Few-Shot Adaptation for Multimedia Semantic Indexing

Nakamasa Inoue
Tokyo Institute of Technology
inoue@ks.cs.titech.ac.jp

Koichi Shinoda
Tokyo Institute of Technology
shinoda@c.titech.ac.jp

ABSTRACT

We propose a few-shot adaptation framework, which bridges zero-shot learning and supervised many-shot learning, for semantic indexing of image and video data. Few-shot adaptation provides robust parameter estimation with few training examples, by optimizing the parameters of zero-shot learning and supervised many-shot learning simultaneously. In this method, first we build a zero-shot detector, and then update it by using the few examples. Our experiments show the effectiveness of the proposed framework on three datasets: TRECVID Semantic Indexing 2010, 2014, and ImageNET. On the ImageNET dataset, we show that our method outperforms recent few-shot learning methods. On the TRECVID 2014 dataset, we achieve 15.19 % and 35.98 % in Mean Average Precision under the zero-shot condition and the supervised condition, respectively. To the best of our knowledge, these are the best results on this dataset.

KEYWORDS

Semantic Indexing, Word Vectors, Zero-Shot Learning, Few-Shot Learning, Many-Shot Learning

1 INTRODUCTION

With advances in information technologies, the amount of multimedia data such as video, image, audio, and text data has been increasing rapidly. Detecting semantic concepts is known to be a fundamental technology to improve the performance of many multimedia applications including search [1, 2], summarization [3, 4], and surveillance [5, 6]. Here, semantic concepts are objects, actions, and scenes.

How to bridge the semantic gap [7], the lack of correspondence between low-level features and high-level semantic concepts? This is the most important problem to be solved in semantic concept detection. Previous studies have proved that supervised learning with many examples, i.e., supervised many-shot learning [1], is a straightforward way to find a mapping from low-level features to high-level semantics. For example, support vector machines (SVMs) [8, 9] and deep neural networks [10–12] have been shown to be effective in video semantic indexing [2, 13] and object recognition [10–12, 14]. These methods require large-scale training data in which positive and negative labels of semantic concepts for each image/video are given. However, the cost of collecting training data increases as the number of target semantic concepts increases, since manual annotation is needed.

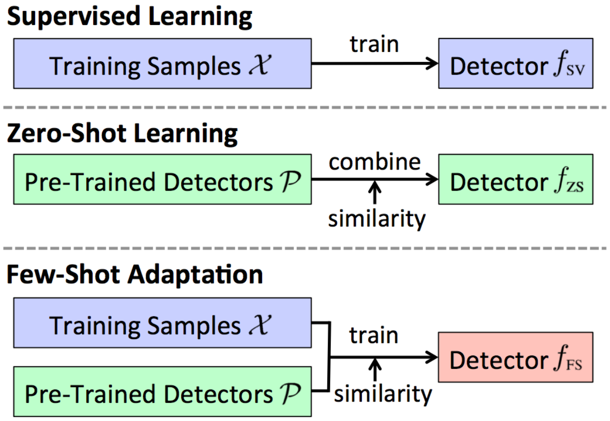
Supervised Learning
Training Samples $\mathcal{X}$
train
Detector f_{SV}
Zero-Shot Learning
Pre-Trained Detectors $\mathcal{P}$
combine
similarity
Detector f_{ZS}
Few-Shot Adaptation
Training Samples $\mathcal{X}$
Pre-Trained Detectors $\mathcal{P}$
train
similarity
Detector f_{FS}

Figure 1: Few-Shot Adaptation Framework. Few-shot adaptation combines supervised many-shot learning and zero-shot learning. To train a detector, few-shot adaptation accepts two inputs: a set of training samples $\mathcal{X}$ and a set of pre-trained detectors $\mathcal{P}$ from supervised and zero-shot learning frameworks, respectively.

To reduce such costs, some researchers are focusing on techniques to train statistical models with few training samples. For example, few-shot learning in [15] has introduced artificially generated features, namely hallucinating features, which can be used as additional samples for network training with few images. Matching networks [16] for one-shot learning has provided a framework to adapt an image embedding network to a given training image with attention mechanism. Domain adaptation [17, 18] and Bayesian estimation [19, 20] are also known to be effective. However, these methods rely on the assumption that given few training samples are high-quality, i.e., an object is often at the center of an image without occlusion or noise. Thus, they are not always effective for concept detection from images/videos in the wild with low-quality training samples.

Another effective approach, which does not rely on the high quality of images, is to utilize semantic relation among concepts obtained from large-scale text data. Recent zero-shot learning studies [21–26] have shown that combining pre-trained detectors based on semantic relation is effective. For example, convex combination of pre-trained detectors for 1,000 objects has been proposed to make effective detectors for other unseen objects [21, 22] or actions [24, 27–29]. In these methods, word vectors, which represent a word by a real-valued vector, e.g., *word2vec* [30, 31], are often introduced to measure similarity among objects and/or actions, and are used to determine weights for the convex combination. Some recent studies have focused on applications of zero-shot learning to the other learning frameworks. For example, prior knowledge

[1] hereinafter it is simply referred to as supervised learning

from zero-shot learning is introduced to active learning in [32]. In few-shot learning, we believe that techniques for zero-shot learning and supervised learning benefit from each other, because their inputs are different and complementary.

In this paper, we propose a few-shot adaptation framework, which bridges supervised learning and zero-shot learning for image and video semantic indexing. It optimizes the parameters of supervised learning and zero-shot learning, simultaneously, under an assumption that a set of training samples and a set of pre-trained detectors are given (Figure 1). In our experiments, the proposed framework is evaluated on three datasets: TRECVID Semantic Indexing 2010, 2014, and ImageNet. We achieve 15.19 % and 35.98 % in Mean Average Precision under the zero-shot condition and the official supervised condition, respectively. To the best of our knowledge, these are the best results on the TRECVID 2014 dataset.

The rest of this paper is organized as follows. Section 2 summarizes related studies. Section 3 defines the notations for supervised learning and zeros-shot learning for preparation. Section 4 presents the proposed few-shot adaptation framework. Section 5 reports the results of experimental evaluations, and Section 6 describes conclusion and future work.

2 RELATED WORK

2.1 Supervised Learning and Adaptation

Supervised learning is a straightforward way to obtain detectors of semantic concepts from training samples. It requires positive and negative samples for training. Recent studies have shown the effectiveness of deep learning using convolutional neural networks (CNNs) on large-scale datasets. AlexNet [10] with 8 layers is their typical example. It accepts raw image data as input to train object classifiers at the final softmax layer. GoogLeNet [12], VGGNet [11], ResNet [33], and DenseNet [34] are its extension to deeper networks. These networks are often trained on a large-scale image dataset such as the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) dataset [35] and the Places 365 dataset [36].

Some recent studies have focused on techniques to train statistical models with a small number of training samples. Examples include one-/few-shot learning [15, 16, 37, 38], model transformation [39], domain adaptation [17, 18], and Bayesian estimation [19, 20].

Hariharan *et al.* [15] proposed a few-shot learning method, which uses artificially generated features, called hallucinating features, for additional training data. They introduced a network to predict and generate features based on analogies on the ImageNet dataset. For example, from a given image of a bird with a sky background, it predicts features of bird images with other backgrounds such as forest, by using images of other objects with these backgrounds in the ImageNet dataset. Matching networks in [16] tackled a one-shot learning problem by introducing attention mechanism to adapt an image embedding network to given one example.

For generative models such as Gaussian Mixture Models (GMM), Bayesian estimation [19, 20, 40] is known to be effective. Perronnin *el al.* [20] proposed vocaburary adaptation using GMMs. It is extended to image representation called Fisher vectors [41]. Mensink *el al.* [19] applied maximum a posterior estimation to metric learning for image classification. They showed the effectiveness of distance-based classifiers in image recognition with a small set of training samples. For recent network-based discriminative models such as the above CNNs, fine-tuning is the most promising approach. It is effective for many tasks including event detection [43, 44] and action recognition [45, 46].

For video semantic indexing, updating or replacing only the final classification layer is one of the best ways in fine-tuning. For example, the final softmax layer is replaced by support vector machines (SVMs) to detect semantic concepts from video data in [42, 43]. This performed the best at the semantic indexing competition in the TRECVID workshop [42, 47, 48, 67].

However, with few training examples, these methods are often sensitive to the selection of training examples. Reducing the sensitivity is known to be one of the challenging tasks in training using few examples.

2.2 Zero-Shot Learning

Zero-shot learning [21, 25, 49], for the case when a set of semantic concepts for training and that for testing are disjoint, has been receiving attention in recent years. Compared with supervised learning, the performance of zero-shot learning is rather low, because image/video samples are not given for training. A recent trend is to build detectors by a weighted combination of pre-trained detectors, in which weights are determined based on similarity between concepts [24, 25, 27, 28, 50]. Here, how to use the relation among concepts to measure the similarity is a key factor to improve the performance.

For object recognition, the object attributes are useful to measure similarity between objects. For example in animal recognition, attributes such as color, texture, and shape are known to be useful for zero-shot learning [49, 51, 52]. The more detailed the attributes are, the more precise similarity between concepts will be. However, it is often given to a small set of objects, and is difficult to manually prepare the attributes for a large number of objects.

Another approach to measure the similarity between concepts is to utilize word vectors, which represent a word by a real-valued vector, obtained from large-scale text data. For example, word vectors extracted by the skip-gram model [31] are introduced to zero-shot object recognition [50]. Here, the skip-gram model is a neural network to extract vector representation of words. It is often trained on a large-scale text corpus such as Wikipedia [31]. Since words in a text corpus include not only nouns but also verbs, these methods have wide-range applications related to actions [24, 27, 28], events [53–56], and semantic concepts [57]. Some recent studies focus on improving word embedding methods with joint learning [58–61].

Compared with supervised learning, zero-shot learning effectively introduces knowledge from text data. This is the reason why we believe that supervised learning and zero-shot learning benefit from each other.

3 SUPERVISED LEARNING AND ZERO-SHOT LEARNING

This section briefly summarizes the supervised and zero-shot learning methods we employ as components of our proposing method. In these methods, the goal is to build a detector f for each semantic concept $c \in C$, where C is a set of concepts. We define

Symbols	Meaning
$c \in C$	a concept to be detected
$x \in \mathbb{R}^d$	a feature vector for testing
$\mathcal{X} = \{(x_i, y_i)\}_{i=1}^N$	a set of training samples
$x_i \in \mathbb{R}^d$	a feature vector for training
$y_i \in \{-1, +1\}$	a label for training
$f_{\text{SV}}(\cdot)$	a supervised detector
$\mathcal{P} = \{g_j(\cdot)\}_{j=1}^M$	a set of pre-trained detectors
$g_j(\cdot)$	a pre-trained detector for d_j
$d_j \in D$	a concept where $D \cap C = \emptyset$
$\text{sim}(\cdot, \cdot)$	similarity between two concepts
$f_{\text{ZS}}(\cdot)$	a zero-shot detector

Table 1: Summary of notations.

notations to be used in this section and the next section as in Table 1.

Supervised Learning. Supervised learning assumes that a set of training samples $\mathcal{X} = \{(x_i, y_i)\}_{i=1}^N$ is given for each concept $c \in C$, where $x_i \in \mathbb{R}^d$ is a feature vector of an image or video, $y_i \in \{+1, -1\}$ is a positive or negative label for x_i, and N is the number of training samples. A supervised detector f_{SV} is then trained from these samples. Its simplest example is a linear detector given by

$$f_{\text{SV}}(x) = \sum_{i=1}^{N} \alpha_i x_i^T x + \gamma, \tag{1}$$

where x is a testing sample, and α_i and γ are the model parameters. Recent studies extend it to non-linear detectors by introducing kernel tricks [62, 63] and/or deep neural networks [10, 12] . Note that, by introducing explicit feature maps [64] corresponding to kernel functions or by viewing deep neural networks as a feature extractor, these non-linear detectors often can be re-formulated as linear detectors in a high-dimensional feature space.

Zero-Shot Learning. Zero-shot learning assumes that a set of pre-trained detectors $\mathcal{P} = \{g_j(\cdot)\}_{j=1}^M$ is given instead of a set of training samples. Here, g_j is a pre-trained detector for a concept $d_j \in D$, where D is another set of concepts disjoint to C, and M is the number of pre-trained detectors. To build a detector for a concept $c \in C$ ($c \notin D$), recent zero-shot learning methods [21, 24, 25] combine given detectors by

$$f_{\text{ZS}}(x) = \sum_{j=1}^{M} \beta_j g_j(x) + \gamma', \tag{2}$$

where β_j and γ' are weighting and bias parameters, respectively. Since β_j is a weight which relates the concept d_j to the target concept c, the similarity measure between d_j and c, $\text{sim}(d_j, c)$, is often used as β_j, i.e., $\beta_j = \text{sim}(d_j, c)$. Its example is cosine similarity between word vectors [31] given by

$$\text{sim}(d_j, c) = \frac{\psi(d_j)^T \psi(c)}{\|\psi(d_j)\|_2 \|\psi(c)\|_2}, \tag{3}$$

where $\psi(\cdot) \in \mathbb{R}^{d'}$ is a word vector of a concept. A word vector, which is a word representation by a real-valued vector, is obtained from semantic embedding methods such as skip-gram [31]. The bias parameter γ' is often experimentally optimized.

4 FEW-SHOT ADAPTATION

4.1 Overview

Our basic idea of few-shot adaptation is to optimize the parameters of supervised learning and zero-shot learning, simultaneously. As shown in Figure 1, the proposed framework accepts a pair of the following two sets as inputs:

(1) a set of training samples $\mathcal{X}$,
(2) a set of pre-trained detectors $\mathcal{P}$,

which are from supervised and zero-shot learning frameworks, respectively.

To bridge supervised learning and zero-shot learning, we impose the following two constraints:

(C1) few-shot adaptation outputs a supervised detector f_{SV} if the set of pre-trained detectors is empty ($\mathcal{P} = \emptyset$).
(C2) few-shot adaptation outputs a zero-shot detector f_{ZS} if the set of training samples is empty ($\mathcal{X} = \emptyset$).

To simultaneously optimize the parameters of supervised learning and zero-shot learning, few-shot adaptation linearly combines a supervised detector f_{SV} and a zero-shot detector f_{ZS}, i.e., we define a detector in few-shot adaptation by

$$f_{\text{FS}}(x) = f_{\text{SV}}(x) + f_{\text{ZS}}(x). \tag{4}$$

For example, with a linear supervised detector in Eq. (1) and a zero-shot detector in Eq. (2), we have

$$f_{\text{FS}}(x) = \sum_{i=1}^{N} \alpha_i x_i^T x + \sum_{j=1}^{M} \beta_j g_j(x) + \gamma''. \tag{5}$$

The goal is to optimize α_i, β_j, and γ'', where the two bias parameters are unified into $\gamma'' = \gamma + \gamma'$.

We believe that this is a straightforward way to unify two learning frameworks, and expect that few-shot adaptation will be effective in cases where the number of training samples is small, because zero-shot learning and supervised learning benefit from each other.

4.2 Introducing an Objective Function from Supervised Learning

To satisfy the constraint (C1), an objective function should be imported from supervised learning to optimize the parameters in Eq. (5). However, since supervised learning is a mapping from a set of training samples $\mathcal{X}$ to a detector f_{SV}, it can not be directly applied to the input $(\mathcal{X}, \mathcal{P})$ of few-shot adaptation, as inputs and outputs are summarized in Table 2.

To solve this problem, our idea is to generate a set of *pseudo* training samples $\mathcal{X}_{\mathcal{P}}$ from pre-trained detectors, and to apply supervised learning to a union set $\mathcal{U} = \mathcal{X} \cup \mathcal{X}_{\mathcal{P}}$. In this way, many types of supervised learning techniques can be introduced to our framework without modifying their objective function. Note that by simply defining $\mathcal{X}_{\emptyset} = \emptyset$, we have $\mathcal{U} = \mathcal{X} \cup \mathcal{X}_{\emptyset} = \mathcal{X}$ when $\mathcal{P} = \emptyset$. This shows that the constraint (C1) is satisfied. The definition of $\mathcal{X}_{\mathcal{P}}$ is given in the following subsection.

Method	Input	# Training Samples	# Pre-trained Detectors	Output	Parameters
Supervised Learning	$\mathcal{X} = \{(x_i, y_i)\}_{i=1}^{N}$	N	0	f_{SV}	α_i, γ
Zero-Shot Learning	$\mathcal{P} = \{g_j(\cdot)\}_{j=1}^{M}$	0	M	f_{ZS}	β_j, γ'
Few-Shot Adaptation	$(\mathcal{X}, \mathcal{P})$	N	M	f_{FS}	$\alpha_i, \beta_j, \gamma''$

Table 2: Summary of assumptions and parameters of each detector.

4.3 Generating Pseudo Training Samples from Zero-Shot Detectors

Our next focus is on the constraint (C2) for a zero-shot detector. If $\mathcal{X} = \emptyset$, few-shot adaptation applies a supervised learning method to a set $\mathcal{U} = \mathcal{X} \cup \mathcal{X}_{\mathcal{P}} = \mathcal{X}_{\mathcal{P}}$ as described above. We focus on *how to generate pseudo training samples that give a zero-shot detector as a result of supervised learning.*

Let us start from the simplest example using linear function as the supervised detector and the zero-shot detector in Eq. (5). Let the pre-trained detectors for zero-shot learning be given by $g_j(x) = w_j^T x$. Then,

$$f_{\text{FS}}(x) = \sum_{i=1}^{N} \alpha_i x_i^T x + \sum_{j=1}^{M} \beta_j w_j^T x + \gamma''. \tag{6}$$

Here, w_j ($\|w_j\| = 1$) is the normal vector to the decision boundary of $g_j(x) = 0$. In this equation, we see that its two terms on the right-hand side share a common structure that each term is a product of a parameter (α_i and β_j) and an inner product of two vectors ($x_i^T x$ and $w_j^T x$). Since x_i can be understood as a training sample to optimize α_i, this one-to-one correspondence implies to us that w_j can be used as a pseudo training sample to optimize β_j.

How to make pseudo training samples? To obtain a function $g(x) = w^T x$ as a result of supervised learning, the easiest way is to have a pair of the normal vector w and its mirrored vector $-w$ for training with positive and negative labels, respectively, as shown in Figure 2 (b). In this case, a set of pseudo training samples is given by $\mathcal{X}_{\mathcal{P}} = \{(\lambda w, +1), (-\lambda w, -1)\}$, where $\lambda > 0$ is a scaling coefficient.

This can be extended to a zero-shot detector f_{ZS}, a weighted sum of functions g_j. By multiplying weight values $\text{sim}(d_j, c)$ given from the zero-shot learning framework, e.g., Eq. (3), a set $\mathcal{X}_{\mathcal{P}} = \{(\tilde{x}_k, \tilde{y}_k)\}_{k=1}^{K}$ is defined by

$$\tilde{x}_{2j} = +\lambda \text{sim}(d_j, c) w_j, \tag{7}$$

$$\tilde{x}_{2j-1} = -\lambda \text{sim}(d_j, c) w_j, \tag{8}$$

with $\tilde{y}_{2j} = +1, \tilde{y}_{2j-1} = -1$ for $j = 1, 2, \cdots, M$, where $K = 2M$ is the number of pseudo training samples.

Finally, by applying supervised learning to a union set $\mathcal{U} = \mathcal{X} \cup \mathcal{X}_{\mathcal{P}}$, the detector of few-shot adaptation is reformulated by

$$f_{\text{FS}}(x) = \sum_{i=1}^{N} \alpha_i x_i^T x + \sum_{k=1}^{K} \tilde{\beta}_k \tilde{x}_k^T x + \gamma'', \tag{9}$$

where $\alpha_i, \tilde{\beta}_k$, and γ'' are parameters. Figure 2 shows how few-shot adaptation works with the minimum sets of training data.

To exactly obtain a zero-shot detector, the objective function imported from supervised learning for parameter estimation is required to give $\tilde{\beta}_{2j} = +1$ and $\tilde{\beta}_{2j-1} = -1$ when $\mathcal{X} = \emptyset$. In practice, this is often a trivial solution of parameter estimation since $\mathcal{U} = \emptyset \cup \mathcal{X}_{\mathcal{P}} = \mathcal{X}_{\mathcal{P}}$ only has pairs of symmetric samples. In this case, we have

$$\begin{aligned} f_{\text{FS}}(x) &= \sum_{k=1}^{K} \tilde{\beta}_k \tilde{x}_k^T x + \gamma'' && (10) \\ &= \sum_{j=1}^{M} (\tilde{\beta}_{2j} \tilde{x}_{2j} + \tilde{\beta}_{2j-1} \tilde{x}_{2j-1})^T x + \gamma'' && (11) \\ &= \sum_{j=1}^{M} \lambda (\tilde{\beta}_{2j} - \tilde{\beta}_{2j-1}) \text{sim}(d_j, c) w_j^T x + \gamma'' && (12) \\ &= 2\lambda f_{\text{ZS}}(x), && (13) \end{aligned}$$

and thus by setting $\lambda = \frac{1}{2}$, f_{FS} is equal to f_{ZS} as required in (C2). Note also that if the dimension of feature vector x is larger than the number of pre-trained detectors M, $\mathcal{X}_{\mathcal{P}}$ becomes linearly separable. This supports many supervised learning methods to satisfy the requirement, by introducing recent high-dimensional feature extractor including deep convolutional networks.

4.4 Extensions to Non-Linear Functions

This subsection presents three methods to introduce nonlinearity to few-shot adaptation. The first method extends linear few-shot adaptation in Eq. (9) to kernelized few-shot adaptation. The second method introduces a deep convolutional network to a zero-shot detector in our framework. The third method extends our framework to multi-class classification using neural networks.

4.4.1 Kernelized Few-Shot Adaptation. To introduce nonlinearity into our framework, we apply kernel tricks [62, 63] to Eq. (9), by replacing dot products with a kernel $\kappa(\cdot, \cdot)$. The detector is then given by

$$f_{\text{FS}}(x) = \sum_{i=1}^{N} \alpha_i \kappa(x_i, x) + \sum_{k=1}^{K} \tilde{\beta}_k \kappa(\tilde{x}_k, x) + \gamma''. \tag{14}$$

Note that we keep to use a set $\mathcal{U} = \mathcal{X} \cup \mathcal{X}_{\mathcal{P}}$ for training. With this kernelization, a linear zero-shot detector f_{ZS} will not be obtained exactly when $\mathcal{X} = \emptyset$. Instead, it provides a kernelized zero-shot detector, which can be viewed as an extended method for zero-shot learning.

4.4.2 Pre-trained Detectors with Neural Networks. In practice, many recent studies have proved the effectiveness of neural networks trained on large-scale datasets. Most of these networks have a softmax classifier at the final layer. Here, we present a way to introduce them to our framework by defining w_j in Eq. (7) and (8) by a concatenation of network parameters.

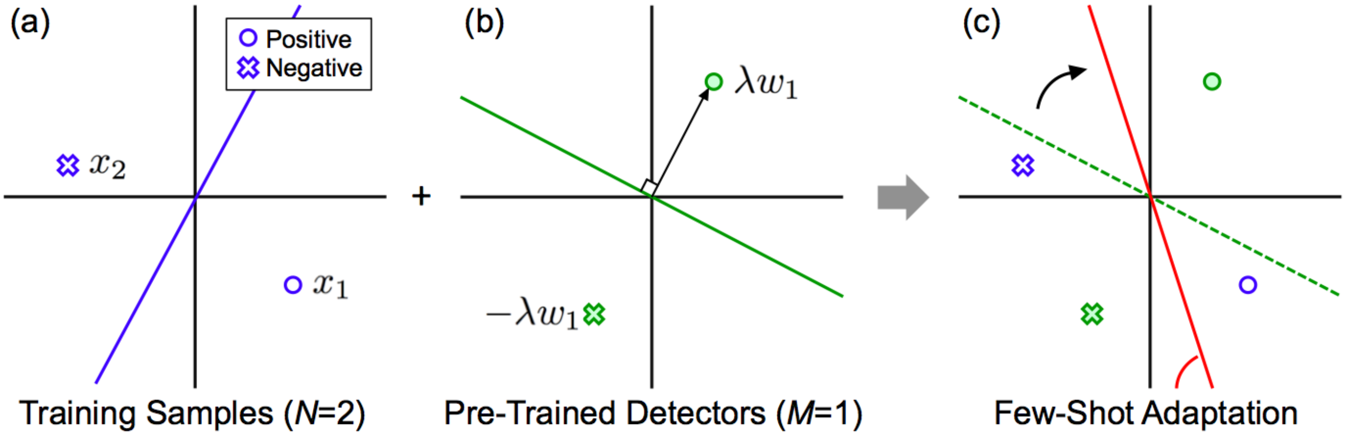

Figure 2: Example of few-shot adaptation. (a) Training samples from a supervised learning framework. x_1 and x_2 are positive and negative training samples, respectively. (b) Pre-trained detectors from zero-shot learning framework. An example using a single pre-trained detector $g_1(x) = w_1^T x$ is given with illustration of the normal vector λw_1. (c) Few-shot adaptation with a decision boundary, in which supervised learning and zero-shot learning are combined.

Let h be an input of a softmax layer, which has outputs for M concepts. The output value (posterior probability) for the j-th concept is given by

$$p_j(h) = \frac{\exp(a_j^T h + b_j)}{\sum_{j=1}^{M} \exp(a_j^T h + b_j)} \tag{15}$$

where a_j and b_j are parameters on the layer. To extract pseudo training samples, we focus on its linear calculation $a_j^T h + b_j$, and define

$$w_j = \left(\begin{array}{c} a_j \\ b_j \end{array} \right), \tag{16}$$

with a feature vector

$$x = \left(\begin{array}{c} h \\ 1 \end{array} \right). \tag{17}$$

Note that exp function and the normalization process (with a denominator of the sum of exp values) in Eq. (15) are omitted with these definition. They can be again introduced by utilizing Gaussian kernel in kernelized few-shot adaptation in Eq. (14), and by applying score normalization to values of $f_{\text{FS}}(x)$ if they are needed.

4.4.3 Extension to Multi-Class Classification. Our framework presented above is for binary classification to train concept detectors in a one-versus-all manner. This is effective to detection tasks in the wild, e.g., TRECVID Semantic Indexing Task [67], in which a video shot can have multiple labels. On the other hand, a number of recent studies on few-shot learning [15, 16, 39] have focused on multi-class classification using neural networks, for exmaple, object recognition on the ImageNet dataset. To compare our approach with them, we extend our framework to multi-class classification by utilizing only positive pseudo training samples, i.e., a set of positive pseudo training samples $\mathcal{X}_{\mathcal{P}}^c$ for each concept c is added to training samples for multi-class classification.

5 EXPERIMENTS

Our few-shot adaptation framework is evaluated on three datasets, TRECVID 2010, TRECVID 2014, and ImageNet.

5.1 Evaluation on TRECVID datasets

5.1.1 Experimental Settings. The TRECVID 2010 and 2014 datasets consist of Internet videos used in the TRECVID Semantic Indexing Competition [67]. Here, we use the whole ImageNet images [35] and Places 365 images [36] for pre-training. We believe this is one of the best choices to report results by increasing the number of training samples from zero to many, and to show the versatility of ImageNet and Places 365 datasets with our proposed framework.

The task is to detect semantic concepts from each video shot. Shot boundaries are provided in the datasets. The number of video shots for training and testing are listed in Table 3. Each dataset has 30 types of semantic concepts to be detected. The evaluation measure is Mean Average Precision (Mean AP), which is calculated by using the official toolkit and annotations.

Evaluation results are reported on three training conditions: zero-shot, few-shot, and many-shot. The zero-shot condition does not use TRECVID videos for training. The few-shot condition limits the number of training video shots to N ($0 < N \leq 100$) by random sampling. The many-shot condition use all TRECVID videos for training. In the many-shot condition, we can compare our results with official submissions using supervised learning methods at the competition. Note that our main focus is on the few-shot condition.

For pre-trained detectors, three types of GoogLeNets [12] are used: ImageNet-1K, Places-365, and ImageNet-Shuffle13K. The Goog-LeNet is a convolutional neural network with 23 layers. ImageNet-1K uses 1.2 million images of 1,000 objects in ILSVRC 2012 for training [12]. Places-365 uses 1.8 million images of 365

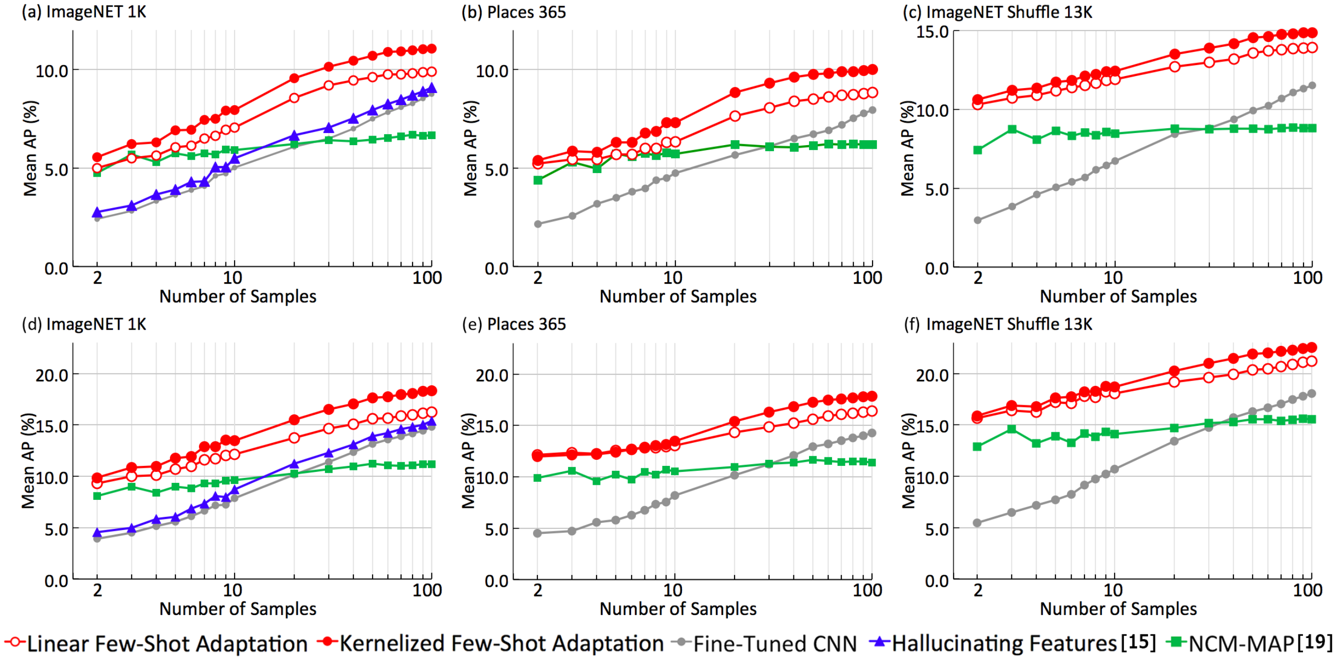

Figure 3: Few-Shot Evaluation on the TRECVID 2010 and 2014 datasets. Linear/Kernelized Few-Shot Adaptation: our proposed framework. Fine-Tuned CNN: baseline using GoogLeNet features. Hallucinating Features [15]: artificially-generated samples using neural networks trained using analogy among ImageNet objects. NCM-MAP [19]: Nearest class mean classifier with Maximum a Posteriori estimation using zero-shot priors. Results are reported for three types of pre-trained networks: (a,d) ImageNet-1K, (b,e) Places-365, and (c,f) ImageNet-Shuffle13K, on two datasets: (a,b,c) TRECVID 2010 and (d,e,f) TRECVID 2014. All experiments are repeated for 10 times, and the average results are reported.

scenes in Places 2 dataset [36]. ImageNet-Shuffle13K uses the ImageNet Shuffle method [43] for training, which provides 12,988 object classifiers trained on the whole ImageNet dataset. For word vectors, 300 dimensional vectors obtained from the Skip-gram model in [31] are used. They are used to measure similarity between concepts in Eq. (3). The average similarity between a TRECVID concept and its closest concept in pre-training is 0.557 for ImageNet-1K, 0.606 for Places-365 and 0.781 for ImageNet-Shuffle13K. For objective function, the SVM loss (Hinge loss with L_2 regulalization) is used for parameter optimization.

5.1.2 Few-Shot Evaluation. Figure 3 shows evaluation results on TRECVID 2010 and 2014 datasets under the few-shot condition. As a baseline, results using CNN+SVM with supervised fine-tuning are reported, where the final softmax layer of GoogLeNet is replaced by an SVM. Note that this has been one of the best ways to apply neural networks to the semantic indexing task at TRECVID. For comparison, evaluation results of Nearest Class Mean Classifier [19] with Maximum a Posteriori estimation using zero-shot priors (NCM-MAP) and artificially-generated hallucinating features (HF) in [15] are also reported. HF is applied only for ImageNet-1K because it uses a network trained with analogy among the 1K objects on ILSVRC 2012.

We see from the results that kernelized few-shot adaptation performs the best with all networks and on both datasets. This shows the effectiveness of the proposed framework, and confirms that zero-shot learning and supervised learning benefit from each other when the number of training samples is small ($0 < N \leq 100$). We also see that few-shot adaptation approaches to supervised fine-tuning as N increases. This shows our framework straightforwardly bridges zero-shot learning and supervised learning.

TRECVID Year	2010	2014
Training video shots	119,685	547,634
Positive in training per concept	735	1,657
Testing video shots	144,988	107,806

Table 3: The number of video shots on TRECVID 2010 and 2014 datasets. Each dataset has 30 types of semantic concepts for evaluation.

If we compare linear and kernelized few-shot adaptation, the kernelized one is always better. This shows that the kernel trick in supervised learning is effective. Utilizing the other types of kernels is interesting as a next step in future.

5.1.3 Zero-Shot Evaluation. Table 4 reports our results on the zero-shot condition. Our method performs the best among methods in [47, 48] from the non-annotation track at TRECVID 2014 and zero-shot methods [21, 57]. Here, non-annotation track is a training condition which requires not to use TRECVID videos but to use the other resources such as web images for training. Note that our zero-shot method can be viewed as a modification of ConSE [21],

Zero-Shot Methods	Mean AP
ConSE [21] (ImageNet-1K)	6.39
Inoue et al. [57] (ImageNet-1K)	8.31
Ours (ImageNet-1K)	8.49
Ours (Places-365)	11.63
Ours (ImageNet-Shuffle13K)	14.89
Ours (3-Net Fusion)	**15.19**

Webly Supervised Methods	Mean AP
Jiang et al.[47]	1.21
McGuinness et al. [48]	7.97

Table 4: Zero-Shot Evaluation. Our proposed method is compared with zero-shot learning methods and webly supervised learning methods.

Methods	Mean AP
Ours (with pseudo samples)	**35.89**
Ours (without pseudo samples)	35.73
Snoek [13] (8 CNNs)	33.19
Laaksonen [70] (2 CNNs + Hard Negative Mining)	29.36
Inoue [71] (CNN + Temporal N-Gram Model)	28.12
Safadi [72] (CNN + Re-Ranking Model)	26.59
Ballas [73] (CNN + Audio-Visual Features)	25.90

Table 5: Many-Shot Evaluation with and without pseudo samples. The top 5 official submissions at TRECVID 2014 are also reported.

where a normalization step for multi-class classification is omitted. Since normalization assumes that each video shot has one of concept labels, it is not suitable for our concept detection task, in which a video shot can have multiple concept labels with unbalanced positive and negative samples. To further improve the performance in the zero-shot condition, modifying and introducing recent zero-shot multi-class classification methods such as manifold learning [50] is promising.

The results also show that late fusion of detection scores obtained from three networks improves the performance. This suggests that adding other types of pre-trained detectors is also needed in future work.

5.1.4 *Many-Shot Evaluation.* Table 5 compares our results with the official submissions in TRECVID 2014. We achieved 35.89 % and 35.73% in Mean AP with and without pseudo samples, respectively. To the best of our knowledge, this is the best performance on this dataset. This confirms that our pseudo training samples for few-shot adaptation do not affect the performance in the many-shot condition, and means that our few-shot adaptation successfully unifies zero-shot learning and supervised learning.

5.2 Evaluation on Imagenet dataset

5.2.1 *Experimental Settings.* The ImageNet Large Scale Visual Recognition Competition (ILSVRC) dataset consists of 1.2 million images with 1,000 object categories. For few-shot evaluation, we follow the evaluation setting proposed in [15], which divides the 1,000 categories into 389 base categories and 611 novel categories. All examples from base categories are used for pre-training, and few examples ($N = 1, 2, 5, 10$, and 20) for novel categories are used for few-shot adaptation. Evaluation measure is Top-5 accuracy. For a fair comparison, ResNet-10 used in [15] is applied to this multi-class classification problem.

Few-Shot Methods	$N = 1$	2	5	10	20
Baseline	43.0	54.3	67.2	72.8	75.9
Hallucinating Features [15]	54.3	62.1	71.3	75.8	**78.1**
Matching Network [16]	55.0	61.5	69.3	73.4	76.2
Model Regression [39]	46.4	56.7	66.8	70.4	72.0
Ours	**55.2**	**63.3**	**71.8**	**76.0**	78.0

Table 6: Evaluation on ImageNet dataset. Top-5 accuracy is reported. *N* is the number of training examples per category. The ResNet-10 network architecture in [15] is used for all experiments.

(a) Novel Categories

Few-Shot Methods	$N = 1$	2	5	10	20
Hallucinating Features [15]	32.8	46.4	61.7	69.7	**73.8**
Matching Network [16]	**41.3**	**51.3**	62.1	67.8	71.8
Ours	35.0	49.7	**62.6**	**70.1**	73.7

(b) Base Categories

Few-Shot Methods	$N = 1$	2	5	10	20
Hallucinating Features [15]	**88.4**	**87.1**	**86.6**	**85.5**	**85.0**
Matching Network [16]	76.7	77.8	80.6	82.2	83.3
Ours	87.5	84.8	86.4	85.2	84.8

Table 7: Analysis with novel and base categories. (a) Top-5 accuracy for novel categories. (b) Top-5 accuracy for base categories.

5.2.2 *Comparison with Other Few-Shot Learning Methods.* Table 6 compares our method with state-of-the-art few-shot learning methods for multi-class classification: Matching Network (MN) for one-shot learning [16], Model Regression [39], and Hallucinating Features (HF) [15]. We see our method performs the best for $N = 1, 2, 5, 10$, and the second best for $N = 20$ among these methods. To analyze results, Table 7 separately reports accuracy on novel and base sets of categories. Note that they are in a trade-off relation. We see a tendency that MN and HF are effective for novel and base categories, respectively, and that our method provides well-balanced performance on both. This experiment uses only 389 pre-trained detectors for a fair comparison. Increasing the number of base categories is promising to further improve the accuracy of our method.

5.3 Limitations and Discussions

Here we show some example video shots detected on the TRECVID dataset. Figure 4 shows the top-ranked video shots. They are mostly true positive shots even with 10 training examples. Figure 5 shows Average Precisions (APs) by semantic concepts. Few-shot adaptation is effective for various concepts. However, it is difficult to detect actions such as Walking and SittingDown from video with few examples. To further improve the overall performance, pre-trained detectors closely related to the domain of the TRECVID task are required. For example, introducing 3D CNNs pre-trained on action video datasets such as Kinetics [68, 69] would be interesting as a promising next step.

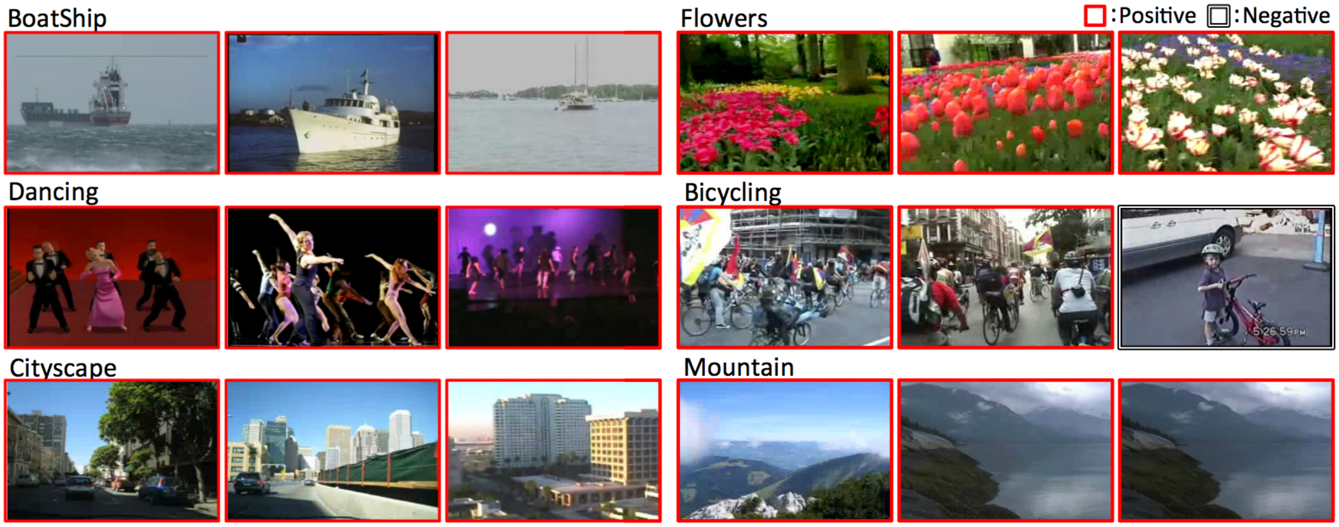

Figure 4: Top three video shots for six types of semantic concepts. Kernelized few-shot adaptation with $N = 10$ is applied.

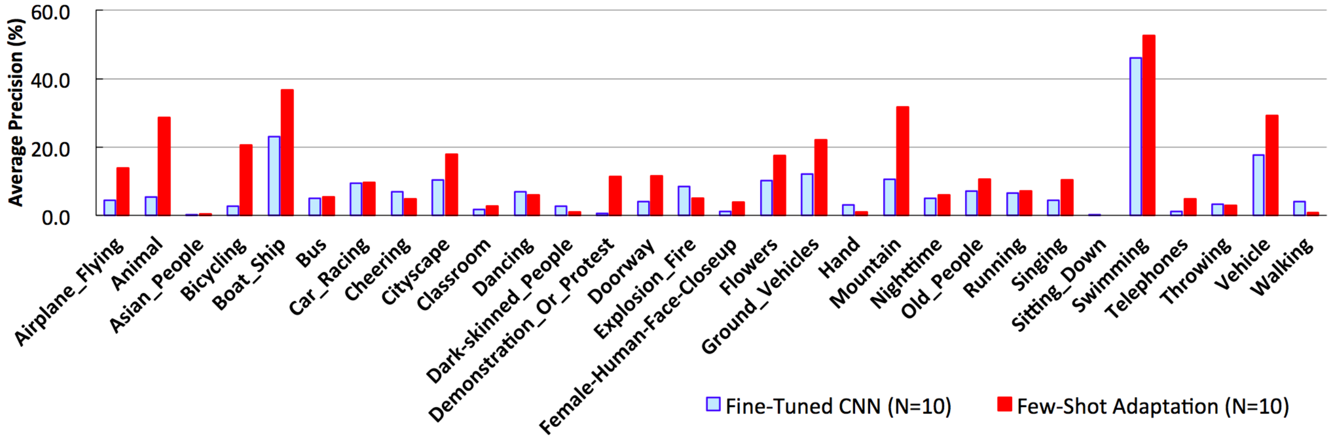

Figure 5: Average Precision by concepts on the TRECVID 2010 dataset.

Another limitation of our framework is in its assumption that word vectors are given by a semantic embedding method. New types of word vectors, such as those obtained by joint training of text and image representation [60, 61] or manifold learning [50], can be introduced to our framework. However, our few-shot adaptation can not update these embeddings in its training phase. The three joint training of whole system including semantic embeddings is needed in future work.

6 CONCLUSION

We proposed a few-shot adaptation framework, which combines zero-shot learning and supervised learning. It provided robust parameter estimation with few training examples, by optimizing the parameters of zero-shot learning and supervised learning simultaneously. Our experiments showed the effectiveness of the proposed framework on TRECVID 2010, 2014, and ImageNet datasets. Our future work will be focusing on audio and text analysis to detect actions and events from video data.

ACKNOWLEDGEMENT

This work was supported by JST ACT-I Grant Number JPMJPR16U5 and JST CREST Grant Number JPMJCR1687, Japan. Part of this work is conducted as research activities of AIST-TokyoTech Real World Big-Data Computation Open Innovation Laboratory (RWBC-OIL).

REFERENCES

[1] C. G. M. Snoek and M. Worring. Multimodal Video Indexing: A Review of the State-of-the-Art. *In Springer Multimedia Tools and Applications*, vol. 25, no. 1, pp. 5–35, 2005.
[2] G. Awad, *et al.* TRECVid Semantic Indexing of Video: A 6-Year Retrospective. *In ITE Trans. on Media Technology and Applications*, vol. 4, no. 3, pp. 187–208, 2016.
[3] B. Zhao, *et al.* Hierarchical Recurrent Neural Network for Video Summarization. *ACM Multimedia*, pp. 863–871, 2017.
[4] T. Mei, *et al.* Near-Lossless Semantic Video Summarization and Its Applications to Video Analysis. *In ACM Trans. on Multimedia Computing, Communications,*

and Applications, vol. 9, no. 3-16, pp. 1–23, 2013.

[5] Y. Xian, *et al.* Evaluation of Low-Level Features for Real-World Surveillance Event Detection. *In IEEE Trans. on Circuits and Systems for Video Technology*, vol. 27, no. 3, pp. 624–634, 2017.

[6] X. Zhang, *et al.* Background-Modeling-Based Adaptive Prediction for Surveillance Video Coding. *In IEEE Trans. on Image Processing*, vol. 23, no. 2, pp. 769–784, 2014.

[7] A. W. M. Smeulders, *et al.* Content-Based Image Retrieval at the End of the Early Years. *In IEEE Trans. on PAMI*, vol. 22, no. 12, pp. 1349–1380, 2000.

[8] H. Drucker, *et al.* Support Vector Regression Machines. *NIPS*, pp. 155–161, 1997.

[9] V. Vapnik. The Nature of Statistical Learning Theory. *Springer, New York*, 1995.

[10] A. Krizhevsky, *et al.* ImageNet Classification with Deep Convolutional Neural Networks. *NIPS*, pp.1–9, 2012.

[11] K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *ICLR*, 2015.

[12] C. Szegedy, *et al.* Going Deeper with Convolutions. *CVPR*, 2015.

[13] C. G. M. Snoek, *et al.* MediaMill at TRECVID 2014: Searching Concepts, Objects, Instances and Events in Video. *TRECVID workshop*, 2014.

[14] F. Perronnin, *et al.* Improving the Fisher Kernel for Large-Scale Image Classification. *ECCV*, 2010.

[15] B. Hariharan and R. Girshick. Low-shot Visual Recognition by Shrinking and Hallucinating Features. *ICCV*, 2017.

[16] O. Vinyals, *et al.* Matching Networks for One Shot Learning. *NIPS*, 2016.

[17] M. Oquab, *et al.* Learning and Transferring Mid-Level Image Representations using Convolutional Neural Networks. *CVPR*, 2014.

[18] A. Babenko, *et al.* Neural Codes for Image Retrieval. *ECCV*, pp. 584–599, 2014.

[19] T. Mensink, J. Verbeek, F. Perronnin, and G. Csurka. Distance-Based Image Classification: Generalizing to New Classes at Near-Zero Cost. *IEEE Trans. on PAMI*, vol. 35, no. 11, pp. 2624–2637, 2013.

[20] F. Perronnin, *et al.* Adapted Vocabularies for Generic Visual Categorization. *ECCV*, pp. 464–475, 2006.

[21] M. Norouzi, *et al.* Zero-shot Learning by Convex Combination of Semantic Embeddings. *ICLR*, 2014.

[22] A. Frome, *et al.* Devise: A Deep Visual-semantic Embedding Model. *NIPS*, 2013.

[23] T. Mensink, *et al.* Costa: Co-occurrence Statistics for Zero-shot Classification. *CVPR*, 2014.

[24] M. Jain, *et al.* Objects2action: Classifying and Localizing Actions without Any Video Example. *ICCV*, 2015.

[25] Y. Xian, *et al.* Zero-Shot Learning - the Good, the Bad and the Ugly. *CVPR*, 2017.

[26] S. Cappallo and C.G.M. Snoek. Future-Supervised Retrieval of Unseen Queries for Live Video. *ACM Multimedia*, pp. 28–36, 2017.

[27] J. Qin, *et al.* Zero-Shot Action Recognition With Error-Correcting Output Codes *CVPR*, 2017.

[28] X. Xu, *et al.* Multi-Task Zero-Shot Action Recognition with Prioritised Data Augmentation. *ECCV*, 2016.

[29] C. Gan, *et al.* Concepts not Alone: Exploring Pairwise Relationships for Zero-Shot Video Activity Recognition. *AAAI*, pp. 3487–3493, 2016.

[30] T. Mikolov, *et al.* Efficient Estimation of Word Representations in Vector Space. *ICLR*, 2013.

[31] T. Mikolov, *et al.* Distributed Representations of Words and Phrases and their Compositionality. *NIPS*, 2013.

[32] E. Gavves, *et al.* Active Transfer Learning with Zero-Shot Priors: Reusing Past Datasets for Future Tasks. *ICCV*, 2015.

[33] K. He, *et al.* Deep Residual Learning for Image Recognition. *CVPR*, 2016.

[34] G. Huang, *et al.* Densely Connected Convolutional Networks *CVPR*, 2017.

[35] O. Russakovsky, *et al.* ImageNet Large Scale Visual Recognition Challenge. *In IJCV*, vol.115, no.3, pp.211–252, 2015.

[36] B. Zhou, *et al.* Places: A 10 million Image Database for Scene Recognition. *In IEEE Trans. on PAMI*, in press, 2017.

[37] A. Santoro, *et al.* Meta-Learning with Memory Augmented Neural Network. *ICML*, 2016.

[38] R. Kwitt, *et al.* One-Shot Learning of Scene Locations via Feature Trajectory Transfer. *CVPR*, 2016.

[39] Y.X. Wang and M. Hebert. Learning to Learn: Model Regression Networks for Easy Small Sample Learning. *ECCV*, 2016.

[40] N. Inoue and K. Shinoda. A Fast and Accurate Video Semantic-Indexing System Using Fast MAP Adaptation and GMM Supervectors. *IEEE Trans. on Multimedia*, vol. 14, no. 4, pp. 1196–1205, 2012.

[41] F. Perronnin, *et al.* Improving the fisher kernel for large-scale image classification. *ECCV*, pp. 143–156, 2010.

[42] C.G.M. Snoek, *et al.* Qualcomm Research and University of Amsterdam at TRECVID 2015: Recognizing Concepts, Objects, and Events in Video. *TRECVID workshop*, 2015.

[43] P. Mettes, *et al.* The ImageNet Shuffle: Reorganized Pre-training for Video Event Detection. *ICMR*, 2016.

[44] A. Habibian, *et al.* Video2vec Embeddings Recognize Events when Examples are Scarce. *In IEEE Trans. on PAMI*, in press, 2017.

[45] C. Feichtenhofer, *et al.* Convolutional Two-Stream Network Fusion for Video Action Recognition. *CVPR*, 2016.

[46] L. Sun, *et al.* Human Action Recognition Using Factorized Spatio-Temporal Convolutional Networks. *ICCV*, 2015.

[47] L. Jiang, *et al.* CMU-Informedia at TRECVID Semantic Indexing. *TRECVID workshop*, 2014.

[48] K. McGuinness, *et al.* Insight Centre for Data Analytics at TRECVid 2014: Instance Search and Semantic Indexing. *TRECVID workshop*, 2014.

[49] C.H. Lampert, *et al.* Attribute-Based Classification for Zero-Shot Visual Object Categorization. *In IEEE Trans. on PAMI*, vol. 36, no. 3, pp. 453–465, 2013.

[50] S. Changpinyo, *et al.* Synthesized Classifiers for Zero-Shot Learning. *CVPR*, 2016.

[51] S. Huang, *et al.* Learning Hypergraph-Regularized Attribute Predictors. *CVPR*, 2015.

[52] X. Yu and Y. Aloimonos. Attribute-based Transfer Learning for Object Categorization with Zero or One Training Example. *ECCV*, 2010.

[53] Q. Yu, *et al.* Multimedia event recounting with concept based representation. *ACM Multimedia*, 2012.

[54] L. Jiang, *et al.* Easy Samples First: Self-Paced Reranking for Zero-Example Multimedia Search. *ACM Multimedia*, 2014.

[55] A. Habibian, *et al.* Composite Concept Discovery for Zero-shot Video Event Detection. *ICMR*, 2014.

[56] S. Wu, *et al.* Zero-shot Event Detection using Multi-modal Fusion of Weakly Supervised Concepts. *CVPR*, pp.2665–2672, 2014.

[57] N. Inoue and K. Shinoda. Adaptation of Word Vectors using Tree Structure for Visual Semantics. *ACM Multimedia*, pp. 277–281, 2016.

[58] Z. Zhang and V. Saligrama. Zero-Shot Learning via Joint Latent Similarity Embedding. *CVPR*, 2016.

[59] Y. Xian, *et al.* Latent Embeddings for Zero-shot Classification *CVPR*, 2016.

[60] L. Zhang, *et al.* Learning a Deep Embedding Model for Zero-Shot Learning. *CVPR*, 2017.

[61] E. Kodirov, *et al.* Semantic Autoencoder for Zero-Shot Learning *CVPR*, 2017.

[62] T. Hofmann, *et al.* Kernel Methods in Machine Learning. *In The Annals of Statistics*, vol. 36, no. 3, pp. 1171–1220, 2008.

[63] M.A. Aizerman, *et al.* Theoretical Foundations of the Potential Function Method in Pattern Recognition Learning. *In Automation and Remote Control*, vol. 25, pp. 821–837, 1964.

[64] A. Vedaldi and A. Zisserman. Efficient Additive Kernels via Explicit Feature Maps. *In IEEE Trans. on PAMI*, vol. 34, no. 3, pp. 480–492, 2012.

[65] Y. Jason, *et al.* How Transferable are Features in Deep Neural Networks?. *NIPS*, pp. 3320–3328, 2014.

[66] L. Shao, *et al.* Transfer Learning for Visual Categorization: A Survey. *In IEEE Trans. on Neural Networks and Learning Systems*, vol. 26, no. 5, pp. 1019–1034, 2015.

[67] G. Awad, *et al.* TRECVID 2015 – An Overview of the Goals, Tasks, Data, Evaluation Mechanisms and Metrics. *TRECVID workshop*, 2015.

[68] W. Kay, *et al.* The Kinetics Human Action Video Dataset. *arXiv preprint*, arXiv:1705.06950, 2017.

[69] K. Hara, *et al.* Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet?. *CVPR*, 2018.

[70] J. Laaksonen, *et al.* PicSOM Experiments in TRECVID 2014 Semantic Indexing Task. *TRECVID workshop*, 2014.

[71] N. Inoue and K. Shinoda. n-gram Models for Video Semantic Indexing. *ACM Multimedia*, 2014.

[72] B. Safadi, *et al.* Descriptor Optimization for Multimedia Indexing and Retrieval. *In Springer Multimedia Tools and Applications*, vol. 74, no. 4, pp. 1267–1290, 2015.

[73] N. Ballas, *et al.* Irim at TRECVID 2014: Semantic indexing and Instance Search. *TRECVID workshop*, 2014.