Anticipating the future by watching unlabeled video

Carl Vondrick, Hamed Pirsiavash, Antonio Torralba
Massachusetts Institute of Technology
{vondrick,hpirsiav,torralba}@mit.edu

Abstract

In many computer vision applications, machines will need to reason beyond the present, and predict the future. This task is challenging because it requires leveraging extensive commonsense knowledge of the world that is difficult to write down. We believe that a promising resource for efficiently obtaining this knowledge is through the massive amounts of readily available unlabeled video. In this paper, we present a large scale framework that capitalizes on temporal structure in unlabeled video to learn to anticipate both actions and objects in the future. The key idea behind our approach is that we can train deep networks to predict the visual representation of images in the future. We experimentally validate this idea on two challenging "in the wild" video datasets, and our results suggest that learning with unlabeled videos significantly helps forecast actions and anticipate objects.

1. Introduction

What will the man do next in Figure 1? Although predicting the future is difficult, you can find several clues in the image if you look carefully. The young couple seems to be at an open house, the real estate agent is holding paperwork, and the man's arm is starting to move. You might wager that this couple has bought a house, and therefore, to finalize the deal, the man will soon shake hands.

While computer vision systems can now recognize objects, actions, and scenes with astounding accuracy [34, 40, 51, 16], an unsolved problem is how to create machines that predict what will happen in the *future*. Developing the capability for machines to anticipate events before they start would enable several real-world applications. For example, to make plans or interact with humans, robots will require predictions of the future [19]. Recommendation systems can suggest products or services based on what they anticipate a person will do. Predictive models can also find abnormal situations in surveillance videos, and alert responders.

Developing an algorithm to anticipate the future is challenging. Humans can rely on extensive knowledge accumulated over their lifetime to infer that the man will soon shake hands in Figure 1. How do we give prediction models access to this commonsense knowledge?

Figure 1: **What will happen next?** By learning from massive amounts of unlabeled video, we train models that anticipate the future.

Our insight is to exploit the regularities inherent in video to train models that predict the future. Videos come with the temporal ordering of frames "for free", which is a valuable asset because machines can look forward in time to learn to predict the future. However, annotating videos with commonsense knowledge is likely too expensive [47, 43] and difficult to define.

We believe that a promising resource to train these models are abundantly available unlabeled videos. Although lacking ground truth annotations, they are attractive for training vision systems because they are cheap to obtain at massive scales and still contain rich signals. Foundational work in computer vision has explored the challenging task of visualizing the future using unlabeled videos [29, 45, 48]. However, in many applications, we are interested in predicting the presence of concepts in the future, which may be easier than predicting low level signals.

Rather than predicting pixels, the main idea in this paper is to predict the visual representations of future frames. Recent progress in computer vision has built rich visual representations for recognition. Although going from these abstract representations to exact pixel values is not easy

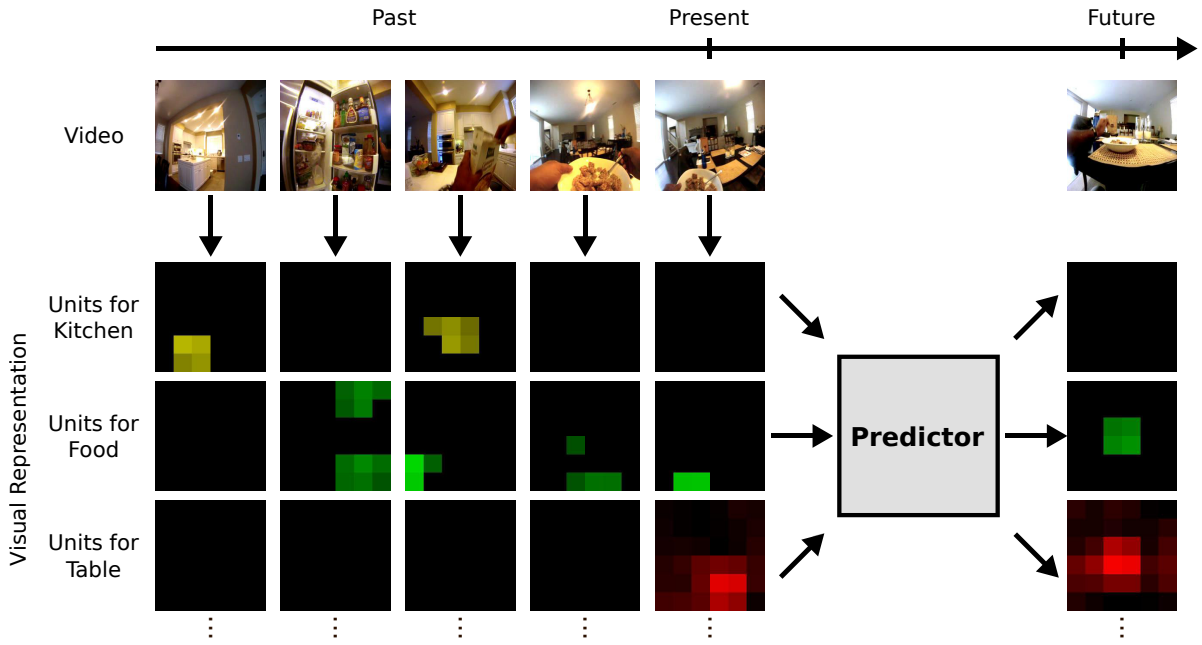

Figure 2: **Predicting Future Representations:** The egocentric video above shows a person walking into a kitchen, preparing cereal, and starting to sit down [30]. Below each frame, we visualize an example hidden representation with [52]. While predicting the exact pixel values in the future may be too difficult in the wild, often predicting the representation is easier because it encodes images at a more abstract level. In this paper, we anticipate objects and actions seconds before they appear by learning to predict visual representations in the future.

[24, 49, 42], they capture the signals necessary for recognizing several high level concepts. Pioneering work has forecasted high level semantics before [17, 21, 13], but they have required supervision, which makes it difficult to scale. Representations have the advantage that they scale to unlabeled videos because they are automatic to compute. Moreover, these representations have been empirically shown to work on a large variety of tasks [33, 5, 50]. In this paper, we anticipate both actions and objects seconds before they appear by applying recognition models on predicted future representations, illustrated in Figure 2.

In our largest experiment, we collect and learn a prediction model with over 600 hours of unlabeled video downloaded from the web. To handle data at these magnitudes, we propose to use deep networks to construct our models. Deep networks are well suited for this problem because their capacity can grow with the size of data available and be trained efficiently with large-scale optimization algorithms. Although we are still far from human performance at this task, our experiments suggest that learning with large amounts of unlabeled videos can significantly help machines predict the future. We validate our approach on forecasting actions and objects on two challenging in-the-wild datasets of human actions in television shows [26] and egocentric videos for activities of daily living [30].

The primary contribution of this paper is showing that extracting signals from massive amounts of unlabeled video can help machines anticipate concepts in the future. The remainder of the paper discusses this contribution in detail. In section 2, we first review related work. In section 3, we then present our deep network to predict visual representations in the future. Since anticipating the future is an ambiguous task, we extend our network architecture to produce multiple predictions. In section 4, we show several experiments to forecast both actions and objects. In our most difficult setting, we forecast objects five seconds before they appear with reasonable performance.

2. Related Work

The problem of predicting the future in images and videos has received growing interest in the computer vision community, and our work builds upon this foundation:

Prediction with Unlabeled Videos: Perhaps the ideas most similar to this paper are the ones that capitalize on the wide availability of big video collections. In early work, Yuen and Torralba [48] propose to predict motion in a single image by transferring motion cues from visually similar videos in a large database. Building on the rich potential of large video collections, Walker et al. [45] demonstrate a compelling data-driven approach that animates the trajectory of objects from a single frame. Ranzato et al. [32] and

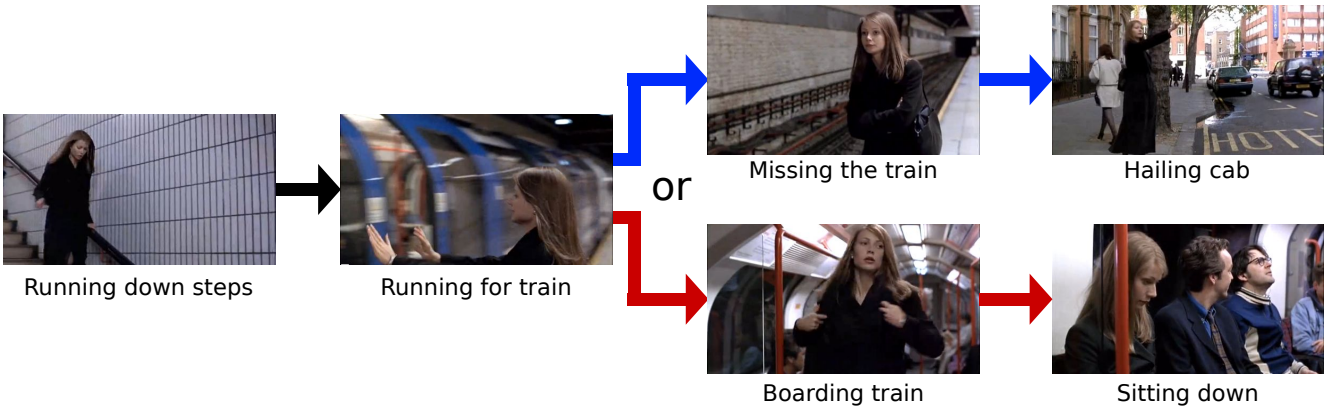

Figure 3: **Multiple Futures:** Since predicting the future is ambigious, models need to anticipate multiple possibilities.

Srivastava et al. [38] also learn predictive models from large unlabled video datasets to predict pixels in the future. In this paper, we also use large video collections. However, unlike previous work that predicts low-level pixels or motions, we develop a system to predict high-level concepts such as objects and actions by learning from unlabeled video.

Predicting Actions: There have been a few promising works on predicting future action categories. Lan et al. [21] propose a hierarchical representation to predict future actions in the wild. Ryoo [35] and Hoai and De la Torre [11] propose models to predict actions in early stages. Vu et al. in [44] learn scene affordance to predict what actions can happen in a static scene. Pei et al. [27] and Xie et al. [46] infer people's intention in performing actions which is a good clue for predicting future actions. We are different from these approaches because we use large-scale unlabeled data to predict a rich visual representation in the future, and apply it towards anticipating both actions and objects.

Predicting Human Paths: There have been several works that predict the future by reasoning about scene semantics with encouraging success. Kitani et al. [17] use concept detectors to predict the possible trajectories a person may take in surveillance applications. Lezema et al. [23], Gong et al. [9] and Kooij et al. [18] also predict the possible future path for people in the scene. Koppula and Saxena [19] anticipate the action movements a person may take in a human robot interaction scenario using RGB-D sensors. Our approach extends these efforts by predicting human actions and objects in images in the wild.

Predicting Motions: One fundamental component of prediction is predicting short motions, and there have been some investigations towards this. Pickup et al. in [28] implicitly model causality to understand what should happen before what in a video. Fouhey and Zitnick [8] learn from abstract scenes to predict what objects may move together. Pintea et al. [29] predict the optical flow from single images by predicting how pixels are going to move in future. We are hoping that our model learns to extrapolate these motions automatically in the visual representation, which is helpful if we want to perform recognition in the future rather than rendering it in pixel space.

Big Visual Data: We are inspired by recent work that leverages the large wealth of visual data readily available online. Torralba et al. [41] use millions of Internet images to build object and scene recognition systems. Chen et al. [2] and Divvala et al. [3] build object recognition systems that have access to common sense by mining visual data from the web. Doersch et al. [4] use large repositories of images from the web to tease apart visually distinctive elements of places. Zhou et al. [51] train convolutional neural networks on a massive number of scene images to improve scene recognition accuracy. In our work, we also propose to mine information from visual media on the web, however we do it for videos with the goal of learning a model to predict the future.

Unsupervised Learning in Vision: To handle large-scale data, there have been some efforts to create unsupervised learning systems for vision. Ramanan et al. [31] uses temporal relationships in videos to build datasets of human faces. Ikizler-Cinbis et al. [14] propose to use images from the web to learn and annotate actions in videos without supervision. Le et al. [22] show that machines can learn to recognize both human and cat faces by watching an enormous amount of YouTube videos. Chen and Grauman [1] propose a method to discover new human actions by only analyzing unlabeled videos., and Mobhai et al. [25] similarly discover objects. Fouhey et al. [7] propose to watch people with minimal supervision in order to learn room geometry. This paper also proposes to use unlabeled data, but we differ from previous works because we do it for predicting visual representations.

3. Anticipating Visual Representations

We now present our large-scale learning framework for predicting visual representations in the future.

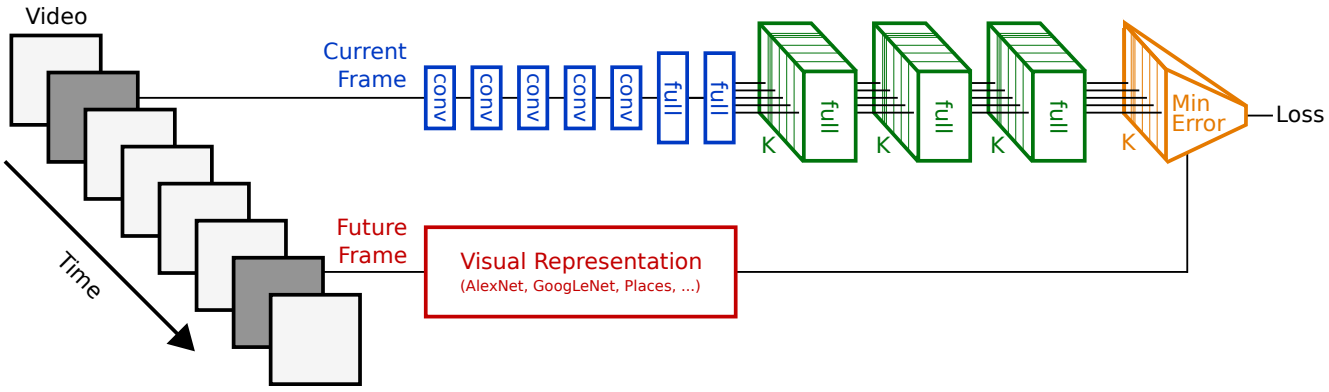

Figure 4: **Network for Predicting Future Representations:** We visualize the network architecture of our full model. During training, our network uses videos to learn to predict the representation of frames in the future. Since predicting the future is ambiguous, our network predicts K futures. Blue layers are the same for each output while green layers have hidden units that are interleaved between the K outputs. During inference, we only input the current frame, and the network estimates K representations for the future. Please see section 3 for full details.

3.1. Self-supervised Learning

The key idea of our method is that we can use large amounts of unlabeled video to train models that predict the visual representation in the future. The temporal structure of video automatically provides self-supervision to train models to anticipate events.

Given a video frame x_t^i at time t from video i, our goal is to predict the visual representation for the future frame $x_{t+\Delta}^i$. Let $\phi(x_{t+\Delta}^i)$ be the representation in the future. Using videos as training data, we wish to estimate a function $g(x_t^i)$ that closely predicts $\phi(x_{t+\Delta}^i)$:

$$\omega^* = \underset{\omega}{\operatorname{argmin}} \sum_{i,t} \left|\left| g\left(x_t^i; \omega\right) - \phi\left(x_{t+\Delta}^i\right) \right|\right|_2^2 \qquad (1)$$

where our prediction function $g(\cdot)$ is parameterized by ω.

Although our method is general to any visual representation, in our experiments we train $g(\cdot)$ to predict the popular `fc7` hidden layer of AlexNet [20]. We chose this layer because it empirically obtains state-of-the-art performance on several recognition tasks [33, 5, 50].

3.2. Deep Regression Network

Since we do not require data to be labeled for learning, we can collect massive amounts of training data. In order to use large-scale data, we need a model with a large learning capacity and efficient training algorithm. We propose to use deep regression networks for this task because their model complexity can easily expand and can be trained with large scale data efficiently with stochastic gradient descent.

Our network architecture is five convolutional layers followed by five fully connected layers. The last layer is the output vector, which makes the prediction for the future representation. In training, we use the Euclidean loss in Equation 1 to minimize the distance between our predictions $g(x_t)$ and the representation of the future frame $\phi(x_{t+\Delta})$.

Our choice of architecture is inspired by the recent successes of the AlexNet architecture for visual recognition [20, 51]. Note, however, that our architecture differs by having a regression loss function and three more fully connected layers (because we have more training data than typical supervised problems).

Our method focuses on predicting the future given only a single frame. However, we note that our method can be optionally extended with most sequence models in order to forecast the future given a video clip. In experiments, we explored replacing the last three layers with long short term memory [12] since it is a state-of-the-art sequence model for language tasks [39].

3.3. Regression with Multiple Outputs

Given an image, there are often multiple plausible futures, illustrated in Figure 3. To model scenarios like these, our network should produce multiple outputs given one input. While classification networks naturally produce multiple outputs with confidence values, standard regression networks typically produce only one output vector.

We propose to extend deep regression networks to produce multiple outputs. Suppose that there are K possible output vectors for one input frame. One way to support multiple outputs is to train a mixture of K networks, one for each output. Given input x_t^i, network k will produce one of the outputs $g_k(x_t^i)$.

However, in learning, there are two problems with this approach. Firstly, videos only show one of the possible futures (videos like Figure 3 are rare). Secondly, we do not know to *which* of the K mixtures each frame belongs. We can overcome both problems by treating the mixture assignment for a frame as latent. Let $z_t^i \in \{1, \ldots, K\}$ be a latent

variable indicating this assignment for frame t in video i.

We could train each network independently, however the number of parameters would scale linearly with K, which makes learning difficult and prone to overfitting. To efficiently train our model, we interleave the hidden units between the K networks. We randomly commit each hidden unit to a network with probability $p = \frac{1}{2}$, which controls the amount of sharing between networks. Note that we do this assignment once, and do not change it during learning. In contrast to dropout [37] that uses an ensemble of an exponential number of networks to solve a single task, we use K fixed interleaved networks each tuned for their own task.

3.4. Learning

If we knew the ground truth z for which future frames belong, then we would be able to train our networks with standard methods. Since we do not have access to this information, we can instead treat it as latent during learning. We first initialize z uniformly at random. Then, we alternate between two steps. First, we solve for the network weights w end-to-end using backpropagation assuming z is fixed:

$$\omega^* = \underset{\omega}{\operatorname{argmin}} \sum_{i,t} \left|\left| g_{z_t^i}\left(x_t^i;\omega\right) - \phi\left(x_{t+\Delta}^i\right)\right|\right|_2^2 \quad (2)$$

Then, we re-estimate z using the new network weights:

$$z_t^i = \underset{k\in\{1,\ldots,K\}}{\operatorname{argmin}} \left|\left| g_k\left(x_t^i;\omega\right) - \phi\left(x_{t+\Delta}^i\right)\right|\right|_2^2 \quad (3)$$

We alternate between these two steps until the mixture assignments z and network weights w sufficiently converge. We illustrate this type of loss function in Figure 4.

Although we train our network offline in our experiments, we note our network can be also trained online with streaming videos. Online learning is attractive because the network can continuously learn how to anticipate the future without storing frames. Additionally, the model can adapt in real time to the environment, which may be important for some applications such as robotics.

3.5. Predicting Categories

Since our network uses unlabeled videos to predict an abstract representation in the future, we need a way to attach semantic category labels to our predicted representations. To do this, we use a relatively small set of labeled examples from the target task to indicate the category of interest. As the representation that we predict is the same that is used by state-of-the-art recognition systems, we can then just apply standard recognition algorithms to our predicted representation in order to forecast a category.

We explore two strategies for using recognition algorithms on our forecasted representations. The first strategy uses a visual classifier trained on the standard features (we use `fc7`) from frames containing the category of interest, but applies it on our predicted representation. The second strategy trains the visual classifier on our predicted representations as well. The second strategy has the advantage that it can adapt to the structure in our predictions.

Figure 5: **Unlabeled Video Repository:** We collected more than 600 hours of unlabeled video from YouTube. We show a sample of the frames above. We use this data to train deep networks that predict visual representations in the future.

During inference, our model will predict multiple representations of the future. By applying category classifiers to each predicted representation, we will obtain a distribution for how likely categories are to happen in each future representation. We marginalize over these distributions to obtain the most likely category in the future.

3.6. Implementation Details

Our network architecture consists of 5 convolutional layers followed by 5 fully connected layers. We use ReLU nonlinear activations throughout the network. The convolutional part follows the popular AlexNet architecture [20]. The number of hidden units for each convolutional layer are: 96, 256, 384, 384, 256. There is pooling after layer 1, layer 2, and layer 5. After the convolutional layers, we have 5 fully connected layers each with 4096 hidden units. For multiple output mixtures, the last three fully connected layers have interleaved hidden units. The rest of the layers are tied between networks.

We trained the networks jointly with stochastic gradient descent. We used a Tesla K40 GPU and implemented the network in Caffe [15]. We modified the learning procedure to handle our interleaved hidden units with latent variables. We initialized the first seven layers of the network with the Places-CNN network weights [51], and the remaining layers with Gaussian white noise and the biases to a constant. During learning, we also used dropout [37] with a dropout ratio of $\frac{1}{2}$ on every fully connected layer. We used a fixed learning rate throughout the experiments.

Method	Accuracy
Chance	25.00
Human	71.77 ± 4.26
Identity (Middle of Action)	28.78 ± 6.09
Identity (Start of Action)	36.25 ± 4.93
SVM	35.89 ± 4.39
MMED [11]	34.03 ± 7.03
Nearest Neighbor [48], Off-the-shelf	29.98 ± 4.63
Nearest Neighbor [48], Adapted	34.99 ± 4.73
Linear, Off-the-shelf	32.87 ± 6.14
Linear, Adapted	34.10 ± 4.81
Ours: Deep K=3, Adapted	**43.38** ± 4.70

Table 1: **Single Frame Action Prediction:** Classification accuracy for predicting actions one second before they begin given only a single frame. The standard deviation across cross-validation splits is next to the accuracy. Our results suggest that training deep networks to predict the future from unlabeled video can improve action forecasts.

4. Experiments

In this section, we quantify how well our predicted representations can forecast both actions and objects before they appear. We present several experiments to evaluate the performance of our framework on these tasks.

4.1. Unlabeled Repository

In order to train our network to predict features, we need a large amount of unlabeled video. We downloaded over 600 hours of publicly available videos from YouTube by querying for videos of television shows and movies. The videos generally consist of people performing a large variety of everyday actions, such as eating or driving, as well as interactions with objects and other people. We show a few example frames of these videos in Figure 5. This dataset is unique for its size as, to our knowledge, it is the largest available unlabeled training set for action forecasting.

4.2. Forecasting Actions

To evaluate how well our method can forecast actions, we use the TV Human Interactions dataset [26], which consists of people performing four different actions (hand shake, high five, hug, and kissing). There are a total of 300 videos, with each clip ranging from 1 to 20 seconds. As the videos are collected from television shows, the unconstrained nature of the videos makes this dataset challenging to anticipate people's actions. Since the starting time of actions are annotated, we can run our predictor on the frames before the action begins. We use the provided train-test splits with 25-fold cross validation. We evaluate classification accuracy (averaged across cross validation folds) on making predictions one second before the action has started.

Method	Accuracy
Deep K=1, ActionBank [36]	34.08 ± 6.16
Deep K=3, ActionBank [36]	35.78 ± 6.25
Deep K=1, `fc7`, Off-the-shelf	36.13 ± 6.45
Deep K=3, `fc7`, Off-the-shelf	35.44 ± 5.23
Deep K=1, `fc7`, Adapted	40.01 ± 4.92
Deep K=3, `fc7`, Adapted	**43.38** ± 4.70

Table 2: **Breakdown Diagnostic for Actions:** We show performance of our model with different components turned on. By learning to produce multiple predictions, we are able to improve performance by 3%.

Method	Accuracy
Mean Pooling	28.99 ± 4.93
Mean Pooling, Adapted	32.38 ± 6.20
Max Pooling	36.72 ± 4.95
Max Pooling, Adapted	33.72 ± 6.93
Conv Max Pooling, Adapted	38.50 ± 4.39
Deep Feedforward K=1, Off-the-shelf	36.13 ± 6.45
Deep Feedforward K=1, Adapted	40.01 ± 4.92
Deep LSTM K=1, Off-the-shelf	33.20 ± 5.18
Deep LSTM K=1, Adapted	**43.11** ± 5.20

Table 3: **Multiple Frame Action Prediction:** Classification accuracy for predicting actions one second before they begin given a four second video clip. The standard deviation across cross-validation splits is next to the accuracy, and chance is 25%. By learning recurrent networks that fuse signals across frames, we improve action forecasts.

We train our full network on our unlabeled video repository to predict the future representation one second into the future. Since several works have empirically shown `fc7` features to perform well on diverse visual recognition tasks [33, 5, 50], we predict the `fc7` of the future. To attach semantic meaning to our representation, we use the labeled examples from the training set in [26]. As we make multiple predictions, for evaluation purposes we consider a prediction to be correct only if the ground truth action is the most likely prediction under our model.

We compare the performance of our model against several baselines. **Identity:** One baseline is to assume that the visual features of the present are the same as the future. We can train a classifier to recognize actions using the frames containing the actions, but apply it on the frames before the action. We explored two strategies: training this classifier on the middle frame of the action, and on the starting frame. **SVM:** A fully supervised approach can train a classifier on the frames before the action starts to anticipate the category label in the future. This baseline is able to adapt to contextual signals that may suggest the onset of an action. **MMED:** We can extend the SVM to handle sequential data

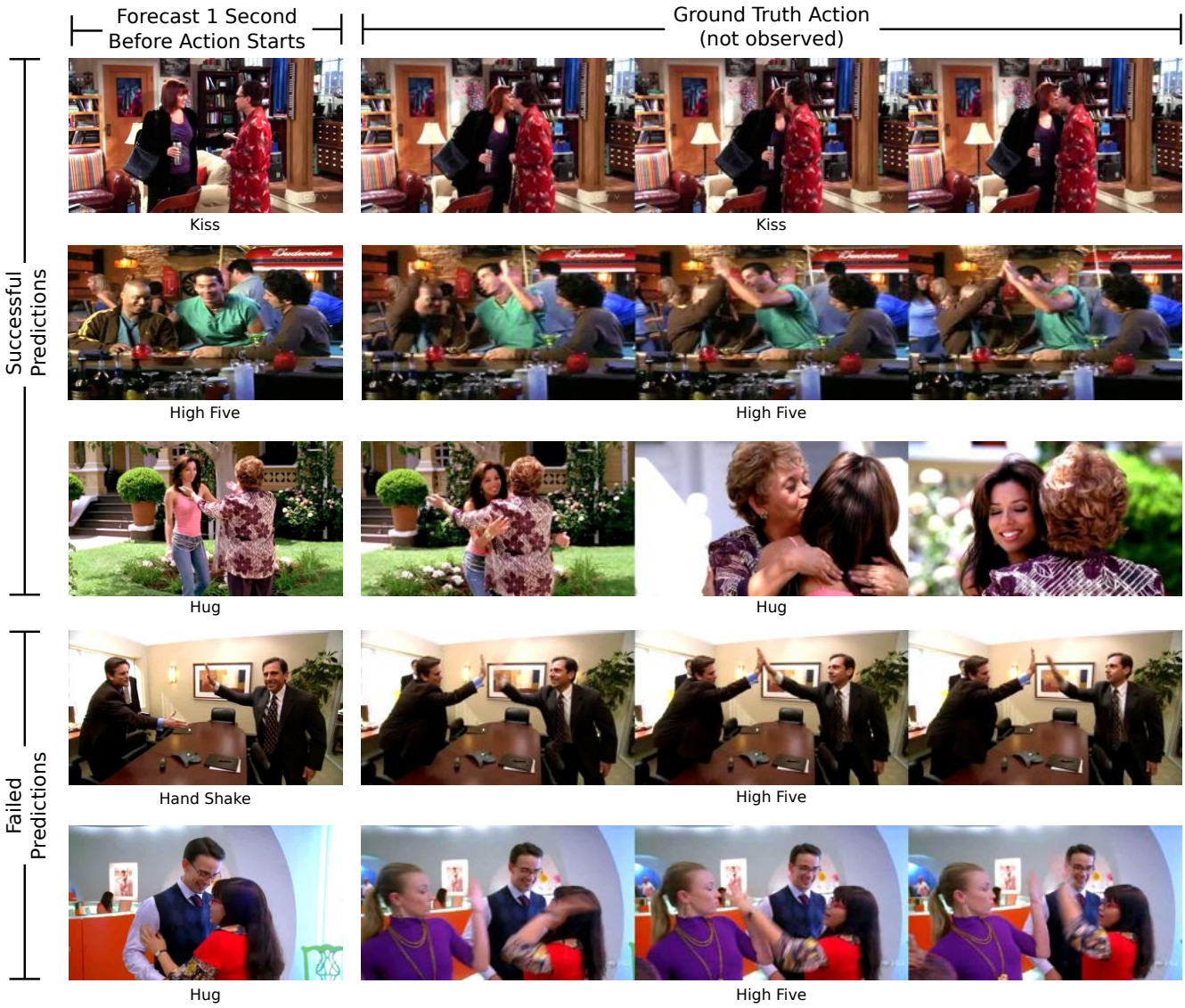

Figure 6: **Forecasting Actions:** We show some examples of our forecasts of actions one second before they begin. The left most column shows the frame before the action begins, and our forecast is below it. The right columns show the ground truth action. Note that our model does not observe the action frames during inference.

in order to make early predictions. We use the code out-of-the-box provided by [11] for this baseline. Note MMED is not designed for this task because it predicts actions as they are starting. Rather, our task is forecasting an action before it even starts. **Nearest Neighbor:** Since we have a large unlabeled repository, one reasonable approach is to search for the nearest neighbor, and use the neighbor's future frame, similar to [48]. **Linear:** Rather than training a deep network to make the prediction, we can also train a linear regression on our unlabeled repository to predict `fc7` in the future. **Adaptation:** We also examine two strategies for training the final classifier. One way is to train the classifier only on frames containing the action and test it on our inferred representations. The second way is to adapt to the predictions by training the visual classifier on the prediction output. Unfortunately, we are unable to compare to Lan et al. [21] because their method uses an annotated bounding box around people. In contrast, our approach is designed to be mostly automatic for situations where bounding boxes are not available.

Table 1 shows the classification accuracy of our full method versus the baselines for predicting the future action one second into the future given only a single frame. Our results show that training deep models to predict future representations with massive amount of unlabeled videos are able to significantly help machines forecast actions, beating

Method	dish	door	utensil	cup	oven	person	soap	tap	tbrush	tpaste	towel	trashc	tv	remote	mean
Chance	1.2	2.8	1.1	2.4	1.6	0.8	1.5	2.1	0.2	0.3	0.6	1.1	0.5	0.3	1.2
Identity	2.6	15.4	2.9	5.0	9.4	6.9	11.5	17.6	1.6	1.0	1.5	6.0	2.0	5.9	6.4
SVM	3.0	8.2	5.2	3.6	8.3	12.0	6.7	11.7	3.5	1.5	4.9	1.3	0.9	4.1	5.3
Scene, Off-the-shelf	3.3	18.5	5.6	3.6	18.2	10.8	9.2	6.8	8.0	8.1	5.1	5.7	2.0	10.3	8.2
Scene, Adapted	4.6	9.1	6.1	5.7	15.4	13.9	5.0	15.7	13.6	3.7	6.5	2.4	1.8	1.7	7.5
Linear, Off-the-shelf	7.5	9.3	7.2	5.9	2.8	1.6	13.6	15.2	3.9	5.6	2.2	2.9	2.3	7.8	6.3
Linear, Adapted	2.8	13.5	3.8	3.6	11.5	11.2	5.8	4.9	5.4	3.3	3.4	1.6	2.1	1.0	5.3
Deep K=1, Off-the-shelf	4.4	17.9	3.0	14.8	11.9	9.6	17.7	15.1	6.3	6.9	5.0	5.0	1.3	8.8	9.1
Deep K=1, Adapted	3.5	11.0	9.0	6.5	16.7	16.4	8.4	22.2	12.4	7.4	5.0	1.9	1.6	0.5	8.7
Deep K=3, Off-the-shelf	4.1	22.2	5.7	16.4	17.5	8.4	19.5	20.6	9.2	5.3	5.6	4.2	8.0	2.6	**10.7**
Deep K=3, Adapted	3.5	14.7	14.2	6.7	14.9	15.8	8.6	29.7	12.6	4.6	10.9	1.8	1.4	1.9	**10.1**

Table 4: **Object Prediction:** We show average precision for forecasting objects five seconds before they appear in egocentric videos. For most categories, our method significantly improve performance. The last column is the mean across all categories.

Figure 7: **Multiple Predictions:** Given a single input frame (left), our model predicts multiple representations in the future that can each be classified into actions (middle). When the future is ambiguous, each network can predict a different representation, allowing for multiple action forecasts. To obtain the most likely future action, we can marginalize the distributions from each network (right).

the best baseline by over 7%. To establish an upper expectation for the performance on this task, we also had 12 human volunteers study the training sets and make predictions on our testing set. Human accuracy is high, but not perfect due to the ambiguous nature of the task, which motivates the need for models to make multiple predictions.

We breakdown the performance of our method in Table 2. If we train our model with only $K = 1$ output, performance drops by a few percent. We also tried to train a deep network to forecast ActionBank [36] in the future instead of `fc7`, which performed worse. Representations are richer than action labels, which provides more constraints during learning that may help build more robust models [10].

We also experimented in Table 3 with forecasting given video clips instead of a static frame by changing the last three layers of our network to long short term memory units [12, 6]. We train and tested this recurrent network on 8 frames sampled at 2 frames per second. We compared this approach with a few baselines. **Mean Pooling:** Given a sequence of frames, we take the *mean* of the `fc7` activations to obtain a 4096 dimensional vector, and train an SVM. **Max Pooling:** Instead of averaging features, we instead take *max* of the `fc7` activations, and train an SVM. **Conv Max Pooling:** Both the previous baselines lose the arrow of time. To capture this directionality, we take the max of `fc7` over a sliding temporal window. Each window produces a 4096 dimensional vector. We then concatenate adjacent windows. **Feedforward Network:** We compare against a network that makes a prediction given only a single frame for $K = 1$. Overall, our results suggest that our model can be improved by observing several frames, likely because the extra frames provide some motion cues.

We qualitatively show some of our predictions in Figure 6, which in many cases are reasonable. For example, our model correctly predicts that a man and woman are about to kiss or hug or that men in a bar will high five. Moreover, the failures are sensible. The second to last row shows a comic scene where one man is about to handshake and the other is about to high five, which our model confuses. In the last row of Figure 6, our model incorrectly forecasts a hug because a third person unexpectedly enters the scene.

Since we learn a mixture of networks, our model will make diverse predictions when the future is ambigious. To analyze this, Figure 7 shows a scene and a distribution of possible future actions. For example, consider the first row where the man and woman are about to embrace, however whether they will kiss or hug is ambiguous. In our model, two of the networks predict a representation where kissing is the most likely future action, but one network predicts a representation where the most likely action is hugging.

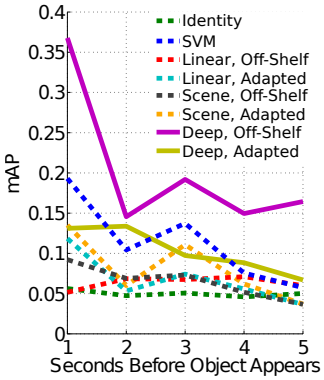

Figure 8: **AP vs Δ:** We plot performance at forecasting an object category (mugs) versus the seconds before they appear. As Δ decreases, performance tends to improve.

4.3. Forecasting Objects

Since our method predicts a visual representation in the future that is generic to several recognition tasks, we wish to understand how well we can anticipate concepts other than actions. We experimented with forecasting objects in egocentric videos five seconds before the object appears. We use the videos from Activities of the Daily Living dataset [30], which is one of the largest in-the-wild datasets of egocentric videos from multiple people. The dataset is very challenging due to the unconstrained nature of the videos, and even recognizing objects in this dataset is difficult [30].

In order to train our deep network on egocentric videos, we reserved three fourths of the dataset as our repository for self-supervised learning. While collecting orders of magnitude more unlabeled egocentric video is easy (just wear a camera and hit record), we use existing datasets. We evaluate on the remaining one fourth videos, performing leave-one-out to learn future object category labels. Since multiple objects can appear in a frame, we evaluate the average precision for forecasting the occurrence of objects five seconds before they appear, averaged over leave-one-out splits.

We compare against baselines that are similar to our action forecasting experiments. However, we add an additional baseline that uses scene features [51] to anticipate objects. Since most objects are strongly correlated with their scene, recognizing the scene may be a good cue for predicting objects. We use an SVM trained on state-of-the-art scene features to create this baseline.

Table 4 shows average precision for our method versus the baselines on forecasting objects five seconds into the future. For the most of the object categories, our model consistently outperforms the baselines at anticipating objects, occasionally with large margins. Moreover, model with multiple outputs improves over a single output network, suggesting that handling ambiguity in learning is important. The adapted and off-the-shelf networks perform similarly to each other in the average. Finally, as shown in Figure 8, performance generally improves as we make predictions closer to the onset of an object, and our model outperforms the baselines for most temporal offsets.

Figure 9: **Forecasting Objects:** We show examples of our high scoring forecasts for objects in egocentric videos. The left most frame is five seconds before the object appears.

We show some of our high scoring predictions in Figure 9. These qualitative results seem to suggest that our model anticipates objects using several different cues. For example, some objects can be predicted largely with scene signals (outdoors is correlated with doors) while other objects require some action understanding to forecast (such as cleaning implies soap or trash cans).

5. Conclusion

We hypothesize that in order to build models that reliably forecast the future, machines will need access the same knowledge that humans accumulate over their lifetimes. We believe abundantly available unlabeled videos are an effective resource we can use to acquire this knowledge. Our hope is that this paper will spur progress on building new methods to anticipate the future with unlabeled video.

The capability for machines to anticipate future events before they begin is a key problem in computer vision that will enable many real-world applications in robotics, recommendation systems, and healthcare. Our insight in this paper has been to capitalize on the temporal regularities of unlabeled video to learn to predict visual representations of the future. By training on truly massive amounts of unlabeled data, we can capture some commonsense knowledge about the world that is valuable for anticipating the future.

Acknowledgements: We thank members of the MIT vision group for predicting the future on our test set. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Tesla K40 GPU used for this research. This work was supported by ONR MURI N000141010933 to AT and the NSF GRFP to CV.

References

[1] C.-Y. Chen and K. Grauman. Watching unlabeled video helps learn new human actions from very few labeled snapshots. *CVPR*, 2013.

[2] X. Chen, A. Shrivastava, and A. Gupta. Neil: Extracting visual knowledge from web data. *ICCV*, 2013.

[3] S. K. Divvala et al. Learning everything about anything: Webly-supervised visual concept learning. *CVPR*, 2014.

[4] C. Doersch, S. Singh, A. Gupta, J. Sivic, and A. A. Efros. What makes paris look like paris? *ACM Trans. Graph.*, 2012.

[5] J. Donahue et al. Decaf: A deep convolutional activation feature for generic visual recognition. *arXiv*, 2013.

[6] J. Donahue et al. Long-term recurrent convolutional networks for visual recognition and description. *arXiv*, 2014.

[7] D. F. Fouhey, V. Delaitre, et al. People watching: Human actions as a cue for single view geometry. *ECCV*, 2012.

[8] D. F. Fouhey and C. L. Zitnick. Predicting object dynamics in scenes. *CVPR*, 2014.

[9] H. Gong, J. Sim, M. Likhachev, and J. Shi. Multi-hypothesis motion planning for visual object tracking. *CVPR*, 2011.

[10] G. E. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. *NIPS*, 2014.

[11] M. Hoai and F. De la Torre. Max-margin early event detectors. *IJCV*, 2014.

[12] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 1997.

[13] D.-A. Huang and K. M. Kitani. Action-reaction: Forecasting the dynamics of human interaction. *ECCV*, 2014.

[14] N. Ikizler-Cinbis, R. G. Cinbis, and S. Sclaroff. Learning actions from the web. *ICCV*, 2009.

[15] Y. Jia, E. Shelhamer, et al. Caffe: Convolutional architecture for fast feature embedding. *arXiv*, 2014.

[16] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. *CVPR*, 2014.

[17] K. M. Kitani, B. D. Ziebart, J. A. Bagnell, and M. Hebert. Activity forecasting. *ECCV*, 2012.

[18] J. F. P. Kooij, N. Schneider, F. Flohr, and D. M. Gavrila. Context-based pedestrian path prediction. *ECCV*, 2014.

[19] H. S. Koppula and A. Saxena. Anticipating human activities using object affordances for reactive robotic response. *RSS*, 2013.

[20] A. Krizhevsky et al. Imagenet classification with deep convolutional neural networks. *NIPS*, 2012.

[21] T. Lan, T.-C. Chen, and S. Savarese. A hierarchical representation for future action prediction. *ECCV*, 2014.

[22] Q. V. Le et al. Building high-level features using large scale unsupervised learning. *ICML*, 2013.

[23] J. Lezama et al. Track to the future: Spatio-temporal video segmentation with long-range motion cues. *CVPR*, 2011.

[24] A. Mahendran and A. Vedaldi. Understanding deep image representations by inverting them. *arXiv*, 2014.

[25] H. Mobahi, R. Collobert, and J. Weston. Deep learning from temporal coherence in video. *ICML*, 2009.

[26] A. Patron-Perez, M. Marszalek, A. Zisserman, and I. Reid. High five: Recognising human interactions in tv shows. *BMVC*, 2010.

[27] M. Pei, Y. Jia, and S.-C. Zhu. Parsing video events with goal inference and intent prediction. *ICCV*, 2011.

[28] L. C. Pickup, Z. Pan, D. Wei, Y. Shih, C. Zhang, A. Zisserman, B. Scholkopf, and W. T. Freeman. Seeing the arrow of time. *CVPR*, 2014.

[29] S. L. Pintea et al. Déjà vu. *ECCV*, 2014.

[30] H. Pirsiavash and D. Ramanan. Detecting activities of daily living in first-person camera views. *CVPR*, 2012.

[31] D. Ramanan, S. Baker, and S. Kakade. Leveraging archival video for building face datasets. *ICCV*, 2007.

[32] M. Ranzato et al. Video (language) modeling: a baseline for generative models of natural videos. *arXiv*, 2014.

[33] A. S. Razavian et al. Cnn features off-the-shelf: an astounding baseline for recognition. *arXiv*, 2014.

[34] O. Russakovsky et al. Imagenet large scale visual recognition challenge. *IJCV*, 2014.

[35] M. Ryoo. Human activity prediction: Early recognition of ongoing activities from streaming videos. *ICCV*, 2011.

[36] S. Sadanand and J. J. Corso. Action bank: A high-level representation of activity in video. *CVPR*, 2012.

[37] N. Srivastava et al. Dropout: A simple way to prevent neural networks from overfitting. *JMLR*, 2014.

[38] N. Srivastava et al. Unsupervised learning of video representations using lstm. *arXiv*, 2015.

[39] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In *NIPS*, 2014.

[40] C. Szegedy, W. Liu, et al. Going deeper with convolutions. *arXiv*, 2014.

[41] A. Torralba, R. Fergus, and W. T. Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *PAMI*, 2008.

[42] C. Vondrick et al. Visualizing object detection features. *arXiv*, 2015.

[43] C. Vondrick, D. Patterson, and D. Ramanan. Efficiently scaling up crowdsourced video annotation. *IJCV*, 2013.

[44] T.-H. Vu, C. Olsson, I. Laptev, A. Oliva, and J. Sivic. Predicting actions from static scenes. *ECCV*, 2014.

[45] J. Walker, A. Gupta, and M. Hebert. Patch to the future: Unsupervised visual prediction. *CVPR*, 2014.

[46] D. Xie, S. Todorovic, and S.-C. Zhu. Inferring" dark matter" and" dark energy" from videos. *ICCV*, 2013.

[47] J. Yuen et al. Labelme video: Building a video database with human annotations. In *ICCV*, 2009.

[48] J. Yuen and A. Torralba. A data-driven approach for event prediction. *ECCV*, 2010.

[49] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *ECCV*. 2014.

[50] S. Zha, F. Luisier, W. Andrews, N. Srivastava, and R. Salakhutdinov. Exploiting image-trained cnn architectures for unconstrained video classification. *arXiv*, 2015.

[51] B. Zhou et al. Learning deep features for scene recognition using places database. *NIPS*, 2014.

[52] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Object detectors emerge in deep scene cnns. *arXiv*, 2014.