InFi: End-to-End Learning to Filter Input for Resource-Efficiency in Mobile-Centric Inference

Mu Yuan, Lan Zhang, Member, IEEE, Fengxiang He, Member, IEEE, Xueting Tong, Miao-Hui Song, Zhengyuan Xu, Senior Member, IEEE, and Xiang-Yang Li, Fellow, IEEE

Abstract—Mobile-centric AI applications have high requirements for the resource-efficiency of model inference. Input filtering is a promising approach to eliminate redundancy so as to reduce the cost of inference. Previous efforts have tailored effective solutions for many applications, but left two essential questions unanswered: (1) theoretical filterability of an inference workload to guide the application of input filtering techniques, thereby avoiding the trial-and-error cost for resource-constrained mobile applications; (2) robust discriminability of feature embedding to allow input filtering to be widely effective for diverse inference tasks and input content. To answer them, we first formulate the input filtering problem and theoretically compare the hypothesis complexity of inference models and input filters to understand the optimization potential. Then we propose the first end-to-end learnable input filtering framework that covers most state-of-the-art methods and surpasses them in feature embedding with robust discriminability. We design and implement InFi that supports different input modalities and mobile-centric deployments. Comprehensive evaluations confirm our theoretical results and show that InFi outperforms strong baselines in applicability, accuracy, and efficiency. InFi can achieve 8.5× throughput and save 95% bandwidth, while keeping over 90% accuracy, for a video analytics application on mobile platforms.

Index Terms—Input filtering, model inference, mobile computing, multimodal applications.

I. INTRODUCTION

The increased computing power of mobile devices and the growing demand for real-time sensor data analytics have created a trend of mobile-centric artificial intelligence (AI) [2], [3], [4], [5]. It is estimated that over 80% of enterprise IoT projects will incorporate AI by 2022. The on-device inference of computer vision models brings us increasingly rich real-time AR applications on mobile devices [6]. A judicious combination of on-device and edge computing can analyze videos taken by drones in real-time [7]. The resource efficiency of model inference is critical for AI applications, especially for resource-limited mobile devices and latency-sensitive tasks. However, many AI models with state-of-the-art accuracy [8], [9], [10] are too computationally intensive to perform high-throughput inference, even when they are offloaded to edge or cloud servers [11].

For resource-efficient inference, one direct and popular way is to eliminate the redundancy of the deep model itself via accelerating and compressing techniques [12], [13], [14], [15], [16], [17], [18]. In this work, we follow another series of approaches [1], [19], [20], [21], [22], [23], [24] that attempt to filter the redundancy in the input data. Fig. 1 shows four examples of input redundancy in mobile-centric AI applications. We call this series of approaches input filtering and classify them into two categories: SKIP and REUSE. (1) SKIP methods [19], [24] aim to filter input data that will bring useless inference results, e.g., images without faces for a face detector (Fig. 1(a)) and audios without a valid command for a speech recognizer (Fig. 1(b)). FilterForward [19] trains a binary classifier and sets a threshold on classification confidence to filter input images. (2) REUSE methods [21], [22] attempt to filter input whose results can reuse the previous inference results, e.g., motion signals of the same action (Fig. 1(c)) and video frames with the same vehicle count (Fig. 1(d)). FoggyCache [21] maintains a cache of feature embedding and inference results of previous inputs and searches reusable results in the cache for newly arrived data. Input filtering usually works as a necessary prelude to inference for under-resourced mobile systems. Moreover, compared with model optimizations, input filtering provides more flexible trade-offs between accuracy and efficiency, e.g., FilterForward can adjust the threshold in SKIP and FoggyCache can adjust the cache size in REUSE. Although prior efforts have designed effective input filters for a range of applications, two important and challenging questions remain unanswered:
To answer these questions, we first provide a generic formulation of the input filtering problem and conditions of valid filters. Then we theoretically define filterability and analyze the filterability of the two most common types of inference workloads (namely, classification and regression) by comparing the hypothesis complexity [26], [27] of the inference model and its input filter. Instead of designing bespoke solutions for narrowly-defined tasks, we propose the first end-to-end learnable framework which unifies both SKIP and REUSE approaches [19], [20], [21]. The end-to-end learnability provides feature embedding with robust discriminability in a workload-agnostic manner, thus significantly broadening the applicability. Based on the unified framework, we design an input filtering system, named InFi, which supports both SKIP and REUSE functions. In addition to image, audio, and video inputs, InFi complements existing techniques in supporting text, sensor signal, and feature map inputs. Previous methods are typically designed for a certain deployment, e.g., inference offloading [20], [21]. InFi flexibly supports common deployments in mobile systems, including on-device inference, offloading, and model partitioning [28]. In summary, our main contributions are as follows:

- We formulate the input filtering problem and provide validity conditions of a filter. We present the analysis based on complexity comparisons between hypothesis families of inference workloads and input filters, which can guide and explain the application of input filtering techniques.
- We propose the first end-to-end learnable input filtering framework that unifies SKIP and REUSE methods. Our framework covers most existing methods and surpasses them in feature embedding with robust discriminability, thus supporting more input modalities and inference tasks.
- We design and implement an input filtering system InFi. Comprehensive evaluations on workloads with 8 input modalities, 14 inference tasks, and 3 types of mobile-centric deployments show that InFi has wider applicability and outperforms strong baselines in accuracy and efficiency. For a video analytics application on a mobile platform, InFi can achieve up to 8.5× throughput and save 95% bandwidth compared with the naive vehicle counting workload, while keeping over 90% accuracy.

II. INPUT FILTERING

This section formulates the input filtering problem and provides the conditions of a “valid” input filter for resource-efficient mobile-centric inference.

A. Problem Definition

An input filtering problem needs to determine what input is redundant and should be filtered for a given inference model. First, the definition of an input filtering problem is based on its target inference model. Let $X$, $Y$ denote the input space and the label space of the target model, respectively. Define $c : X \to Y$ named the target concept [29], which provides the ground-truth label for each input. Then training a target model is to search for a function $h$ from a hypothesis family [29] $H$ using a set of training samples $S = \{(x_i, y_i)\}_{i=1}^m$, where $(x_1, \ldots, x_m)$ are sampled...
independently from $\mathcal{X}$ with an identical distribution $D$ and $y_i = c(x_i)$. Using the above notations, we define the learning problem of the target inference model $h$ by $\langle \mathcal{X}, \mathcal{Y}, c, \mathcal{H}, D, S \rangle$. Step 0 in Fig. 2 shows the original inference workflow of a trained model $h$, which takes input from $\mathcal{X}$ and returns an inference result $y \in \mathcal{Y}$.

Next, given a trained inference model $h$, its redundancy measurement function can be defined as:

**Definition 2.1 (Redundancy Measurement):** A redundancy measurement $f_h : \mathcal{Y} \rightarrow \mathbb{R}$ of a model $h$ is a function that takes only the output of $h$ as input and returns a score that indicates whether the inference computation is redundant.

Such measurements are common in practice. For example, based on the output of a face detector the inference computation that returns no detected face is redundant and can be skipped, and we can set the score $z = 0$; Otherwise, $z = 1$. Formally, $y \mapsto 1(|y| > 0)$, where $y$ is the output set of detected faces, $1(\cdot)$ is the indicator function. For REUSE cases, if the inference result of an action classifier on a new query is the same as previously cached, the computation is redundant, and we can define $f_h(y) = 1(y \notin Y_{\text{cached}})$. Note that, this definition of redundancy measurement does not depend on ground-truth labels, since our focus is not the accuracy but to optimize the resource efficiency of a deployment-ready target model with trusted accuracy by eliminating its redundant inference. Step 1 in Fig. 2 shows how redundancy measurement works.

Given the inference workload $h$ and redundancy measurement $f_h$, as Step 2 in Fig. 2, learning an input filter is defined as searching for a function $g$ from a hypothesis family $\mathcal{G}$ using a set of training samples $S' = \{(x_i, z_i)\}_{i=1}^n$, where $(x_1, \ldots, x_n)$ are sampled independently with distribution $D'$ and $z_i = f_h(h(x_i))$. This learning problem is denoted by $\langle \mathcal{X}, \mathcal{Z}, f_h \circ h, \mathcal{G}, D', S' \rangle$, i.e., $g$’s target concept is the composite function of $f_h$ and $h$.

**Inference With an Input Filter:** Once an input filter $g$ is trained, the inference workflow changes from Step 0 to Step 3 in Fig. 2. The input filter $g$ becomes the entrance of the workflow, which predicts the redundancy score of each input $x$. If not redundant, the inference model $h$ will be directly executed on the input.

### B. Validity Conditions

After defining an input filter, we now give the conditions that a “valid” input filter needs to meet for resource-efficient mobile inference. The input filter is designed to balance the resource and accuracy: filtering more inputs can save more resources, but it also brings a higher risk of incorrect inference results.

**Inference Accuracy:** With an input filter, the inference result $y$ for input $x$ is returned either by executing $h(x)$ or applying $\hat{h}(x)$. Following previous work [19], [20], [21], the correctness of the result $y$ refers to its consistency with the exact inference result by $h(x)$, rather than the ground-truth label. An input filter’s inference accuracy $\text{Acc}$ is defined as the ratio of correct results obtained by the inference workload with the filter.

**Filtering Rate:** The filtering rate, denoted by $r$, is defined as the ratio of filtered inputs (i.e., the ratio of results obtained by applying $h$), which is also an important performance metric considered in previous work [19], [20], [21].

**Overall Cost:** The overhead of an inference workload with an input filter needs to take $g$, $h$ and $\hat{h}$ into consideration. Let $C(\cdot)$ denote the cost of a certain function. For the cost of computation (e.g., runtime), the average cost per input changes from $C(h)$ into $C(g) + (1 - r) C(h) + r C(\hat{h})$. The communication cost (e.g., bandwidth) depends on the deployment of the mobile-centric inference workload. On-device inference does not involve communication, while the overall bandwidth cost of offloading [19], [20] and model partitioning [28] deployments becomes the original cost multiplied by $(1 - r) < 1$.

Based on the above metrics, we define an input filter as “valid” if it satisfies two conditions: 1) **Accurate Enough:** $\text{Acc} > T_{\text{Acc}}$, where $T_{\text{Acc}}$ is the threshold of acceptable inference accuracy. 2) **Reduced overhead:** the overall cost with an input filter is lower. If we aim to reduce the computation cost, we need $(C(g) + (1 - r) C(h) + r C(\hat{h}))/C(h) < 1$, i.e., $r > C(g)/(C(h) - C(\hat{h}))$; If we aim to reduce the communication cost, we only need $r > 0$.

### III. Filterability Analysis

As mentioned in Section I, not all inference workloads have the optimization potential by using input filtering techniques. Given an inference workload in a mobile-centric AI application, is there a valid input filter? To answer this question, based on our formulation of the input filtering problem, we first define the filterability of an inference workload. Then we analyze filterability in three typical inference cases in SKIP settings and discuss uncovered cases.

#### A. Definition of Filterability

Given the learning problem $\langle \mathcal{X}, \mathcal{Y}, c, \mathcal{H}, D, S \rangle$ of an inference model and the learning problem $\langle \mathcal{X}, \mathcal{Z}, f_h \circ h, \mathcal{G}, D', S' \rangle$ of its input filter, to simplify the analysis, we make assumptions as follows: (1) $D = D'$, i.e., the training samples follow the identical distribution; (2) $S' = \{(x_i, z_i)\}_{i \in S}$, i.e., the two learning problems share the same inputs in their training samples. But they are supervised under different labels. The inference model $h$ is supervised by $y_i = c(x_i)$, while the input filter $g$ is supervised by $z_i = (f_h \circ h)(x_i)$. Our intuitive idea for filterability is that, if an inference workload is filterable, the learning problem of
its input filter should have lower complexity than the learning problem of its inference model. Formally, we define filterability as follows:

**Definition 3.1 (Filterability):** Let \( \text{Complex}(\cdot) \) denote the complexity measurement of a hypothesis family. We say that the inference workload is filterable, if \( \text{Complex}(\mathcal{G}) \leq \text{Complex}(\mathcal{H}) \), where \( h \in \mathcal{H} \) and \( (f_h \circ h) \in \mathcal{G} \).

Since the hypothesis family cannot be determined based only on input and output spaces, we use the family of the input filter’s target concept \( f_h \circ h \) as \( \mathcal{G} \).

Now we can characterize the theoretically achievable accuracy and overhead of the input filter for a given inference model by leveraging computational learning theory [26]. It has been proven that the more complex the hypothesis family is, the worse the bounds of generalization error. On the other hand, the hypothesis complexity of neural networks has a positive correlation with the number of parameters. For example, let \( W, L \) denote the number of weights and the number of layers in a deep neural network. The VC-dimension [30] (a measurement of the hypothesis complexity) is \( O(WL \log(W)) \) [31]. In the case of the same layer structure, the more parameters the higher the inference overhead of neural networks. The generalization error bound and the number of parameters correspond to the accuracy and efficiency metrics in validity conditions (§ II-B), respectively, although they are not strict quantification. Therefore, if an inference workload is filterable, whose input filter has lower hypothesis complexity, we are confident to obtain a valid filter with sufficiently high accuracy and lower overhead than the inference model. Next, we will analyze the complexities of the hypothesis family of inference workload \( h \) and its input filter \( g \) in different cases.

### B. Low-Confidence Classification as Redundancy

Considering an inference workload, where the inference model is a binary classifier \( h \) that returns the classification confidence, and the redundancy measurement regards the classification result with a confidence lower than a threshold \( t \) as redundant, i.e., \( f_h(y) = \text{sign}(y > t) \). Confidence-based classification is very common in mobile AI applications, such as speaker verification. We adopt the empirical Rademacher complexity [27], denoted by \( \hat{\mathcal{R}}_S(\cdot) \), as the complexity measurement, which derives the following generalization bounds [26]:

**Theorem 1 (Rademacher complexity bounds):** Let \( \mathcal{H} \) be a family of hypothesis taking values in \( \{-1, 1\} \). Then for any \( \delta > 0 \), with probability at least \( 1 - \delta \), the following holds for all \( h \in \mathcal{H} \):

\[
R(h) \leq \hat{R}(h) + \hat{\mathcal{R}}_S(\mathcal{H}) + 3\sqrt{\frac{\log(2/\delta)}{2m}} \tag{1}
\]

where \( R(h) \) and \( \hat{R}(h) \) denote the empirical and generalization errors, and \( m \) is the number of training samples.

This theorem shows that the higher a hypothesis family’s empirical Rademacher complexity, the worse the bounds of its generalization error. The classification confidence-based redundancy measurement creates two hyperplanes parallel to \( h = 0 \): points between them are considered redundant, and points outside are considered not redundant. Thus, the hypothesis family of the input filter’s target concept has the form: \( \mathcal{G} = \{\text{sign}(h(x)(h(x) + b))\} \), where \( h \in \mathcal{H} \) and \( b \in \mathbb{R} \). Then we have proven the following lemma, which shows that the discussed inference workload is not filterable.

**Lemma 2:** Let \( \mathcal{H} \) be a family of binary classifiers taking values in \( \{-1, +1\} \). For \( \mathcal{G} = \{\text{sign}(h(h + b))\} \) where \( h \in \mathcal{H}, b \in \mathbb{R} \):

\[
\hat{\mathcal{R}}_S(\mathcal{G}) \geq \hat{\mathcal{R}}_S(\mathcal{H}). \tag{2}
\]

**Proof:** By definition,

\[
\hat{\mathcal{R}}_S(\mathcal{H}) = E_{\sigma} \left[ \sup_{h \in \mathcal{H}} \left( \frac{1}{m} \sum_{i=1}^{m} \sigma_i h(x_i) \right) \right]
\]

and

\[
\hat{\mathcal{R}}_S(\mathcal{G}) = E_{\sigma} \left[ \sup_{h \in \mathcal{H}, b \in \mathbb{R}} \left( \frac{1}{m} \sum_{i=1}^{m} \sigma_i \text{sign}(h(x_i)(h(x_i) + b)) \right) \right],
\]

where Rademacher variables \( \sigma_i \in \{-1, +1\} \). Fixing \( b = 2 \),

\[
\hat{\mathcal{R}}_S(\mathcal{G}) \geq E_{\sigma} \left[ \sup_{h \in \mathcal{H}, x \in S} \left( \frac{1}{m} \sum_{i=1}^{m} \sigma_i \text{sign}(h(x_i)(h(x_i) + 2)) \right) \right]
\]

\[
= E_{\sigma} \left[ \sup_{h \in \mathcal{H}, x \in S} \left( \frac{1}{m} \sum_{i=1}^{m} \sigma_i \text{sign}(h(x_i)) \right) \right] = \hat{\mathcal{R}}_S(\mathcal{H}),
\]

where we used the fact that \( \text{sign}(h(x_i) + 2) = 1 \). \( \blacksquare \)

Multi-class classifiers can be treated as a set of confidence scoring functions, one for each class. The above lemma can also be applied to derive that multi-class classifiers using such a confidence-based redundancy measurement are not filterable either.

### C. Class Subset as Redundancy

Considering the inference model \( h \) as a multi-class mono-label classifier and \( \mathcal{Y} = \{y_1, \ldots, y_l\} \). Then its hypothesis family \( \mathcal{H} \) has the form: \( \mathcal{H} = \{\max(h_1, \ldots, h_l) : h_i \in \mathcal{H}_i, i \in [1, l]\} \), where \( h_i \) returns the probability of the \( i \)-th class. The redundancy measurement checks whether the predicted class belongs to a specific subset, i.e., \( f_h(y) = 1(y \in \mathcal{Y}') \), where \( \mathcal{Y}' \subseteq \mathcal{Y} \). It is common in mobile applications to select only a subset of labels for use. For example, when deploying a pre-trained common object detector [32] on a drone for traffic monitoring, we only care about the labels of vehicles and pedestrians, while considering other labels like animals and trees as redundancy. With the class subset-based redundancy measurement, the hypothesis family of the input filter’s target concept has the form: \( \mathcal{G} = \{\max(h_i) : y_i \in \mathcal{Y}'\} \). We have proven the following lemma, which shows that the discussed inference workload is filterable:

**Lemma 3:** Let \( \mathcal{H}_1, \ldots, \mathcal{H}_l \) be \( l \) hypothesis sets in \( \mathbb{R}^X \), \( l \geq 1 \), and let \( \mathcal{H} = \{\max(h_1, \ldots, h_l) : h_i \in \mathcal{H}_i, i \in [1, \ldots, l]\} \). For \( \mathcal{G} = \{\max(h_i) : i \in J\} \), where \( J \subseteq [1, \ldots, l]\):

\[
\hat{\mathcal{R}}_S(\mathcal{G}) \leq \hat{\mathcal{R}}_S(\mathcal{H}). \tag{3}
\]
Proof: For any \( j = 1, \ldots, l \):

\[
\hat{R}_S(\mathcal{H}) = \frac{1}{m} \mathbb{E}_\sigma\left[ \sup_{x_i \in \mathcal{S}} \sigma_i \max_{h_k \in \hat{R}_k} (h_k(x_i)) \right] \\
\geq \frac{1}{m} \mathbb{E}_\sigma\left[ \sup_{x_i \in \mathcal{S}} \sigma_i \max_{j \in J} (h_j(x_i)) \right] = \hat{R}_S(\mathcal{G}).
\]

The equation holds only if the max-value scoring function is in the selected subset for all \( x_i \in \mathcal{S} \), which means that without loss of inference accuracy, the optimal filterable ratio in the data is 0. Except in this extreme case, we can think that the complexity of learning the input filter is strictly lower.

D. Regression Bound as Redundancy

Considering a bounded regression model \( h \), whose outputs are bounded by \( M \in \mathbb{R} \) that \( |h(x) - c(x)| \leq M \) (recall that \( c \) is the target concept) for all \( x \in X \). The redundancy measurement checks whether the returned value is larger than a threshold, i.e., \( f_h(y) = 1(y > T) \). As an example, face authentication on mobile devices usually requires the coordinates of the detected face to be within the specified range. Then learning the target concept of the input filter becomes learning a regression model whose outputs are bounded by \( T \), where \( T < M \). We also adopt the empirical Rademacher complexity and have the following theorem [26]:

**Theorem 4:** Let \( p \geq 1 \) and \( \mathcal{H} = \{ x \mapsto |h(x) - c(x)|^p : h \in \mathcal{H} \} \). Assume that \( |h(x) - c(x)| \leq M \) for all \( x \in X \) and \( h \in \mathcal{H} \). Then the following inequality holds: \( \hat{R}_S(\mathcal{H}) \leq pM^{p-1}\hat{R}_S(\mathcal{H}) \).

Since \( M > T \), this theorem shows that the upper bound of \( \hat{R}_S(\mathcal{G}) \) is tighter than the upper bound of \( \hat{R}_S(\mathcal{H}) \). So we can be confident that the bounded regression inference workload discussed is filterable.

E. Discussions

**Other Inference Tasks:** Classification and regression are the most common inference tasks, and the three redundancy measurements discussed are widely adopted [19], [33], [34]. However, there are some inference tasks that are challenging to measure the hypothesis complexity, like reinforcement learning [35] and structured learning [36]. Besides, their redundancy measurements are typically ill-defined. We believe that our problem formalization and analysis approach are general, based on which we will analyze the filterability of other tasks in future work.

**Characteristics of REUSE:** For the REUSE approach, we cannot determine the hypothesis family of the input filter’s target concept. Here we only give one necessary condition: the inference result is discrete or can be discretized. For example, classification and counting models return discrete outputs. But continuous localization coordinates of detection models cannot be reused directly unless reusing detection results with high IoU are regarded as correct, which is equivalent to discretizing the outputs.

IV. FRAMEWORK

In this section, we first propose a novel input filtering framework that unifies SKIP and REUSE approaches. Then we discuss how existing approaches are covered by our framework and their limitations. Finally, we present the key design, end-to-end learnability, and advantages it brings.

A. Skip as Reuse

We unify SKIP and REUSE approaches based on the idea that:

**SKIP equals to REUSE the NONE output of \( h(\vec{0}) \).**

Suppose we have an all-zero input \( \vec{0} \) and apparently its inference result can be interpreted as NONE. Then given a new input \( x \), if it is similar to \( \vec{0} \) in the feature space, we can REUSE the cached NONE result, i.e., we SKIP the inference computation. The key to reuse is to measure the semantic similarity between the current input and previously cached ones. However, it is difficult to accurately measure semantic similarity directly based on the raw input. As Step 1 in Fig. 3 illustrated, our framework first computes the feature embedding of each raw input. Taking a pair of inputs \( x, x’ \), then our framework applies a difference function \( d \) on their corresponding embeddings \( e, e’ \) and feeds the result into a classifier that predicts a single scalar \( z \). Under this framework, for SKIP, we fix \( x’ \) as an all-zero input \( \vec{0} \), then the process degenerates to a binary classification task that takes \( x \) as input and returns the prediction \( z \). In this way, our framework unifies SKIP and REUSE approaches, with only a difference in interpretation of the value \( z \). For REUSE, we interpret \( z \) as the distance between two inputs. For SKIP, we interpret \( z \) as the probability that input \( x \) is not redundant.

B. Inference With an Input Filter

For the inference phase, as shown in Step 2 in Fig. 3, SKIP and REUSE filters only differ in the inputs of the difference function \( d \). (1) **SKIP:** Inference with a SKIP filter is the same as serving a binary classifier. We can set a threshold on the predicted redundancy score \( z \) to determine whether to skip. (2) **REUSE:** Inference with a REUSE filter needs to maintain a key-value table, where a key is a feature embedding and its

---

**Fig. 3.** Unified and end-to-end learnable framework for both SKIP and REUSE input filtering.
value is the corresponding inference result. For an arrived input \( x \), the trained feature embedding network returns its embedding \( e \) and the distances \( z \) between \( e \) and cached keys are computed by the difference function \( d \) and the trained classifier. Then we can leverage classification algorithms, e.g., KNN, to obtain the reusable cached results.

C. Sub-Instance Approaches

Here we explain how our framework covers three state-of-the-art input filtering methods [19], [20], [21] that will be used for comparison in our evaluations.

Sub-instance 1: FilterForward (FF) [19] is a SKIP method for image input. FF uses a pre-trained MobileNet’s intermediate output as the feature embedding. Then it trains a “micro-classifier” that consists of convolution blocks to make the binary decision for filtering.

Sub-instance 2: FoggyCache (FC) [21] is a REUSE method for image and audio input. FC uses low-level features (SIFT for image, MFCC for audio) and applies locality-sensitive hashing (LSH) for embedding. Then FC uses L2 norm as the difference function and applies KNN to get the reusable inference results from previously cached ones.

Sub-instance 3: Reducto [20] is a variant of the SKIP method for video input. It measures low-level features (pixel, edge, corner, area) differences between successive frames. If they are similar enough, Reducto skips the current frame and returns the latest result. Formally, let \( x \) be the current frame and \( x' \) be the previous frame. Reducto defines \( d(e, e') = (e - e')/e' \), where \( e, e' \) are low-level features of \( x, x' \). It uses a threshold function as the classifier, i.e., \( 1(d(e, e') > T) \).

D. End-to-End Learnability

To obtain features with robust discriminability for diverse data modalities and inference tasks in mobile applications, a key design principle of our framework is end-to-end learnability. End-to-end learning system casts complex processing components into coherent connections in deep neural networks [37] and optimizes itself by applying gradient-based back-propagation algorithms all through the networks. Deep end-to-end models have shown state-of-the-art performance on various tasks including autonomous driving [38] and speech recognition [39]. As aforementioned, a main component of our unified framework is to measure the semantic similarity between two inputs. To make our framework end-to-end learnable, we leverage the metric learning paradigm, whose goal is to learn a task-specific distance function on two objects. The metric learning paradigm turns the fixed difference function \( d \) (e.g., Euclidean distance and L2 norm) used by existing methods into an end-to-end learnable network. Within the metric learning paradigm, we adopt Siamese network structure [40] for feature embedding to support two inputs and flexible input modalities. The Siamese network uses the same weights while working on two different inputs to compute comparable output vectors, and has been successfully applied in face verification [41], pedestrian tracking [42], etc. We can flexibly implement the Siamese feature embedding by incorporating different neural network blocks to learn modality-specific features in an end-to-end manner, instead of tailoring handcrafted or pre-trained feature modules. Our experimental results show that the end-to-end learned features have robust discriminability to diverse inference workloads in mobile-centric AI applications.

V. DESIGN OF INF

Based on our input filtering framework, in this section, we present the concrete design of INF (InFilter), which supports both SKIP and REUSE functions, named INF-Skip and INF-Reuse. The design of INF has four key components: feature embedding, classifier, training mechanism, and inference algorithm. We also discuss diverse deployments of INF in AI applications on mobile, edge, and cloud devices.

A. Feature Networks for Diverse Input Modalities in Mobile-Centric AI

INF supports filtering inference workloads with six typical input modalities in mobile applications: text, image, video, audio, sensor signal, and feature map. We develop a collection of modality-specific feature networks as building blocks for learning feature embedding. Our major consideration in designing these feature networks is resource efficiency on mobile devices.

Text Modality \( g_{text} \): Text is tokenized into a sequence of integers, where each integer refers to the index of a token. We adopt the word-embedding layer to map the sequence to a 3528-dimensional vector. The word-embedding layer uses the same weights as the feature embedding. Then it trains a “micro-classifier” that consists of convolution blocks to make the binary decision for filtering.

Image Modality \( g_{image} \): We use depth-wise separable convolution [43], denoted by SepConv, to learn visual features. SepConv is a parameter-efficient and computation-efficient variant of the traditional convolution which performs a depth-wise spatial convolution on each feature channel separately and a point-wise convolution mixing all output channels. Then we build residual convolution blocks [44] ConvRes as follows:

\[
\text{ConvStep}(x) = \text{LN}(\text{SepConv}(\text{ReLU}(x))),
\]

\[
c_1(x) = \text{ConvStep}(x), c_2(x) = \text{ConvStep}(c_1(x)),
\]

\[
\text{ConvRes}(x) = \text{MaxPool2D}(c_2(x)) + \text{ConvStep}(x),
\]

where \( \text{ReLU} \) denotes the rectified linear unit, \( \text{LN} \) denotes the layer normalization and \( \text{MaxPool2D} \) denotes the 2D max-pooling layer. Finally, we build the image feature network with two residual blocks followed by a global max-pooling layer and a Sigmoid-activated dense layer.

Video Modality \( g_{video} \): For video modality, we need to represent not only the spatial but also the temporal features. Given a window of frames, we stack one residual block for each frame and then concatenate their resulting feature maps. Except for the first residual block, the video feature network performs the same operation as the image feature network.
Audio Modality ($g_{\text{audio}}$): We consider audio inputs in the form of either a 1D waveform or a 2D spectrogram and use the same structure as image feature networks to learn features from audio.

Sensor Signal and Feature Map Modality ($g_{\text{vec}}$): Motion sensors are widely used in mobile devices and play a key role in many smart applications, e.g., gyroscope for augmented reality [45] and accelerator for activity analysis [46]. Feature maps refer to the intermediate outputs of deep models and need to be transmitted in workloads that involve model partitioning [28]. We consider these two types of input as a vector with a fixed shape and use two densely connected layers to learn the feature embedding from the flattened vector.

Flexible Support for Input Modalities: Our design provides flexible support for diverse input modalities in mobile-centric AI applications. We can easily integrate a modality-specific neural network from advanced machine learning research as the feature network block into our framework, so as to learn feature embeddings in an end-to-end way.

B. Task-Agnostic Classifier

Each feature network $g_{\text{modality}}$, where modality belongs to {text, image, video, audio, vec}, takes $x$ as input and outputs the embedding $\text{emb}$. We add a dropout layer after the last dense layer of feature networks to reduce overfitting. Following the previous design of Siamese network [40], we use the absolute difference as the function $d$. Let $\text{emb}_1, \text{emb}_2$ denote the embedding outputs of two inputs $x_1, x_2$. The classifier is defined as $g_{\text{cls}} = \sigma(\sum w_j (\text{emb}_1(j) - \text{emb}_2(j)) + b)$, where $\text{emb}_j(j)$ denotes the $j$-th element in the embedding vector and $\sigma$ is the Sigmoid function. To sum up, the input filter function $g: \mathcal{X} \rightarrow \mathcal{Z}$ can be defined as $g(x) = (g_{\text{cls}} \circ g_{\text{modality}})(x)$. With proper implementation, the modality of input data can be automatically detected without manually setting.

C. Multi-Task Extension

The above design is described for single-task workloads, however, it is common to concurrently run multiple AI models in real applications. We will show that the design of InFi can be flexibly extended to multi-modal and multi-task inference workloads.

Multi-Modality Single-Task: Multi-modal learning aims to learn AI models given multiple inputs with different modalities, which is receiving increasing attention in areas such as autonomous driving [38]. Our designs of modality-specific feature networks and task-agnostic classifier naturally support multi-modal extension: For each modality $\text{mod} \in \{\text{text, image, video, audio, vec}\}$, we build the corresponding feature network $g_{\text{mod}}$ to learn its embedding. Then we concatenate the resulting embeddings and feed it to the classifier $g_{\text{cls}}$.

Single-Modality Multi-Task: It is common to deploy multiple AI models to analyze the same input, e.g., detecting vehicles and classifying traffic conditions on the same video stream. For input filtering, we simply extend the length of the last dense layer in the classifier $g_{\text{cls}}$, one dimension per task. Existing work on multi-task learning [47] demonstrates that the cross-task representation improves learning performance. Formally, given $t$ tasks, it has been proven that the sample complexity needed [48] is as follows:

$$\text{Complex}(g_{\text{mod}}) + t \cdot \text{Complex}(g_{\text{cls}}).$$

That is, we can save $(t - 1) \cdot \text{Complex}(g_{\text{mod}})$ sample complexity, compared with learning a filter for each task separately. And our experimental results (Fig. 14) also show that the cross-task representation is beneficial for input filtering.

Multi-Modality Multi-Task: Considering a general case where we need to filter inputs for multi-modality and multi-task workloads, we can combine the above two extensions, as shown in Fig. 4: For each input modality, we build the corresponding feature network and concatenate the resulting embeddings; And for each task, we build a multi-dimension classifier, one dimension per task, which takes the concatenated embedding as input. Compared with the naive way that deploys independent InFi for different inference workloads, our proposed extension saves computation and leverages potential advantages of cross-task and cross-modality representation.

D. End-to-End Training

InFi-Skip and InFi-Reuse share the same model architecture, but have different formats of training data. 1) Learning an InFi-Skip filter uses the same paradigm as training a binary classifier. Thus its training samples are $(x_i, f_h(h(x_i)))_{i=1}^n$ and we use the binary cross-entropy loss function. In practice, we can use the original training set of $h$ or data collected during serving $h$. Since $f_h$ only depends on the inference result, the supervision labels can be collected automatically. 2) InFi-Reuse filters are
trained using the contrastive loss [49] with a margin parameter of one. Given a set of input and their discrete inference results, the redundancy measurement is defined as the distance metric between a pair of inputs. Formally, a training sample consists of a pair of inputs and their distance label \((x_i, x_j, 1(y_i \neq y_j))\). We can optimize all trainable parameters end-to-end, using standard back-propagation algorithms.

**Online Active Update:** Unlike benchmark datasets, the distribution of real-world inputs, e.g., the video streams captured by surveillance cameras, is much narrower and changes online [50]. In a video-based vehicle counting application (see Section VI-A for detailed setup), we explored the shifted distribution of frames over wall-clock time. As shown in Fig. 5(a), the vehicle count varies with time. There are two distinct count peaks in the morning and evening and the night-time results remain stable at low values. We noticed that the captured frames switched between infrared (IR) and RGB images when the lighting condition changed. The RGB-IR technology provides day and night vision capability for cameras and is supported by popular commercial sensors. We split all frames into infrared and RGB subsets and plot their distribution over the number of detected vehicles in Fig. 5(b). We can see a clear difference in the distribution. Therefore, a vanilla offline training policy that selects initial samples (e.g., frames in the first hour) for training the input filter results in sub-optimal performance quickly. To overcome the poor adaptability of offline training, we adopt the least confidence [51] strategy to actively select samples for updates on the fly. Existing work [52] proved that the sample complexity of active learning is asymptotically smaller than passive learning. Specifically, we set a period length and a sampling ratio \(\beta\%\). Then we execute the input filter on all inputs within a period and select \(\beta\%\) samples with the least confidence \((g(x) - 0.5)\). For InFi-Reuse, we treat \(1 - \theta\) as the confidence score. Our experimental results (Fig. 15) show that, given the same budget for the number of training samples, this active strategy significantly outperforms the offline one.

**F. Mobile-Centric Deployments**

Unlike existing work tailored for specific deployment, e.g., inference offloading [19], [20], [21], InFi supports diverse mobile-centric deployments: (1) *On-device:* both inference model and input filter are deployed on one device; (2) *Offloading:* the input filter is deployed on one device, and the inference model is deployed on another device. (3) *Model Partitioning (MP)* [28]: the inference model is partitioned across two devices, and the input filter is deployed with the first part. MP is a promising approach to collaboratively make use of the computing resources of mobile and edge devices [53], [54] and better protect the privacy of mobile data [55]. For MP deployment, the filter’s input is the feature map, so existing filtering approaches [19], [20], [21] cannot be applied. Due to the support of feature map modality, InFi is the first input filter that can be applied in model partitioning workloads. Note that InFi is not limited to systems with a single mobile and edge node. For example, training one filter per server, or changing one filter’s binary classifier into a multi-category one (one bit per server), InFi-Skip can be used in the multi-tenancy context [19].
that first computes low-level embedding for activity recognition and deploy a ResNet-based model for vehicle counting. We build all feature models using K-nearest neighbors vote (K=10) for both SKIP and REUSE cases. We also deployed YOLOv3-tiny [12] model for vehicle counting and a lightweight pose estimation model [13] to compare input filtering and model compression techniques.

VI. EVALUATION
A. Implementation and Configurations
We implemented InFi 1 in Python. We build all feature networks and classifiers with TensorFlow 2.4. The learning rate is set as 0.001, the batch size is 32, and the number of training epochs is 20. In the text feature network, the output dimension of the embedding layer is 32. For image, video, and audio feature networks, we use 32 and 64 convolution kernels in the two residual blocks. We use 128 units in the first dense layer in vector feature networks. The last dense layer of all feature networks has 200 units and 0.5 dropout probability.

Datasets and Inference Models: To evaluate InFi’s wide applicability, we choose 14 inference workloads that cover 8 input modalities and three deployments (see Table I). Seven datasets are used: (1) We reprocessed a standard video dataset, Hollywood2 [56], to create four different input modalities: video clip, image, audio, and text. An action classification model [57] is deployed on the original video clips. Images are sampled from the video clips and a face detection [58], a pose estimation [59], and a gender classification [58] models are deployed. Audio is extracted from each video clip and we deploy a speech recognition model [39]. Text is the caption generated on sampled images from each video clip and we deploy an image captioning model [60]. A named entity recognition model (spaCy [61]) and a sentiment classification model [62] are deployed. (2) We use the ESC-10 dataset [63] for audio anomaly detection and deploy an transformer-based model [64]. (3) We use the UCI HAR dataset [46] for motion signal-based human activity recognition and deploy an LSTM-based model. (4) We use the MoCap dataset [65] for training a motion signal-based user identification (12 users) model, using an LSTM-based architecture, and deploying it as the inference workload. (5) We use the DeepSig dataset [66] and deploy a ResNet-based model for modulation recognition of radio signals. (6) We use the WiFiHAR dataset [67] for activity recognition and deploy an LSTM-based model. (7) We collected a video dataset, named City Traffic, from a real city-scale video analytics platform. We collected 48 hours of videos (1FPS) from 10 cameras at road intersections and deploy YOLOv3 re-implemented with TensorFlow 2.0 to count the number of vehicles in video frames. All deployed inference models load publicly released pre-trained weights. And we split each dataset for training and testing by 1:1 (Hollywood2 and UCI HAR are split randomly, while City Traffic is split by time on each camera).

Devices and Deployments: We use an edge server with one NVIDIA 2080Ti GPU and three mobile platforms: (1) NVIDIA JETSON TX2, (2) XIAOMI Mi 5, and (3) HUAWEI WATCH. All device-independent metrics are tested on the edge. For vehicle counting, we test three deployments: on-device, offloading, and model partitioning (see Section V-F).

Baselines: We adopt three strong baselines: FilterForward (FF) [19], Reducto [20], and FoggyCache (FC) [21]. See Section IV-C for details of baselines. For workloads with no existing method presented (to our best knowledge), we tested a method dubbed Low-level that first computes low-level embedding for inputs (MFCC for audio, Bag-of-Words for text, raw data for motion signal and feature map). Then Low-level uses K-nearest neighbors vote (K=10) for both SKIP and REUSE cases.

B. Inference Accuracy Vs. Filtering Rate
First, we test two device-independent metrics (inference accuracy and filtering rate) on the ten inference workloads. We adjust the confidence threshold in FF, Reducto, and InFi-Skip, and the ratio of cached inputs in FC and InFi-Reuse, from 0 to 1 with 0.01 interval.

Redundancy Measurements: (1) SKIP: For FD (PE), outputs with no detected face (person keypoints) are redundant. For GC (SC), outputs with classification confidence less than a threshold, CONF (0.9), are redundant. For AC, outputs that are not in a subset of classes, Sub, are redundant. For SR, outputs that do not belong to the first 6 categories are redundant. For HAR, outputs that are not “LAYING” are redundant. For UI, outputs that do not belong to the first 6 users are redundant. For AD, outputs that are not in {“Cry, Sneeze, Firing”} (anomaly events) are redundant. For MR, we randomly select half of the radio modulation types as redundant. For WAR, outputs with “NO PERSON” are redundant. For VC and VC-MP, outputs with zero count are redundant. (2) REUSE: Experimental results show that cache miss happens rarely, so the homogeneity threshold is set as 0.5. We regard inputs that hit the cache as redundant. For the VC (-MP), since we have 86 K images from each camera, a fixed cache ratio can lead to serious inefficiency in the KNN algorithm. We fix the cache size as 1000 and reinitialize the cache every 5000 frames. For other inference workloads, we set a fixed cache size according to the cache ratio.

Overview of Results: Tables II and III summarize the results of the SKIP and REUSE methods. Following related work [20].

\[\text{https://github.com/yuanmu97/infi}\]
TABLE II
FILTERING RATE (%) @ 90% INFERENCE ACCURACY OF SKIP METHODS

| Method  | FD  | PE  | GC  | AC  | VC  | AD  | WAR |
|---------|-----|-----|-----|-----|-----|-----|-----|
| FF      | 0   | 14.5| 0.0 | 0.0 | 48.0| /   | /   |
| Reducto | /   | /   | /   | /   | 48.6| /   | /   |
| InFi-Skip | 36.1 | 18.9| 33.1| 56.0| 66.5| 75.4| 11.6|
| Optimal | 64.8| 34.4| 71.8| 91.2| 77.7| 86.8| 31.1|

| Method  | SR  | NER | HAR | UI  | SC  | VC-MP | MR   |
|---------|-----|-----|-----|-----|-----|-------|------|
| InFi-Skip | 44.1| 26.8| 91.2| 72.4| 22.5| 70.7  | 40.9 |
| Optimal  | 59.9| 34.4| 91.8| 79.8| 63.8| 77.7  | 59.9 |

TABLE III
FILTERING RATE @ 90% INFERENCE ACCURACY OF REUSE METHODS

| Method  | GC  | AC  | HAR | SC  | VC-MP | VC   |
|---------|-----|-----|-----|-----|-------|------|
| FC      | 66.1%| 13.2%| /   | /   | 59.4%  |      |
| InFi-Reuse | 98.8%| 32.1%| 98.3%| 43.4%| 95.0%  | 91.1% |

Fig. 6. Comparison between FF and InFi-Skip filters on visual detection workloads.

we report the filtering rates at 90% inference accuracy. The optimal results are computed by $(1-0.9)+r_N$ where $r_N$ denotes the ratio of redundant inputs in the test dataset. Results show that InFi-Skip outperforms FF and Reducto on all 10 workloads with significantly higher filtering rates and wider applicability. Similarly, InFi-Reuse significantly outperforms FC on all 6 applicable workloads. InFi-Skip can filters 18.9%-91.2% inputs and InFi-Reuse can filters 32.1%-98.8% inputs, while keeping more than 90% inference accuracy. For all workloads, Low-level method cannot achieve 90% inference accuracy unless no input is filtered (i.e. 0.0% filtering rate), and we omit these results in the tables.

Feature Discriminability: By comparing FF and InFi on FD, PE, GC, and AC workloads, we evaluate the discriminability of our end-to-end learned features. As shown in Fig. 6, FF works on the pose estimation workload, but not on the face detection workload. The “Worst” case is calculated by $r = 1 - \text{Acc}$. The reason may be that there is a “person” label in the ImageNet dataset, so the pre-trained feature embedding in FF is discriminative for determining whether there is a human pose. However, on other tasks (e.g., FD, GC, and AC), the pre-trained features are not discriminative and FF can only provide two extreme filtering policies: either filtering all input or filtering nothing, which is useless in practice. On the contrary, InFi-Skip learns feature embedding with robust discriminability and performs well on all four workloads. With over 90% inference accuracy, InFi-Skip can filter 18.9% and 36.1% inputs for PE and FD workloads, respectively.

Transferability: One interesting question is, how transferable is the trained filter to workloads with a looser or tighter redundancy measurement? We set the minimal number of recognized words, N, as 0 and 2 and train two InFi-Skip filters. Then we test the two filters on two test sets with different N. As shown in Fig. 7, the performance of InFi-Skip (N=2) is close to InFi-Skip (N=0) when tested with N=0, however, the performance of InFi-Skip (N=0) is apparently worse when tested with N=2. An intuitive explanation is that the learned feature with a looser redundancy measurement covers the one with a tighter redundancy measurement, while the opposite is not true.

Sensitivity to Class Subset Size: For the class-subset redundancy measurement, we set different subset sizes to test the sensitivity. As shown in Fig. 8, for action classification workload, setting a smaller class subset brings more redundant samples. And InFi-Skip robustly provides smooth accuracy-efficiency trade-off curves in both cases, which significantly outperforms FF (only two extreme points are provided).

Sensitivity to Training Size: We further divide training splits into sets with different sizes. As shown in Fig. 9, using only 10% samples from the training set, InFi can still achieve near-optimal performance on HAR workload. Let R denote the ratio of training samples used for training. When achieving over 95% inference accuracy, InFi-Skip (R=1) filters 86.4% inputs while InFi-Skip (R=0.1) still filters 81.1%. For high-accuracy reuse, the impact of training size is relatively greater. When filtering 90% inputs, InFi-Reuse (R=1) can achieve 95.9% inference accuracy.
Fig. 9. InFi filters on HAR inference workloads. R denotes the ratio of training samples used. The “Random” case labels each input randomly.

Fig. 10. InFi-Skip on UI workload. NDense is the number of dense units. EmbLen is the length of embedding.

Fig. 11. Comparison of FC and InFi-Reuse on visual classification workloads. K is the parameter in KNN.

accuracy, while the accuracy of InFi-Reuse (R=0.1) decreases to 88.1%.

Sensitivity to Model Complexity: To explore the relationship between the complexity and performance of input filters, we trained InFi-Skip filters for the UI workload using the different lengths of embedding (1, 16, 32, 64, 128, 256) and the number of dense units (1, 100, 200, 400) in the classifier. And we measure the performance by the maximum filtering rate when achieving 90% inference accuracy. As shown in Fig. 10(b), except for extreme cases (e.g., single dense or embedding unit), the filtering performance is relatively robust.

Sensitivity to K in KNN: The parameter K in KNN affects the classification accuracy. We vary K from 1 to 20 and test the REUSE filters’ performance. As shown in Fig. 11, on GC workload, InFi-Reuse is robust to varied K parameters, while FC suffers serious performance degradation. For example, with 90% inference accuracy, FC (K=5) can filter 68.4% inputs, while FC (K=1) can only filter 27.3% which is slightly higher than the random guess (20%). On the contrary, InFi-Reuse (K=1,5) can all achieve a 94.3% filtering rate with more than 95% inference accuracy. For the AC workload, the results show that the handcrafted feature SIFT is not discriminative, and all tested K parameters lead to similar performance with random labeling. InFi-Reuse can learn an action-related discriminative feature, it can filter 18.6% inputs and keep more than 90% inference accuracy (K=10).

Comparisons on VC(-MP) Workloads: Unlike other datasets, the video frames arrive in time order rather than randomly. For VC-MP, we partition the YOLOv3 model to mobile-side (the first 39 layers) and edge-side (the rest layers). As shown in Fig. 12, InFi outperforms FF, Reducto, and FC, and also is the only applicable method for the VC-MP workload. With over 90% inference accuracy, InFi-Skip achieves 66.5% filtering rate, while FF and Reducto achieve 48.0% and 48.6%, respectively; InFi-Reuse filters 31.7% more inputs than FC when K=10. The results show the superiority of end-to-end learned features over handcrafted and pre-trained ones.

Mobile-Featured Modalities: Radio signals and WiFi CSI (channel side information) are featured modalities for mobile AI applications. Fig. 13 shows the experimental results of applying InFi on two mobile-featured tasks: radio modulation recognition and WiFi CSI-based action recognition. Note that, existing methods cannot be applied to both tasks. With 90% target accuracy, InFi-Skip saves 40.9% and 11.6% computations for MR and WAR workloads, respectively.

Multi-Task Extension: In Section V-C, we present how to extend InFi to multi-task workloads. We use the Hollywood2 dataset and corresponding inference tasks to evaluate the multi-task extension of InFi-Skip. First, we select three inference tasks on image modality: FD, GC, and PE. We build InFi-Skip filters
using a single task, two tasks, and three tasks and evaluate their performance on each task. As shown in Fig. 14(a), the multi-task filter outperforms single-task ones on all three tasks, improving the filtering rate up to 7.6% (for the GC task) when achieving 90% inference accuracy. Next, we select three inference tasks on different modalities: PE on images, NER on texts, and SR on audio. And we build InFi-Skip filters using a single modality, two modalities, and three modalities and evaluate them. As shown in Fig. 14(b), fusing these multi-modality tasks into one filter results in a slight decrease in the filtering rate. But note that the overall efficiency is improved due to the shared parameters among different tasks.

**Online Active Update:** To evaluate the active strategy for online adaptation of InFi, we select the VC workload and compare three training methods: (1) *Offline:* selects the first 10% frames of a day to train; (2) *Periodic:* selects the first 10% frames of each hour to train and update; (3) *Active:* see Section V-D. For a fair comparison, we set the same threshold (0.5) for the three methods. As shown in Fig. 15, our proposed active strategy significantly improves the online adaptability of InFi-Skip. The offline policy’s performance seriously degrades when the input distribution changes, mainly because the frames change from infrared to RGB images. On average, the offline policy achieves only 56.4% inference accuracy. The periodic update policy alleviates this problem to some extent, improving the average accuracy to 87.0%, but still suffers from performance fluctuations. The active strategy’s performance only drops at the 7th time segment, as it does not see any RGB images before that. And the active strategy effectively selects informative samples to fit the new distribution and performs accurate filtering robustly. On average, our active policy achieves 94.8% inference accuracy, which is 38.4% higher than the offline policy.

**C. Filterability**

In Section III, we compare the hypothesis complexity of the inference and filter models. Let “Conf. > T” denote the low-confidence classification case (§ III-B), “Class Subset” denote the redundant class subset case (§ III-C), and “Reg. > T” denote the bounded regression case (§ III-D). GC and SC belong to the “Conf. > T” case, where T is 0.9. AC, NER, and HAR belong to the “Class Subset” case, where AC selects 2 action labels, NER selects the “PERSON” label, and HAR selects the “LAYING” label. FD, PE, and VC(MP) belong to the ‘Reg. > T” case, where T is 0. SR is a sequence-to-sequence model, which cannot perfectly fit any of these three cases. We compute the ratio of the resulting filtering rate to the optimal filtering rate at 90% inference accuracy to compare the filterability of different cases. From a practical perspective, we evaluate the overall throughput with and without InFi-Skip filters. As shown in Fig. 16, the “Conf. > T” case in which we proved that the filter’s complexity is not less than the inference model achieves obviously lower filtering ratio (0.41 median), while other cases in which we proved that the filter tends to be less complex achieve apparently higher ratios (0.71/0.78 medians). On the other hand, the overall throughput improved by InFi-Skip filters on filterable cases is more significant than the non-filterable cases. In the non-filterable cases, GC and SC, InFi achieves around 1.3× throughput, while in the filterable cases, it can improve the throughput up to 5.92× and achieves 1.8 and 2.25 medians for regression and subset-class cases, respectively. These results show the guiding significance of our proven filterability in real applications.

**D. Computation and Resource Efficiency**

As we discussed in Section II-B, a “valid” filter should be both accurate and lightweight. The above results have shown that InFi can filter a significant amount of inputs while keeping accurate inference. In the training phase, InFi (image modality) takes around 710 ms per batch (batch size is 32) and requires 5337 MB GPU memory which most commercial GPUs can meet. InFi for other input modalities requires far fewer resources, e.g., InFi (vector) tasks 3 ms per batch and needs only 435 MB memory. We test the latency and energy in the inference phase on mobile platforms. As a fair comparison, we chose the TFLite-optimized MobileNetV1, which is one of the most
efficient CNNs on mobile devices. As shown in Fig. 17, on three mobile platforms, InFi with the image feature network costs only 12-25% runtime of MobileNetV1. The average energy costs of InFi are 14.4/79.7 mJ per frame, which are much lower than MobileNetV1 (410.4/803.8 mJ per frame) on the phone/smartwatch. We implement InFi with MindSpore and the results show that InFi’s low-energy consumption and low-latency execution do not depend on the implementation framework.

On-Device Online Update: Based on the Chaquopy library, we tested the overhead of training on a mobile phone (XIAOMI Mi 5) and a smartwatch (HUAWEI WATCH). We randomly generated images with the shape (224, 224, 3) and set the batch size as 16. For InFi (image modality), experiments show that it takes around 20 s and 50 s per batch to online update weights on the phone and the watch, respectively. And for NVIDIA JETSON TX2, training the input filter with the same configurations takes around 1 s per batch.

E. Different Mobile-Centric Deployments

Now we evaluate the overall performance of inference workloads in real systems with three ways of deployments.

Vehicle Counting: First, we consider the vehicle counting workload: 1) on-device: InFi (image) and YOLOv3 model on TX2; 2) offload: InFi (image) on TX2 and YOLOv3 model on edge; 3) model partitioning (MP): first 39 layers (10 convolution blocks) of YOLOv3 and InFi (feature map) on TX2, rest of YOLOv3 on the edge server. The average throughput of the YOLOv3 model on TX2 and edge is 3.2 FPS and 22.0 FPS, respectively. For MP deployment, the edge-side model serves 24.5 FPS. We report the average throughput and the bandwidth saving of using InFi-Skip and InFi-Reuse, with over 90% inference accuracy, in Table IV. As a fair comparison, we test the throughput of YOLOv3-tiny [12] model, a compressed version for YOLOv3. The inference accuracy of YOLOv3-tiny is only 67.9% which does not meet the 90% target. Breaking down the overheads, InFi’s inference costs around 3 ms per frame, and the average latency of KNN is 6 ms per frame with K=10 and cache size=1000. Achieving over 90% inference accuracy, InFi-Skip improves the throughput to 9.3/55.2/39.0 FPS for on-device/offload/MP deployments, respectively. Apparently, in vehicle counting workloads, there are more filtering opportunities for InFi-Reuse. InFi-Reuse improves the throughput to 27.2/77.2/46.0 FPS for these three deployments. Except for the on-device deployment that does not involve cross-device data transmission, InFi-Skip / InFi-Reuse also save 66.5% / 91.1% and 70.7% / 95.0% bandwidth for offloading and MP workloads. Unlike YOLOv3-tiny which trades a significant and fixed loss of accuracy for efficiency, InFi provides a flexible trade-off between the inference accuracy and overheads.

Pose Estimation: Second, we evaluate the pose estimation workload: 1) on-device: InFi (image) and OpenPose model on TX2; 2) offload: InFi (image) on TX2 and OpenPose model on edge; 3) model partitioning (MP): first 39 layers (10 convolution blocks) of OpenPose and InFi (feature map) on TX2, rest of OpenPose on the edge server. Also, we test the throughput of OpenPose-light [13] model, a lightweight version of OpenPose. Experimental results are shown in Table V. Similar to the vehicle counting workload, the lightweight model cannot achieve our target 90% inference accuracy, although its throughput boosts significantly. InFi-Skip can flexibly balance the inference accuracy and throughput. For example, for the on-device deployment, the throughput improves to 1.17× after using InFi-Skip and the inference accuracy keeps over 90%.

Natural Language Processing Workloads: Third, we test two NLP workloads, NER and SC, with different deployments. Note that, model partitioning is not applicable to our NER workload due to black-box APIs. As shown in Table VI, InFi-Skip effectively improves throughput and saves 22.5% and 26.8% offloading communication for the two tasks, respectively.
VII. RELATED WORK

Frame Filtering: NoScope [24] trains task-specific difference detectors to choose necessary frames for object queries in the video database. FilterForward [19] leverages MobileNet and trains a binary micro-classifier on the intermediate output of a selected layer to determine whether to transmit the input image to the server with the offloaded model. Reducto [20] performs on-device frame filtering by thresholding the difference of low-level features between successive frames. Through elaborate selection for different tasks, low-level features can efficiently and accurately measure the difference.

Inference Caching: Potluck [22] stores and shares inference results between augmented reality applications. It dynamically tunes the threshold of input similarity and manages cache based on the reuse opportunities. FoggyCache [21] is more general and can be applied to both image and audio inputs. It designs adaptive LSH and homogenized KNN algorithms to address practical challenges in inference caching. Instead of caching the final inference results, DeepCache [68] stores the intermediate feature maps to achieve more granular reuse. For object recognition, Glimpse [23] maintains a cache of video frames on mobile devices. It uses cached results to perform on-device object tracking and sends only trigger frames to the server with offloaded recognition model.

Approaches Tailored for Specific Pipelines: Focus [69] is designed for querying detected objects in a video database and uses compressed CNN to index possible object classes at ingest stage and reduces the query latency by clustering similar objects. Blazet [70] develops neural-network-based methods to optimize approximate aggregation queries of detected objects in video databases. Focusing on object detection in video streams, Chameleon [71] proposes to adaptively select a suitable pipeline configuration including the resolution and frame rate of videos, backbone neural networks for inference, etc. Elf [11] is designed for mobile video analytics where the input data is pre-processed by a lightweight on-device model and then offloaded in parallel to multiple servers with the same subsequent inference functionality.

Our proposed input-filtering framework unifies the frame filtering and inference caching approaches. And we complement existing work in theoretical analysis and flexible supports for more input modalities and deployments.

VIII. CONCLUSION

In this paper, we study the input filtering problem and provide theoretical results on complexity comparisons between the hypothesis families of inference models and their input filters. We propose the first end-to-end learnable framework that unifies both SKIP and REUSE methods and supports multiple input modalities and deployments. We design and implement an input filter system InFi based on our framework. Comprehensive evaluations confirm our proven results and show that InFi has wider applicability and outperforms strong baselines on accuracy and efficiency.

REFERENCES

[1] M. Yuan, L. Zhang, F. He, X. Tong, and X.-Y. Li, “InFi: End-to-end learnable input filter for resource-efficient mobile-centric inference,” in Proc. 28th Ann. Int. Conf. Mobile Comput. Netw., 2022, pp. 228–241.
[2] C. Liu, L. Zhang, Z. Liu, K. Liu, X. Li, and Y. Liu, “Lasagna: Towards deep hierarchical understanding and searching over mobile sensing data,” in Proc. 22nd Ann. Int. Conf. Mobile Comput. Netw., ser. MobiCom ’16, New York, NY, USA, 2016, pp. 334–347.
[3] N. Mazumder et al., “A survey on the optimization of neural network accelerators for Micro-Al on-device inference,” IEEE Trans. Emerg. Sel. Topics Circuits Syst., vol. 11, no. 4, pp. 532–547, Dec. 2021.
[4] E. Li, L. Zeng, Z. Zhou, and X. Chen, “Edge AI: On-demand accelerating deep neural network inference via edge computing,” IEEE Trans. Wireless Commun., vol. 19, no. 1, pp. 447–457, Jan. 2020.
[5] X. Wang, Y. Han, C. Wang, Q. Zhao, X. Chen, and M. Chen, “In-edge AI: Intelligentizing mobile edge computing, caching and communication by federated learning,” IEEE Netw., vol. 33, no. 5, pp. 156–165, Sep./Oct. 2019.
[6] D. Chatzopoulos, C. Bermejo, Z. Huang, and P. Hui, “Mobile augmented reality survey: From where we are to where we go,” IEEE Access, vol. 5, pp. 6917–6950, 2017.
[7] Y. Wang et al., “Bandwidth-efficient live video analytics for drones via edge computing,” in Proc. IEEE/ACM Symp. Edge Comput., 2018, pp. 159–173.
[8] G. Ghiasi et al., “Simple copy-paste is a strong data augmentation method for instance segmentation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2021, pp. 2918–2928.
[9] A. Bulat, J. Kossaifi, G. Tzimiropoulos, and M. Pantic, “Toward fast and accurate human pose estimation via soft-gated skip connections,” in Proc. IEEE 15th Int. Conf. Autom. Face Gesture Recognit., 2020, pp. 8–15.
[10] H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and T. Zhao, “Smart: Robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization,” in Proc. 35th Ann. Meeting Assoc. Comput. Linguistics, 2020, pp. 2177–2190.
[11] W. Zhang et al., “Elf: Accelerate high-resolution mobile deep vision with content-aware parallel offloading,” in Proc. 27th Ann. Int. Conf. Mobile Comput. Netw., 2021, pp. 201–214.
[12] P. Adarsh, P. Rathi, and M. Kumar, “YOLO v3-tiny: Object detection and recognition using one stage improved model,” in Proc. IEEE 6th Int. Conf. Adv. Comput. Commun. Syst., 2020, pp. 687–694.
[13] D. Osokin, “Real-time 2D multi-person pose estimation on CPU: Lightweight openpose,” in Proc. 8th Int. Conf. Pattern Recognit. Appl. Methods, 2019, pp. 744–748.
[14] Z. Sun, H. Yu, X. Song, R. Liu, Y. Yang, and D. Zhou, “MobileBERT: A compact task-agnostic BERT for resource-limited devices,” in Proc. 35th Ann. Meeting Assoc. Comput. Linguistics, 2020, pp. 2158–2170.
[15] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, “MobileNetV2: Inverted residuals and linear bottlenecks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 4510–4520.
[16] M. Tan et al., “Mnasnet: Platform-aware neural architecture search for mobile,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2019, pp. 2820–2828.
[17] S. Han, H. Shen, M. Philosof, S. Agarwal, A. Wolman, and A. Krizhevsky, “MCDNN: An approximation-based execution framework for deep stream processing under resource constraints,” in Proc. 14th Ann. Int. Conf. Mobile Syst. Appl. Serv., 2016, pp. 123–136.
[18] Y. He, J. Lin, Z. Liu, H. Wang, L.-J. Li, and S. Han, “AMC: AutoML for model compression and acceleration on mobile devices,” in Proc. Eur. Conf. Comput. Vis., 2018, pp. 784–800.
[19] C. Canel et al., “Scaling video analytics on constrained edge nodes,” in Proc. Mach. Learn. Syst., A. Talwalkar, V. Smith, and M. Zaharia, Eds., 2019, pp. 406–417.
[20] Y. Li, A. Padmanabhan, P. Zhao, W. Yang, G. H. Xu, and R. Netravali, “Reductor: On-camera filtering for resource-efficient real-time video analytics,” in Proc. Annu. Conf. ACM Special Int. Group Data Commun. Appl. Technol. Archit. Protoc. Comput. Commun., ser. SIGCOMM ’20, New York, NY, USA, 2020, pp. 359–376.
[21] P. Guo, B. Hu, R. Li, and W. Hu, “FoggyCache: Cross-device approximate computation reuse,” in Proc. 24th Annu. Int. Conf. Mobile Comput. Netw., 2018, pp. 19–34.
[22] P. Guo and W. Hu, “Potluck: Cross-application approximate deduplication for computation-intensive mobile applications,” in Proc. 23rd Int. Conf. Archit. Support Program. Lang. Operating Syst., 2018, pp. 271–284.
K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. Eur. Conf. Comput. Vis., Springer, 2016, pp. 499–515.

M. J. Kearns, U. V. Vazirani, and U. Vazirani, An Introduction to Computational Learning Theory. Cambridge, MA, USA: MIT Press, 1994.

L. Zhou, H. Wen, R. Teodorescu, and D. H. Du, “Distributing deep neural networks with containerized partitions at the edge,” in Proc. 2nd USENIX Workshop Hot Top. Edge Comput., 2019.

L. Valiant, Probably Approximately Correct: Nature’s Algorithms for Learning and Prospering in a Complex World. New York, NY, USA: Basic Books (AZ), 2013.

V. Vapnik and A. Y. Chervonenkis, “On the uniform convergence of relative frequencies of events to their probabilities,” Theory Probability Appl., vol. 16, no. 2, pp. 264–280, 1971.

N. Hariyanto, C. Liaw, and A. Mehrabian, “Nearly-tight VC-dimension bounds for piecewise linear neural networks,” in Conf. Learn. Theory, 2017, pp. 1064–1068.

T.-Y. Lin et al., “Microsoft coco: Common objects in context,” in Proc. Eur Conf. Comput. Vis., Springer, 2014, pp. 740–755.

M. Yuan, L. Zhang, X.-Y. Li, and H. Xiong, “Comprehensive and efficient data labeling via adaptive model scheduling,” in Proc. IEEE 36th Int. Conf. Data Eng., 2020, pp. 1858–1861.

M. Yuan, L. Zhang, X.-Y. Li, L.-Z. Yang, and H. Xiong, “Adaptive model scheduling for resource-efficient data labeling,” ACM Trans. Knowl. Discov. Data, vol. 16, no. 4, pp. 1–22, Jan. 2022.

Y. Duan, C. Jin, and Z. Li, “Risk bounds and rademacher complexity in batch reinforcement learning,” in Proc. Int. Conf. Mach. Learn., PMLR, 2021, pp. 2892–2902.

C. Cortes, V. Kuznetsov, M. Mohri, and S. Yang, “Structured prediction theory based on factor graph complexity,” in Adv. Neural Inf. Process. Syst., vol. 29, pp. 2514–2522, 2016.

T. Glasmachers, “Limits of end-to-end learning,” in Proc. Asian Conf. Mach. Learn., PMLR, 2017, pp. 17–32.

A. Prakash, K. Chitta, and A. Geiger, “Multi-modal fusion transformer for end-to-end autonomous driving,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2021, pp. 7077–7087.

D. Amodei et al., “Deep speech 2: End-to-end speech recognition in English and mandarin,” in Proc. Int. Conf. Mach. Learn., PMLR, 2016, pp. 173–182.

G. Koch et al., “Siamese neural networks for one-shot image recognition,” in Proc. Int. Conf. Deep Mach. Learn. Workshop, 2015.

Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, “DeepFace: Closing the gap to human-level performance in face verification,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2014, pp. 1701–1708.

L. Leal-Taixé, C. Cunton-Ferrer, and K. Schindler, “Learning by tracking: Siamese CNN for robust target association,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops, 2016, pp. 33–40.

F. Hollett, “Xception: Deep learning with depthwise separable convolutions,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 1251–1258.

K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 770–778.

I. Indrawan, I. Hayupati, and D. P. S. Putri, “Markerless augmented reality utilizing gyroscope to demonstrate the position of Dewata Nawa Sanga,” Int. J. Interactive Mobile Technol., vol. 12, no. 1, 2018, Art. no. 19.

D. Anguita et al., “A public domain dataset for human activity recognition using smartphones,” in Proc. Eur. Symp. Artif. Neural Netw., 2013, Art. no. 3.

A. Maurer, M. Pontil, and B. Romera-Paredes, “The benefit of multitask representation learning,” J. Mach. Learn. Res., vol. 17, no. 81, pp. 1–32, 2016.

N. Tripuraneni, M. Jordan, and C. Jin, “On the theory of transfer learning: The importance of task diversity,” in Proc. Adv. Neural Inf. Process. Syst., 2020, pp. 7852–7862.
Lan Zhang (Member, IEEE) received the bachelor’s and PhD degrees from Tsinghua University, China. She is currently a professor with the School of Computer Science and Technology, University of Science and Technology of China. Her research interests include mobile computing, privacy protection, and data sharing and trading.

Fengxiang He (Member, IEEE) received the BSc degree in statistics from the University of Science and Technology of China, the MPhil and PhD degrees in computer science from the University of Sydney. He is currently algorithm scientist with JD Explore Academy leading its trustworthy AI team. His research interest is in the theory and practice of trustworthy AI, including deep learning theory, privacy-preserving ML, algorithmic game theory, and decentralized learning. He publish in prominent venues, including ICML, NeurIPS, ICLR, CVPR, and ICCV. He is the area chair of prestigious conferences, AISTATS, BMVC, and ACML. He is the leading author of several standards.

Xuetong Tong received the bachelor’s degree in computer science and technology from Henan University. She is currently working toward the master’s degree with the Institute of Advanced Technology, University of Science and Technology of China. Her research interests include model reasoning and optimization.

Miao-Hui Song received the bachelor’s degree in computer science and technology from Chongqing University. She is currently working toward the master’s degree with the School of Computer Science and Technology, University of Science and Technology of China. Her research interests include active learning and data-labeling systems.

Zhengyuan Xu (Senior Member, IEEE) received the BS and MS degrees from Tsinghua University, China, and the PhD degree from the Stevens Institute of Technology, USA. He was a tenured full professor with the University of California at Riverside and later with Tsinghua University before he joined University of Science and Technology of China (USTC). He was founding Ddirector of the multi-campus Center for Ubiquitous Communication by Light (UC-Light), University of California, and Founding Director of Wireless-Optical Communications Key Laboratory of Chinese Academy of Sciences. He was a distinguished expert and chief scientist of the National Key Basic Research Program of China. His research focuses on Petahertz communications, optical wireless communications, mobile networking, artificial intelligence, wireless Big Data, sensing, ranging and localization. He has published more than 400 international journal and conference papers, and co-authored a book titled Visible Light Communications: Modulation and Signal Processing which has been selected by IEEE Series on Digital & Mobile Communication and published by Wiley-IEEE Press. He has been on the Elsevier annual list of Most Cited Chinese Researchers since 2014. He has served as an associate editor for different IEEE/OSA journals and was a founding co-chair of IEEE Workshop on Optical Wireless Communications in 2010.

Xiang-Yang Li (Fellow, IEEE) received the bachelor’s degree from the Department of Computer Science from Tsinghua University, China, in 1995 and the MS and PhD degrees from the Department of Computer Science from University of Illinois at Urbana-Champaign, in 2000 and 2001, respectively. He is a professor and executive dean with the School of Computer Science and Technology, USTC. He is an ACM Fellow (2019) and an ACM distinguished scientist (2014). He was a full professor with Computer Science Department of IIT and co-Chair of ACM China Council. His research interests include Artificial Intelligence of Things (AIoT), privacy and security of AIoT, and data sharing and trading.