

# Fast Video Classification via Adaptive Cascading of Deep Models

Haichen Shen, Seungyeop Han\*, Matthai Philipose†, Arvind Krishnamurthy

University of Washington, Rubrik, Inc.\*, Microsoft Research†

{haichen, arvind}@cs.washington.edu, seungyeop.han@rubrik.com\*, matthaip@microsoft.com†

## Abstract

*Recent advances have enabled “oracle” classifiers that can classify across many classes and input distributions with high accuracy without retraining. However, these classifiers are relatively heavyweight, so that applying them to classify video is costly. We show that day-to-day video exhibits highly skewed class distributions over the short term, and that these distributions can be classified by much simpler models. We formulate the problem of detecting the short-term skews online and exploiting models based on it as a new sequential decision making problem dubbed the Online Bandit Problem, and present a new algorithm to solve it. When applied to recognizing faces in TV shows and movies, we realize end-to-end classification speedups of 2.4-7.8×/2.6-11.2× (on GPU/CPU) relative to a state-of-the-art convolutional neural network, at competitive accuracy.*

## 1. Introduction

Consider recognizing entities such as objects, people, scenes and activities in every frame of video footage of day-to-day life. Such footage may come, for instance, from the media, wearable cameras, movies, or surveillance cameras. In principle, these entities could be drawn from thousands of classes: many of us encounter hundreds to thousands of distinct people, objects, scenes and activities through our life. Over short intervals such as minutes, however, we tend to encounter a very small subset of classes of entities. For instance, a wearable camera may see the same set of objects from our desk at work for an hour, a movie may focus only on cooking-related activities through a five-minute kitchen sequence, and media footage of an event may focus on only those celebrities participating in the event. In this paper, we characterize and exploit such *short-term class skew* to significantly reduce the latency of classifying video using Convolutional Neural Networks (CNNs).

Since the seminal work of Viola and Jones [23] on face detection, one of the best-known techniques to speed

up classification has been to structure the classifier as a cascade (or tree [27]) of simple classifiers such that “easy” examples lead to early exits and are therefore classified faster. Cascaded classifiers require that training and test data are strongly (and identically) biased toward a small number of easy to detect classes. In the (binary) face detection task, for example, the class “not a face” is both (i) by far more common than “face”, and (ii) often quite easy to classify via a small number of comparisons of inexpensive Haar-style features. In fact, traditional cascades are most applicable in *detection* tasks [4], where the background is both much more common and easier to classify than the foreground.

Distinct from the (two-class) detection setting in traditional cascading, recent advances in convolutional neural networks (CNNs) [15, 21, 30] have opened up the possibility of using a single, pre-trained “oracle” classifier to *recognize* thousands of classes such as people, objects and scenes. When training such oracle classifiers, such as GoogLeNet [21] or VGGFace [15]), a small number of classes do not usually dominate the training set: for broad applicability, the classifier is trained assuming that all classes are more or less equally likely. Even if such a skew toward such small classes existed, there is no *a priori* reason that these dominant classes are fast to classify. It may seem therefore that cascading is not a promising optimization for improving the speed of entity recognition in video via CNNs.

We demonstrate, however, that for many recognition tasks, *day-to-day video often exhibits significant short-term skews* in class distribution. We present measurements on a diverse set of videos that show, for instance, that in over 90% of 1-minute windows, at least 90% of objects interacted with by humans belong to a set of 25 or fewer objects. The underlying ImageNet-based recognizer, on the other hand, can recognize up to 1000 objects. We show that similar skews hold for faces and scenes in videos.

Even if such skew exists, to our knowledge, it has not been shown that distributions skewed toward small sets of classes can be classified accurately by simpler

CNNs than uniformly distributed ones. We therefore also demonstrate that *when class distribution is highly skewed, “specialized” CNNs trained to classify inputs from this distribution can be much more compact* than the oracle classifier. For instance, we present a CNN that executes  $200\times$  fewer FLOPs than the state-of-the-art VGGFace [15] model, but has comparable accuracy when over 50% of faces come from the same 10 or fewer people. We present similar order-of-magnitude faster specialized CNNs for object and scene recognition.

Given the ability to produce fast, accurate versions of CNNs specialized for particular test-time skews, we seek to estimate the (possibly non-stationary) skew at test-time, produce a specialized model if appropriate, exploit the model as long as the skew lasts, detect when the skew disappears and then revert to the oracle model. As with standard “bandit”-style sequential decision-making problems, the challenge is in balancing exploration (i.e., using the expensive oracle to estimate the skew) with exploitation (i.e., using a model specialized to the current best available estimate of the skew). We formalize this problem as the Oracle Bandit Problem and propose a new exploration/exploitation-based algorithm we dub Windowed  $\epsilon$ -Greedy (WEG) to address it.

Using a combination of synthetic data and real-world videos, we empirically validate the WEG algorithm. In particular, we show that WEG can reduce the end-to-end classification overhead of face recognition on TV episodes and movies by  $2.4\text{--}7.8\times$  relative to unspecialized classification using the VGGFace classifier on a GPU ( $2.6\text{--}11.2\times$  on a CPU). We show via synthetic data that similar gains are to be had on object and scene recognition as well. We provide a detailed analysis of WEG’s functioning, including an accounting of how much its key features contribute. To our knowledge our system is the first to use test-time sequential class skews in video to produce faster classifiers.

## 2. Related work

There is a long line of work on cost-sensitive classification, the epitome of which is perhaps the cascaded classification work of Viola and Jones [23]. The essence of this line of work [26, 28] is to treat classification as a sequential process that may exit early if it is confident in its inference, typically by learning sequences that have low cost in expectation over training data. Recent work [14] has even proposed cascading CNNs as we do. All these techniques assume that testing data is i.i.d. (i.e., not sequential), that all training happens before any testing, and rely on skews in *training* data to capture cost structure. As such, they are not equipped to exploit short-term class skews in test data.

Traditional sequential models such as probabilis-

tic models [3, 16, 25] and Recurrent Neural Networks (RNNs) [5, 11] are aimed at classifying instances that are not independent of each other. Given labeled sequences as training data, these techniques learn more accurate classifiers than those that treat sequence elements as independent. However, to our knowledge, none of these approaches produces classifiers that yield *less expensive* classification in response to favorable inputs, as we do.

Similar to adaptive cascading, online learning methods [9, 12, 22] customize models at test time. For training, they use labeled data from a sequential stream that typically contains both labeled and unlabeled data. As with adaptive cascading, the test-time cost of incrementally training the model in these systems needs to be low. A fundamental difference in our work is that we make no assumption that our input stream is partly labeled. Instead, we assume the availability of a large, resource-hungry model that we seek to “compress” into a resource-light cascade stage.

Estimating distributions in sequential data and exploiting it is the focus of the multi-armed bandit (MAB) community [1, 13]. The Oracle Bandit Problem (OBP) we define differs from the classic MAB setting in that in MAB the set of arms over which exploration and exploitation happen are the same, whereas in OBP only the oracle “arm” allows exploration whereas specialized models allow exploitation. Capturing the connection between these arms is the heart of the OBP formulation. Our Windowed  $\epsilon$ -Greedy algorithm is strongly informed by the use of windows in [6] to handle non-stationarities and the well-known [20]  $\epsilon$ -greedy scheme to balance exploration and exploitation.

Finally, much recent work has focused on reducing the resource consumption of (convolutional) neural networks [2, 7, 8, 17]. These techniques are oblivious to test-time data skew and are complementary to specialization. We expect that even more pared-down versions of these optimized models will provide good accuracy when specialized at test-time.

## 3. Class skew in day-to-day video

Specialization depends on skew (or bias) in the temporal distribution of classes presented to the classifier. In this section, we analyze the skew in videos of day-to-day life culled from YouTube. We assembled a set of 30 videos of length 3 minutes to 20 minutes from five classes of daily activities: socializing, home repair, biking around urban areas, cooking, and home tours. We expect this kind of footage to come from a variety of sources such as movies, amateur productions of the kind that dominate YouTube and wearable videos.

We sample one in three frames uniformly from these videos and apply state-of-the-art face (derived from

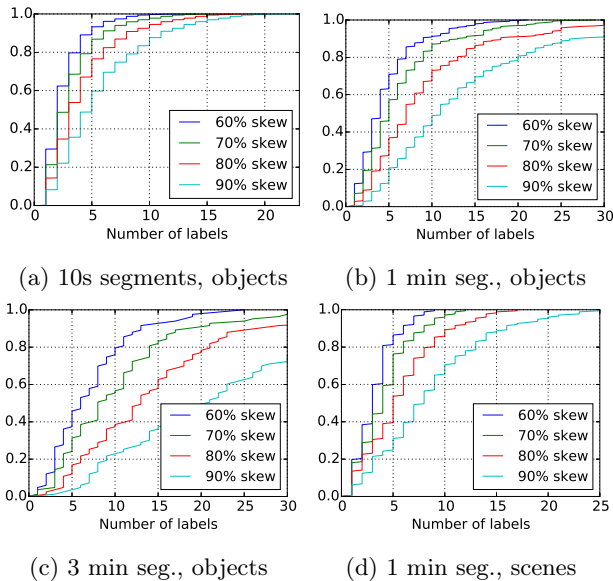


Figure 1: Temporal skew of classes in day-to-day video.

[15]), scene [30] and object recognizers [18] to every sampled frame. Note that these “oracle” recognizers can recognize up to 2622 faces, 205 scenes and 1000 objects respectively. For face recognition, we record the top-scoring label for each face detected, and for the others, we record only the top-scoring class on each frame. For object recognition in particular, this substantially undercounts objects in the scene; our count (and specialization) applies to applications that identify some distinctive subset of objects (e.g., all objects “handled” by a person). We seek to compare these numbers to the number of distinct recognized faces, scenes and objects that dominate “epochs” of  $\tau = 10$  seconds, 1 minute and 3 minutes.

Figure 1 shows the results for object recognition and scene recognition. We partition the sequence of frames into segments of length  $\tau$  and show one plot per segment length. Each line in the plot corresponds to percentage skew  $s \in \{60, 70, 80, 90\}$ . Each line in the plots shows the cumulative distribution representing the fraction of all segments where  $n$  labels comprised more than  $s$  percent of all labels in the segment. For instance, for 10-second segments (Figure 1(a)), typically roughly 100 frames, 5 objects comprised 90% of all objects in a segment 60% of the time (cyan line), whereas they comprise 60% of objects 90% of the time (dark blue).

In practice, detecting skews and training models to exploit them within 10 seconds is often challenging. As figures (b) and (c) show, the skew is less pronounced albeit still very significant for longer segments. For instance, in 90% of 3-minute segments, the top 15 objects comprise 90% of objects seen. The trend is similar with faces and scenes, with the skew significantly more pronounced, as is apparent from comparing figures (b)

Task	Model	Acc.(%)	FLOPs	CPU lat.(ms)	GPU lat.(ms)
Object (1000 classes)	[21]	68.9	3.17G	779.3	11.0
	O1	48.9	0.82G	218.2 ( $\times 3.6$ )	4.4 ( $\times 2.5$ )
	O2	47.0	0.43G	109.1 ( $\times 7.1$ )	2.8 ( $\times 3.9$ )
Scene (205)	[29]	58.1	30.9G	2570	28.8
	S1	48.9	0.55G	152.2 ( $\times 16.9$ )	3.36 ( $\times 8.6$ )
	S2	40.8	0.43G	141.5 ( $\times 18.2$ )	2.44 ( $\times 11.8$ )
Face (2622)	[15]	95.8	30.9G	2576	28.8
	F1	84.8	0.60G	90.1 ( $\times 28.6$ )	2.48 ( $\times 11.6$ )
	F2	80.9	0.13G	40.4 ( $\times 63.7$ )	1.93 ( $\times 14.9$ )

Table 1: Oracle classifiers versus compact classifiers in top-1 accuracy, number of FLOPs, and execution time. Execution time is feedforward time of a single image without batching on Caffe [10], a Linux server with a 24-core Intel Xeon E5-2620 and an NVIDIA K20c GPU.

and (d); e.g. the cyan line in (d) dominates that in (b). We expect that if we ran a hand-detector and only recognized objects in the hand (analogously to recognizing detected faces), the skew would be much sharper.

Specialized models must exploit skews such as these to deliver appreciable speedups over the oracle. Typically, they should be generated in much less than a minute, handle varying amounts of skew gracefully, and deliver substantial speedups when inputs belong to subsets of 20 classes or fewer out of a possible several hundred in the oracle.

## 4. Specializing Models

In order to exploit skews in the input, we cascade the expensive but comprehensive *oracle model* with a (hopefully much) less expensive “compact” model. This *cascaded classifier* is designed so that if its input belongs to the frequent classes in the incoming distribution it will return early with the classification result of compact model, else it will invoke the oracle model. Thus if the skew dictates that  $n$  frequent classes, or *dominant classes*, comprise percentage  $p$  of the input, or *skew*, model execution will cost the overhead of just executing compact model roughly  $p\%$  of the time, and the overhead of executing compact model and oracle sequentially the rest of the time. When  $p$  is large, the lower cost compact model will be incurred with high probability.

To be more concrete, we use state of the art convolutional neural networks (CNNs) for oracles. In particular, we use the GoogLeNet [21] as our oracle model, for object recognition; the VGG Net 16-layer version for scene recognition [30]; and the VGGFace network [15] for face recognition. The compact models are also CNNs. For these, we use architectures derived from the corresponding oracles by systematically (but manually) removing layers, decreasing kernel sizes, increasing kernel strides,

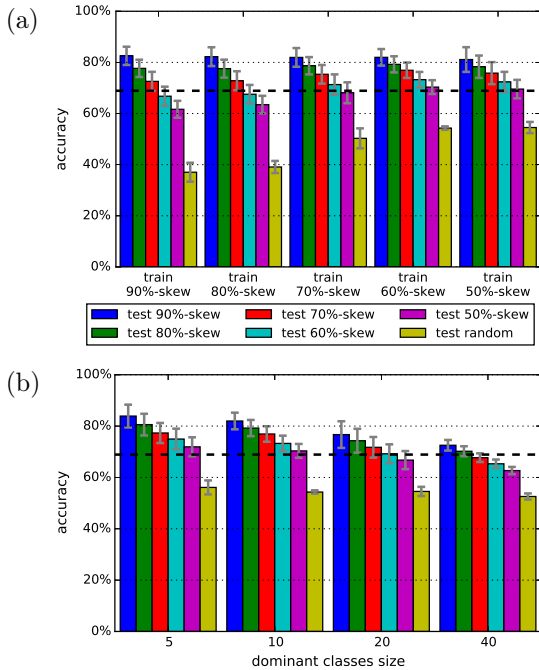


Figure 2: (a) When compact model O1 is trained by various skews and cascaded with the oracle, accuracy of the cascaded classifier tested by various skews for 10 dominant classes; (b) accuracy of O1 trained by 60% skew and tested by various skews for different number of dominant classes. The dashed line shows the accuracy of GoogLeNet as a baseline. All experiments repeated 5x with randomly selected dominant sets.

and reducing the size of fully-connected layers. The end results are architectures (O[1|2] for objects, S[1|2] for scenes and F[1|2] for faces) that use noticeably less resources (Table 1), but also yield significantly lower average accuracy when trained and validated on unskewed data, i.e., the same training and validation sets for oracle models. For instance, O1 requires roughly 4× fewer FLOPs to execute than VGGFace, but achieves roughly 70% of its accuracy.

However, in our approach, we train these compact models to classify *skewed* distributions observed during execution, denoted by *specialized classifier*, and their performance on skewed distributions is the critical measure. In particular, to generate a specialized model, we create a new training dataset with the data from the  $n$  dominant classes of the original data, and a randomly chosen subset from the remaining classes with label “other” such that the dominant classes comprise  $p$  percent of the new data set. We train the compact architecture with this new dataset.

Figure 2 shows how compact models trained on skewed data and cascaded with their oracles perform

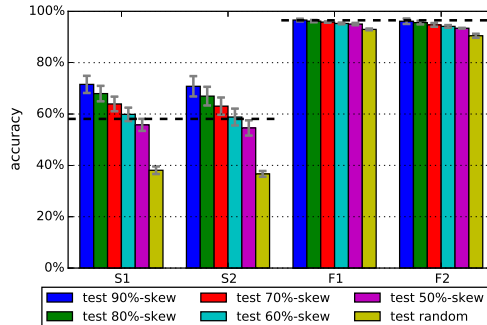


Figure 3: Accuracy of scene classifiers trained by 70% fixed skew and 10 dominant classes and face classifiers trained by 50% fixed skew and 10 dominant classes. Dashed lines show the accuracy of the oracle classifier for scene and face recognition task.

on validation data of different skews. Figure 2(a) analyzes the case where  $n = 10$ , for various combinations of training and validation skews for model O1. Recall from Table 1 that O1 delivers only 70% of its accuracy on unskewed inputs. However, when training and testing is on skewed inputs, the numbers are much more favorable. When O1 is trained on  $p=90\%$  skewed data with  $n=10$  dominant classes, it delivers over 84% accuracy on average (the left-most dark-blue bar). This is significantly *higher* than the oracle’s average of 68.9% (top-1 accuracy), denoted by the horizontal black line. We also observed from Figure 2(a) that when O1 is trained on 60% skewed data, the cascaded classifier maintains high accuracy across a wide range of testing skews from 90% to 50%. Therefore, in what follows, we use 60% skew as *fixed training skew* to specialize object compact models in the rest of paper (similarly 70% fixed skew for scene and 50% for face). Figure 2(b) shows that, where  $n$  is varied for O1, the cascaded classifier degrades very gracefully with  $n$ . Finally, Figure 3, which reports similar measurements on compact models S[1|2] and F[1|2] shows that these trends carry over to scene and face recognition.

Finally, we note that since skews are only evident at test-time, specialized models must be trained extremely fast (ideally a few seconds at most). We use two techniques to accomplish this. First, before we begin processing any inputs, we train all model architectures on the full, unskewed datasets of their oracles. At test time, when the skew  $n, p$  and the identity of dominant classes is available, we only retrain the top (fully connected and softmax) layers of the compact model. The lower layers, being “feature calculation” layers do not need to change with skew. Second, as a pre-processing step, we run all inputs in the training dataset through the lower feature-calculation layers, so

that when re-training the top layers at test time, we can avoid doing so. This combination of techniques allows us to re-train the specialized model in roughly 4s for F1 and F2 and 14s for O1/O2, many orders of magnitude faster than fully re-training these models.

## 5. Sequential Model Specialization

### 5.1. The Oracle Bandit Problem (OBP)

Let  $x_1, x_2, \dots, x_i, \dots \in X = \mathbb{R}^n$  be a stream of images to be classified. Let  $y_1, y_2, \dots, y_i, \dots \in Y = [1, \dots, k]$  be the corresponding classification results. Let  $\pi : \mathbb{I}^+ \rightarrow \mathbb{I}^+$  be a partition over the stream. Associate the distribution  $T_j$  with partition  $j$ , so that each pair  $(x_i, y_i)$  is sampled from  $T_{\pi(i)}$ . Intuitively,  $\pi$  partitions, or segments,  $\dots, x_i, \dots$  into a sequence of “epochs”, where elements from each epoch  $j$  are drawn independently from the corresponding stationary distribution  $T_j$ . Thus, for the overall series, samples are drawn from an abruptly-changing, piece-wise stationary distribution. At test time, neither results  $y_i$  nor partitions  $\pi$  are known.

Let  $h^* : X \rightarrow Y$  be a classifier, designated the “oracle” classifier, trained on distribution  $T^*$ . Intuitively  $T^*$  is a mixture of all distributions comprising the oracle’s input stream:  $T^* = \sum_j T_j$ . Let  $R^*$  be the cost (e.g., number cycles), assumed invariant across  $X$ , needed to execute  $h^*$  on any  $x \in X$ . At test time, on each input  $x_i$ , we can always consult  $h^*$  at cost  $R^*$  to get a label  $y_i$  with some (high) accuracy  $a^*$ .

Let  $m_1, \dots, m_M$  be *model architectures*, such as those of O1, O2, S1, S2, F1 and F2 in [Table 1](#). Suppose each architecture  $m_k$  is trained *offline* on  $T^*$  to obtain a “template” classifier  $h_k$ . We assume that re-targeting template  $h_k$  to a new size- $j$  set of dominant classes has a flat cost  $R_0$ .

Finally, for each set of dominant classes  $D$ , the corresponding *specialized classifier*  $h_D$  is trained by re-targeting some template  $h_k$ , using a dataset that draws half its examples from classes in  $D$  and the rest (with a single label “other”) from  $Y - D$ . Let the cost of executing  $h_d$  be  $R_{h_d}$ . Chaining  $h_D$  with  $h^*$  gives a *cascaded classifier*  $\hat{h}_D(x) \triangleq$  if  $y = h_D(x) \in D$ , return  $y$ , otherwise return  $h^*(x)$ . Note that executing  $\hat{h}_D$  will either cost  $R_{h_d}$  (in the case that the condition is true), or  $R_{h_d} + R^*$  in the case that it is false. Given that  $R_{h_d} \ll R^*$ , developing and using specialized classifiers  $h_D$  can thus reduce costs significantly. Since  $x_i$  is drawn from some distribution  $T$ , each classifier  $\hat{h}_D$  also has *cost* that belongs to a corresponding distribution, which we write as  $R_{T\hat{h}_D}(x_i)$ .

Now consider a policy (or algorithm)  $P$  that, for each incoming image  $x_i$  belonging to stationary distribution  $T_{\pi(i)}$  as above, selects a classifier  $\hat{h}_D^{(i)}$  (for some set choice

---

### Algorithm 1 Windowed $\epsilon$ -Greedy (WEG)

---

```

1:  $j, S_0 \leftarrow 1, []$ 
    $\triangleright$  Note:  $\tau_r, \tau_a, \tau_{FP}$  and  $\epsilon$  below are hyper-parameters.
2: Window Initialization Phase
3: Repeat  $w_{min}$  times
4:    $y_t \leftarrow h^*(x_t)$ 
5:    $S_j \leftarrow S_j \oplus [y_t]$   $\triangleright$  Append new sample
6:   if  $||\text{DOMCLASSES}(S_{j-1}), \text{DOMCLASSES}(S_j)|| \leq \tau_r$  then
7:      $\triangleright$  dominant classes match sufficiently, old epoch continues
8:      $S_j \leftarrow S_{j-1} \oplus S_j$ 
9:    $w \leftarrow |S_j|$  and go to Line 8

10: Template Selection Phase
11:  $D \leftarrow \text{DOMCLASSES}(\text{last } w \text{ elements in } S_j)$ 
12: Estimate acc.  $a_{\hat{h}_D}$  of  $\hat{h}_D$ ; use  $p^*$  derived from  $S_j$  (Equation 1)
13: if  $a_{\hat{h}_D} \geq a^* + \tau_a$  then
14:   train specialized classifier  $h_D$  on dominant classes  $D$ 
15:   go to Line 16  $\triangleright$  Exploit cascaded classifier  $\hat{h}_D$ 
16:  $y_t \leftarrow h^*(x_t)$   $\triangleright$  Else, continue exploring with oracle
17:  $S_j \leftarrow S_j \oplus [y_t]$ 
18: go to Line 8

19: Specialized Classification Phase
20:  $n_c, n^*, S \leftarrow 0, 0, S_j$ 
21:  $y_t, c \leftarrow \hat{h}_D(x_t)$   $\triangleright$  exploit;  $c = 0|1$  iff-if-not cascaded to oracle
22:  $n^* \leftarrow (c \text{ or } \text{rand}() \geq \epsilon) ? n^* : n^* + (h^*(x_t) \neq y_t)$ 
23:  $n_c \leftarrow n_c + c$   $\triangleright$  Increment if  $\hat{h}_D$  did not use oracle
24: Estimate acc.  $a_{\hat{h}_D}$  of  $\hat{h}_D$ ; use  $p^*$  derived from  $S_j$  (Equation 1)
25: if  $a_{\hat{h}_D} < a^* + \tau_a$  or  $\frac{n^*}{n_c - \epsilon} > \tau_{FP}$  then
26:    $j \leftarrow j + 1$   $\triangleright$  Exit specialized classification
27:   go to Line 2  $\triangleright$  Potentially start new epoch  $j$ 
28: else
29:   go to Line 2  $\triangleright$  Go back to check if distribution has changed
30:  $S \leftarrow S \oplus [y_t]$ ; go to Line 17

```

---

of  $D$ ), and applies it to  $x_i$ . The classifier selected could also include the oracle. The expected total cost of this policy,  $R_p = |\cup_i \{\hat{h}_D^{(i)}\}|R_0 + \sum_i E_{x_i \sim T_{\pi(i)}}(R_{T_{\pi(i)}\hat{h}_D^{(i)}}(x_i))$ . We seek a minimal-cost policy:  $P^* = \arg \min_P R_P$  that maintains average accuracy within a threshold  $\tau_a$  of oracle accuracy  $a^*$ .

### 5.2. The Windowed $\epsilon$ -Greedy (WEG) Algorithm

A close look at the policy cost above provides some useful intuition on what good policies should do. First, given the high fixed cost  $R_0$  of re-targeting models as opposed to just running them, re-targeting should be infrequent. We expect re-targeting to occur roughly once an epoch. Second, the cost of running the cascade is much lower than that of running the oracle *if the input  $x_i$  is in the dominant class set  $D$*  and higher otherwise. It is important therefore to identify promptly when a dominant set  $D$  exists, produce a specialized model  $h_D$  that does not lose too much accuracy, and revert back to the oracle model when the underlying distribution changes and  $D$  is no longer dominant. We provide a heuristic exploration-exploitation based algorithm ([Algorithm 1](#)) based on these intuitions.

The algorithm runs in three phases.

1. **Window Initialization** [lines 2 - 7] identifies the

dominant classes of the current epoch. To do so, we run the oracle on a fixed number  $w_{min}$  ( $= 30$  in our implementation) of examples. The `DOMCLASSES` helper identifies the dominant classes in the window as those that appear at least twice in the window. If the dominant classes are each within  $\tau_r$  ( $= 2$ ) of those of the previous epoch, we conclude the previous epoch is continuing and fold information collected on it into that for the current epoch  $S_j$ .

2. **Template Selection** [lines 8 - 15] Given a candidate set of dominant classes  $D$ , we estimate (Equation 1 below details precisely how) the accuracy of the cascaded classifier  $\hat{h}_D$  for various template classifiers  $h_i$  when specialized to  $D$  and their current empirical probability skew  $p^*$  derived from measured data  $S_j$ . Estimating these costs instead of explicitly training the corresponding specialized classifiers  $h_D$  is significantly cheaper. If the estimate is within a threshold  $\tau_a$  ( $= 0.05$  for object and scene recognition, and  $-0.05$  for face recognition since the accuracy of oracle is higher) of the oracle, we produce the specialized model and go to the specialized classification phase. If not, we continue running the oracle on inputs and collecting more information on the incoming class skew.

3. **Specialized Classification** [lines 16 - 17] The specialized classification phase simply applies the current cascaded model  $\hat{h}_D$  to inputs (Line 17) until it determines that the distribution it was trained on (as represented by  $D$ ) does not adequately match the actual current distribution. This determination is non-trivial because in the specialization phase, we wish to avoid consulting the oracle in order to reduce costs. However, the oracle is (assumed to be) the only unbiased source of samples from the actual current distribution.

We therefore run the oracle in addition to the cascaded model with probability  $\epsilon$  ( $= 0.01$ ), as per the standard  $\epsilon$ -greedy policy for multi-arm bandits. Given the resulting empirical estimate  $p^*$  of skew, we can again estimate the accuracy of the current cascade  $\hat{h}_D$  as per Equation 1. If the estimated accuracy of the cascade is too low, or if the classification results of  $\hat{h}_D$  cascade are different from the oracle too often (we use a threshold  $\tau_{FP} = 0.5$ ), we assume that the underlying distribution may have shifted and return to the Window Initialization phase.

Finally, we focus on estimating the expected accuracy of the cascaded classifier given the current skew  $p$  of its inputs (i.e., the fraction of its inputs that belong to the

dominant class set). The accuracy of cascaded classifier  $\hat{h}_D$  can be estimated by:

$$a_{\hat{h}_D} = p \cdot a_{in} + p \cdot e_{in \rightarrow out} \cdot a^* + (1 - p) \cdot a_{out} \cdot a^* \quad (1)$$

where  $a_{in}$  is the accuracy of specialized classifier  $h_D$  on  $n$  dominant classes,  $e_{in \rightarrow out}$  is the fraction of dominant inputs that  $h_D$  classifies as non-dominant ones, and  $a_{out}$  is the fraction of non-dominant inputs that  $h_D$  classifies as non-dominant (note that these inputs will be cascaded to the oracle). We have observed previously (Section 4) that these parameters  $a_{in}$ ,  $e_{in \rightarrow out}$ ,  $a_{out}$  of specialized classifier  $h_D$  are mainly affected only by the size of the dominant class  $D$ , not the identity of elements in it. Thus, we pre-compute these parameters for a fixed set of values of  $n$  (averaging over 10 samples of  $D$  for each  $n$ ), and use linear interpolation for other  $n$ s at test time.

## 6. Evaluation

We implemented the WEG algorithm with a classification runtime based on Caffe [10]. The system can be fed with videos to produce classification results by recognizing frames. Our goal was to measure both how well the large specialized model speedups of Table 1 translated to speedups in diverse settings and on long, real videos. Further we wished to characterize the extent to which elements of our design contributed to these speedups.

### 6.1. Synthetic experiments

First, we evaluate our system with synthetically generated data in order to study diverse settings. For this experiment, we generate a time-series of images picked from standard large validation sets of CNNs we use. Each test set comprises of one or two segments where a segment is defined by the number of dominant classes, the skew, and the duration in minutes. For each segment, we assume that images appear at a fixed interval (1/6 seconds) and that each image is picked from the testing set based on the skew of the segment. For an example of a segment with 5 dominant classes and 90% skew, we pre-select 5 classes as dominant classes and pick an image with 90% probability from the dominant classes and an image with 10% probability from the other classes at each time of image arrival over 5 minutes duration. Images in a class are picked in a uniform random way. We also generate traces with two consecutive segments with different configurations to study the effect of moving from one context to the other.

Table 2 shows the average top-1 accuracies and per-image processing latencies using GPU for the recognition tasks with and without the specializer enabled. The results are averaged over 5 iterations for each experiment. The specializer was configured to use the

Segments	Object				Scene				Segments	Face			
	disabled		enabled		disabled		enabled			disabled		enabled	
	acc(%)	lat(ms)	acc(%)	lat(ms)	acc(%)	lat(ms)	acc(%)	lat(ms)	acc(%)	lat(ms)	acc(%)	lat(ms)	
(n=5,p=.8)	69.5	11.6	77.0	6.0	57.6	28.9	65.2	12.0	(n=5,p=.8)	95.2	28.7	95.1	9.2
(n=10,p=.8)	66.7	11.7	72.5	7.4	57.2	28.9	57.8	18.6	(n=5,p=.9)	97.0	28.6	96.2	6.7
(n=10,p=.9)	71.8	11.6	78.0	5.9	59.1	28.8	63.5	15.4	(n=10,p=.9)	95.4	28.5	94.3	11.0
(n=15,p=.9)	68.7	11.6	68.9	9.1	57.8	28.8	57.2	22.6					
(random)	68.1	12.1	68.1	11.5	59.1	28.9	59.1	28.8	(random)	95.9	28.5	95.9	28.5
(n=10,p=.9) +(random)	67.9	11.8	70.2	9.1	57.0	28.8	56.0	22.6	(n=5,p=.9) +(random)	96.2	28.5	96.2	17.6
(n=15,p=.9) +(n=5,p=.8)	70.6	11.6	73.9	7.8	61.1	28.7	63.0	17.1	(n=10,p=.9) +(n=10,p=.8)	95.8	28.5	95.2	10.4

Table 2: Average accuracy and GPU latency of recognition over segments. For the segment column, each parenthesis indicates a segment of 5 minutes with the number of dominant classes and the skew.

video	length (min)	oracle			WEG			special rate(%)	cascade rate(%)	trans. special	dom. size	window size
		acc(%)	CPU lat	GPU lat	acc(%)	CPU lat	GPU lat					
Friends	24	93.2	2576	28.97	93.5	538( $\times 4.8$ )	7.0( $\times 4.1$ )	88.0	7.5	51	2.8	41.8
Good Will Hunting	14	97.6	2576	28.84	95.1	231( $\times 11.2$ )	3.7( $\times 7.8$ )	95.9	3.4	4	3.5	37.5
Ellen Show	11	98.6	2576	29.26	94.6	325( $\times 7.9$ )	4.7( $\times 6.2$ )	93.7	4.8	19	1.7	47.4
The Departed	9	93.9	2576	29.18	93.5	508( $\times 5.1$ )	6.9( $\times 4.2$ )	92.0	10.3	9	2.4	40.0
Ocean's Eleven / Twelve	6	97.9	2576	28.97	96.0	1009( $\times 2.6$ )	12.3( $\times 2.4$ )	80.1	18.0	23	2.0	52.2

Table 3: Accuracy and average processing latency per frame on videos with oracle vs. WEG (latencies are shown in ms). For additional insight, the last 5 columns show key statistics from WEG usage.

compact classifiers O2 for objects, S2 for scenes, and F2 for face recognition from Table 1.

The following points are worth noting. (i) (Row 1 and it’s sub-rows) *WEG is able to detect and exploit skews over 5-minute intervals and get significant speedups over the oracle while preserving accuracy.* For the single segment cases, the GPU latency speedup per-image was  $1.3\times$  to  $2.0\times$ ,  $1.3\times$  to  $2.4\times$ , and  $2.6\times$  to  $4.3\times$ , for object, scene, and face, respectively. However, due to WEG’s overhead these numbers are noticeably lower than the raw speedups of specialized models (Table 1). When the number of dominant classes increase, the specializer latency increases because it alternates between exploration and exploitation to recognizes more dominant classes. The latency also increases when the skew of dominant classes decreases because specializer cascades more times to oracle model when using the cascaded classifier. (ii) (Row 2) *WEG is quite stable in handling random inputs*, essentially resorting to the oracle so that accuracy and latency are unchanged. (iii) (Rows 3 and 4) *WEG is able to detect abrupt input distribution changes* as the accuracy remains comparable to oracle accuracy, but with significant speedups when the distribution is skewed (Row 4).

To further understand the limit of the WEG algorithm, we studied how frequently class distributions can change before our technique stops showing benefit. We evaluated face recognition with synthetic traces, changing the distribution every 10/20/30 sec. The WEG algorithm then yields speedups of 0.95/1.19/1.48 $\times$  with

roughly unchanged accuracy. For face recognition therefore, WEG stops gaining benefit for distributions that lasts less than 20 secs.

## 6.2. Video experiments

We now turn to evaluating WEG on real videos. However, we were unable to find a suitable existing dataset to show off specialization. We need several minutes or more of video that contains small subsets of (the oracle model’s) classes that may change over time. In videos of real-world activity, this happens naturally; in popular benchmarks, not so much. For example, the videos in YouTube Faces [24] are short (average 6 sec, max 200 sec) and typically only contain one person. Similarly, clips in UCF-101 [19] have mean length of 7.2 sec (max 71 sec) and focus on classifying *actions* for which no oracle model exists. Finally, the sports 1M dataset [11] assigns labels *per video* instead of *per frame*.

As a consequence, we hand-labeled video clips from three movies, one TV show, and an interview and manually labeled the faces in the videos<sup>1</sup>. The names of video clips with lengths are listed in Table 3. Note that we used the entire videos for Friends and Ellen Show, while we used a video chunk for the movies. For these experiments, we used F2 as the compact classifier.

Table 3 shows the average accuracies and average latencies for processing a frame for 5 videos. We generated these by first extracting all faces from the videos

<sup>1</sup>The dataset is released at <http://syslab.cs.washington.edu/research/neurosys>.

to disk using the Viola Jones detector. We then ran WEG on these faces and measured the total execution time. Dividing by the number of faces gave the average numbers shown here. The most important point is that *even on real-world videos, WEG is able to achieve very significant speedups over the oracle*, ranging from  $2.6\times$ - $11.2\times$  (CPU) and  $2.4\times$ - $7.8\times$  (GPU).

To understand the speedup, we summarize the statistics of WEG execution in Table 3. “Special rate” indicates the percentage of time that specializer exploits cascaded classifier to reduce the latency, while “cascade rate” reveals the percentage of time that a cascaded classifier cascades to the oracle classifier, thus hurting performance. Higher special rate and lower cascade rate yield more speedup. The cascade rate of “Ocean’s Eleven” is significantly higher than that of other videos. We investigated this and found that the specialized compact CNN repeatedly made mistakes on one person in the video, which led to a high cascade rate. “Trans. special” counts the number of times WEG needed to switch between specialized and unspecialized classification to handle the distribution changes and insufficient exploration. The average dominant classes sizes (“dom. size”) show that the real videos are skewed to fewer dominant classes than the configurations used in the synthetic experiments. This explains why our system achieved higher speedup on real videos than on synthetic data. Overall, the statistics show that *the dataset exercise WEG features such as skew estimation, cascading and specialization*.

To understand better the utility of WEG’s features, we performed an ablation study: (a) We disable the adaptive window exploration (Line 5-7 in Algorithm 1), and use a fixed window size of 30 and 60. (b) We use the skew of dominant classes in the input distribution as the training skew for specializing compact CNNs instead of using the fixed (50%) training skew suggested in Section 4. (c) We apply a simple (but natural) criterion to exit from the specialized classification phase: WEG now exits when the current skew is lower than the skew when it entered into specialized classification phase instead of using the estimated accuracy as soft indicator.

Figure 4 shows the comparison between these variants and WEG algorithm in accuracy and CPU / GPU speedups when recognizing faces on Friends video. In the figure we show the absolute differences in accuracy and relative differences in CPU / GPU speedup. (a) Fixed window size (30 and 60) variants achieve similar accuracy but lower speedup. As table 3 (“window size” column) shows, the adaptively estimated size for the window is between 30 and 60. In general, too small a window fails to capture the full dominant classes, yielding specializers that exit prematurely. Too large a

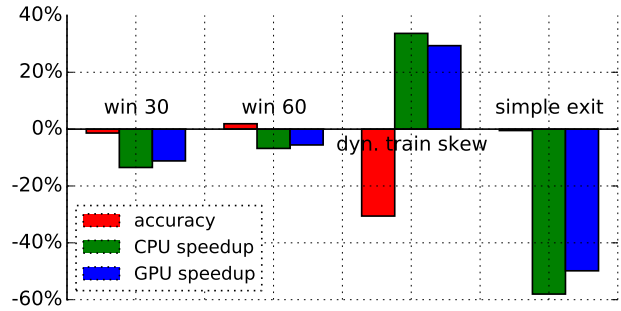


Figure 4: Change in accuracy (absolute difference) and speedup (relative) when individual features are disabled.

window requires more work by the oracle to fill up the window. (b) Using variable rather than fixed skew for training achieves more speedup, but suffers from 30% loss in accuracy. This is because the training skew is usually very high. As discussed in Section 4, training on highly skewed data produces models vulnerable to false positives in “other” classes. (c) The simple exit variant achieves almost comparable accuracy while the latency is more than 50% higher than our system. It demonstrates the value of our accuracy estimate in modeling the accuracy of cascaded classifiers and to prevent premature exit from the specialized classification phase. *In summary, the key design elements of WEG each have a role in producing fast and accurate results.*

## 7. Conclusion

We characterize class skew in day-to-day video and show that the distribution of classes is often strongly skewed toward a small number of classes that may vary over the life of the video. We further show that skewed distributions are well classified by much simpler (and faster) convolutional neural networks than the large “oracle” models necessary for classifying uniform distributions over many classes. This suggests the possibility of detecting skews at runtime and exploiting them using dynamically trained models. We formulate this sequential model selection problem as the Oracle Bandit Problem and provide a heuristic exploration/exploitation based algorithm, Windowed  $\epsilon$ -Greedy (WEG). Our solution speeds up face recognition on TV episodes and movies by  $2.4$ - $7.8\times$  on a GPU ( $2.6$ - $11.2\times$  on a CPU) with little loss in accuracy relative to a modern convolutional neural network.

## Acknowledgements

We thank anonymous reviewers for their helpful comments. This work was funded partially by Google, Huawei, and NSF (grant: CNS-1614717).



## References

- [1] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Mach. Learn.*, 47(2-3), May 2002. 2
- [2] J. Ba and R. Caruana. Do deep nets really need to be deep? In *Advances in neural information processing systems*, pages 2654–2662, 2014. 2
- [3] C. Desai and D. Ramanan. Detecting actions, poses, and objects with relational phraselets. In *Proceedings of the 12th European Conference on Computer Vision - Volume Part IV, ECCV’12*, 2012. 2
- [4] P. Dollár, R. Appel, S. Belongie, and P. Perona. Fast feature pyramids for object detection. *PAMI*, 2014. 1
- [5] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, 2015. 2
- [6] A. Garivier and E. Moulines. On upper-confidence bound policies for non-stationary bandit problems. In *Proceedings of the 22nd International Conference on Algorithmic Learning Theory (ALT)*, 2011. 2
- [7] S. Han, H. Mao, and W. J. Dally. Deep compression: Compressing deep neural network with pruning, trained quantization and Huffman coding. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2016. 2
- [8] S. Han, J. Pool, J. Tran, and W. Dally. Learning both weights and connections for efficient neural network. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Proceedings of the Twenty-ninth Annual Conference on Neural Information Processing Systems (NIPS)*, pages 1135–1143. 2015. 2
- [9] S. Hare, A. Saffari, and P. H. S. Torr. Struck: Structured output tracking with kernels. In *2011 International Conference on Computer Vision*, 2011. 2
- [10] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia (MM)*, 2014. 3, 6
- [11] A. Karpathy et al. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014. 2, 7
- [12] B. Kveton and M. Valko. Learning from a single labeled face and a stream of unlabeled data. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, 2013. 2
- [13] T. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Adv. Appl. Math.*, 6(1), Mar. 1985. 2
- [14] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua. A convolutional neural network cascade for face detection. In *CVPR*, 2015. 2
- [15] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *BMVC*, 2015. 1, 2, 3
- [16] H. Pirsiavash and D. Ramanan. Detecting activities of daily living in first-person camera views. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 2012. 2
- [17] M. Rastegari, V. Ordonez, J. Redmon, and A. Farhadi. XNOR-Net: ImageNet Classification Using Binary Convolutional Neural Networks. *arXiv:1603.05279*, Mar. 2016. 2
- [18] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015. 3
- [19] K. Soomro et al. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv:1212.0402*, 2012. 7
- [20] R. Sutton and A. Barto. *Reinforcement Learning, an introduction*. MIT Press/Bradford Books, 1998. 2
- [21] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. *CoRR*, abs/1409.4842, 2014. 1, 3
- [22] M. Valko, B. Kveton, L. Huang, and D. Ting. Online semi-supervised learning on quantized graphs. In *Proceedings of the Twenty-Sixth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-10)*, pages 606–614, Corvallis, Oregon, 2010. AUAI Press. 2
- [23] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2001. 1, 2
- [24] L. Wolf et al. Face recognition in unconstrained videos with matched background similarity. *IEEE*, 2011. 7
- [25] J. Wu, A. Osuntogun, T. Choudhury, M. Philipose, and J. M. Rehg. A scalable approach to activity recognition based on object use. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8, 2007. 2
- [26] Z. Xu, M. Kusner, M. Chen, and K. Q. Weinberger. Cost-sensitive tree of classifiers. In *ICML*, 2013. 2
- [27] Z. E. Xu, M. J. Kusner, K. Q. Weinberger, M. Chen, and O. Chapelle. Classifier cascades and trees for minimizing feature evaluation cost. *Journal of Machine Learning Research*, 15:2113–2144, 2014. 1
- [28] Q. Yang, C. X. Ling, X. Chai, and R. Pan. Test-cost sensitive classification on data with missing values. *IEEE Trans. Knowl. Data Eng.*, 18(5):626–638, 2006. 2
- [29] B. Zhou, A. Khosla, A. Lapedriza, A. Torralba, and A. Oliva. Places: An image database for deep scene understanding. *arXiv preprint arXiv:1610.02055*, 2016. 3
- [30] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *Proceedings of the Twenty-eighth Annual Conference on Neural Information Processing Systems (NIPS)*, 2014. 1, 3