^s(\text{H}_{t-1}) \).

Using Eq. (7) we iterative sample from distributions \( p(\text{H}_t|s_t, \text{H}_0) \) for \( t = 1, \ldots, T \), and for each \( t \) we compute samples for \( s_t = t, \ldots, T \). An illustration of the algorithm is shown in Figure 4. The computational complexity of a complete pass in this iterative process is \( O(T(T + 1)/2) \), since for each \( t = 1, \ldots, T \), we compute \( T - t + 1 \) samples, each of which is computed in \( O(1) \) from the samples already computed for \( t - 1 \). This is roughly equivalent to the cost of evaluating a single network with dense layer connectivity of depth \( T \) [20], which has a total of \( O(T(T - 1)/2 \) connections implemented by the functions \( f_i^s \).

**Sampling outputs from networks of bounded depth.** Using the described algorithm, \( \tilde{\text{h}}_T^s \sim p(\text{H}_T|s_T = T, \text{H}_0) \) correspond to output tensors \( \text{H}_T \) sampled from the mixture defined in Eq. (2). Moreover, for any \( t \), samples from \( p(\text{H}_t|s_t = T, \text{H}_0) \) are output feature maps generated by networks with depth bounded by \( t \). For instance, in Figure 2, samples \( \tilde{\text{h}}_3^2 \) are generated with one of the networks coded by the sequences 01444 and 04444.

### 3.3. Training and inference

We use \( \psi \) to collectively denote the parameters of the convolutional blocks \( f_i^s \) and the parameters \( \pi_{s_{t-1}}^s \) defining the mixing weights via Eq. (5). Moreover, the parameters of the classifier that predict the image label(s) from the
output tensor \( \mathbf{H}_T \) are denoted as \( \theta \). Given a training set \( \mathcal{D} = \{ \{\mathbf{X}_1, y_1\}, \ldots, \{\mathbf{X}_N, y_N\} \} \) composed of images \( \mathbf{X}_n \) and labels \( y_n \), we optimize the parameters by minimizing

\[
\mathcal{L}_{\text{single}}(\psi, \theta) = \sum_{n=1}^{N} \mathbb{E}_{p(\mathbf{H}_T | \mathbf{H}_0 = \mathbf{X}_n; \psi)} \left[ \mathcal{L}(y_n, \mathbf{H}_T, \theta) \right],
\]

where \( \mathcal{L}(y_n, \mathbf{H}_T, \theta) \) is the cross-entropy loss comparing the label \( y_n \) with the class probabilities computed from \( \mathbf{H}_T \).

**Evaluating intermediate classifiers.** In practice, we replace the expectation over \( \mathbf{H}_T \) in each training iteration with samples from \( p(\mathbf{H}_T | \mathbf{H}_0 = \mathbf{X}_n; \psi) \).

**Learning from subsets of networks.** As discussed in Section 3.2, samples from the distribution \( p(\mathbf{H}_s | s_t = T, \mathbf{H}_0) \) correspond to outputs of CNNs in the mixture with depth at most \( t \). In order to improve performance of models with reduced inference time, we explicitly emphasize the loss for such efficient relatively shallow networks. Therefore, we sum the above loss function over the outputs sampled from networks of increasing depth:

\[
\mathcal{L}(\theta, \psi) = \sum_{n=1}^{N} \sum_{t=1}^{T} \mathbb{E}_{p(\mathbf{H}_t | s_t = T, \mathbf{X}_n; \psi)} \left[ \mathcal{L}(y_n, \mathbf{H}_t, \theta_t) \right],
\]

where we use a separate classifier for each \( t \). In practice, we balance each loss with a weight increasing linearly with \( t \).

**Relaxed binary variables with concrete distributions.** The recurrence in Eq. (7) requires sampling from \( p(s_{t-1} | s_t) \), defined in Eq. (5). The sampling renders the parameters \( \pi_{t-1}^s \) non-differentiable, which prevents gradient-based optimization for them. To address this limitation, we use a continuous relaxation by modelling \( p(s_{t-1} | s_t) \) as a binary “concrete” distribution [42]. In this manner, we can use the re-parametrization trick [27, 48] to back-propagate gradients w.r.t. samples \( \tilde{\pi}_{t-1}^s \) in Eq. (7) and, thus, to compute gradients for the parameters \( \pi_{t-1}^s \).

**Efficient inference by expectation propagation.** Once the CNMM is trained, the predictive distribution on \( y \) is given by \( p(y | \mathbf{X}; \theta) = \mathbb{E}_{p(\mathbf{H}_T | \mathbf{X})} [p(y | \mathbf{H}_T; \theta)] \). The expectation is intractable to compute exactly, contrary to our goal of efficient inference. A naive Monte-Carlo sample approximation is still required multiple evaluations of the full CNMM. Instead, we propose an alternative approximation by propagating expectations instead of samples in Eq. (7), i.e. using the approximation \( \mathbf{H}_T \approx p(\mathbf{H}_T | \mathbf{X}) \), where \( \mathbf{H}_T \) is obtained by running the sampling algorithm replacing the samples \( \tilde{\pi}_{t-1}^s \) with their expectations \( \pi_{t-1}^s \).

### 3.4. Accelerating CNMMs

CNMMs offer two complementary mechanisms in order to accelerate inference. We describe both in the following.

**Evaluating intermediate classifiers.** The different classifiers \( \theta_t \) learned by minimizing Eq. (9) operate over the outputs of a mixture of networks with maximum depth \( t \). Therefore, at each iteration \( t \) of the inference algorithm in Eq. (7) we can already output predictions based on classifier \( \theta_t \). This strategy is related with the one employed in multi-scale dense networks (MSDNets) [17], where “early-exit” classifiers are used to provide predictions at various points in time during the inference process.

**Network pruning.** A complementary strategy to accelerate CNMMs is to remove networks from the mixture. The computational cost of the inference process is dominated by the evaluation of the CNN blocks \( f_{t-1}^{s_{t-1}}(\mathbf{H}_{t-1}^{-1}) \) in Eq. (7). However, these functions do not need to be computed when the variable \( \tilde{\pi}_{t-1}^s = 0 \). Therefore, a natural approach to prune CNMMs is to set certain \( \pi_{t-1}^s \) to zero, removing all the CNNs from the mixture that use \( f_{t-1}^{s_{t-1}} \). We use the learned distribution \( p(s_{t-1} | T) \) in order to remove networks with a low probability. Note that for a given value of \( s_t \), the pairwise marginal \( p(s_t, s_{t-1} = t - 1) \) is exactly the sum of probabilities of all the networks involving the function \( f_{t-1}^{s_{t-1}} \). Using this observation, we use an iterative pruning algorithm where, at each step, we compute all pairwise marginals \( p(s_t, s_{t-1} = t - 1) \) for all possible values of \( s_t \) and \( t \). We then set \( \pi_{t-1}^s = 0 \) where \( (s_t, t^*) = \arg \min_{(s_t, t)} p(s_t, s_{t-1} = t - 1) \). Finally, the marginals are updated, and we iterate.

In this manner, we achieve different pruning levels by progressively removing convolutional blocks that will not be evaluated during inference. This process does not require any re-training of the model, allowing to dynamically set different pruning ratios. Note that this process is complementary to the use of intermediate classifiers, as discussed above. The reason for this is that our pruning strategy may be used to remove functions \( f_{t} \) for any “early” prediction step \( t < T \). Finally, it is interesting observe that the proposed pruning mechanism can be regarded as a form of neural architecture search [37, 63], where the optimal network connectivity for a given pruning ratio is automatically discovered by taking into account the learned probabilities \( p(s_{t-1} = t - 1 | s_t) \).
4. Experiments

We perform experiments over two different tasks: image classification and semantic segmentation. Following previous work, we measure the computational cost in terms of the number of floating point multiply and addition operations (FLOPs) required for inference. The number of FLOPs provides a metric that correlates very well with the actual inference wall-time, while being independent of implementation and hardware used for evaluation.

4.1. Datasets and experimental setup

CIFAR-10/100 datasets. These datasets [29] are composed of 50k train and 10k test images with a resolution of 32×32 pixels. The goal is to classify each image across 10 or 100 classes, respectively. Images are normalized using the means and standard deviations of the RGB channels. We apply standard data augmentation operations: (i) a 4-pixel zeros padding followed by 32×32 cropping. (ii) Random horizontal flipping with probability 0.5. Performance is evaluated in terms of the mean accuracy across classes.

CityScapes dataset. This dataset [6] contains 1024×2048 pixel images of urban scenes with pixel-level labels across 19 classes. The dataset is split into training, validation and test sets with 2,975, 500 and 1,525 samples each. The ground-truth annotations for the test set are not public, and test sets with 2,975, 500 and 1,525 samples each. The dataset is split into training, validation and test sets with 2,975, 500 and 1,525 samples each. The ground-truth annotations for the test set are not public, and the number of channels is doubled. Feature maps are connected by functions \( f_j \) if the output feature map \( H_j \) has the same or half the resolution of the input feature map \( H_i \). Finally, we consider the output tensor \( H_T \) to have different connectivity and spatial resolution depending on the task.

Implementation for image classification. We implement the convolutional layers as the set of operations (BN-ReLU-DConv-BN-ReLU-Conv-BN), where BN refers to batch normalization, DConv is a \((3 \times 3 \times C \times C_4)\) depth-wise separable convolution [16], and Conv is a \((1 \times 1 \times C \times C)\) convolution. In order to reduce computation, for a given tensor \( H_i \), the different functions \( f_j \) share the parameters of the initial operations (BN-ReLU-DConv) for all \( j \). Moreover, when the resolution of the feature map is reduced, we use average pooling after these three initial operations. In all our experiments, the number of initial channels in \( H_1 \) is set to \( C = 64 \). This is achieved by using a \((3 \times 3 \times 3 \times C_4)\) convolution over the input image \( H_0 \) and then apply a \((1 \times 1 \times C \times C)\) convolutional block. Finally, all the tensors \( H_i \) with the lowest spatial resolution are connected to the output \( H_L \). Concretely, \( H_T \) is a vector \( \mathbb{R}^{512} \) obtained by applying the operations (BN-ReLU-GP-FC-BN) to the input tensors, where GP refers to global average pooling, and FC corresponds to a fully-connected layer. The classifier \( \theta \) maps \( H_T \) linearly to a vector of dimension equal to the number of classes. When using Eq. (9) to train the CNMM, we connect a classifier \( \theta_i \) with the end of each block.

Implementation for semantic segmentation. We use the same setup as for image classification, but replace the ReLU activations with parametric ReLUs as in [44]. Moreover, we use max instead of average pooling to reduce the spa-
4.2. Pruning and intermediate classifiers

We evaluate the proposed pruning and intermediate classifiers strategies to reduce the inference time of trained CNMMs. For CIFAR-10/100 we learn a CNMM with \( B = 5 \) blocks and \( S = 5 \) scales. For each dataset, we train one model that uses a single classifier \( \theta \), optimized using Eq. (8). In addition, we train a second model with intermediate classifiers \( \theta_t \), minimizing the loss function in Eq. (9). In the following, we will refer to the first and second variant as CNMM-single and CNMM respectively.

In Figure 6 we report prediction accuracy vs. FLOPs for inference. Each model is represented as a curve, traced by pruning the model to various degrees. Across the three datasets, the CNMM model with intermediate classifiers achieves higher accuracy in fast inference settings than the CNMM-single model. Recall that all the operation point across the different CNMM curves are obtained from a single trained model. Therefore, this single model can realize the upper-envelope of the performance curves. As expected, the maximum performance of the intermediate classifiers increases with the step number. The accuracy of CNMM at the final step is comparable to the level obtained by the CNMM-single model: slightly worse on CIFAR-10, and slightly better at CIFAR-100 and CityScapes. This is because the minimized intermediate losses provide additional supervisory signal which is particularly useful to encourage accurate prediction for shallow, but fast, CNNs. In conclusion, the CNMM model with intermediate classifiers is to be preferred, since it provides a better trade-off between accuracy and computation at a wider range of FLOP counts.

By analysing the operating points along each curve, we can observe the effectiveness of the proposed pruning algorithm. For the CIFAR datasets we can reduce the FLOP count by a factor two without significant loss in accuracy. For CityScapes, about 25% pruning can be achieved without a significant loss. In general, if several exit points can observe the effectiveness of the proposed pruning algorithm. For the CIFAR datasets we can reduce the FLOP count by a factor two without significant loss in accuracy. For CityScapes, about 25% pruning can be achieved without a significant loss. In general, if several exit points can achieve the same FLOP count by applying varying amounts of pruning, best performance is obtained pruning less for an earlier classifier, rather than pruning more for a later exit.

4.3. Comparison with the state of the art

Image classification. We compare our model with dif-
We propose to address model pruning by using Convolutional Neural Mixture Models (CNMMs), a novel probabilistic framework that embeds a mixture of an exponential number of CNNs. In order to make training and inference tractable, we rely on massive parameter sharing across the models, and use concrete distributions to differentiate across the discrete sampling of mixture components. To achieve efficient inference in CNMM we use an early-exit mechanism that allows prediction after evaluating only a subset of the networks, and multiple models need to be trained. Therefore, unlike our approach, these methods do not allow for a fine-grained control over the accuracy-computation trade-off, and multiple models need to be trained. Figure 8 shows qualitative results using different operating points from a single CNMM.

5. Conclusions

We proposed to address model pruning by using Convolutional Neural Mixture Models (CNMMs), a novel probabilistic framework that embeds a mixture of an exponential number of CNNs. In order to make training and inference tractable, we rely on massive parameter sharing across the models, and use concrete distributions to differentiate across the discrete sampling of mixture components. To achieve efficient inference in CNMM we use an early-exit mechanism that allows prediction after evaluating only a subset of the networks. In addition, we use a pruning algorithm to remove CNNs that have low mixing probabilities. Our experiments on image classification and

Semantic segmentation. State-of-the-art methods for real-time semantic segmentation have mainly focused on the manual-design of efficient network architectures. By employing highly optimized convolutional modules, ESPNet [44] and ESPNetv2 [45] have achieved impressive accuracy-computation trade-offs. Other methods, such as [5, 59], offer higher accuracy but at several orders of magnitude higher inference cost, limiting their application in resource constrained scenarios.
semantic segmentation tasks show that CNMMs achieve excellent trade-offs between prediction accuracy and computational cost. Unlike most of previous works, a single CNMM model allows for a large number and wide range of accuracy-compute trade-offs, without any re-training.

Acknowledgements
This work is supported by ANR grants ANR-16-CE23-0006 and ANR-11-LABX-0025-01.

References


A. Supplementary material

We provide further results on CIFAR100 in order to show the importance of all components of our proposed CNMMs. Moreover, we provide additional qualitative results of semantic segmentation on the CityScapes dataset.

A.1. Ablative study of CNMM

Using sampling during training. During learning, CNMMs generate a set of samples \( \tilde{\pi}_{t-1}^s \) using Eq. (7). In contrast, during inference we use the expectations \( \pi_{t-1}^s \) instead. In order to evaluate the importance of sampling during learning, we have optimized a CNMM by using the aforementioned expectations instead of samples. Figure 9 shows the results obtained by the model using this approach, denoted as “Training with expectations”. We observe that, compared to the CNMM using sampling, the accuracy decreases faster when different pruning ratios are applied. We attribute this to the fact that our sampling procedure can be regarded as a continuous-relaxation of dropout, where a subset of functions \( f_{t-1}^s(h_{t-1}^s) \) are randomly removed when computing the output tensor \( h_{t-1}^{s+1} \). As a consequence, the learned model is more robust to the pruning process where some of the convolutional blocks are removed during inference. This is not the case when deterministic expectations are used in Eq. (7) rather than samples.

Comparison with a deterministic model. We compare the performance of our CNMM with a deterministic variant using the same architecture. Concretely, in Eq. (7) we ignore samples \( \tilde{\pi}_{t-1}^s \) and simply sum the feature maps \( \tilde{h}_{t-1}^s \) and \( \tilde{h}_{t-1}^{s+1} \). Note that the resulting model is analogous to a MSDNet [17] using early-exit classifiers. We report the results in Figure 9, denoted as “Deterministic with early-exits”. We observe that our CNMM model obtains better performance than its deterministic counterpart. Moreover, same as MSDNets, accelerating the deterministic model is only possible by using the early-exits. In contrast, the complementary pruning algorithm available in CNMM allows for a finer granularity to control the computational cost.

Expectation approximation during inference. In order to validate our approximation of \( p(y|X; \theta) = E_{p(H_T|X)}[p(y|H_T; \theta)] \) during inference, we evaluate the performance obtained by using a Monte-Carlo procedure for the same purpose. In particular, we generate \( N \) samples from the output distribution \( p(H_T|H_0) \). Then, we compute the class probabilities \( p(y|H_T; \theta) \) for each sample and average them. Table 1 shows the results obtained by varying the number of samples. We observe that our approach offers a similar performance as the Monte-Carlo approximation using \( N = 5 \). For a higher number of samples, we observe slight improvements in the results. However, note that a Monte-Carlo approximation is very inefficient since it requires \( N \) independent evaluations of the model.

A.2. Additional Qualitative Results

In Figure 10 we provide additional qualitative results for semantic segmentation obtained by a single trained CNMM model, using various operating points with different number of FLOPs during inference.

![CIFAR100](image)

**Figure 9.** FLOPs vs. performance curves for our model (CNMM), and the variants described in Section A.1. As in Figure 7 of the main paper, the curves are obtained by using the optimal combination of early-exits and pruning when possible. In this manner, results for CNMM represent upper-envelope of all the different curves depicted in Fig. 6 of the main paper.

In particular, the last row in Table 1 is 30 times more costly to obtain than the two first rows. The minimal gain obtained with more samples could probably be more efficiently obtained by using a larger model.

<table>
<thead>
<tr>
<th>Approximation</th>
<th>FLOPs</th>
<th>Top-1 Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expectation (used)</td>
<td>93M</td>
<td>74.4</td>
</tr>
<tr>
<td>Sampling N=1</td>
<td>93M</td>
<td>71.2</td>
</tr>
<tr>
<td>Sampling N=5</td>
<td>463M</td>
<td>74.4</td>
</tr>
<tr>
<td>Sampling N=15</td>
<td>1390M</td>
<td>74.5</td>
</tr>
<tr>
<td>Sampling N=30</td>
<td>2780M</td>
<td>74.6</td>
</tr>
</tbody>
</table>

**Table 1.** Comparison of the results obtained in CIFAR100 by approximating the CNMM output using our approach or a Monte-Carlo procedure with different number of samples \( N \).
Figure 10. Pixel-level predictions for a single CNMM adapting the number of FLOPs required during inference.