

DeepCache: Principled Cache for Mobile Deep Vision

Mengwei Xu, Mengze Zhu

Key Lab of High-Confidence Software Technologies
(Peking University), MoE, Beijing, China
{xumengwei,zhumz}@pku.edu.cn

Felix Xiaozhu Lin

Purdue ECE
West Lafayette, Indiana, USA
xzl@purdue.edu

Yunxin Liu

Microsoft Research
Beijing, China
yunxin.liu@microsoft.com

Xuanzhe Liu

Key Lab of High-Confidence Software Technologies
(Peking University), MoE, Beijing, China
xzl@pku.edu.cn

ABSTRACT

We present DeepCache, a principled cache design for deep learning inference in continuous mobile vision. DeepCache benefits model execution efficiency by exploiting temporal locality in input video streams. It addresses a key challenge raised by mobile vision: the cache must operate under video scene variation, while trading off among cacheability, overhead, and loss in model accuracy. At the input of a model, DeepCache discovers video temporal locality by exploiting the video’s internal structure, for which it borrows proven heuristics from video compression; into the model, DeepCache propagates regions of reusable results by exploiting the model’s internal structure. Notably, DeepCache eschews applying video heuristics to model internals which are not pixels but high-dimensional, difficult-to-interpret data.

Our implementation of DeepCache works with unmodified deep learning models, requires zero developer’s manual effort, and is therefore immediately deployable on off-the-shelf mobile devices. Our experiments show that DeepCache saves inference execution time by 18% on average and up to 47%. DeepCache reduces system energy consumption by 20% on average.

Xuanzhe Liu is the paper’s corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

MobiCom ’18, October 29–November 2, 2018, New Delhi, India

© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-5903-0/18/10...\$15.00
<https://doi.org/10.1145/3241539.3241563>

CCS CONCEPTS

- Human-centered computing → Ubiquitous and mobile computing;
- Computing methodologies → Computer vision tasks;

KEYWORDS

Deep Learning; Mobile Vision; Cache

ACM Reference Format:

Mengwei Xu, Mengze Zhu, Yunxin Liu, Felix Xiaozhu Lin, and Xuanzhe Liu. 2018. DeepCache: Principled Cache for Mobile Deep Vision. In *The 24th Annual International Conference on Mobile Computing and Networking (MobiCom ’18), October 29–November 2, 2018, New Delhi, India*. ACM, New York, NY, USA, 16 pages. <https://doi.org/10.1145/3241539.3241563>

1 INTRODUCTION

With ubiquitous cameras on mobile and wearable devices, *continuous mobile vision* emerges to enable a variety of compelling applications, including cognitive assistance [29], life style monitoring [61], and street navigation [27]. To support continuous mobile vision, Convolutional Neural Network (CNN) is recognized as the state-of-the-art algorithm: a software runtime, called deep learning engine, ingests a continuous stream of video images¹; for each input frame the engine executes a CNN model as a cascade of *layers*, produces intermediate results called *feature maps*, and outputs inference results. Such CNN executions are known for their high time and space complexity, stressing resource-constrained mobile devices. Although CNN execution can be offloaded to the cloud [2, 34], it becomes increasingly compelling to execute CNNs on device [27, 44, 52], which ensures fast inference, preserves user privacy, and remains unaffected by poor Internet connectivity.

To afford costly CNN on resource-constrained mobile/wearable devices, we set to exploit a mobile video stream’s *temporal locality*, i.e., rich information redundancy among consecutive video frames [27, 51, 52]. Accordingly, a deep learning

¹We refer to them as a *mobile video stream* in the remainder of the paper.

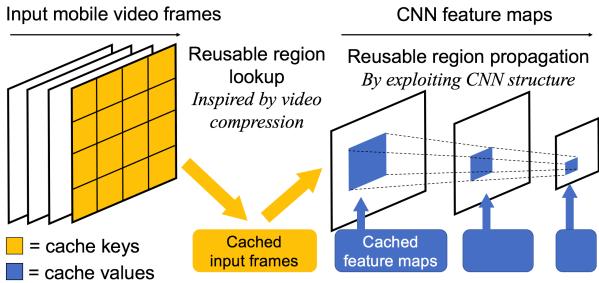


Figure 1: The overview of DeepCache.

engine can *cache* results when it executes CNN over a mobile video, by using input frame contents as cache keys and inference results as cache values. Such caching is expected to reduce the engine’s resource demand significantly.

Towards effective caching and result reusing, we face two major challenges. 1) *Reusable results lookup*: Classic caches, e.g., the web browser cache, look up cached values (e.g., web pages) based on key *equivalence* (e.g., identical URLs). This does not apply to a CNN cache: its keys, i.e., mobile video contents, often undergo moderate scene variation over time. The variation is caused by environmental changes such as user/camera motion, object appearance, and illumination changes [57]. A CNN cache must systematically tolerate the variations and evaluate key *similarity*. In doing so, the engine must trade off among cacheability, overhead, and model accuracy. 2) *Fine-grained reuse within a CNN*: In a CNN model, expensive computations spread across multiple layers. Besides caching the CNN’s final inference outputs, the engine should cache the intermediate results (i.e., feature maps) produced by the internal layers. Furthermore, the engine should reuse the cached feature maps at fine spatial granularity. However, feature maps are high-volume, high-dimensional, barely interpretable data. It can be both expensive to inspect them and difficult to assess their similarity.

Few deep learning engines address the two challenges simultaneously. Commodity engines [6, 11, 17] process video frames in independent inference tasks with no reuse in between. A few recent research prototypes [24, 51] incorporate ad-hoc cache designs: they either look up reusable results based on pixel-wise equivalence of image regions, or perform expensive cache lookup over feature maps at all layers inside a CNN. As a result, they often suffer from low cacheability and high lookup overhead, leaving much caching benefit untapped.

To this end, we advocate a principled cache design called DeepCache. The key ideas of DeepCache, as shown in Figure 1, are that i) it discovers reusable image regions by exploiting *the input video’s internal structure*, for which it borrows the wisdom from decades of video research [21, 59, 68]; ii)

it propagates the discovered reusable regions within a CNN by exploiting *the CNN’s internal structure*.

As shown in Figure 1, DeepCache stores recent input frames as cache keys and stores recent feature maps for individual CNN layers as cache values. To manage the cache, it provides two core mechanisms.

- At the engine input, DeepCache performs cache key lookup: it partitions each video frame into fine-grained regions and searches for similar regions in (cached) recent input frames. It does so by running its region matcher. Inspired by video compression [68], the matcher searches neighboring regions in specific patterns guided by video motion heuristics. DeepCache keeps merging adjacent discovered regions in order to tackle *cache erosion*, i.e., diminishing reusability at deeper layers. In contrast to ad-hoc image comparison used by prior CNN caches [24, 51], our matcher is more robust to the aforementioned scene variations; the matcher runs fast to process more than 1,000 227×227 frames per second.
- Into the CNN execution, DeepCache maps the matched regions on input images to *reusable* regions on feature maps. It propagates the reusable regions across the feature maps of all CNN layers. At each layer, DeepCache transforms the reusable region boundaries based on the operators of this layer; it fills the reusable regions with cached feature map values in lieu of actual CNN execution. During the process, DeepCache weaves cache queries into CNN computations, keeping the cache queries transparent to CNN models.

With these two mechanisms, DeepCache runs its region matcher only *once* per video frame at the input; it then loads cached feature maps at *all* layers inside CNN. This contrasts to ad-hoc approaches that repeat matching processes over both images and feature maps, in and out of CNN. Our rationale is that, while humans have reliable heuristics on similarity of image contents (which allows DeepCache to assess cache key similarity), they still lack knowledge on evaluating similarity of CNN’s internal feature maps that are in disparate dimensions. By always treating feature maps as cache values not keys, DeepCache eschews high-cost, low-return searches over them, while still harvesting substantial caching benefit.

We implement DeepCache in *ncnn* [11], a popular deep learning engine, atop Android 6.0. DeepCache executes standard, unmodified CNN models such as ResNet-50 [35]. We evaluate DeepCache on Nexus 6 with five popular CNN models over two large, real-world video datasets. Compared to a baseline engine version without enabling cache, DeepCache reduces the inference time by 18% on average and up to 47%. The reduction in inference time by DeepCache is 2 \times of

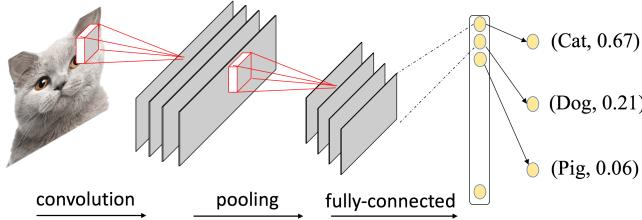


Figure 2: A typical CNN model structure.

the reduction achieved by existing CNN caches design [51]. DeepCache reduces system energy consumption by around 20%. Its incurred accuracy loss is no more than 3%. Across all the models, DeepCache uses 2.5 MB – 44 MB of memory, less than 2% of the total system DRAM.

To summarize, we make the following contributions.

- We present DeepCache, a principled cache for executing CNN over mobile videos (Section 3). DeepCache exploits temporal locality in input mobile videos with proven video heuristics (Section 4), propagates cacheable regions across CNN layers with the CNN knowledge (Section 5), and eschews applying video heuristics to CNN internals.
- We implement DeepCache in a commodity engine. The resultant prototype runs unmodified CNN models, requires zero effort from developers, and is immediately deployable on off-the-shelf Android devices (Section 6).
- We evaluate DeepCache on popular CNN models with real-world datasets (Section 7). The results show that DeepCache can reduce model inference time and energy consumption effectively.

The full source code of DeepCache is at:

<https://github.com/xumengwei/DeepCache>

2 BACKGROUND AND CHALLENGES

In this section, we present CNN background and identify the major challenges to cache for continuous mobile vision.

2.1 Convolutional Neural Network

Convolutional Neural Network (CNN) is the state-of-the-art algorithm in many computer vision tasks, and is recently adopted in many mobile scenarios [3, 18, 52, 53, 64, 66]. As shown in Figure 2, a typical CNN model repeatedly uses convolution and pooling layers to extract features from the whole image, and then applies fully-connected layers (fc) to finalize the vision tasks. Convolutional layers (conv) apply kernels on the input data to extract embedded visual characteristics and generate output data (called *feature map*). For continuous mobile vision, CNN inference operates only on one single segment of data (i.e., an RGB image) at a time.

Model	Lib	conv	fc	pl	act	rest
AlexNet [43]	TF	79.2%	6.4%	11.1%	2.7%	0.6%
	ncnn	77.9%	7.1%	12.1%	1.8%	1.1%
GoogLeNet [58]	TF	80.2%	0.1%	7.5%	8.1%	4.3%
	ncnn	78.8%	0.7%	8.6%	9.3%	2.6%
ResNet-50 [35]	TF	91.8%	5.8%	0.5%	1.7%	0.2%
	ncnn	93.7%	4.9%	0.8%	0.4%	0.2%
YOLO [54]	TF	82.4%	12.8%	2.1%	1.8%	0.9%
	ncnn	84.1%	12.2%	2.6%	0.9%	0.2%
Dave-orig [22]	TF	58.8%	28.6%	4.8%	2.9%	5.2%
	ncnn	62.7%	25.9%	5.8%	3.7%	1.9%

Table 1: Processing time breakdown of popular CNN models, showing that convolutional layers dominate the time. Layer types: convolutional (conv); fully-connected (fc); pooling (pl), activation (act). Hardware: Nexus 6. Engines: Tensorflow (TF) [17]; ncnn [11].

Convolutional layers are hotspots Among all layer types, convolutional layers are the primary performance hotspots. We summarize the latency breakdown of five popular CNN models in Table 1. We use two libraries that support deep learning inference on Android to run these models on a Nexus 6 device: TensorFlow [17] and ncnn [11]. It should be noted that each layer type (e.g., a convolutional layer) can have multiple instances in a model. In the breakdown, convolutional layers dominate the processing time, contributing at least 60% and even up to 90% (ResNet-50). This observation motivates us to focus on caching for convolutional layers in this work.

2.2 Objective and Challenges

Our overall approach to reduce CNN execution workloads is exploiting temporal locality on a mobile video stream. That is, consecutive video frames often have substantial similar or overlapped regions. In general, temporal locality in videos has been known for decades and widely exploited for video compression standards [46, 55]. It is particularly pronounced in mobile videos: mobile devices (e.g., smartphones and glasses), when performing continuous vision tasks [27, 51], capture similar but non-identical image regions continuously. To this end, a deep learning engine can cache the CNN execution outcome from processing earlier frames for reuse in processing a later frame. Of the cache, the *keys* are input image contents and the *values* are the corresponding inference results, i.e., feature maps. This objective, while simple, raises a few unique challenges.

• **Cache lookup under scene variations** In general, cache stores key-value pairs. Classic caches, e.g., for web browsers or disks, look up cached values (e.g., web pages or disk blocks) by evaluating the *equivalence* of keys (e.g., web URLs or block IDs). However, to look up reusable CNN execution results,

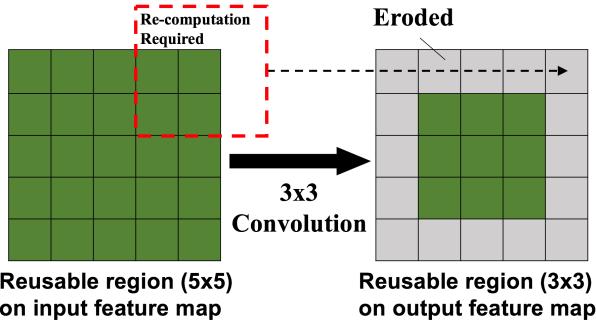


Figure 3: An example showing cache erosion at a convolutional layer (kernel=3x3, stride=1, padding=1).

the cache should evaluate the *similarity* of keys (i.e., input image contents). Images consecutively captured in real world can have various aspects of differences for the presence of large variations in camera motion, object appearance, object scale, illumination conditions, etc. Those complicated conditions make it non-trivial to find out “*what should be reused and what should not*”.

- **Fine-grained reuse of intermediate results** The computation cost of a CNN model spreads over a cascade of internal layers, which produce feature maps as intermediate results. An effective CNN cache should store these feature maps and reuse them at fine spatial granularity whenever possible. However, deciding reusability for feature maps is challenging: since the data volume of feature maps is large, it incurs high overhead for the engine to inspect them; since feature maps consist of data points in higher dimension spaces, it is difficult for the engine to interpret their semantics.

- **Balancing cacheability, model accuracy, and cache overhead** In using cache, the engine will lose CNN model accuracy: it will have to reuse cached values for similar, yet nonidentical, image regions. This entails a complex trade-off. First, while relaxing the criteria for image similarity boosts cacheability, it also reduces model accuracy. Second, while more thorough cache lookup improves cacheability, its additional overhead must be justified by sufficient performance gain.

- **Battling cache erosion** Of a CNN, reusability tends to diminish at its deeper layers, a behavior we dubbed *cache erosion*. More specifically, given an input image region deemed as similar (reusable) to an existing region of previous frame, the amount of reusable results on each layer’s feature map shrinks as execution progresses into the model. Figure 3 shows an example of a convolutional layer, for which the input is a cached region of 5x5 pixels. However, the peripheral pixels (in gray) in the output cannot be loaded from cache as the central ones (in green), and must be exhaustively computed. This is because these peripheral pixels are derived

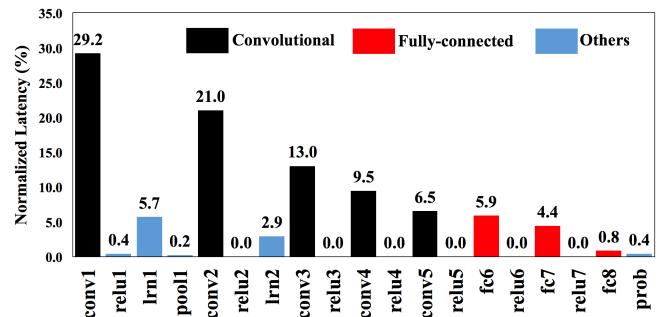


Figure 4: Latency breakdown at layer granularity for AlexNet [43]. Layers are presented at the order of execution: left-side layer will be executed first and the output will be fed to the right-side layer as input.

from both reusable and non-reusable results in the input feature map. As a result, the reusable region has eroded.

Among various CNN layers, convolution, pooling, and LRN erode cache as above; fully-connected layer may completely destroy reusability, since each its output value depends on all its input values, which can hardly be all cached.

Fortunately, in most CNN models, early layers contribute most of the computation cost and also suffer less cache erosion. Fully-connected layers come last in a CNN and contribute minor cost. These are exemplified in Figure 4, which breaks down the execution latency of a popular CNN model. Of the total latency, only 11.5% is contributed by fully-connected layers, while the remaining 88.5% is contributed by earlier layers that can benefit from cache. To further tackle cache erosion, we merge reusable regions into the largest possible ones, as will be discussed in Section 4.

3 SYSTEM OVERVIEW

DeepCache reduces CNN execution workloads by computation reuse. The key advantages of DeepCache include: 1) **No cloud offloading**: DeepCache completely runs on a mobile/wearable device without any offloading onto the cloud. 2) **Widely deployable**: DeepCache works well with popular CNN models. 3) **Transparency and zero developer-effort**: DeepCache caches inference results for *unmodified* CNN models, without requiring the developers to re-train the models or tuning the parameters. This contrasts to disruptive CNN cache designs [24]. In addition, DeepCache exposes optional APIs for apps to fine-control cache behaviors (Section 6), analogous to that a browser cache exposes various policy knobs to web apps [9]. 4) **Minor accuracy loss**: DeepCache minimizes the model accuracy loss, which it trades for cacheability.

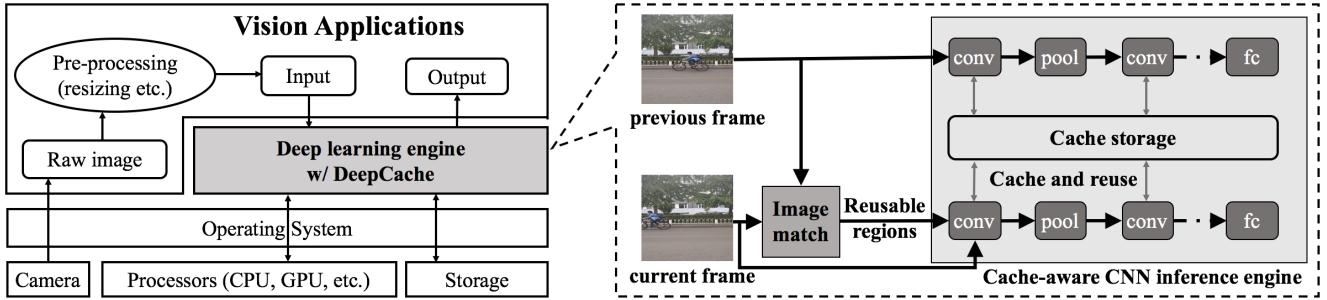


Figure 5: Overall workflow of DeepCache. Black arrows: data flows in CNN execution; green arrows: the reuse of cached data between two consecutive frames.

Figure 5 shows the architecture of DeepCache. DeepCache works as a lightweight extension to a commodity deep learning inference engine. It augments existing model inference with cache, while keeping all other engine components unchanged, including loading CNN model file, ingesting video from the camera, pre-processing video frames, executing CNN models on CPU/GPU, and emitting the final output.

DeepCache in a nutshell For a CNN model, DeepCache maintains a cache, covering the model’s input as well as its internal layers. The cache stores recent video frames for the model input, and recent feature maps for the internal layers. The *cache keys* are equal-sized, fine-grained regions on the cached input frames. The *cache values* are the cached feature maps produced by the layers.

For a new input frame, DeepCache does one-time key lookup by searching for *similar* regions in cached input images. Upon match, DeepCache supplies the engine with corresponding cache values, i.e., feature map regions directly derived from the matched image regions. It further propagates these regions to deeper CNN layers: between layer L_n and layer L_{n+1} , DeepCache maps the reusable regions on L_n ’s feature map to L_{n+1} ’s feature map. It fills these regions with cached feature maps without further key lookup, i.e., search, over these feature maps.

Key lookup: image region matcher (Section 4) Prior to executing CNN over a newly captured video frame, DeepCache partitions the frame into equal-sized, fine-grained regions (default 10x10 pixels in each region). For each region, DeepCache searches for similar regions in recent input frames. It does so based on diamond search [68], a famous algorithm in motion estimation and video compression. Compared to ad-hoc, roll-your-own image match, a mature algorithm is not only proven by decades of practice but also may enjoy pervasive hardware acceleration, e.g., hardware video encoders on mobile SoCs [20]. The matching results are a set of rectangle-to-rectangle mappings, e.g., $(x_i, y_i, w, h) \rightarrow (x'_i, y'_i, w, h)$, where (x_i, y_i) ((x'_i, y'_i)) is the

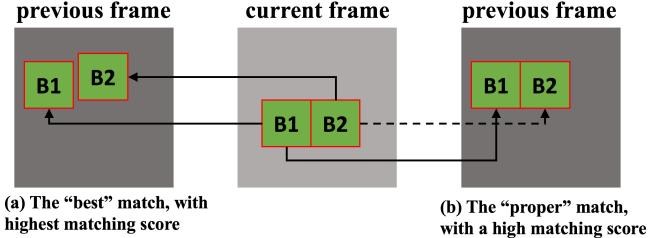


Figure 6: Two matching examples, showing that the best matched block are not always desirable.

left-top point in the current (previous) frame and w (h) is the width (height) of a certain rectangle.

Value mapping: propagating regions across layers (Section 5) After matching image regions on a new input frame, DeepCache sends the frame and the discovered reusable regions into the CNN model. DeepCache augments normal CNN execution with three functions. First, it propagates the mappings between reusable regions (on the new frame) and cached regions (on an old frame), alongside the input data. Second, the spatial convolution operation skips computation for the reusable regions and instead directly loads from cached feature maps. Third, DeepCache caches the output feature map at each convolutional layer for future inference.

4 IMAGE BLOCK MATCHING

Now we present the detailed design of our region matcher and how it deals with *cache erosion*. The goal of our image matching algorithm is to find “similar” regions (rectangles) between two images. There are two ways to match: *block-wise matching* and *pixel-wise matching*. Theoretically, identifying each pixel’s matching level (pixel-wise matching) and reusing its cached results can be more fine-grained and minimize the model accuracy loss. However, we have observed that even similar scenes in two sequential images can have relatively low matching scores of corresponding pixels (pixel

mutation), due to barely unnoticeable environment variations such as light and moving objects. Those “unmatched” pixels can lead to significant reduction of cache reuse due to the *cache erosion* mentioned in Section 2.2. Thus, we use block-wise matching rather than pixel-wise matching, taking a block (e.g., 10x10 pixels) as the basic unit to tell if it’s successfully matched to a corresponding block in the previous image. In this way, a mutated pixel will not affect the block-wise matching decision if other surrounding pixels in the block are well matched.

Two principles should be considered into the design of our block-wise matching algorithm. First, the matching algorithm should run fast, keeping the processing overhead negligible compared to the improvement gained via cache reuse. Second, we want the resulted blocks to be likely merged into larger blocks. The second principle is exemplified by the case shown in Figure 6: match(a) might have the highest matching scores for block B1 and B2, but it’s not suitable in our cache mechanism since these small reusable blocks will quickly vanish after several layers due to *cache erosion* (Section 2.2). Imagine that B1 and B2 have size 5x5, and the convolutional kernel is 3x3. After the *cache erosion*, the reusable regions become two 3x3 rectangles, 18 pixels in total. By contrast, match(b) finds two adjacent blocks in current frame that are similar to the blocks in previous frame, so that these two blocks can be merged into a larger one. In this case, the reusable region becomes one 3x8 rectangle after convolution, 24 pixels in total.

The overall flow of our matching algorithm is as follows.

- **Step 1.** The current frame (image) is divided into an NxN grid, where each grid block contains certain number of pixels.
- **Step 2.** For each divided grid block we find the most matched same-size block in previous frame. Here, we denote the left-top point of i -th block ($i = 1$ to N^2) in current frame as (x_i, y_i) , and the corresponding matched block position in previous frame as (x'_i, y'_i) . We leverage the *diamond search* [68] algorithm which is widely used in video compression to quickly identify the most matched block. The matching level (similarity) between two image blocks is represented by the *PSNR* [68] metric: higher *PSNR* indicates that two blocks are more similar.
- **Step 3.** We calculate the average block movement (M_x, M_y) as the mean movement of the matched blocks whose *PSNR* is larger than the given threshold \mathcal{T} .

$$(M_x, M_y) = \left(\frac{\Sigma(x'_i - x_i)}{K}, \frac{\Sigma(y'_i - y_i)}{K} \right), \langle (x_i, y_i), (x'_i, y'_i) \rangle \in \mathcal{S}$$

where \mathcal{S} is the collection of matched block pair whose *PSNR* is larger than \mathcal{T} , and K is the cardinality of \mathcal{S} .



Figure 7: Matched rectangles in two consecutive images via our proposed algorithm.

- **Step 4.** For each block (x_i, y_i) in the current frame, we calculate its *PSNR* with block $(x_i + M_x, y_i + M_y)$ in the previous frame. If *PSNR* is larger than \mathcal{T} , these two blocks are considered to be properly matched.

- **Step 5.** We merge the small blocks that are properly matched in last step to larger ones. For example, if (x_i, y_i) and (x_j, y_j) in current frame are adjacent, then their matched blocks in Step 4 should also be adjacent since they share the same offset (M_x, M_y) . Thus, we can directly merge them into a larger rectangle as well as their matched blocks.

Figure 7 shows an output example of applying our matching algorithm on two consecutively captured images. As observed, the second frame image is different from the first one in two aspects. First, the camera is moving, so the overall background also moves in certain direction. This movement is captured in Step 3 by looking into the movement of each small block and combining them together. Second, the objects in sight are also moving. Those moved objects (regions) should be detected and marked as non-reusable. This detection is achieved in Step 4.

Our experiments show that most of the processing time of the above matching algorithm is spent at Step 2 and Step 4. In Step 2, we need to explore the previous frame to identify the most matched block for every block in current image. We can accelerate this step by skipping some blocks in current frame, e.g., only matching blocks at $(i * k)$ -th row and $(j * k)$ -th column $(i * k, j * k \leq N)$. Theoretically, a 2-skip ($k=2$) can save 75% of the computation time in this step, and a higher k can even achieve better improvements. However, a higher k might also result in inappropriately calculated (M_x, M_y) , resulting in fewer blocks to be properly matched at the last step. We can further accelerate the computation of Step 4 by reusing the results in Step 2 since both of them need to calculate *PSNR* between two blocks. More specifically, if the *PSNR* between (x_i, y_i) (current frame) and $(x_i + M_x, y_i + M_y)$ (previous frame) is already calculated in Step 2, we simply reuse the result. We demonstrate the efficiency of our proposed algorithm as well as these acceleration approaches in Section 7.6.

Layer Type	Layer Parameters	Output(\mathcal{D}_t)
Convolution	kernel=k x k stride=s, padding=p	$x' = \lceil (x+p)/s \rceil, y' = \lceil (y+p)/s \rceil$ $w' = \lfloor (w-k)/s \rfloor, h' = \lfloor (h-k)/s \rfloor$
Pooling		
LRN [10]	radius=r	$x' = x + r, y' = y + r$ $w' = w - 2 * r, h' = h - 2 * r$
Concat [7]	input number=N	overlapped region of these N rectangles
Fully-connected	/	$(x', y', w', h') = (0, 0, 0, 0)$
Softmax [16]	/	$(x', y', w', h') = (x, y, w, h)$
Others	/	

Table 2: Transformation of reusable region boundaries for layer type (\mathcal{D}_t). Input region is a rectangle (x, y, w, h) .

5 CACHE MECHANISMS INSIDE MODEL EXECUTION

To cache a model’s internal inference results, DeepCache provides two facilities: propagation and reuse.

Propagation To reuse the computation results inside CNN inference, DeepCache needs to identify which regions can be reused and where they are mapped to for each layer’s output. As previously explained in Section 2.2, the mappings obtained by matching raw images (Section 4) need to be dynamically adjusted at inference runtime. This adjusting should also be performed on the corresponding cached blocks of previous frame. Obviously, the strategy about how reusable regions are adjusted is based on the forward operation of different layer types. More specifically, a caching mapping $(x_i, y_i, w, h) \rightarrow (x'_i, y'_i, w, h)$ will be adjusted to $\mathcal{D}_t(x_i, y_i, w, h) \rightarrow \mathcal{D}_t(x'_i, y'_i, w, h)$ after operation t , where function \mathcal{D}_t indicates how a reusable region should be adjusted after going through a layer type t . We show the design details of \mathcal{D}_t for every layer type in Table 2. There are three main types of layers in consideration of how they affect the reusable regions: 1) Locally coherent layers that computes each pixel based on a part of input such as convolutional and pooling layers. These layers will diminish the reusable regions. 2) Fully-connected and softmax layers that connect each neuron in input and output. These layers totally destroy the data localization so that there will be no reusable regions. 3) Activation layers such as ReLu, Sigmoid that produce each output neuron based on the corresponding input neuron. These layers have no effect on the reusable regions.

Figure 8 shows an illustrating example about how a reusable region is propagated among different layers. The current image has been matched to previous image, and a block $(100, 100, 100, 40)$ (left-top= $(100, 100)$, width=100, height=40) is identified to be similar to the block $(120, 120, 100, 40)$ of last frame. This image is the input of a convolutional layer, with kernel=11x11, stride=2, and padding=5. The reusable

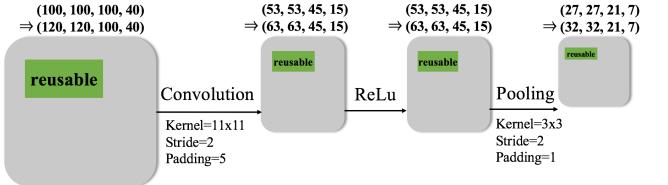


Figure 8: An example showing how the mappings obtained by image matching are adjusted during the CNN inference. The grey rectangles represent the data flow among CNN layers, while the black arrows represent the operations performed on the data.

region of computational output of this layer can be calculated as $(53, 53, 45, 15)$. This output is passed to an activation layer (ReLu) as input, but the reusable region is not changed since the activation layer performs just a certain activation function on every single input unit. Then, the output of ReLu is consumed by a pooling layer, with kernel=3x3, stride=2, padding=1. Similar to the convolutional layer, the reusable region becomes smaller due to the kernel padding.

Reuse After knowing which regions can be reused, DeepCache customizes the convolutional forward so that the computations of these reusable regions are skipped. Instead, they will be directly copied from the corresponding cached region from previous frame. When customizing convolution operations, it’s important to achieve good data locality since data movement is one of the computational bottlenecks [28] during convolution processing. To this end, DeepCache splits the convolution operation into three steps. First, reusable regions are directly copied from cache of last frame. Second, a boolean (bit) map is created to specify whether a pixel (x, y) is already cached. Third, kernel travels on the input feature map and performs convolution only on the non-reusable pixels but skips reusable ones.

DeepCache caches and reuses the computational results only in convolutional layers for two reasons. 1) As mentioned in Section 2, convolution is usually the dominant layer in CNN inference time (e.g., 79.2% for AlexNet). 2) Caching the intermediate output for other layer types (e.g., pooling) requires additional memory overhead. In other words, DeepCache supports caching reuse only in convolutional layers to make proper trade-off among latency improvement and memory overhead. But it’s worth mentioning that we can easily extend our cache mechanism to other layer types.

Though DeepCache reuses only the computation of similar image blocks, there is still accuracy loss since the matched blocks may not be numerically identical. For two consecutive frames, the output disparity can be negligible. However, if the caching goes on for more frames, the accuracy loss might be non-negligible. To mitigate the superposition of accuracy

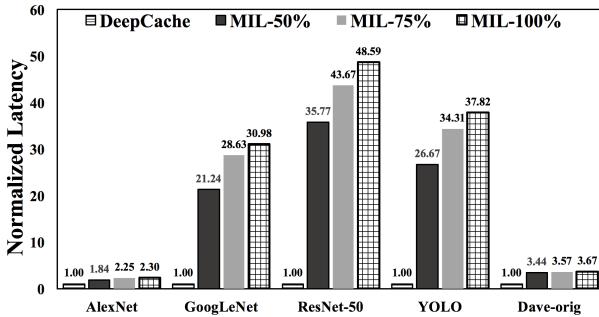


Figure 9: Comparison of match cost (i.e., cache lookup) between DeepCache and “matching internal layers” (MIL), an alternative that attempts to match on all internal feature maps. MIL-N%: matching feature maps at the first N% convolutional layers in the model. Compared to DeepCache, the overhead of MIL is prohibitive.

loss caused by caching, DeepCache periodically cleans its cache and calculates a whole new frame every N frames (default to 10).

Compared to matching internal layers Cache erosion hurts reusability (Section 2.2). An ad-hoc approach to mitigating cache erosion would be aggressively *searching* for reusable regions on feature maps [24] cached for all layers, as we call “matching internal layers” (abbreviated as MIL). Hence, this approach not only matches regions on input frames as DeepCache does, but also matches regions on feature maps that are generated during inference. By doing so, it essentially treats feature map regions as cache keys and looks them up in cache.

Conceptually, MIL may help reusability. Yet, we deem it impractical for the following reasons.

1) *High cost.* Cached feature maps are in high volume. Scanning them for each input frame is expensive. Figure 9 compares the latency in match (i.e., cache lookup) with DeepCache’s approach (propagation of regions) with that of MIL. We thoroughly test MIL by varying the number of convolutional layers it attempts to match on. The results show that MIL incurs much higher latency than DeepCache, even when MIL only covers 50% of the total convolutional layers. This performance gap can be as large as 35×! (e.g., for ResNet)

2) *Low return.* Decades of image research have yielded reliable heuristics on image similarity estimation [59, 68]. By contrast, we know much less about evaluating similarity among CNN feature maps. Hence, when feature maps are used as keys, evaluating their similarity for reuse is fundamentally difficult. One might, for example, devise numerical thresholds for feature map differences [24]. However, our experiences suggest this as intractable: good thresholds, if

exist at all, are specific to models, layers, or even inputs. In other words, MIL inevitably requires extra efforts from application/model developers to identify a good threshold for every single layer of a given CNN model. In comparison, our design of key lookup doesn’t need such efforts from developers.

6 IMPLEMENTATION

We implement our image matcher (Section 4) in RenderScript [14], the Android’s counterpart of CUDA. Thanks to RenderScript, the image matcher execution can be offloaded to GPU for acceleration. Since RenderScript is a generic Android API, our image matcher is portable across Android devices.

We prototype the engine feature of DeepCache atop *ncnn* [11], an open-source deep learning engine optimized for mobile (Android and iOS). *ncnn* works with standard CNN models. DeepCache are directly compatible with those models without requiring extra model changes.

For each supported layer type, *ncnn* provides a function `forward(top_blob, bottom_blob)`, where `top_blob` and `bottom_blob` encapsulate the output and input of this forward step, respectively. We replace `forward()` with our customized `c_forward(top_blob, bottom_blob, c_blob, c_regions)`, where `c_blob` stores the computation results of current layer from the last frame, and `c_regions` specifies which parts can be reused. `c_forward` calculates the output just as `forward` does, except that `c_forward` skips the calculation of cached regions but copies from `c_blob` directly. Before `c_forward` invoked, `c_regions` will be propagated from last layer. As mentioned in Section 5, cached regions will erode (conv, pooling) or vanish (full-connected) during the inference process, thus we use another function named `reg_forward` which calculates how cached regions are propagated among different layers. We also implement some custom layers such as `atan` that are unsupported in the current *ncnn* but necessary for our benchmark models. Overall, our new implementation contains 4,030 lines of code.

DeepCache is fully compatible with *ncnn* APIs. Any existing vision applications built on *ncnn* will work with DeepCache out of box. In addition, DeepCache exposes a few cache parameters, e.g., threshold \mathcal{T} (Section 4), block size N (Section 4), and cache expiration time (Section 5) for developers to optionally fine control DeepCache behavior. This is analogous to a browser cache exposing various policy knobs to web apps [9].

7 EVALUATION

We thoroughly evaluate DeepCache using five typical CNN models on two real-world, large-scale datasets. In summary, DeepCache saves the execution time of CNN models by

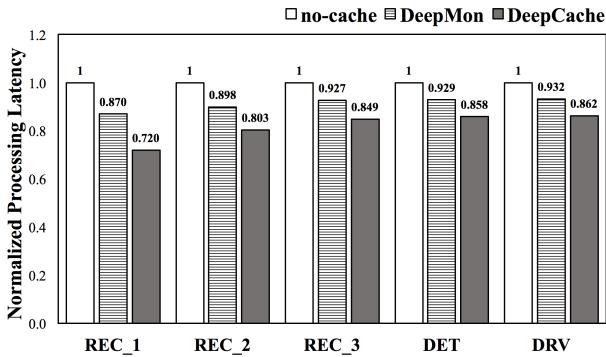


Figure 10: Average processing time of all five CNN models over their test scenarios.

18.2% on average and up to 47.1%, while incurring accuracy loss as low as 3%. In addition, we directly compare DeepCache with the cache mechanism presented in DeepMon [51], a cutting-edge deep learning engine, and the results show that DeepCache outperforms DeepMon on all models and all datasets.

7.1 Experimental Setup

Test Platform We use Nexus 6 (Qualcomm 2.7 GHz quad-core CPU; Adreno 420 GPU) with Android 6.0 as the test platform.

Benchmark Datasets We use two kinds of datasets to evaluate our framework. **UCF101 dataset** [57] contains 101 types of human activities and 13,421 short videos (< one minute) created for activity recognition. We randomly select 10 types from these activities and evaluate DeepCache across them: *Basketball (T1)*, *ApplyEyeMakeup (T2)*, *CleanAndJerk (T3)*, *Billiards (T4)*, *BandMarching (T5)*, *ApplyLipstick (T6)*, *CliffDiving (T7)*, *BrushingTeeth (T8)*, *BlowDryHair (T9)*, and *BalanceBeam (T10)*. In total, **55,680** images have been processed in our evaluation for each selected CNN model. **Nvidia driving dataset** [22] is collected by driving on a wide variety of roads and in a diverse set of lighting and weather conditions. It contains 45,568 static images captured at 10 FPS and the corresponding steering angles made by the driver. We randomly select 10 scenarios² (100 images for each) as the testing set. We use ffmpeg [8] tool to extract raw images from the above video datasets and feed the images to DeepCache sequentially, mimicking video ingestion in real-world continuous vision applications.

Workloads We use a variety of five CNN models to verify DeepCache as shown in Table 3. For activity recognition, our models (REC_1, REC_2, and REC_3) are pre-trained on

²A scenario, in video dataset lingo, refers to a video recorded at specific scene, location, and time.

ILSVRC 2012 dataset [56], and then transferred learned on UCF101. The architectures of those models are initially used for image classification. In our case, we use them to run each single image in the video and average the final result [41]. For object detection, the model (DET) is trained via Pascal VOC 2007 dataset [19]. The model (DRV) used for self-driving is trained and tested on the Nvidia driving dataset mentioned above. It is worth mentioning that these CNN models are quite generalized and can be used in many different tasks with few customization efforts.

Metrics We use accuracy, processing latency, and power consumption to evaluate the performance of our framework. To report the **accuracy** results, we use different metrics to fit into different applications. We report the top-k accuracy for our activity recognition models, and MSE (Mean Squared Error) as the accuracy for object detection and self-driving tasks because their outputs are continuous values. Since the dataset used (UCF101) has no labels for object detection, we treat the output of exhaustively running complete model without cache mechanism as ground truth (observed values). For **latency**, we log the starting time when DeepCache receives the image and the ending time when DeepCache outputs the inference result. The subtracted duration is reported as the processing latency, including the time spent on image matching and CNN inference. Finally, we measure the **energy consumption** via Qualcomm Snapdragon Profiler [15]. The baseline of phone’s idle state is always subtracted.

DeepCache Configuration If not otherwise specified, we use a default block size of 10x10, the matching threshold \mathcal{T} of 20 in our image matching algorithm (Section 4), and the expiration time N of cache is set as 10 (Section 3).

Comparison to Alternatives We experimentally compare the performance of DeepCache to two alternative approaches: *no-cache*: exhaustively running the complete model without cache reuse (ground truth used in measuring accuracy); *DeepMon* [51]: the cache mechanism in a state-of-the-art deep learning engine. To make the comparison fair, we have carefully ported DeepMon’s cache to the ncnn engine executed on the CPU of our test platform, where DeepCache also runs. Note that we have contrasted the design of DeepMon cache with DeepCache (Section 1), and will present more details in related work discussion (Section 9).

7.2 Latency Improvement

Figure 10 summarizes the achieved improvements via applying cache mechanism on average. Our primary observation is that applying DeepCache can have substantial latency reduction compared to *no-cache*, i.e., **18.2%** on average, while *DeepMon* has only **8.9%**. This improvement varies across different CNN models. For a relatively small model

Application	Model Name	Model Architecture	# of Conv	Model Output	Dataset
Activity recognition	REC_1	AlexNet [43]	5	human activity type	UCF101 [57]
	REC_2	GoogLeNet [58]	57		
	REC_3	ResNet-50 [35]	53		
Object detection	DET	YOLO [54]	8	object types and positions	
Self-driving	DRV	Dave-orig [5, 22]	5	steering angle	Nvidia driving dataset [12]

Table 3: List of CNN models used to evaluate DeepCache.

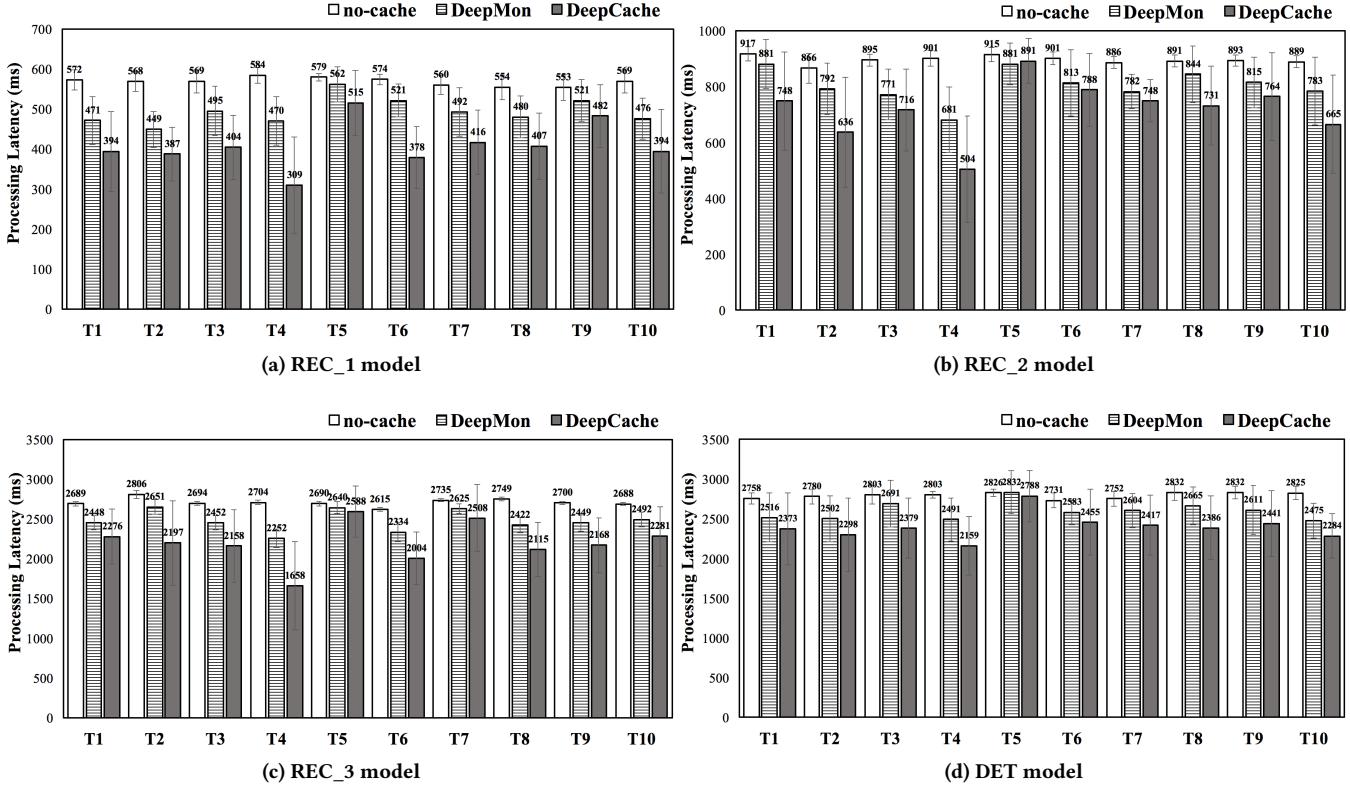


Figure 11: Per-scenario processing time of four CNN models. (For each model, the average time across all scenarios is shown in Figure 10.)

REC_1 (5 convolutions, 25 layers in total), DeepCache results in **28.1%** saving up of total processing time on average, while *DeepMon* only has **13.1%** improvement. For a deeper model REC_2 (57 convolutions, 153 layers in total), the benefit from DeepCache reduces to **19.7%**, while *DeepMon* has only **10.2%**. For DET, DeepCache can have only **14.2%** latency improvement. The reason is that, differently from other classification models, DET is applied in object detection applications and outputs location-sensitive information. Thus, many computation-intensive fully-connected layers reside at the end of DET, making the benefit from convolution-layer cache smaller. The similar situation also applies for the DRV model.

We further illustrate the results under different video benchmarks (UCF101) in Figure 11. We observe that the performance of DeepCache can differ a lot under different benchmarks. Taking REC_1 as an instance, DeepCache saves up **47.1%** processing time under *Billiards* (T4) scenarios. We manually check the dataset videos and identify the reasons of such high speedup as following: 1) camera is in slow motion, 2) most objects are still except the player and balls, 3) indoor lighting is stable. In comparison, DeepCache has only **11.0%** latency improvement when processing *BandMarching* (T5) videos because the camera and most objects (people) in view are moving brokenly. Similarly, for REC_3, DeepCache saves 38.7% processing time when dealing with T4 but only

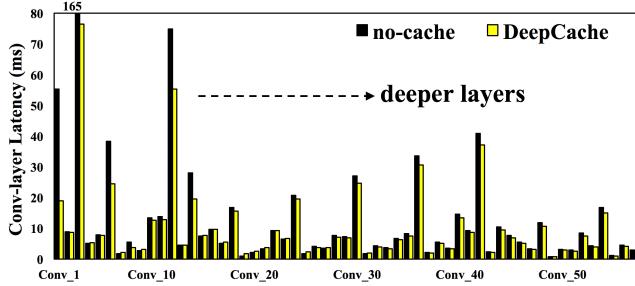


Figure 12: Processing time for individual convolutional layers in model REC_2.

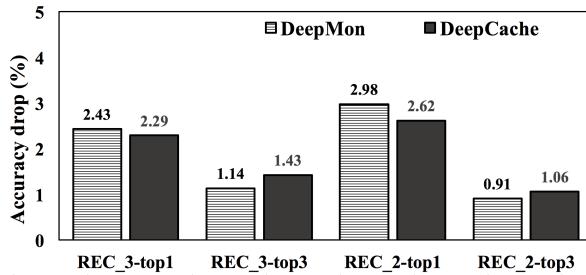


Figure 13: Top-k accuracy drop of DeepCache.

3.8% under T5. Importantly, we observe that DeepCache consistently beats DeepMon for each scenario.

We further dig into the achieved improvement at each individual convolutional layer. As shown in Figure 12, the latency improvement mainly comes from the first few layers due to the *cache erosion* mentioned previously. Fortunately, these layers often contribute to the majority of overall latency, indicating that the benefit remains meaningful when models grow deeper. For example, the third convolutional layer takes 165ms to run, which contributes around 18.4% to the total model. DeepCache is able to save 90.2ms from this single layer since this layer resides at the beginning of the overall model.

7.3 Accuracy Loss

We then investigate how much accuracy DeepCache compromises in return for the latency benefits. The top-k accuracy drop for our activity recognition is shown in Figure 13. In overall, DeepCache and DeepMon both have very small accuracy drop ($\leq 3\%$ for top-1 and $\leq 1.5\%$ for top-3). These loss is acceptable given the observation that our baseline (no-cache) can achieve round 62.8% top-1 accuracy and 76.1% top-3 accuracy. We have even observed cases where the baseline wrongly classifies the image while our DeepCache does it correctly. This is because that we have designed our image matching algorithm to carefully choose which part of computations to reuse, and this reusable information is properly

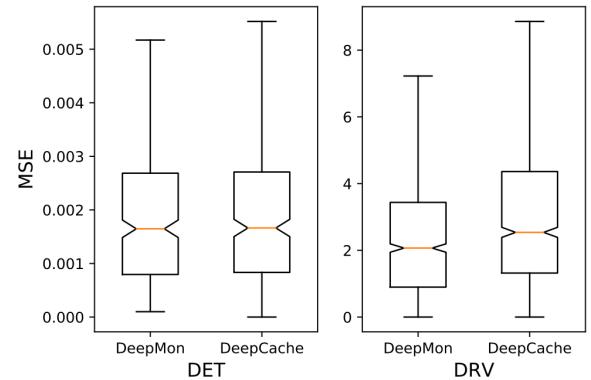


Figure 14: MSE between the output of caching approaches (DeepCache, DeepMon) and ground truth (no-cache).

propagated during inference, thus minimizing the impact on the recognition output.

Figure 14 shows the MSE between ground truth (no-cache) and other cache approaches (DeepCache and DeepMon) when running DET and DRV models. As observed, the median MSE of DeepCache is **0.00166** and **2.617** for DET and DRV respectively, quite similar to the results of DeepMon with **0.00164** and **2.017**. For the DRV case, the results can be interpreted that DeepCache leads to 2.6 degrees offset from the decision made by human driver. Considering that DeepCache will periodically run the total image without cache reuse, as mentioned in Section 3, this offset will not be accumulated. To be compared, our above latency experiment shows that DeepCache can accelerate CNN models two times as DeepMon, e.g., **18.2%** vs. **8.9%** on average across all models and benchmarks.

7.4 Energy Saving

We now investigate the energy consumption of DeepCache across all selected benchmarks, and illustrate results in Figure 15. It is observed that DeepCache can save **19.7%** of energy consumption on average and up to **28.6%** (REC_1), while DeepMon only has **8.0%** on average. This saving is mostly from the reduced processing time. Considering that vision tasks are very energy-intensive, this saving up is able to substantially lengthen battery life. For example, applying DeepCache on REC_3 to classify 10 images can help spare 66.8J energy, enough to support 40 seconds of video playing on Nexus 6 phone according to our measurement.

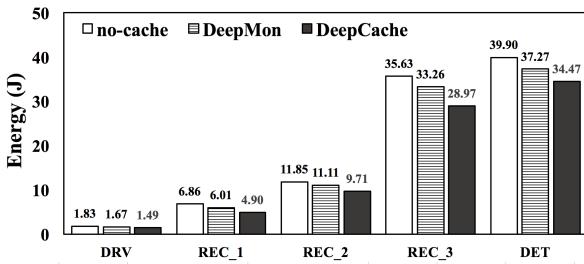


Figure 15: Energy consumption of DeepCache.

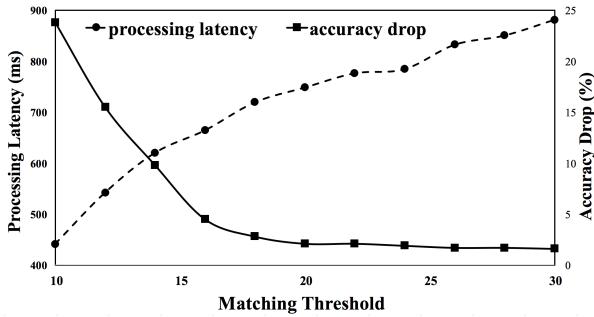


Figure 16: Effect of varied matching threshold \mathcal{T} on processing latency and top-1 accuracy drop of REC_2 model.

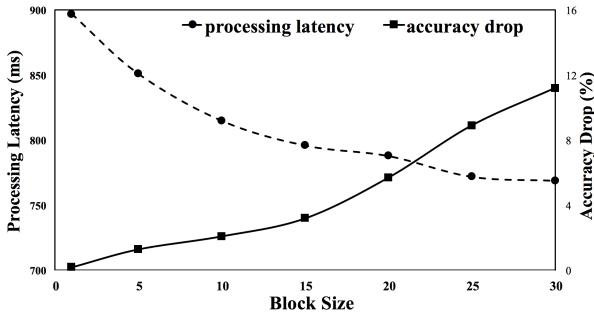


Figure 17: Effect of varied block size on processing latency and top-1 accuracy drop of REC_2 model.

7.5 Choices of Parameters

In our matching algorithm mentioned in Section 4, some variables can be used to make trade-off between latency improvement and accuracy drop. Matching threshold \mathcal{T} is the key to decide whether two image blocks are similar enough to be reused. Figure 16 illustrates how \mathcal{T} can affect the latency and accuracy (REC_2 + T1). As expected, higher \mathcal{T} indicates fewer blocks can be matched, thus leading to less top-1 accuracy drop, but also higher processing latency. In our default setting ($\mathcal{T} = 20$), DeepCache can achieve considerable latency improvement, e.g., **18.3%** (from **917ms** to **748ms**), with acceptable accuracy loss (**2.1%**). This setting

	LATENCY (ms)	MATCH RATIO (%)
DeepMon	4.7 ± 0.7	46.1
ES	33.3 ± 13.16	71.5
TSS	24.7 ± 9.41	70.8
DS	19.5 ± 6.53	71.2
DS + optimization	9.7 ± 2.55	69.5

Table 4: A comparison of image matching algorithms between DeepMon [51] (row 1), which uses histogram-based matching, and DeepCache (the remaining rows) that uses different block matching algorithms in combination with optimization techniques mentioned in Section 4. DeepCache achieves much higher match ratios with minor increase in latency.

aligns with the fact that the acceptable values for wireless transmission quality loss are commonly considered to be about 20 to 25 [13]. However, the threshold can also be set by application developers to adapt to task-specific requirements. For applications that are not very sensitive to the output accuracy, developers can aggressively use a smaller \mathcal{T} to achieve higher latency improvement.

Another configurable parameter in our image matching algorithm is the block size. As observed from Figure 17, a larger block size results in more latency improvement but also higher accuracy loss. This result is reasonable since splitting an image into large blocks indicates more coarse-grained matching. As an extreme case, when block size equals to 1, the accuracy loss is very small (**0.2**) but the latency improvement is also very low (**2.19%**). This is actually the *pixel-wise* approach discussed previously in Section 3, and the result is consistent with our discussion. Our empirical suggestion is setting block size around 10 for 227x227 images.

7.6 Image Matching Performance

Finally, we report the performance of our renderscript-based implementation of image matching algorithm individually. Our current matching algorithm mentioned in Section 4 is based on the diamond search (DS), i.e., DS as an “algorithm unit” (used in Step 2). In addition to the DS, there are several other block matching algorithms that can be plugged into our image matching algorithm to replace DS, such as the Exhaustive Search (ES) and the Three Step Search (TSS). The details and differences of these algorithms are summarized in the survey effort [21]. In this part of evaluation, we also implement the ES-based and the TSS-based image matching to compare. We run preceding algorithms on 10,000 images that are randomly picked from UCF101 and resized to 227x227, and log the processing time (*latency*) and the proportion of matched regions (*match ratio*).

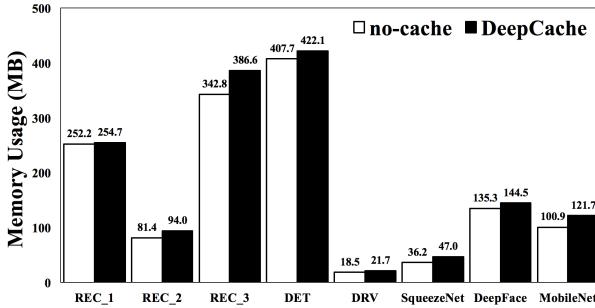


Figure 18: Memory overhead of DeepCache.

As shown in Table 4, our image matching algorithm can achieve around 70% match ratio. The use of different block matching algorithms has minor impacts on the match ratio, but the DS-based implementation is much faster than the ES-based and TSS-based implementation, i.e., 19.5ms vs. 33.3ms & 24.7ms. Another important observation is that the acceleration techniques mentioned in Section 4, i.e., k-skip and reusing, can significantly improve the processing latency from **19.5ms** to **9.7ms** on average, with only **2.4%** loss in the match ratio. These results indicate that our image matching algorithm works well for our CNN cache mechanism, as it occurs quite negligible overhead ($\leq 10\text{ms}$) compared to the benefit gained from cache reusing. To be compared, the histogram-based matching algorithm used in *DeepMon* matches only 46.1% of image areas, while only runs 5ms faster.

In our above experiments, we treat the image matching and CNN inference as two sequential stages so that the time consumed on the image matching diminishes the benefits gained from cache-reuse. Though the matching algorithm is accelerated, it still has non-trivial impacts on the performance of DeepCache especially when the model is relatively small such as DRV. But in practice, these two stages can often be carried out asynchronously when the images can be captured at a higher rate than our CNN inference. More specifically, DeepCache can run the image matching algorithm on i -th image and CNN inference on $(i + 1)$ -th image at the same time. In our case, since we implement these two stages on different mobile processors (GPU and CPU), their processing should not interfere each other, therefore DeepCache can further improve the overall performance.

7.7 Memory Overhead

Figure 18 shows the memory overhead of DeepCache. Besides the 5 models used above, we also test on other three popular CNN models: MobileNet [37], SqueezeNet [39], and DeepFace [62]. Here we assume that all model parameters

are read into memory once initialized without I/O transmission during the inference. We report the memory peak usage during the inference here. As observed, the memory overhead occurred by DeepCache ranges from **2.5MB** to **43.8MB** depending on the internal structure of models. This overhead is quite trivial since nowadays mobile devices are usually equipped with large size of memory, e.g., 3GB in Nexus 6. Note that we only cache and reuse the computation results of convolutional layers so that no extra memory usage will be wasted on other computation-light layers.

8 DISCUSSION

Applicability to other CNN models This paper reports only the results of DeepCache on five typical CNN models. Yet, we expect that DeepCache applies to emerging CNN models, such as the SqueezeNet [39], the MobileNet [37], and the DenseNet [38]. Intuitively, the new models, with their innovated inter-layer organizations and intra-layer optimizations, still preserve the temporal locality that DeepCache hinges on. Furthermore, our observation on the dominating cost of early convolutional layers (Section 2) is true for these new models.

Implementation on accelerators While we prototype the inference stage of DeepCache on CPU, we expect that it can be ported to and benefit from hardware accelerators. Taking GPU as an example, DeepCache is capable of reducing the redundant processing by avoiding GPU kernels for computing output feature maps. For FPGA, we expect that our caching mechanism can be implemented as the hardware logic for further speedup.

Applicability to other video types The idea and high-level design of DeepCache can be applied on other non-mobile videos as long as there's redundancy between adjacent frames. Currently DeepCache is optimized for the mobile vision, because (1) mobile videos contain much richer temporal locality than other video types such as edited movies, and (2) mobile devices are much more sensitive to the latency and the energy consumption in deep vision as compared to other platforms such as desktops or servers.

9 RELATED WORK

Convolutional Layer Caching As most related efforts, DeepMon [51] and CBinfer [24] incorporate CNN caches that we deem ad-hoc. First, they match the image blocks (or pixels) in only the same positions, therefore are unable to tolerate the scene variation as we highlighted in Section 1. By contrast, DeepCache retrofits proven video techniques to systematically search for nearby similar image blocks. Second, they execute cache lookup over feature maps at all layers. Such each-layer matching strategy not only incurs too much

runtime overhead, but also requires extra efforts from application/model developers to manually set a “proper” matching threshold for each layer. By contrast, DeepCache runs lookup only once at the input raw images, and propagates the reusable region boundaries across all the layers. In a concurrent project, *EVA*² [23] proposes hardware optimization for exploiting temporal redundancy in live computer vision. By contrast, DeepCache is designed and implemented to run on general-purpose processors that are widely available on commodity mobile devices. Besides, *EVA*² requires a model to be manually separated into two parts, and the output of the prefix part will be saved and reused while the suffix part will be fully executed. In DeepCache, such manual efforts are naturally avoided by our propagation mechanism mentioned in Section 5. Potluck [32] enables the cross-application cache reuse of computations on a similar video input. However, unlike DeepCache that identifies which parts of image regions shall be reused, the cache mechanism of Potluck is rather coarse-grained since it can reuse **only** the whole output.

Continuous Mobile Vision Emerging mobile vision systems span from commercial products [1, 4] to research prototypes [30, 36, 40, 53, 66, 69]. To optimize mobile vision tasks, [48, 49] made the early energy characterization and optimization towards continuous mobile vision. Starfish [50] allows concurrent vision applications to share computation and memory objects. RedEye [47] reduces image sensor energy by offloading CNN layers to analog domain. DeepEye [52] enables rich analysis of images in near real-time via novel, small form factor wearable camera. Such high interest in mobile vision motivates DeepCache.

Optimizing Deep Learning Execution for Mobile Extensive work is done on making deep learning affordable on mobile devices. The approaches include making models much smaller to fit mobile devices [25, 37, 45, 60], specializing hardware to deep learning algorithms [26, 28, 33, 67], compressing existing models [31, 34, 42, 44, 63, 65], etc. Complementary to these techniques, DeepCache speeds up mobile deep vision through systematically exploiting temporal locality in input data, across multiple inference tasks. DeepCache can coexist with these techniques in one engine.

10 CONCLUSIONS

To conclude our paper, we have proposed DeepCache, a principled cache design, to accelerate the execution of CNN models via leveraging video temporal locality for continuous vision tasks. At the beginning of model input, DeepCache discovers temporal locality by exploiting the video’s internal structure, for which it borrows proven heuristics from video compression; into the model, DeepCache propagates reusable result regions by exploiting the model’s internal structure. We have implemented a prototype of DeepCache

to run unmodified CNN models on commodity Android device, and comprehensively evaluate its effectiveness via a set of experiments on typical CNN models.

ACKNOWLEDGMENT

This work was supported by the National Key R&D Program under the grant number 2018YFB1004801, the National Natural Science Foundation of China under grant numbers 61725201, 61528201, 61529201, and a Google Faculty Award.

REFERENCES

- [1] 2015. How Google Translate squeezes deep learning onto a phone. <https://research.googleblog.com/2015/07/how-google-translate-squeezes-deep.html>.
- [2] 2016. Apple moves to third-generation Siri back-end, built on open-source Mesos platform. <https://9to5mac.com/2015/04/27/siri-backend-mesos/>.
- [3] 2016. TensorZoom App. <https://play.google.com/store/apps/details?id=uk.tensorzoom&hl=en>.
- [4] 2017. Amazon App. <https://itunes.apple.com/us/app/amazon-app-shop-scan-compare/id297606951?mt=8>.
- [5] 2017. Autopilot-TensorFlow. <https://github.com/SullyChen/Autopilot-TensorFlow>.
- [6] 2017. Caffe2 deep learning framework. <https://github.com/caffe2/caffe2>.
- [7] 2017. Concat Layer. [http://caffe.berkeleyvision.org/tutorial/layers\(concat\).html](http://caffe.berkeleyvision.org/tutorial/layers(concat).html).
- [8] 2017. ffmpeg: a video processing platform. <https://www.ffmpeg.org/>.
- [9] 2017. HTTP Caching. <https://developers.google.com/web/fundamentals/performance/optimizing-content-efficiency/http-caching>.
- [10] 2017. Local Response Normalization (LRN). <http://caffe.berkeleyvision.org/tutorial/layers/lrn.html>.
- [11] 2017. Ncnn: a high-performance neural network inference framework. <https://github.com/Tencent/ncnn>.
- [12] 2017. Nvidia driving dataset. <https://drive.google.com/file/d/0B-KJCaaF7elleG1RbzVPZWV4Tlk/view?usp=sharing>.
- [13] 2017. Peak signal-to-noise ratio. https://en.wikipedia.org/wiki/Peak_signal-to-noise_ratio.
- [14] 2017. RenderScript. <https://developer.android.com/guide/topicsrenderscript/compute.html>.
- [15] 2017. Snapdragon Profiler. <https://developer.qualcomm.com/software/snapdragon-profiler>.
- [16] 2017. Softmax Layer. <http://caffe.berkeleyvision.org/tutorial/layers/softmax.html>.
- [17] 2017. TensorFlow. <https://www.tensorflow.org/>.
- [18] 2017. TensorFlow Android Camera Demo. <https://github.com/tensorflow/tensorflow/tree/master/tensorflow/examples/android>.
- [19] 2017. The PASCAL Visual Object Classes. <http://host.robots.ox.ac.uk/pascal/VOC/>.
- [20] 2018. Android MediaCodec. <https://developer.android.com/reference/android/media/MediaCodec.html>.
- [21] Aroh Barjatya. 2004. Block matching algorithms for motion estimation. *IEEE Transactions Evolution Computation* 8, 3 (2004), 225–239.
- [22] Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, et al. 2016. End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316* (2016).

- [23] Mark Buckler, Philip Bedoukian, Suren Jayasuriya, and Adrian Sampson. 2018. *EVA²*: Exploiting Temporal Redundancy in Live Computer Vision. In *Proceedings of the 45th Annual International Symposium on Computer Architecture, ISCA’18* (2018).
- [24] Lukas Cavigelli, Philippe Degen, and Luca Benini. 2017. CBinfer: Change-Based Inference for Convolutional Neural Networks on Video Data. *arXiv preprint arXiv:1704.04313* (2017).
- [25] Guoguo Chen, Carolina Parada, and Georg Heigold. 2014. Small-footprint Keyword Spotting Using Deep Neural Networks. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP’14)*. 4087–4091.
- [26] Tianshi Chen, Zidong Du, Ninghui Sun, Jia Wang, Chengyong Wu, Yunji Chen, and Olivier Temam. 2014. DianNao: a Small-footprint High-throughput Accelerator for Ubiquitous Machine-Learning. In *Proceedings of the Architectural Support for Programming Languages and Operating Systems (ASPLOS’14)*. 269–284.
- [27] Tiffany Yu-Han Chen, Lenin Ravindranath, Shuo Deng, Paramvir Bahl, and Hari Balakrishnan. 2015. Glimpse: Continuous, Real-Time Object Recognition on Mobile Devices. In *Proceedings of the 13th ACM Conference on Embedded Networked Sensor Systems (SenSys’15)*. 155–168.
- [28] Yu-Hsin Chen, Joel S. Emer, and Vivienne Sze. 2016. Eyeriss: A Spatial Architecture for Energy-Efficient Dataflow for Convolutional Neural Networks. In *Proceedings of the 43rd ACM/IEEE Annual International Symposium on Computer Architecture, (ISCA’16)*. 367–379.
- [29] Zhuo Chen, Wenlu Hu, Junjue Wang, Siyan Zhao, Brandon Amos, Guanhong Wu, Kiryong Ha, Khalid Elgazzar, Padmanabhan Pillai, Roberta Klatzky, Daniel Siewiorek, and Mahadev Satyanarayanan. 2017. An Empirical Study of Latency in an Emerging Class of Edge Computing Applications for Wearable Cognitive Assistance. In *Proceedings of the Second ACM/IEEE Symposium on Edge Computing (SEC ’17)*.
- [30] Anupam Das, Martin Degeling, Xiaoyou Wang, Junjue Wang, Norman M. Sadeh, and Mahadev Satyanarayanan. 2017. Assisting Users in a World Full of Cameras: A Privacy-Aware Infrastructure for Computer Vision Applications. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR’17*. 1387–1396.
- [31] Emily L. Denton, Wojciech Zaremba, Joan Bruna, Yann LeCun, and Rob Fergus. 2014. Exploiting Linear Structure Within Convolutional Networks for Efficient Evaluation. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS’14)*. 1269–1277.
- [32] Peizhen Guo and Wenjun Hu. 2018. Potluck: Cross-Application Approximate Deduplication for Computation-Intensive Mobile Applications. In *Proceedings of the Twenty-Third International Conference on Architectural Support for Programming Languages and Operating Systems*. 271–284.
- [33] Song Han, Xingyu Liu, Huizi Mao, Jing Pu, Ardavan Pedram, Mark A. Horowitz, and William J. Dally. 2016. EIE: Efficient Inference Engine on Compressed Deep Neural Network. In *Proceedings of the 43rd ACM/IEEE Annual International Symposium on Computer Architecture, (ISCA’16)*. 243–254.
- [34] Seungyeop Han, Haichen Shen, Matthai Philipose, Sharad Agarwal, Alec Wolman, and Arvind Krishnamurthy. 2016. MCDNN: An Approximation-Based Execution Framework for Deep Stream Processing Under Resource Constraints. In *Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys’16)*. 123–136.
- [35] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR’16)*. 770–778.
- [36] Steve Hodges, Lyndsay Williams, Emma Berry, Shahram Izadi, James Srinivasan, Alex Butler, Gavin Smyth, Narinder Kapur, and Kenneth R. Wood. 2006. SenseCam: A Retrospective Memory Aid. In *UbiComp 2006: Ubiquitous Computing, 8th International Conference, UbiComp 2006, Orange County, CA, USA, September 17-21, 2006*. 177–193.
- [37] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. 2017. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv preprint arXiv:1704.04861* (2017).
- [38] Gao Huang, Zhuang Liu, Kilian Q Weinberger, and Laurens van der Maaten. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, Vol. 1. 3.
- [39] Forrest N. Iandola, Matthew W. Moskewicz, Khalid Ashraf, Song Han, William J. Dally, and Kurt Keutzer. 2016. SqueezeNet: AlexNet-level Accuracy with 50x Fewer Parameters and <1MB Model Size. *arXiv preprint arXiv:1602.07360* (2016).
- [40] Puneet Jain, Justin Manweiler, and Romit Roy Choudhury. 2015. Overlay: Practical Mobile Augmented Reality. In *Proceedings of the 13th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys’15)*. 331–344.
- [41] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Fei-Fei Li. 2014. Large-Scale Video Classification with Convolutional Neural Networks. In *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR’14)*. 1725–1732.
- [42] Kleomenis Katevas, Ilias Leontiadis, Martin Pielot, and Joan Serrà. 2017. Practical Processing of Mobile Sensor Data for Continual Deep Learning Predictions. In *Proceedings of the 1st International Workshop on Embedded and Mobile Deep Learning (Deep Learning for Mobile Systems and Applications) (EMDL@MobiSys’17)*. 19–24.
- [43] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Proceedings of the 26th Annual Conference on Neural Information Processing Systems (NIPS’12)*. 1106–1114.
- [44] Nicholas D. Lane, Sourav Bhattacharya, Petko Georgiev, Claudio Forlivesi, Lei Jiao, Lorena Qendro, and Fahim Kawzar. 2016. DeepX: A Software Accelerator for Low-power Deep Learning Inference on Mobile Devices. In *15th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN 2016)*. 23:1–23:12.
- [45] Nicholas D. Lane, Petko Georgiev, and Lorena Qendro. 2015. DeepEar: Robust Smartphone Audio Sensing in Unconstrained Acoustic Environments Using Deep Learning. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp’15)*. 283–294.
- [46] Didier Le Gall. 1991. MPEG: A video compression standard for multimedia applications. *Commun. ACM* 34, 4 (1991), 46–58.
- [47] Robert LiKamWa, Yunhui Hou, Yuan Gao, Mia Polansky, and Lin Zhong. 2016. RedEye: Analog ConvNet Image Sensor Architecture for Continuous Mobile Vision. In *Proceedings of the 43rd ACM/IEEE Annual International Symposium on Computer Architecture ISCA’16*. 255–266.
- [48] Robert LiKamWa, Bodhi Priyantha, Matthai Philipose, Lin Zhong, and Paramvir Bahl. 2013. Energy characterization and optimization of image sensing toward continuous mobile vision. In *International Conference on Mobile Systems, Applications, and Services (MobiSys’13)*. 69–82.
- [49] Robert LiKamWa, Bodhi Priyantha, Matthai Philipose, Lin Zhong, and Paramvir Bahl. 2013. Energy proportional image sensors for continuous mobile vision. In *International Conference on Mobile systems, Applications, and Services (MobiSys’13)*. 467–468.
- [50] Robert LiKamWa and Lin Zhong. 2015. Starfish: Efficient Concurrency Support for Computer Vision Applications. In *Proceedings of the 13th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys’15)*. 213–226.

- [51] Huynh Nguyen Loc, Youngki Lee, and Rajesh Krishna Balan. 2017. DeepMon: Mobile GPU-based Deep Learning Framework for Continuous Vision Applications. In *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys'17)*. 82–95.
- [52] Akhil Mathur, Nicholas D. Lane, Sourav Bhattacharya, Aidan Boran, Claudio Forlivesi, and Fahim Kawsar. 2017. DeepEye: Resource Efficient Local Execution of Multiple Deep Vision Models using Wearable Commodity Hardware. In *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys'17)*. 68–81.
- [53] Zhonghong Ou, Changwei Lin, Meina Song, and Haihong E. 2017. A CNN-Based Supermarket Auto-Counting System. In *Proceedings of the 17th International Conference on Algorithms and Architectures for Parallel Processing (ICA3PP'17)*. 359–371.
- [54] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. 2016. You Only Look Once: Unified, Real-Time Object Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'16)*. 779–788.
- [55] Iain E Richardson. 2004. *H. 264 and MPEG-4 video compression: video coding for next-generation multimedia*.
- [56] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Fei-Fei Li. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision* 115, 3 (2015), 211–252.
- [57] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. *arXiv preprint arXiv:1212.0402* (2012).
- [58] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going Deeper with Convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'15)*. 1–9.
- [59] Jo Yew Tham, Surendra Ranganath, Maitreya Ranganath, and Ashraf A Kassim. 1998. A novel unrestricted center-biased diamond search algorithm for block motion estimation. *IEEE transactions on Circuits and Systems for Video Technology* 8, 4 (1998), 369–377.
- [60] Ehsan Variani, Xin Lei, Erik McDermott, Ignacio Lopez-Moreno, and Javier Gonzalez-Dominguez. 2014. Deep Deural Networks for Small Footprint Text-dependent Speaker Verification. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'14)*. 4052–4056.
- [61] Tri Vu, Feng Lin, Nabil Alshurafa, and Wenyao Xu. 2017. Wearable Food Intake Monitoring Technologies: A Comprehensive Review. *Computers* 6 (2017).
- [62] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. 2016. A Discriminative Feature Learning Approach for Deep Face Recognition. In *Proceedings of the 14th European Conference on Computer Vision (ECCV'16)*. 499–515.
- [63] Jiaxiang Wu, Cong Leng, Yuhang Wang, Qinghao Hu, and Jian Cheng. 2016. Quantized Convolutional Neural Networks for Mobile Devices. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, (CVPR'16)*. 4820–4828.
- [64] Shuochao Yao, Shaohan Hu, Yiran Zhao, Aston Zhang, and Tarek F. Abdelzaher. 2017. DeepSense: A Unified Deep Learning Framework for Time-Series Mobile Sensing Data Processing. In *Proceedings of the 26th International Conference on World Wide Web, (WWW'17)*. 351–360.
- [65] Shuochao Yao, Yiran Zhao, Aston Zhang, Lu Su, and Tarek Abdelzaher. 2017. DeepIoT: Compressing deep neural network structures for sensing systems with a compressor-critic framework. In *Proceedings of the 15th ACM Conference on Embedded Network Sensor Systems. ACM*.
- [66] Xiao Zeng, Kai Cao, and Mi Zhang. 2017. *MobileDeepPill: A Small-Footprint Mobile Deep Learning System for Recognizing Unconstrained Pill Images*. In *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys'17)*. 56–67.
- [67] Chen Zhang, Peng Li, Guangyu Sun, Yijin Guan, Bingjun Xiao, and Jason Cong. 2015. Optimizing FPGA-based Accelerator Design for Deep Convolutional Neural Networks. In *Proceedings of the 2015 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays (FPGA'15)*. 161–170.
- [68] Shan Zhu and Kai-Kuang Ma. 1997. A new diamond search algorithm for fast block matching motion estimation. In *Proceedings of the International Conference on Information, Communications and Signal Processing (ICICS'97)*. 292–296.
- [69] Yanzi Zhu, Yuanshun Yao, Ben Y. Zhao, and Haitao Zheng. 2017. Object Recognition and Navigation using a Single Networking Device. In *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys'17)*. 265–277.