Knowledge Distillation for Action Anticipation via Label Smoothing

Guglielmo Camporese*, Pasquale Coscia*, Antonino Furnari†, Giovanni Maria Farinella†, Lamberto Ballan*
*Department of Mathematics "Tullio Levi-Civita", University of Padova, Italy
†Department of Mathematics and Computer Science, University of Catania, Italy

***Abstract*—Human capability to anticipate near future from visual observations and non-verbal cues is essential for developing intelligent systems that need to interact with people. Several research areas, such as human-robot interaction (HRI), assisted living or autonomous driving need to foresee future events to avoid crashes or help people. Egocentric scenarios are classic examples where action anticipation is applied due to their numerous applications. Such challenging task demands to capture and model domain's hidden structure to reduce prediction uncertainty. Since multiple actions may equally occur in the future, we treat action anticipation as a multi-label problem with missing labels extending the concept of label smoothing. This idea resembles the knowledge distillation process since useful information is injected into the model during training. We implement a multi-modal framework based on long short-term memory (LSTM) networks to summarize past observations and make predictions at different time steps. We perform extensive experiments on EPIC-Kitchens and EGTEA Gaze+ datasets including more than 2500 and 100 action classes, respectively. The experiments show that label smoothing systematically improves performance of state-of-the-art models for action anticipation.**

I. Introduction

Human action analysis is a central task in computer vision that has a enormous impact on many applications, such as, video content analysis [1], [2], video surveillance [3], [4], and intelligent transportation [5], [6]. Systems interacting with humans also need the capability to promptly react to the context changes, and plan their actions accordingly. Most previous works focus on the tasks of action recognition [7], [8], [9] or early-action recognition [10], [11], [12], i.e., recognition of an action *after* its observation (happened in the past) or recognition of an *on-going* action from its partial observation (only part of the current action is available). A more challenging task is to predict near future, i.e., to forecast actions that will be performed ahead in time. Predicting future actions before observing the corresponding frames [13], [14] is demanded by many applications which need to anticipate human behaviour. For example, intelligent surveillance systems may support human operators to avoid hazards or assistive robotics may help non-self-sufficient people. Such task requires to analyze significant spatio-temporal variations among actions performed by different people. For this reason, multiple modalities (e.g., appearance and motion) are typically considered to improve the identification of similar actions. Egocentric scenarios provide useful settings to study early-action recognition or action anticipation tasks. Indeed, wearable cameras offer an explicit point-of-view to capture human motion and object interaction.

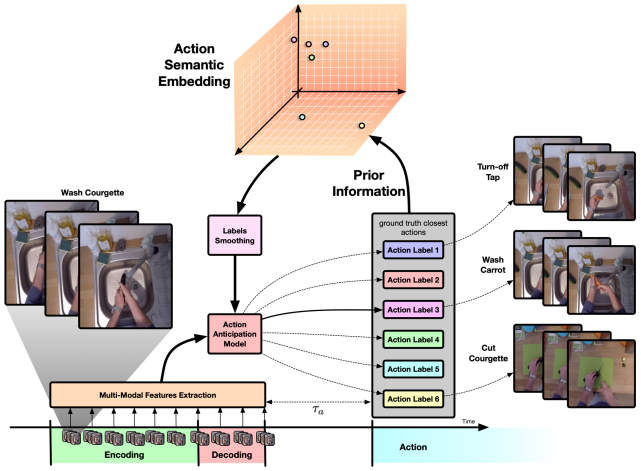

Fig. 1: To anticipate next actions in egocentric scenarios, our framework processes input videos summarizing useful information during the encoding stage. In the decoding stage, our model predicts the next action at different timesteps. Label smoothing techniques are employed to reduce the uncertainty on future predictions.

In this work, we address the problem of anticipating egocentric human actions in an indoor scenario at several time steps. More specifically, we aim to anticipate an action by leveraging its previous video segments. We employ the encoding-decoding scheme of [14]. During the first stage, the model summarizes the video content extracting relevant information to infer the future while, in the second stage, predicts the next action at multiple time-steps (e.g., $1.5s - 1.0s - ... - 0.25s$ before the action to be predicted). Fig. 1 depicts such procedure. We exploit an LSTM-based network to capture temporal correlations among video frames and three different modalities as input representation: *appearance* (RGB), *motion* (optical flow) and *object-based* features.

An important aspect to consider when dealing with human actions is the future's uncertainty since multiple actions may equally occur. For example, the actions "*sprinkle over pumpkin seeds*" and "*sprinkle over sunflower seeds*" may be equally performed when preparing a recipe. Nevertheless, assigning zero-probability to uncorrect yet semantically similar actions

may lead predictions to not capture data structure randomly selecting from a set of plausible targets.

For this reason, to deal with the uncertainty of future predictions, we propose several label smoothing techniques which broaden the set of possible futures and reduce the uncertainty caused by one-hot encoded labels. Label smoothing is introduced in [15] as a form of regularization for classification models since it introduces a positive constant value into wrong classes components of one-hot encoded targets. Compared to the typical knowledge distillation process where soft-labels are provided by an external network (the teacher) to train a smaller network (the student), our model distills knowledge extracted from prior of action labels (e.g., verb-noun). We extensively test our architecture on EPIC-Kitchens and EGTEA Gaze+ datasets which record egocentric scenarios where different people are involved to complete a recipe performing numerous actions and involving different objects within a kitchen. Our results show that label smoothing increases the performance of state-of-the-art models and yields better generalization on test data.

The main contributions of our work are as follows: 1) we generalize the label smoothing idea extrapolating semantic priors from the action labels to capture the multi-modal future component of the action anticipation problem. We show that label smoothing, in this context, can be seen as a knowledge distillation process where the teacher gives semantic prior information for the action anticipation model; 2) we perform extensive experiments on egocentric videos proving that our label smoothing techniques systematically improve results of action anticipation of state-of-the-art models.

II. Related Work

Action anticipation requires the interpretation of the current activity using a number of observations in order to foresee the most likely actions. For this reason, we briefly review three related research areas: *action recognition, early-action recognition* and *action anticipation*.

Action recognition. Action recognition is the task of recognizing the action contained in an observed trimmed video. Classic approaches to action recognition have leveraged hand-designed features, coupled with machine learning algorithms to recognize actions from videos [7], [16], [17]. More recent works have investigated the use of deep learning to obtain representations suitable for action recognition directly from videos in an end-to-end fashion. Among these approaches, a line of works has investigated ways to exploit standard 2D CNNs for action recognition, often relying on optical flow as a mean to represent motion [8], [18], [9], [19], [20], [21]. Other works have focused on the extension of 2D CNNs to 3D CNNs able to process spatio-temporal volumes [2], [22], [23]. Some approaches have used recurrent networks to model the temporal relationships between per-frame observations [24], [25], [14]. All of these works investigate deeply how to represent and leverage the input video but less or even no importance is given to the representation of the action labels.

Early-action recognition. Early action recognition consists in recognizing on-going actions from partial observations of streaming video [12]. Classic works have addressed the task using integral histograms of spatio-temporal features [10], sparse-coding [11], Structured Output SVMs [26], and Sequential Max-Margin event detectors [27]. Another line of research has leveraged the use of LSTMs [28], [29], [30], [14] to account for the sequential nature of the task.

Action anticipation. Action anticipation deals with forecasting actions that will happen in the future. Previous studies have investigated different approaches such as hierarchical representations [31], auto-regressive HMMs [32], regressing future representations [33], encoder-decoder LSTMs [34], and inverse reinforcement learning [35]. Some other approaches proposed to perform long-term predictions only focusing on appearance features [36], [37]. However, differently from this work, very little or even no attention has been payed either to the knowledge distillation of the action semantics or label smoothing for action anticipation.

Knowledge distillation and label smoothing Knowledge distillation [38] is the procedure of transferring the information extracted by a teacher network (with high learning capacity) to a student network (with low learning capacity) in order to allow the latter to reach similar performance. This is usually obtained by training the student via a distillation loss which takes into account both the ground truth and the prediction of the pre-trained teacher. Since this procedure can distill useful information to low learning capacity networks, we perform semantic distillation via label smoothing without employing an external network that supervises the process.

Label smoothing, introduced in [15], is the procedure of softening the distribution of the target labels, reducing the most confident value of the one-hot vector and considering a uniform value for all the zero vector components. Although such procedure improves results for classification problems reducing overfitting, no previous works investigate other design approaches except for the uniform smoothing. In our work, we both generalize this idea and show a systematically improvement to state-of-the-art models.

III. Proposed Approach

Anticipating human actions is essential for developing intelligent systems able to avoid accidents or guide people to correctly perform their actions. Nevertheless, training models on one-hot encoded labels in egocentric videos may led to poor performance when dealing with similar actions. To this end, in the next sections we briefly describe our architecture and then study different label smoothing techniques to address this issue and improve generalization performance.

A. Action Anticipation Architecture

For our experiments, we consider a learning architecture based on recurrent neural networks (RNNs). We process input frames preceding the action that we want to anticipate grouping them into video snippet of length 5. Each video

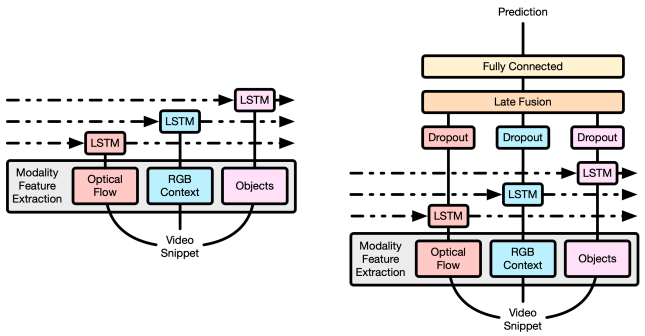

Fig. 2: Encoding (left) and decoding (right) branches of our architecture. The encoding branch processes, using LSTM networks, video snippets of three modalities (optical flow, RGB and objects). The decoding branch simultaneously processes input video snippets and predicts the next action through a fully connected layer with soft-max activation.

snippet is collected every $\alpha = 0.25$ seconds and processed considering three different modalities: *RGB* features computed using a Batch Normalized Inception CNN [39] trained for action recognition, *objects* features computed using Fast-R CNN [40] and *optical flow* computed as in [19], processed through a Batch Normalized Inception CNN trained for action recognition. Our multi-modal architecture processes the above inputs and encompasses two building blocks: an encoder which recognizes and summarises past observations and a decoder which predicts future actions at different anticipation time steps. Fig. 2 depicts the two stages. Firstly, each modality is independently processed by an LSTM layer; then, such streams are merged with late fusion and fed to a fully connected layer to predict the next action.

B. Label Smoothing

Egocentric videos exhibit an inherent uncertainty when dealing with predicting future actions [41]. In fact, given the current state observation of an action there can be multiple, but still plausible, future scenarios that can occur. For this reason, the problem can be reformulated as a multi-label task with missing labels where, from a set of valid future realizations, only one is sampled.

All previous models designed for action anticipation are trained with cross-entropy using one-hot labels, leveraging only one of the possible future scenarios as ground truth. A major drawback of using hard labels is to favour logits of correct classes weakening the importance of other plausible ones. In fact, given the one-hot encoded vector $y^{(k)}$ for a class k, the prediction of the model p and the logits of the model z such that $p(i) = e^{z(i)}/\sum_j e^{z(j)}$, the cross entropy is minimized only if $z(k) \gg z(i)\ \forall i \neq k$. This fact encourages the model to be over-confident about its predictions since during training it tries to focus all the energy on one single logit leading to overfitting and scarce adaptivity [15]. To this purpose, we smooth the target distribution enabling the chance of negative (yet still plausible) classes to be selected. The most common form of label smoothing procedure relies on a uniform positive component added to a set of similar classes without capturing the difference between such actions. To overcome this issue, we design and compare several label smoothing techniques that exploit both semantic and temporal relations among actions. In the following, we firstly model the general idea of label smoothing and then propose multiple techniques to define the corresponding components.

As a form of regularization, [15] introduces the idea of smoothing hard labels by averaging one-hot encoded targets with constant vectors as follows:

$$y^{soft}(i) = (1-\alpha)y(i) + \alpha/K, \tag{1}$$

where $y(i)$ is the one-hot encoding of the i^{th} example, α is the smoothing factor ($0 \leq \alpha \leq 1$) and K represents the number of classes. Since cross entropy is linear w.r.t. its first argument, it can be written as follows:

$$\begin{aligned} CE[y^{soft}, p] &= \sum_i -y^{soft}(i)\log(p(i)) = \\ &= (1-\alpha)CE[y,p] + \alpha CE[1/K, p]. \end{aligned} \tag{2}$$

The optimization based on the above loss can be seen as a distillation knowledge procedure [38] where the teacher randomly predicts the output, i.e., $p^{teacher}(i) = 1/K,\ \forall i$. Hence, the connection with the distillation loss proves that the second term in Eq. (1) can be seen as a prior knowledge, given by an agnostic teacher, for the target y. Taking this into account, we extend the idea of smoothing labels by modeling the second term of Eq. (1), i.e., the prior knowledge of the targets, as follows:

$$y^{soft}(i) = (1-\alpha)y(i) + \alpha\pi(i), \tag{3}$$

where $\pi \in \mathbb{R}^K$ is the prior vector such that $\sum_i \pi(i) = 1$ and $\pi(i) \geq 0\ \forall i$.

Therefore, the resulting cross entropy with soft labels is written as follows:

$$CE[y^{soft}, p] = (1-\alpha)CE[y,p] + \alpha CE[\pi, p] \tag{4}$$

This loss not only penalizes errors related to the correct class but also errors related to the positive entries of the prior. Starting from this formulation, we introduce *Verb-Noun*, *GloVe* and *Temporal* priors for smoothing labels in the knowledge distillation procedure.

Verb-Noun label smoothing. EPIC-KITCHENS [13] contains action labels structured as verbs-noun pairs, like "*cut onion*" or "*dry spoon*". More formally, if we define $\mathcal{A}$ the set of actions, $\mathcal{V}$ the set of verbs, and $\mathcal{N}$ the set of nouns, then an action is represented by a tuple $a = (v, n)$ where $v \in \mathcal{V}$ and $n \in \mathcal{N}$. Let $\mathcal{A}_v(\bar{v})$ the set of actions sharing the same verb $\bar{v}$ and $\mathcal{A}_n(\bar{n})$ the set of actions sharing the same noun $\bar{n}$, defined as follows:

$$\mathcal{A}_v(\bar{v}) = \{(\bar{v}, n) \in \mathcal{A}\ \forall\ n \in \mathcal{N}\}, \tag{5}$$

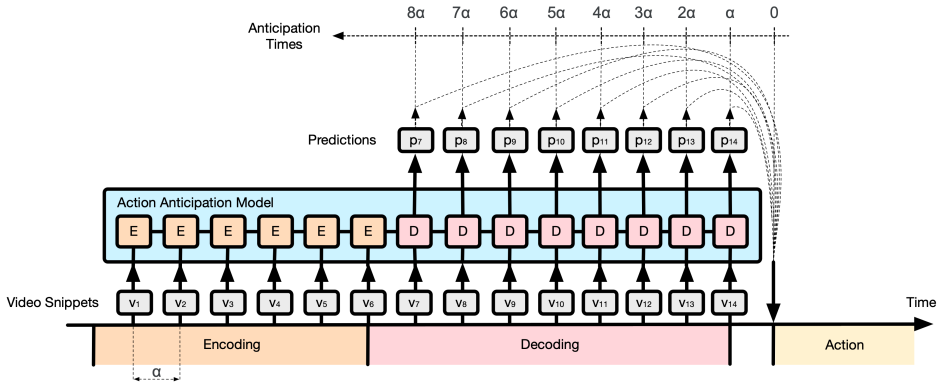

Fig. 3: Action anticipation protocol based on the encoding and decoding stages. We summarize past observations by processing video snippets sampled every $\alpha = 0.25$ seconds in the encoding stage. After 6 steps, we start making predictions every α seconds for 8 steps to anticipate the future action.

$$\mathcal{A}_n(\bar{n}) = \{(v, \bar{n}) \in \mathcal{A} \; \forall \; v \in \mathcal{V}\}, \tag{6}$$

where $\bar{n} \in \mathcal{N}$ and $\bar{v} \in \mathcal{V}$.

We define the prior of the k^{th} ground-truth action class as

$$\pi^{(k)}_{VN}(i) = \mathbb{1}\left[a^{(i)} \in \mathcal{A}_v(v^{(k)}) \cup \mathcal{A}_n(n^{(k)})\right] \frac{1}{C_k}, \tag{7}$$

where $a^{(i)}$ is the i^{th} action, $v^{(k)}$ and $n^{(k)}$ are the verb and the noun of the k^{th} action, $\mathbb{1}[\cdot]$ is the indicator function, and $C_k = |\mathcal{A}_v(v^{(k)})| + |\mathcal{A}_n(n^{(k)})| - 1$ is a normalization term. Using such encoding rule, the cross entropy not only penalizes the error related to the correct class but also the errors with respect to all the other *similar* actions with either the same verb or noun[1].

GloVe based label smoothing. An important aspect to consider when dealing with actions represented by verbs and/or nouns is their semantic meaning. In the *Verb-Noun* label smoothing, we define the prior considering a rough yet still meaningful semantic where actions that share either the same verb or noun are considered similar. To extend this idea, we extrapolate the prior from the word embedding of the actions. One of the most important properties of word embeddings is to put closer words with similar semantic meanings and to move dissimilar ones far away, as opposed to hard labels that cannot capture at all similarity between classes

Using such action representation, we enable the distillation of useful information into the model during training since the cross entropy not only penalizes the error related to the correct class but also the error related to all other similar actions. In order to compute the word embeddings of the actions we use the GloVe model [42] pretrained on the Wikipedia 2014 and Gigaword 5 datasets. We use the GloVe model since it does not only rely on local statistics of words, but also incorporates global statistics to obtain word vectors. Since the model takes as input only single words, we encode the action as follows:

$$\phi^{(k)} = Concat\left[GloVe(v^{(k)}), GloVe(n^{(k)})\right] \tag{8}$$

where ϕ is the obtained action representation of $a^{(k)} = (v^{(k)}, n^{(k)})$ and $GloVe(\cdot) \in \mathbb{R}^{300}$ is the output of the GloVe model. We finally compute the prior probability for smoothing the labels as the similarity between two action representations, which is computed as follows:

$$\pi^{(k)}_{GL}(i) = \frac{|\phi^{(k)T}\phi^{(i)}|}{\sum_j |\phi^{(k)T}\phi^{(j)}|}. \tag{9}$$

Hence, $\pi^{(k)}_{GL}(i)$ in Eq. (9) represents the similarity between the k^{th} and the i^{th} action.

Temporal label smoothing. Some pairs of actions are more likely to occur than other ones. Furthermore, only specific action sequences may be considered valid. For this reason, it could be reasonable to focus on most frequent action sequences since they may reveal possible valid paths in the actions space. In this case, we build the prior probability of their observations by considering two subsequent actions in order to estimate the transition probability from the i^{th} to the k^{th} action as follows:

$$\pi^{(k)}_{TE}(i) = \frac{Occ\left[a^{(i)} \rightarrow a^{(k)}\right]}{\sum_j Occ\left[a^{(j)} \rightarrow a^{(k)}\right]} \tag{10}$$

where $Occ\left[a^{(i)} \rightarrow a^{(k)}\right]$ is the number of times that the i^{th} action is followed by the k^{th} action. Using such representation, we reward both the correct class and most frequent actions that precede the correct class.

IV. Experiments

A. Dataset, Experimental Protocol and Evaluation Measures

Dataset. Our experiments are performed on EPIC-KITCHENS [13] and EGTEA Gaze+ [43] datasets. The

[1]It can be proved that in terms of scalar product two different classes having the same noun or verb and encoded with Verb-Noun label smoothing are closer with respect to classes encoded with hard labels

	Temporal	Uniform	Verb-Noun	GloVe	GLoVe+Verb-Noun
α^*	0.6	0.1	0.45	0.6	0.5

TABLE I: Results of a grid search procedure to detect the best smooth factor α for proposed label smoothing techniques.

former, is a large-scale collection of egocentric videos that contains $39,596$ action annotations divided into 2513 unique actions, 125 verbs, and 352 nouns. We use the same split as [14] producing $23,493$ segments for training and $4,979$ segments for validation. The latter contains $10,325$ action annotations, 19 verbs, 51 nouns and 106 unique actions. In EGTEA Gaze+, methods are evaluated reporting the average performance across the three splits provided by the authors of the dataset [43].

Experimental Protocol. Following [14], our approach uses the protocol depicted in Fig. 3 for anticipating future actions based on encoding and decoding stages. Given the action to predict, we consider the previous 14 timesteps spaced out by 0.25 s. In the encoding phase, our network processes input video snippets of 5 frames to extract relevant information about the past (6 timesteps). During the decoding phase, it predicts the next action at 8 successive timesteps simultaneously processing the corresponding video snippets. This protocol lets us to measure the performance of the network at different anticipation times.

Evaluation Measures. To asses the quality of predictions and compare all methods, we use the Top-k accuracy, i.e. we assume the prediction correct if the label falls into the best top-k predictions. As reported in [41], [44], such measure is one of the most appropriate given the uncertainty of future predictions. More specifically, we use the Top-5 accuracy for methods comparison. For the test set, we also use the Top-1 accuracy, the Macro Average Class Precision and the Macro Average Class Recall. The last two metrics are computed only on many-shot nouns, verbs and actions as explained in [13].

B. Models and Baselines

In the comparative analysis, we exploit the architecture proposed in Sec. III-A employing the different label smoothing techniques defined in Sec. III-B. In our experiments we consider models trained using one-hot vectors (One Hot), uniform smoothing (Smooth Uniform), temporal soft labels (Smooth TE), Verb-Noun soft labels (Smooth VN), GloVe based soft labels (Smooth GL) and GloVe + Verb-Noun soft labels (Smooth GL+VN). We design the last method (Smooth GL+VN) by smoothing hard labels with the average of the two related priors. We also evaluate the effect of the proposed label smoothing techniques on the state-of-the-art approach RU-LSTM [14]. In this case, we choose *GL+VN* as main label smoothing technique.

To implement our architecture, we used TensorFlow 2. Each model is trained for 100 epochs with batch size of 256 using Adam optimizer with learning rate of 0.001. The best model is selected using early stopping by monitoring the Top-5 accuracy at anticipation time $\tau_a = 1$ on the validation set.

Results. For each label smoothing method we select the best smooth factor α^* with a grid search procedure between 0 and 1 and step size $\Delta\alpha = 0.05$. The smooth factors used for the results discussed in this section are shown in Table I.

We notice that our label smoothing procedures such as Verb-Noun, GloVe and Temporal perform well with higher smooth factors ($\alpha^* \simeq 0.5$) compared to Uniform label smoothing ($\alpha^* = 0.1$). This suggests that, when soft labels encodes semantic information, prior information becomes more relevant assuming the same importance of the ground-truth due to the similar smooth factors ($\alpha \simeq 1 - \alpha$). During training, we fix $\alpha = \alpha^*$ for each method and, in order to have a robust performance estimation among trials, we iterate ten times all the experiments.

Table II reports our results on the EPIC-KITCHENS validation set. We notice that all our proposed label smoothing methods improve model performance as compared with training using one-hot encoded labels. Soft labels based on GloVe + Verb-Noun attain best performance improving the Top-5 accuracy from $+1.17\%$ to $+2.13\%$ compared to hard labels. To validate our results, we trained RU-LSTM [14], the state-of-the-art model for action anticipation, considering label smoothing using EPIC-KITCHENS validation and test sets. Similar results are obtained with the RU-LSTM architecture using both one-hot encoding (our baseline) and smoothed labels. We select *GloVe + Verb-Noun* soft labels since they show a higher performance increase. As shown by the experimental results (see Table II), label smoothing improves the Top-5 accuracy for all the anticipation times of a margin from $+\%0.58$ to $+1.59\%$. Such behavior highlights the systematic effect of smoothed labels on the model performance. In Table III we also report the performance of our label smoothing procedures for each model branch. The results show a systematic increase of the accuracy using soft labels for each branch mainly for the GloVe-based method.

Table IV reports our results on the EGTEA Gaze+ test set. All our label smoothing methods improve the baseline trained with one-hot labels. GloVe softening procedure shows the highest accuracy increase ranging from $+1.31\%$ to $+1.89\%$. We also report the performance of RU-LSTM baseline trained using GloVe label smoothing and, in this case, for all the anticipation times, the accuracy is improved by a significant margin, from $+2.34\%$ to $+3.83$. Compared to the EPIC-KITCHENS dataset, the larger improvement could be owed to the smaller dataset showing that label smoothing helps bridge the lack of data in the training set.

In Table V, we report the results obtained considering the test set of EPIC-KITCHENS. Different approaches are considered for the comparison. It is wort noting that the method which consider label smoothing together with RU-LSTM improves the performances obtaining best Top-1 and Top-5 accuracy for anticipating *verb*, *noun* and *action*. Label smoothing helps also to improve Precision for *verb* and

Top-5 Action Accuracy % @ different anticipation times [s]								
	2	1.75	1.5	1.25	1	0.75	0.5	0.25
LSTM One-hot Encoding	27.71 $\pm$0.33	28.69 $\pm$0.34	29.84 $\pm$0.24	30.90 $\pm$0.48	31.93 $\pm$0.45	33.14 $\pm$0.36	34.10 $\pm$0.44	35.16 $\pm$0.35
LSTM TE Smoothing	27.94 $\pm$0.24	28.90 $\pm$0.27	30.06 $\pm$0.24	31.13 $\pm$0.19	32.19 $\pm$0.28	33.21 $\pm$0.36	34.17 $\pm$0.37	35.10 $\pm$0.25
LSTM Uniform Smoothing	28.16 $\pm$0.27	29.06 $\pm$0.26	30.23 $\pm$0.24	31.25 $\pm$0.27	32.41 $\pm$0.28	33.64 $\pm$0.27	34.69 $\pm$0.19	35.75 $\pm$0.13
LSTM VN Smoothing	28.43 $\pm$0.30	29.41 $\pm$0.31	30.68 $\pm$0.28	31.85 $\pm$0.21	33.08 $\pm$0.18	34.35 $\pm$0.19	35.38 $\pm$0.34	36.46 $\pm$0.26
LSTM GL Smoothing	28.61 $\pm$0.26	29.87 $\pm$0.25	30.97 $\pm$0.34	31.94 $\pm$0.34	33.12 $\pm$0.36	34.40 $\pm$0.37	35.51 $\pm$0.37	36.87 $\pm$0.25
LSTM GL+VN Smoothing	**28.88** $\pm$0.20	**29.94** $\pm$0.19	**31.23** $\pm$0.32	**32.54** $\pm$0.31	**33.56** $\pm$0.28	**34.92** $\pm$0.25	**36.06** $\pm$0.33	**37.29** $\pm$0.30
Improv.	+1.17	+1.25	+1.39	+1.64	+1.63	+1.78	+1.96	+2.13
RU-LSTM	29.44	30.73	32.24	33.41	35.32	36.34	37.37	38.39
RU-LSTM GL+VN Smoothing	**30.37**	**31.64**	**33.17**	**34.86**	**35.90**	**37.07**	**38.96**	**39.74**
Improv.	+0.93	+0.91	+0.93	+1.45	+0.58	+0.73	+1.59	+1.35

TABLE II: Top-5 accuracy for action anticipation task at different anticipation time steps on EPIC-KITCHENS validation set. We report results of several label smoothing techniques and show a systematic performance improvement compared to one-hot encoding. These results are confirmed for a simple LSTM-based model and also for the state-of-the-art RU-LSTM [14] model.

Top-5 Action Accuracy % @ 1 [s]					
	One-hot	TE Smoothing	Uniform Smoothing	VN Smoothing	GL Smoothing
LSTM (RGB-only)	29.54	29.60	29.76	29.83	**29.85**
LSTM (Flow-only)	20.43	20.53	20.65	20.69	**20.77**
LSTM (OBJ-only)	29.5	29.64	29.69	29.79	**29.82**
RU-LSTM (RGB-only)	30.83	30.95	31.00	31.05	**31.19**
RU-LSTM (Flow-only)	21.42	21.51	21.51	21.63	**21.73**
RU-LSTM (OBJ-only)	29.89	29.99	30.07	30.04	**30.19**

TABLE III: Top-5 accuracy for action anticipation task at $\tau_a = 1$ second on the EPIK-KITCHENS validation set for each model branch (i.e. RGB, Flow and OBJ). We report results of several label smoothing techniques and show a systematic performance improvement compared to one-hot encoded labels.

Top-5 Action Accuracy % @ different anticipation times [s]								
	2	1.75	1.5	1.25	1	0.75	0.5	0.25
LSTM One-hot Encoding	55.94	58.75	60.94	63.02	65.78	68.04	71.55	73.94
LSTM TE Smoothing	56.13	58.93	61.17	63.24	66.00	68.20	71.75	74.20
LSTM Uniform Smoothing	56.35	59.20	61.37	63.36	66.12	68.41	71.95	74.35
LSTM VN Smoothing	56.85	59.67	61.64	64.03	66.81	68.94	72.56	75.44
LSTM GL Smoothing	**57.34**	**60.11**	**62.25**	**64.42**	**67.21**	**69.56**	**73.02**	**75.83**
Improv.	+1.4	+1.36	+1.31	+1.4	+1.43	+1.52	+1.47	+1.89
RU-LSTM	56.82	59.13	61.42	63.53	66.40	68.41	71.84	74.28
RU-LSTM GL Smoothing	**59.99**	**62.02**	**63.95**	**66.47**	**68.74**	**72.16**	**75.21**	**78.11**
Improv.	+3.17	+2.89	+2.53	+2.94	+2.34	+3.75	+3.37	+3.83

TABLE IV: Top-5 accuracy for action anticipation task at different anticipation time steps on the EGTEA GAZE+ dataset.

obtains comparable results in anticipating *action* and *noun*. Results in terms of Recall point out that label smoothing helps for *noun* anticipation maintaining comparable results for the anticipation of *verb* and *action*.

Finally, Fig. 4 shows some qualitative results obtained with our framework.

V. Conclusion

This study proposed a knowledge distillation procedure which accounts for multi-modality of future predictions in the action anticipation task. We implemented knowledge distillation through label smoothing by relying on priors capturing inter-dependencies between verb and noun labels, past and future actions, as well as semantic relevance between different actions. Experimental results corroborate out findings compared to state-of-the-art models highlighting that label smoothing systematically improves performance when dealing with future uncertainty.

Acknowledgements: Research at the University of Padova is partially supported by MIUR PRIN-2017 PREVUE grant. Authors from the Univ. of Padova gratefully acknowledge the support of NVIDIA for their donation of GPUs, and the UNIPD CAPRI Consortium for its support and access to computing resources. Research at the University of Catania is supported by PTR 2016-2018 Linea 2 of DMI and MIUR AIM - Linea 1 - AIM1893589 - CUP E64118002540007.

References

[1] L. Ballan, M. Bertini, A. Del Bimbo, L. Seidenari, and G. Serra, "Event detection and recognition for semantic annotation of video," *Multimedia tools and applications*, vol. 51, no. 1, pp. 279–302, 2011.

[2] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *IEEE Computer Vision and Pattern Recognition*, Jun 2014.

[3] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 221–231, Jan 2013.

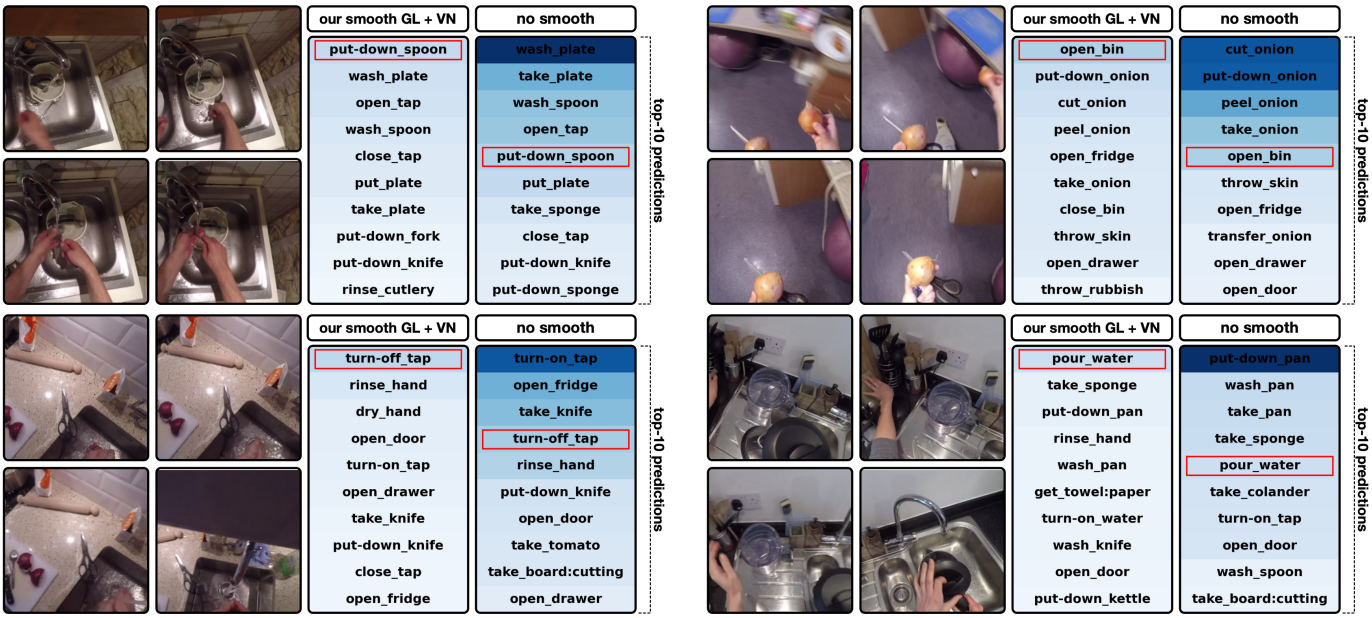

Fig. 4: Qualitative results of our smoothing labels procedure. We show the comparison between the top-10 predictions of our model at $\tau_a = 1$ second trained either with smoothed labels or with one-hot vectors. These examples are from the validation set where both the models predict correctly, under the top-5 accuracy, the upcoming action. The model trained on one-hot labels is more confident and tries to concentrates all the prediction energy on a very restricted set of actions, without capturing the uncertainty of the future. By contrast, smoothing labels shape the prediction distribution considering similar actions to the ground truth.

[4] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in *IEEE Conf. on Computer Vision and Pattern Recognition*, June 2018.

[5] A. Bhattacharyya, M. Fritz, and B. Schiele, "Long-term on-board prediction of people in traffic scenes under uncertainty," in *The IEEE Conf. on Computer Vision and Pattern Recognition*, 2018.

[6] A. I. Maqueda, A. Loquercio, G. Gallego, N. García, and D. Scaramuzza, "Event-based vision meets deep learning on steering prediction for self-driving cars," in *The IEEE Conf. on Computer Vision and Pattern Recognition*, June 2018.

[7] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2008.

[8] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.

[9] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *IEEE Computer Vision and Pattern Recognition*, 2016.

[10] M. S. Ryoo, "Human activity prediction: Early recognition of ongoing activities from streaming videos," in *IEEE Int. Conf. on Computer Vision*, 2011.

[11] Y. Cao, D. Barrett, A. Barbu, S. Narayanaswamy, H. Yu, A. Michaux, Y. Lin, S. Dickinson, J. Mark Siskind, and S. Wang, "Recognize human activities from partially observed videos," in *IEEE Computer Vision and Pattern Recognition*, 2013.

[12] R. De Geest, E. Gavves, A. Ghodrati, Z. Li, C. Snoek, and T. Tuytelaars, "Online action detection," in *European Conf. on Computer Vision*, 2016.

[13] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, and M. Wray, "The epic-kitchens dataset: Collection, challenges and baselines," *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020.

[14] A. Furnari and G. M. Farinella, "Rolling-unrolling lstms for action anticipation from first-person video," *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 2020. [Online]. Available: https://iplab.dmi.unict.it/rulstm

[15] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *IEEE Conf. on Computer Vision and Pattern Recognition*, June 2016.

[16] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *Int. Journal of Computer Vision*, vol. 103, no. 1, pp. 60–79, 2013.

[17] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *IEEE Int. Conf. on Computer Vision*, 2013.

[18] C. Feichtenhofer, A. Pinz, and R. P. Wildes, "Spatiotemporal multiplier networks for video action recognition," in *IEEE Computer Vision and Pattern Recognition*, 2017.

[19] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *European Conf. on Computer Vision*, 2016.

[20] B. Zhou, A. Andonian, A. Oliva, and A. Torralba, "Temporal relational reasoning in videos," in *European Conf. on Computer Vision*, 2018.

[21] J. Lin, C. Gan, and S. Han, "TSM: Temporal shift module for efficient video understanding," in *IEEE Int. Conf. on Computer Vision*, 2019.

[22] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *IEEE Int. Conf. on Computer Vision*, 2015.

[23] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2018.

[24] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2015.

[25] S. Sudhakaran, S. Escalera, and O. Lanz, "LSTA: Long short-term attention for egocentric action recognition," in *IEEE Computer Vision and Pattern Recognition*, 2018.

[26] M. Hoai and F. De la Torre, "Max-margin early event detectors," *Int. Journal of Computer Vision*, vol. 107, no. 2, pp. 191–202, 2014.

[27] D. Huang, S. Yao, Y. Wang, and F. De La Torre, "Sequential max-margin event detectors," in *European Conf. on Computer Vision*, 2014.

[28] M. S. Aliakbarian, F. S. Saleh, M. Salzmann, B. Fernando, L. Petersson, and L. Andersson, "Encouraging LSTMs to anticipate actions very early," in *IEEE Int. Conf. on Computer Vision*, 2017.

		Top-1 Accuracy%			Top-5 Accuracy%			Avg Class Precision%			Avg Class Recall%		
		VERB	NOUN	ACTION	VERB	NOUN	ACTION	VERB	NOUN	ACTION	VERB	NOUN	ACTION
S1	DMR [33]	26.53	10.43	01.27	73.30	28.86	07.17	06.13	04.67	00.33	05.22	05.59	00.47
	2SCNN [13]	29.76	15.15	04.32	76.03	38.56	15.21	13.76	17.19	02.48	07.32	10.72	01.81
	ATSN [13]	31.81	16.22	06.00	76.56	42.15	28.21	23.91	19.13	03.13	09.33	11.93	02.39
	MCE [41]	27.92	16.09	10.76	73.59	39.32	25.28	23.43	17.53	06.05	14.79	11.65	05.11
	ED [34]	29.35	16.07	08.08	74.49	38.83	18.19	18.08	16.37	05.69	13.58	14.62	04.33
	Miech et al. [45]	30.74	16.47	09.74	76.21	42.72	25.44	12.42	16.67	03.67	08.80	12.66	03.85
	RU-LSTM [14]	33.04	22.78	14.39	79.55	50.95	33.73	25.50	24.12	**07.37**	**15.73**	19.81	**07.66**
	RU-LSTM GL+VN Smoothing	**35.04**	**23.03**	**14.43**	**79.56**	**52.90**	**34.99**	**30.65**	**24.29**	06.64	14.85	**20.91**	07.61
S2	DMR [33]	24.79	08.12	00.55	64.76	20.19	04.39	09.18	01.15	00.55	05.39	04.03	00.20
	2SCNN [13]	25.23	09.97	02.29	68.66	27.38	09.35	16.37	06.98	00.85	05.80	06.37	01.14
	ATSN [13]	25.30	10.41	02.39	68.32	29.50	06.63	07.63	08.79	00.80	06.06	06.74	01.07
	MCE [41]	21.27	09.90	05.57	63.33	25.50	15.71	10.02	06.88	01.99	07.68	06.61	02.39
	ED [34]	22.52	07.81	02.65	62.65	21.42	07.57	07.91	05.77	01.35	06.67	05.63	01.38
	Miech et al. [45]	28.37	12.43	07.24	69.96	32.20	19.29	11.62	08.36	02.20	07.80	09.94	03.36
	RU-LSTM [14]	27.01	15.19	08.16	69.55	34.38	21.10	13.69	**09.87**	03.64	**09.21**	11.97	**04.83**
	RU-LSTM GL+VN Smoothing	**29.29**	**15.33**	**08.81**	**70.71**	**36.63**	**21.34**	**14.66**	09.86	**04.48**	08.95	**12.36**	04.78

TABLE V: Results of action anticipation on the test sets of EPIC-Kitchens. The test set is divided into kitchens already seen $S1$ or unseen $S2$ from the model. The table confirms also on the test set that our smoothing labels procedure can improve performances of the state-of-the-art model RU-LSTM.

[29] F. Becattini, T. Uricchio, L. Seidenari, L. Ballan, and A. Del Bimbo, "Am I done? predicting action progress in videos," *ACM Trans. on Multimedia Computing, Communications, and Applications (TOMM)*, vol. in press, 2020.

[30] R. De Geest and T. Tuytelaars, "Modeling temporal structure with lstm for online action detection," in *IEEE Winter Conf. on Applications in Computer Vision*, 2018.

[31] T. Lan, T.-C. Chen, and S. Savarese, "A hierarchical representation for future action prediction," in *European Conf. on Computer Vision*, 2014.

[32] A. Jain, H. S. Koppula, B. Raghavan, S. Soh, and A. Saxena, "Car that knows before you do: Anticipating maneuvers via learning temporal driving models," in *IEEE Int. Conf. on Computer Vision*, 2015.

[33] C. Vondrick, H. Pirsiavash, and A. Torralba, "Anticipating visual representations from unlabeled video," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2016.

[34] J. Gao, Z. Yang, and R. Nevatia, "RED: Reinforced encoder-decoder networks for action anticipation," *British Machine Vision Conf.*, 2017.

[35] K.-H. Zeng, W. B. Shen, D.-A. Huang, M. Sun, and J. C. Niebles, "Visual forecasting by imitating dynamics in natural sequences," in *IEEE Int. Conf. on Computer Vision*, Oct 2017.

[36] T. Mahmud, M. Hasan, and A. K. Roy-Chowdhury, "Joint prediction of activity labels and starting times in untrimmed videos," in *IEEE Int. Conf. on Computer Vision*, Oct 2017.

[37] Y. Abu Farha, A. Richard, and J. Gall, "When will you do what? - anticipating temporal occurrences of activities," in *IEEE Conf. on Computer Vision and Pattern Recognition*, June 2018.

[38] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in *NIPS Deep Learning and Representation Learning Workshop*, 2015.

[39] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Int. Conf. on Machine Learning*, 2015.

[40] R. B. Girshick, "Fast R-CNN," in *IEEE Int. Conf. on Computer Vision*, 2015.

[41] A. Furnari, S. Battiato, and G. Maria Farinella, "Leveraging uncertainty to rethink loss functions and evaluation measures for egocentric action anticipation," in *European Conf. on Computer Vision Workshops*, 2018.

[42] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global vectors for word representation," in *Empirical Methods in Natural Language Processing*, 2014.

[43] Y. Li, M. Liu, and J. M. Rehg, "In the eye of beholder: Joint learning of gaze and actions in first person video," in *ECCV*, 2018.

[44] H. S. Koppula and A. Saxena, "Anticipating human activities using object affordances for reactive robotic response," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 38, no. 1, pp. 14–29, 2016.

[45] A. Miech, I. Laptev, J. Sivic, H. Wang, L. Torresani, and D. Tran, "Leveraging the present to anticipate the future in videos," in *IEEE Conf. on Computer Vision and Pattern Recognition Workshops*, 2019.