Leveraging Uncertainty to Rethink Loss Functions and Evaluation Measures for Egocentric Action Anticipation

Antonino Furnari[0000−0001−6911−0302], Sebastiano Battiato[0000−0001−6127−2470], Giovanni Maria Farinella[0000−0002−6034−0432]

Department of Mathematics and Computer Science - University of Catania
{furnari,battiato,gfarinella}@dmi.unict.it

Abstract. Current action anticipation approaches often neglect the intrinsic uncertainty of future predictions when loss functions or evaluation measures are designed. The uncertainty of future observations is especially relevant in the context of egocentric visual data, which is naturally exposed to a great deal of variability. Considering the problem of egocentric action anticipation, we investigate how loss functions and evaluation measures can be designed to explicitly take into account the natural multi-modality of future events. In particular, we discuss suitable measures to evaluate egocentric action anticipation and study how loss functions can be defined to incorporate the uncertainty arising from the prediction of future events. Experiments performed on the EPIC-KITCHENS dataset show that the proposed loss function allows improving the results of both egocentric action anticipation and recognition methods.

Keywords: egocentric vision, action anticipation, loss functions, first person vision

1 Introduction

Egocentric vision aims at enabling intelligent wearable assistants to understand the user's needs and augment their abilities [15]. Among other tasks to be addressed to allow user behavior understanding from egocentric imagery, the ability to anticipate what is likely to happen in the near future is of great importance. Previous works investigated different egocentric anticipation tasks [39, 26, 29, 24, 28, 37, 7, 11, 4, 25]. Egocentric action anticipation has recently gained attention with the release of the EPIC-KITCHEN dataset and its related challenges [6]. We focus on the egocentric action anticipation challenge, the task of predicting the most likely actions which will be performed by the camera wearer from an egocentric observation of the past.

Humans anticipate future events with natural uncertainty. Consider Fig. 1: what is going to happen after the observations on the left? There are probably more than one likely answers to this question and some answers are clearly

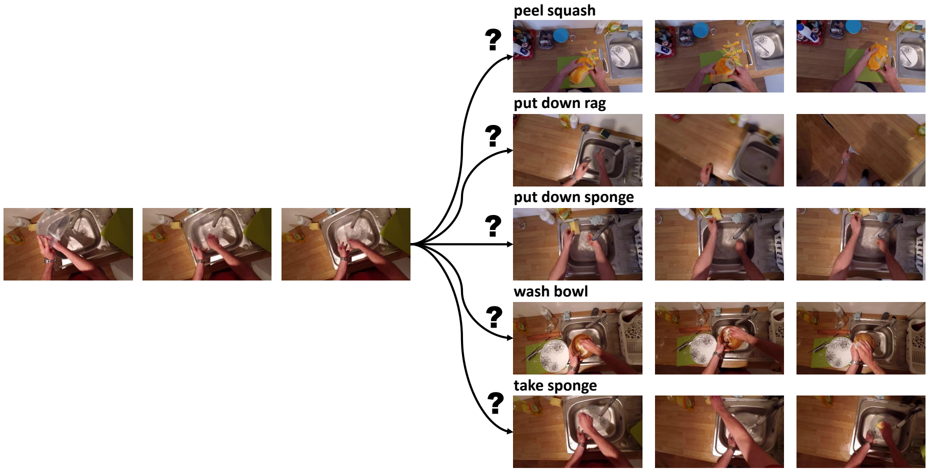

Fig. 1. An observed video sequence (left) along with five possible subsequent action segments (right). Which actions are likely to happen after the observation on the left? Note that some of the actions (e.g., "peel squash" or "put down rag") are less likely to happen than others.

not correct. This simple example highlights the intrinsic multi-modal nature of anticipation tasks[1], i.e., given the observation of the present, multiple predictions about the future are possible.

Even if the anticipation of future events is by nature a multi-modal task, current approaches rarely take into account such uncertainty either when the algorithm is designed or when it is evaluated. One of the main motivations of this lack of exploration is due to the fact that the explicit modeling of the multi-modal dependence between past and future is hard, whereas connecting a past observation to the future action immediately following in a video stream is, on the contrary, straightforward.

In this study, we explore how this multi-modality assumption can be leveraged to design better loss functions and evaluation measures for egocentric action anticipation. We begin by showing that egocentric action anticipation can be regarded to as a special case of multi-label classification with missing labels [36]. We then show that the currently used TOP-1 accuracy is not always appropriate to evaluate action anticipation algorithms, while scores based on TOP-K criteria are more reliable and effective. This leads to the adoption of TOP-K accuracy as a class-agnostic measure and to the definition of TOP-K recall as a class-aware measure. Relying on the common *(verb, noun)* representation of egocentric actions, we investigate how loss functions can be relaxed to penalize less partially correct predictions (i.e., only the *verb* or *noun* is predicted correctly). This is

[1] In this paper, the term "multi-modal" is used to refer to the presence of two or more modes in the distribution of future events. This should not be confused with the multi-modality of the inputs (e.g., images, audio, text, etc.).

done by introducing a Verb-Noun Marginal Cross Entropy (VNMCE) training objective which encourages an action anticipation model to anticipate correct *verbs* or *nouns* alone when correct action anticipation is not possible. The proposed loss is empirically shown to improve the results of both egocentric action anticipation and recognition. Finally, we show that the use of TOP-K losses can improve egocentric action anticipation models. All experiments are performed on the large-scale EPIC-KITCHEN dataset [6], which is a realistic and diverse set of data, with the aim to draw general conclusions which might be useful for future research in the field of egocentric anticipation.

In sum, the contributions of this paper are as follows: 1) we discuss which evaluation measures are appropriate for action anticipation, 2) we introduce a novel loss which improves the training of both egocentric action recognition and anticipation models[2], 3) we investigate the use of TOP-K losses to improve action anticipation results.

2 Related Work

Action Recognition: Our research is related to previous work on action recognition both from third and first person visual data. Wang et al. [33, 34] designed dense trajectories to describe local motion patterns and object appearance. Karpathy et al. [16] evaluated Convolutional Neural Networks (CNN) on large scale video classification. Simonyan et al. [27] designed a Two-Stream CNN (TS-CNN) able to process both motion (optical flow) and appearance (RGB) data for action classification. Feichtenhofer et al. [10] investigated ways to fuse spatio-temorally motion and appearance to improve recognition. Wang et al. [35] proposed Temporal Segment Network (TSN), a framework for video-based action classification. Carreira and Zisserman [5] introduced inflated 3D CNNs to obtain spatio-temporal representation for action recognition.

Egocentric action recognition has also been addressed in literature. Spriggs et al. [30] employed Inertial Measurement Units (IMU) and a wearable camera to segment an egocentric video into action segments. Fathi et al. [8] proposed to model activities, hands and objects jointly. Fathi et al. [9] predicted gaze to recognize activities which require eye-hand coordination. Li et al. [20] benchmarked different egocentric cues for action recognition. Ma et al. [21] designed a CNN architecture to integrate different egocentric cues for the recognition of egocentric actions and activities. Recently, Damen et al. [6] proposed a large-scale dataset of egocentric videos to encourage research on egocentric action recognition.

Anticipation in Third Person Vision: Previous works investigated anticipation from third person visual data. Huang and Kitani [13] explored activity forecasting in the context of dual-agent interactions. Lan et al. [18] proposed a hierarchical representation for human action anticipation. Jain et al. [14] designed a system to anticipate driving maneuvers before they occur by looking at both the

[2] The implementation of the proposed loss is available at the following URL: https://github.com/antoninofurnari/action-anticipation-losses

driver and the road. Walker et al. [32] use variational autoencoders to predict the dense trajectory of pixels from a static image. Koppula and Saxena [17] proposed to leverage object affordances to predict future human actions in order to enable reactive robotic response. Vondrick et al. [31] adapted a deep convolutional network to anticipate multiple future representation from current frames and perform action anticipation. Gao et al. [12] proposed an encoder-decoder LSTM model to anticipate future representations and actions. Mahmud et al. [23] investigated the prediction of labels and starting times of future activities. Abu et al. [1] designed two methods to predict future actions and their durations.

While these methods concentrate on third person vision, we focus on egocentric action anticipation, explicitly designing loss functions to exploit the *(verb, noun)* representation of actions.

Anticipation in First Person Vision: Researchers have considered different egocentric anticipation tasks. Zhou et al. [39] studied computational approaches to recover the correct order of egocentric video segments. Ryoo et al. [26] proposed methods to anticipate human-robot interactions from the robotic perspective. Soran et al. [29] designed a system capable of inferring whether the next action likely to be performed is correct in a given work-flow. Park et al. [24] addressed the prediction of future locations from egocentric video. Zhang et al. [37] proposed to anticipate gaze in future frames. Furnari et al. [11] investigated methods to anticipate user-object interactions. Chenyou et al. [7] designed a method to forecast the position of hands and objects in future frames. Bokhari et al. [4] and Rhinehart and Kitani [25] proposed methods to predict future activities from first person video.

Differently from the aforementioned works, we seize the action anticipation challenge proposed in [6], and investigate the design of evaluation measures and loss functions, with the aim to draw general conclusions useful for future research.

Explicit Modeling of Uncertainty for Anticipation and Early Recognition: Some researchers considered the explicit modeling of uncertainty of future predictions either in the design of the algorithms or in the definition of loss functions. In particular, Vondrick et al. [31] designed a deep multi-modal regressor to allow multiple future predictions. Walker et al. [32] proposed a generative framework which, given a static input image, outputs the space of possible future motions. Rhinehart and Kitani [25] show that explicitly incorporating goal uncertainty improves the results of their method. Park et al. [24] design methods for future localization which find multiple hypotheses of future trajectories at test time. Aliakbarian et al. [2] design a new loss function for early action recognition which softens the penalization of false positives when the action has been only partially observed. Similarly, Ma et al. [22] address early action detection constraining predicted action scores to increase monotonically over time.

Although the aforementioned works considered the inclusion of uncertainty in different ways, action anticipation algorithms have usually been evaluated and compared using standard evaluation measures based on TOP-1 accuracy. Moreover, most of the considered works did not consider the egocentric point of view. In this work, we investigate loss functions and evaluation measures which consider uncertainty of future actions in egocentric vision.

Multi-Label Classification and TOP-K Losses: Past literature investigated multi-label classification [38], which arises when a given instance can be naturally assigned to more than one label. Moreover, in certain cases, some labels can be missing in the training set, which yields the problem of multi-label learning with missing labels [36]. Interestingly, Lapin et al. [19] noted that class ambiguity easily arises in single-label datasets with a large amount of classes. For instance, an image labeled as "Mountain" could be correctly classified as belonging to class "Chalet". In such cases, evaluating classification algorithms using TOP-K accuracy is a natural option. To allow training algorithms to incorporate uncertainty during training and produce better TOP-K results, the authors of [19] define and evaluate a series of TOP-K losses. With similar motivations, Berrada et al. [3] proposed a smooth TOP-K SVM loss specifically designed for training deep network classifiers such as CNNs.

3 Action Anticipation as Multi-Label Learning with Missing Labels

As previously discussed, the prediction of future events is by nature multi-modal, which implies that multiple future actions can naturally follow a given observation. If we could label each observation with its potential set of future actions, it would come clear that action anticipation can be seen as a multi-label learning problem [38]. For instance, the observation on the left of Fig. 1 could be labeled correctly with the following future actions: "wash bowl", "take sponge", "turn-off tap". We should note that train and test data for action anticipation is generally collected by exploiting datasets labeled for action recognition. This is done by associating a video segment (the observation of the past) with the labeled action following in the video of origin. Therefore, while multiple actions can naturally follow a given observation, we systematically observe just one of the possible actions happening. However, if the dataset is large enough, we can expect to find a similar observation followed by a different, possible action. This makes action anticipation an extreme case of multi-label learning with missing labels [36], where each observation is assigned to a single label drawn from the set of possible future actions.

Since action anticipation can be seen as a multi-label learning task, it would be natural to evaluate anticipation algorithms using standard multi-label classification measures such as Example-Based precision and recall [38]. In particular, in a multi-label classification scenario, Example-Based precision estimates the fraction of correctly retrieved labels among all predicted labels for each example. The per-example scores are averaged over the whole test set to obtain a single evaluation score. Example-Based precision is defined as follows:

$$Precision_{EB}(\{Y_i\}_i, \{\hat{Y}_i\}_i) = \frac{1}{p}\sum_{i=1}^{p}\frac{|Y_i \cap \hat{Y}_i|}{|\hat{Y}_i|} \tag{1}$$

where x_i is the i^{th} example in the test set, Y_i is the set of ground truth labels associated to example x_i, $\hat{Y}_i$ is the set of labels predicted for example x_i and p is the total number of examples in the test set. Similarly, Example-Based recall measures the fraction of correctly retrieved labels among the ground truth labels of each example. Example-Based recall is defined as follows:

$$Recall_{EB}(\{Y_i\}_i, \{\hat{Y}_i\}_i) = \frac{1}{p}\sum_{i=1}^{p}\frac{|Y_i \cap \hat{Y}_i|}{|Y_i|} \tag{2}$$

The reader is referred to [38] for a review of multi-label classification measures.

3.1 TOP-K Accuracy as a Class-Agnostic Measure

Since a given observation is associated to only one of the possible ground truth labels, there is no direct way to measure Example-Based recall. On the contrary, we can approximate Example-Based precision by allowing the algorithm to predict multiple labels (e.g., by choosing the K highest scored predictions) and checking that the single ground truth label is among the K predicted ones. Please note that this evaluation criterion corresponds to TOP-K accuracy. Although the approximation of Example-Based precision by TOP-K accuracy is hindered by the limitedness of the test set (i.e., only one of the possible labels are available), in the following we show that, under ideal circumstances, TOP-K accuracy perfectly recovers Example-Based precision, whereas TOP-1 accuracy tends to underestimates it.

Let $\mathcal{S} = \{(x_i, Y_i)\}_i$ be a set of multi-label examples. Let $\hat{\mathcal{S}} = \{(x_j, y_j)\}_j$ be the corresponding set of single-label data where each multi-label observation $(x_i, \{y_{i,1}, y_{i,2}, \ldots, y_{i,n}\})$ has been replaced by n single-label observations $S_i = \{(x_i, y_{i,l})\}_l$, i.e., $\hat{\mathcal{S}} = \cup_i^p S_i$. We refer to S_i as the "expanded set" of the example (x_i, Y_i) and $\hat{\mathcal{S}}$ as the "expanded dataset" of $\mathcal{S}$. Let also assume that the cardinality of Y_i is fixed and equal to K for each example x_i, i.e., $|Y_i| = K \ \forall i$ and that the model predicts exactly K labels for each observation $x_i \in \mathcal{S}$, i.e., $|\hat{Y}_i| = K \ \forall i$. The TOP-K accuracy on the test set $\hat{\mathcal{S}}$ is computed as:

$$TOP_K(\{y_j\}_j, \{\hat{Y}_j\}_j) = \frac{1}{K \cdot p}\sum_{j=1}^{K \cdot p}[y_j \in \hat{Y}_j] =$$

$$\frac{1}{p}\sum_{i=1}^{p}\frac{|Y_i \cap \hat{Y}_i|}{K} = Precision_{EB}(\{Y_i\}_i, \{\hat{Y}_i\}_i) \tag{3}$$

where $\{y_j\}_j$ is the set of ground truth labels contained in $\hat{\mathcal{S}}$, $\hat{Y}_j$ is the set of labels predicted for sample x_j, $[\cdot]$ denotes the Iverson bracket and $K \cdot p$ is the product between the number of predicted labels K and the number of examples in the dataset p. It should be noted that, while TOP-K accuracy recovers Example-Based precision under the considered ideal conditions, standard TOP-1 accuracy tends to underestimate it, as it is illustrated in the example in Fig.2.

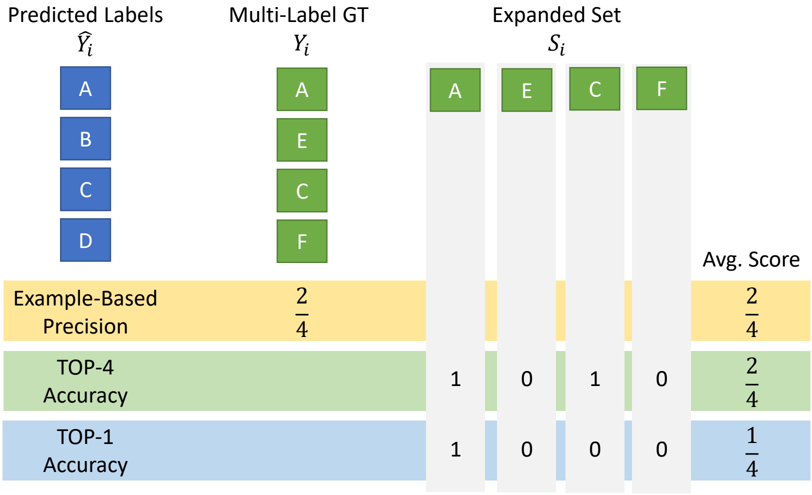

Fig. 2. An example of the differences between Example-Based precision, TOP-K accuracy and TOP-1 accuracy. The illustration reports the 4 labels $\hat{Y}_i$ predicted by the model for a given sample x_i, along with the corresponding set of labels Y_i and the labels of its expanded set S_i. As shown in the example, the TOP-4 accuracy computed over the expanded set of x_i recovers the Example-Based precision related to x_i, while the TOP-1 score underestimates it.

The ideal conditions considered in the previous paragraph may seem to strict for real scenarios. To assess the behavior of the considered measures in a more complex scenario, we performed the following simple experiment. We generated a synthetic multi-label dataset $\mathcal{S}$ of $1,000$ examples. Each example contained in average 5 labels drawn from 50 classes [3]. The expanded dataset $\hat{\mathcal{S}}$ is hence computed from $\mathcal{S}$. To obtain a realistic set of single-label dataset, we drop each sample from $\hat{\mathcal{S}}$ with probability $\frac{1}{2}$. We trained multiple instances of an SVM classifier with an RBF kernel and different choices of the γ parameter to predict multiple labels for each example. Each classifier has been evaluated on the synthetic set $\mathcal{S}$ using Example-Based precision and on its expanded counterpart $\hat{\mathcal{S}}$ using TOP-1 and TOP-5 accuracy. Table 1 reports the results of the experiment. Along the values of each evaluation measure, we also report the induced rank in parenthesis. As can be noted, TOP-5 accuracy can effectively recover the rank induced by Example-Based precision, even if such measure is underestimated due to the non-ideal conditions introduced by the dataset. On the contrary, TOP-1 accuracy induces a different ranking of the algorithms, which points out that such measure is not always appropriate to evaluate multi-label algorithms.

[3] The dataset has been generated using the *make_multilabel_classification* function from the *scikit-learn* library.

Table 1. Performance measures for different multi-label SVMs along with the induced ranks (in parenthesis). As can be noted, the rank induced by TOP-5 accuracy is coherent with the rank induced by Example-Based precision, while TOP-1 accuracy induces a different rank.

γ	$Precision_{EB}\%$	**TOP-5%**	**TOP-1%**
0.10	100.0 (1)	82.14 (1)	20.48 (2)
0.09	100.0 (2)	81.23 (2)	20.71 (1)
0.08	99.80 (3)	79.60 (3)	20.48 (3)
0.07	99.50 (4)	75.79 (4)	20.08 (5)
0.06	98.40 (5)	70.04 (5)	20.20 (4)
0.05	93.80 (6)	58.65 (6)	19.40 (6)
0.04	81.50 (7)	41.67 (7)	16.70 (7)
0.03	51.00 (8)	24.21 (8)	11.39 (8)
0.02	21.80 (9)	13.76 (9)	06.03 (9)
0.01	03.40 (10)	08.97 (10)	02.66 (10)

3.2 TOP-K Recall as a Class-Aware Measure

TOP-K accuracy can be used to the measure the overall performance of an action anticipation method. However, when the dataset is unbalanced, it is often useful to refer to class-aware measures such as per-class precision and recall. Per-class precision is not easy to measure in the case of multi-label learning with missing labels. Indeed, it is not possible to assess if a predicted label which is not in the available set of ground truth labels is correct or not (it might be one of the missing labels). On the contrary, it is much more straightforward to assess per-class recall, i.e., the fraction of cases in which the ground truth class is the list of the K predicted labels. We refer to this measure as TOP-K recall and define it as follows:

$$REC_K^c(\{y_j\}_j, \{\hat{Y}_j\}_j) = \frac{1}{p_c}\sum_{j=1}^{p_c}[y_j \in \hat{Y}_i \wedge y_j = c] \qquad (4)$$

where c denotes the class with respect to which TOP-K recall is computed and $p_c = \sum_j^p [y_j = p]$ is the number of examples belonging to class c.

4 Loss Functions for Egocentric Action Anticipation

As discussed in the previous section, action anticipation algorithms should be able to associate a single observation to a set of possible future actions. However, standard loss functions employed for classification tasks encourage the model to predict a large score for the ground truth class and small scores for all other classes. We explore two different ways to relax this constraint and improve the quality of action anticipation predictions. Specifically, we introduce a novel *verb-noun* marginal cross entropy loss in Section 4.1 and summarize the relevance of TOP-K loss in Section 4.2.

4.1 Verb-Noun Marginal Cross Entropy Loss

Egocentric actions are generally represented as *(verb, noun)* pairs [6]. However, directly anticipating *(verb, noun)* pairs, can be difficult for the following reasons: 1) future actions can be ambiguous and hence anticipating the correct *(verb, noun)* pair can be much more difficult than anticipating the correct *verb* or *noun* alone; 2) egocentric data collected in a natural way can present thousands of unique *(verb, noun)* pairs, the majority of which appear just a few times [6]. Standard classification loss functions would force anticipation algorithms to associate a given observation with the related *(verb, noun)* pair, ignoring for instance that the same observation could be associated to the same *noun* but a different *verb*. To mitigate this effect, previous works proposed to predict *verb* and *noun* separately [6]. However, such approach moves the focus away from *(verb, noun)* pairs and might encourage suboptimal action anticipations as we show in the experiments. We propose a novel loss function which, while maintaining the focus on actions, allows to leverage the uncertainty offered by the *(verb, noun)* representation.

Let $\mathcal{V}$ be the set of *verbs*, $\mathcal{N}$ the set of nouns and $\mathcal{A} \subseteq \mathcal{V} \times \mathcal{N}$ the set of actions. Note that some of the *(verb, noun)* pairs might be not possible, in which case $\mathcal{A} \subset \mathcal{V} \times \mathcal{N}$. Given a *verb* $\overline{v} \in \mathcal{V}$, let $\mathcal{A}_\mathcal{V}(\overline{v})$ be the set of actions including *verb* $\overline{v}$, i.e., $\mathcal{A}_\mathcal{V}(\overline{v}) = \{(v,n) \in \mathcal{A} \mid v = \overline{v}\}$. Similarly, given a *noun* $\overline{n} \in \mathcal{N}$, let $\mathcal{A}_\mathcal{N}(\overline{n}) = \{(v,n) \in \mathcal{A} \mid n = \overline{n}\}$. Let $p(a|x_i)$ be the posterior probability distribution over the set of actions $a = (v,n) \in \mathcal{A}$ given the observation x_i. The posterior probability distributions for *verbs* and *nouns* can be obtained by marginalizing:

$$p(v|x_i) = \sum_{a \in \mathcal{A}_\mathcal{V}(v)} p(a|x_i), \quad p(n|x_i) = \sum_{a \in \mathcal{A}_\mathcal{N}(n)} p(a|x_i) \; . \tag{5}$$

We formulate Verb-Noun Marginal Cross Entropy Loss (VNMCE) for observation x_i as the sum of the Cross Entropy loss computed with respect to the three posterior probability distributions $p(a_i|x_i)$, $p(v_i|x_i)$, $p(n_i|x_i)$:

$$VNMCE(x_i, a_i = (v_i, n_i)) = -\log(p(a_i|x_i)) - \log(p(v_i|x_i)) - \log(p(n_i|x_i)) \tag{6}$$

where $a_i = (v_i, n_i)$ is the ground truth action composed by *verb* v_i and *noun* n_i. We note that:

$$-\log(p(a_i|x_i)) = -\log\Big(\frac{\exp(s^i_{a_i})}{\sum_{a \in \mathcal{A}} \exp(s^i_a)}\Big) = -s^i_{a_i} + \log\Big(\sum_{a \in \mathcal{A}} \exp(s^i_a)\Big) \tag{7}$$

where s^i is the vector of action class scores produced by the model for observation x_i and s^i_a is the score predicted for class a. Analogously, and applying Eq. (5):

$$-\log(p(v_i|x_i)) = -\log\Big(\frac{\sum_{a \in \mathcal{A}_\mathcal{V}(v_i)} \exp(s^i_a)}{\sum_{v \in \mathcal{V}} \sum_{a \in \mathcal{A}_\mathcal{V}(v)} \exp\left(s^i_a\right)}\Big) =$$

$$-\log\Big(\sum_{a \in \mathcal{A}_\mathcal{V}(v_i)} \exp(s^i_a)\Big) + \log\Big(\sum_{a \in \mathcal{A}} \exp(s^i_a)\Big) \; . \tag{8}$$

Similarly for nouns:

$$-\log(p(n_i|x_i)) = -\log\Big(\sum_{a\in\mathcal{A}_\mathcal{N}(n_i)} \exp(s_a^i)\Big) + \log\Big(\sum_{a\in\mathcal{A}} \exp(s_a^i)\Big)\ . \qquad (9)$$

Using Eq. (7)-(9), the VNMCE loss can be re-written as:

$$VNMCE(x_i, a_i) = 3\log\Big(\sum_{a\in\mathcal{A}} \exp(s_a^i)\Big) - s_{a_i}^i$$
$$-\log\Big(\sum_{a\in\mathcal{A}_\mathcal{V}(v_i)} \exp(s_a^i)\Big) - \log\Big(\sum_{a\in\mathcal{A}_\mathcal{N}(n_i)} \exp(s_a^i)\Big)\ . \qquad (10)$$

Note that the proposed $VNMCE$ loss leverages the assumption that verb and noun *are not* conditionally independent with respect to the input sample x, and hence $p((v_i, n_i)|x_i) \neq p(v_i|x_i)p(n_i|x_I)$. In the following sections, we evaluate the proposed loss with respect to standard Cross Entropy Loss in the tasks of action anticipation and recognition.

4.2 TOP-K Losses

As discussed in Section 3, the TOP-1 accuracy is not always suitable to evaluate anticipation methods. However, standard loss functions for classification, such as the cross entropy loss, are designed to penalize all predictions which do not score the ground truth class in the first position, hence forcing the model to concentrate on a single future class for each sample. It is hence natural to exploit loss functions targeted to the optimization of TOP-K scores such as the Truncated TOP-K Entropy Loss proposed in [19] and the Smooth TOP-K SVM loss proposed in [3]. Differently from standard Cross Entropy loss, TOP-K losses are designed to produce a small error whenever the correct class is ranked among the TOP-K predictions. We considered this class of loss functions in our study to point out the relevancy of this aspect.

5 Experimental Settings

Dataset: We perform experiments on the EPIC-KITCHENS dataset [6] to assess the performance of the considered loss functions. Since only the training annotations of the EPIC-KITCHENS dataset are available for the challenge, we randomly split the set of training videos into three parts and consider two folds for training and the remaining fold for testing. The considered split consists of $19,452$ training action annotations, $9,018$ testing action annotations, 2521 different action classes, 125 *verbs* and 352 *nouns*.

Action Anticipation and Classification Baselines: We use the investigated loss functions to train a Temporal Segment Network (TSN) [35] for anticipation and classification following the baselines in [6]. In particular, for action anticipation [6], given an action segment $A_i = [t_{s_i}, t_{e_i}]$, where t_{s_i} and t_{e_i} denote the starting and ending times of the action segment A_i, we train the TSN model to predict the *action/verb/noun* label related to action segment A_i by observing the τ_o long video segment preceding the action start time t_{s_i} by τ_a, that is $[t_{s_i} - (\tau_a + \tau_o), t_{s_i} - \tau_a]$. We follow the settings of [6] and set both the anticipation and observation time to $1s$: $\tau_a = 1s, \tau_o = 1s$. All models are trained for 160 epochs with a starting learning rate equal to 0.001. The learning rate is decreased by a factor of 10 after 80 epochs. At the end of the training, we selected the iteration reporting the best performance. In particular, we selected the best iteration using the TOP-1 accuracy in the case of classification. In the case of anticipation we use TOP-5 accuracy for losses not based on TOP-K criteria and TOP-K accuracy in the case of TOP-K losses. RGB and Flow predictions are fused using weights 0.6 and 0.4 respectively. Testing is performed by averaging the class scores predicted for the center crop of 25 temporal segment sampled from each observation.

Compared Methods We compared the following methods:

- *VN-CE [6]*: the model predicts the posterior probability distributions of *verbs* and *nouns* $p(v|x_i)$, and $p(n|x_i)$ independently. Actions are anticipated by assuming *verbs* and *nouns* to be independent and computing the probability distribution of actions as $p(a = (v, n)|x_i) = p(v|x_i)p(n|x_i)$. The loss function used to train the model is the sum of the Cross Entropy Loss (CE) function computed with respect to *verbs* and *nouns*;
- *A-CE*: the model predicts the posterior probability distribution of actions $p(v|x_i)$ directly. It is trained using Cross Entropy (CE) loss;
- *VNMCE*: action anticipation TSN (same as A-CE) trained using the loss proposed in Eq. (10);
- *TE-TOP3 [19]*: action anticipation TSN trained using the Truncated TOP-K Entropy Loss proposed in [19] with $K = 3$;
- *TE-TOP5 [19]*: same as TE-TOP3 with $K = 5$;
- *SVM-TOP3 [3]*: action anticipation TSN trained using the Smooth TOP-K SVM loss proposed in [3] with $K = 3$;
- *SVM-TOP5 [3]*: same as SVM-TOP3 with $K = 5$;
- *VNCME+T3*: action anticipation TSN trained combining the loss proposed in Eq. (10) and the Truncated TOP-K Entropy Loss proposed in [19] with $K = 3$. TOP-K truncation is only applied to the part of the loss in Eq. (10) dealing with actions;
- *VNCME+T5*: same as VNCME+T3 but with $K = 5$.

6 Results

Table 2 reports the results of the TSN action anticipation baseline trained using different loss functions. TOP-K recalls are averaged over the many shot sets of *verbs*, *nouns* and *actions* provided by [6]. For each method, we evaluate the ability to predict *verbs*, *nouns* and *actions*. For all methods except VN-CE,

Table 2. Action anticipation results of the investigated methods according to different evaluation measures. Best per-column results are reported in bold for each section of the table. Global per-column best results are underlined.

	TOP-1 Accuracy%			TOP-3 Accuracy%			TOP-5 Accuracy%			Avg. TOP-3 Recall%			Avg. TOP-5 Recall%		
LOSS	VERB	NOUN	ACTION	VERB	NOUN	ACTION	VERB	NOUN	ACTION	VERB	NOUN	ACTION	VERB	NOUN	ACTION
VN-CE [6]	**31.77**	**15.81**	05.79	**66.01**	30.20	12.64	**77.67**	**39.50**	17.31	22.55	25.26	**05.45**	34.05	**34.50**	**07.73**
A-CE	24.86	15.01	**09.89**	59.16	29.92	19.37	74.23	38.29	**25.40**	**29.35**	23.87	03.81	**41.40**	31.38	05.52
VNMCE	26.61	15.49	09.74	60.32	**30.82**	**19.61**	73.56	38.91	25.14	27.54	**25.57**	04.45	38.01	34.21	05.34
TE-TOP3 [19]	24.64	15.71	10.59	58.86	30.72	20.06	73.53	39.54	25.25	32.09	**27.85**	**05.06**	**44.14**	**36.69**	**06.14**
TE-TOP5 [19]	23.54	14.75	09.68	**59.46**	**31.41**	20.12	**73.75**	**40.10**	**25.74**	**32.38**	27.52	03.76	43.77	36.38	05.79
SVM-TOP3 [3]	**25.65**	**15.99**	**11.09**	58.87	30.89	**20.51**	72.70	38.41	25.42	31.62	27.75	03.61	41.90	34.69	5.32
SVM-TOP5 [3]	25.01	15.42	10.47	56.50	29.27	19.28	69.17	36.66	24.46	30.72	25.93	03.68	40.27	32.69	05.23
VNMCE+T3	**27.63**	**15.77**	10.39	**61.11**	30.42	19.99	74.05	**39.18**	25.95	**31.38**	25.57	**05.15**	40.17	34.15	05.57
VNMCE+T5	27.18	15.76	**10.65**	60.52	**31.02**	**20.57**	**74.07**	39.10	**26.01**	30.97	**26.85**	04.51	**41.62**	**35.49**	**05.78**

we compute *verb* and *noun* probabilities by marginalization. Best results per-columns are reported in bold numbers.

We begin by comparing VN-CE with respect to A-CE and the method based on the proposed loss VNMCE (top part of Table 2). Putting emphasis on the independent prediction of *verbs* and *nouns*, VN-CE anticipates *verbs* and *nouns* better than its action-based counterpart A-CE (e.g., VN-CE obtains a TOP-3 score of 66.01% for *verbs*, whereas A-CE obtains a TOP-3 score of 59.16%). However, the performance of VN-CE on action anticipation (i.e., independent prediction of *verbs* and *nouns*) is pretty low as compared to A-CE according to all evaluation measures (e.g., 17.31% vs 25.40% in the case of the TOP-5 Accuracy). This suggest that VN-CE is not able to effectively model the relationships between *verbs* and *nouns* (e.g., meaningless *(verb, noun)* combinations such as "wash door" could be predicted). On the contrary, optimizing directly for actions allows for a significant gain in performance. It should be noted that, while this is true for class-agnostic metrics, the same observations do not hold for average TOP-3 and TOP-5 recall, where the VN-CE method seems to outperform the action-based losses. As can be observed from Table 3, this happens consistently also in the case of action recognition and it is probably due to the long tail distribution characterizing actions (some actions appear just once in the whole dataset). The proposed VNMCE loss allows to obtain action recognition results similar to A-CE (e.g., 19.61% vs 19.37% in the case of the TOP-3 accuracy, or 25.14% vs 25.40% in the case of the TOP-5 accuracy), while occasionally allowing to obtain better performance for verb or *noun* prediction alone (e.g., 26.61% vs 24.86% in the case of TOP-1 verb accuracy or 34.21% vs 31.38% for Avg. TOP-5 *noun* recall). However, it should be noted that such gains are not consistent over all the evaluation metrics.

The middle part of Table 2 reports the results obtained using the two investigated TOP-K losses with $K = 3, 5$. As can be noted, TOP-K losses in general allow to improve action anticipation results (e.g., 11.09% vs 09.89% in the case of TOP-1 action accuracy and 44.14% vs 41.40% in the case of Avg. TOP-5 *noun* recall). These results suggest that relaxing the training objective allows models to diversify the predictions and obtain more general anticipations rather than concentrating on the single *(verb, noun)* label associated to a given training sample.

OBSERVED SEGMENT	VN-CE [6]	A-CE	VNMCE	VNMCE+T3 [19]	GT
	wash tap **wash board** place tap close tap wipe sink	put **board** take spoon put box put knife take bowl	put knife **wash board** put **board** take knife take spoon	**wash board** put **board** put knife **wash** knife take spoon	**wash board**
	wash **tap** wash pan place **tap** put pan open **tap**	open **tap** **close tap** turn-on **tap** wash spoon wash pan	open **tap** wash spoon **close tap** take spoon rinse hand	open **tap** **close tap** wash container wash spoon turn on **tap**	**close tap**
	take plastic **put** packet take cookie take knife place cookie	open door **put** knife wash spoon open fridge open packet	open fridge take knife **put towel** open door **put** knife	**put towel** open fridge open door **put** knife take knife	**put towel**
	put plate place tap **put** bowl wash plate open dish	**put** plate open door **put** bowl open tap **put** lid	**put** plate open tap open door **put** bowl take cutlery	**put glass** open tap **put** bowl open door **put** plate	**put glass**

Fig. 3. Example action anticipation predictions obtained by some of the investigated approaches. For each example we report the observed video segment preceding the action by 1 second, the TOP-5 predictions obtained by the algorithms and the ground truth label associated to the segment. Correct *verb* or *noun* predictions are reported in bold, whereas correct action predictions are underlined

We finally assess the effect of combining TOP-K losses with the proposed VNMCE loss in the bottom part of Table 2. The combined VNMCE+T3 loss allows to improve verb accuracy with respect to TOP-K losses in some cases (e.g., 27.63% vs 25.65% in the case of TOP-1 verb accuracy), while performing in general similarly to TOP-K losses.

Fig. 3 reports some qualitative examples of the action anticipation predictions obtained by VN-CE, A-CE, VNMCE, and VNMCE+T3. As can be observed, due to the independent modeling of *verbs* and *nouns*, VN-CE often predicts unfeasible actions such as "wash tap", "place tap" or "open dish". Modeling actions directly, A-CE allows to predict feasible actions, even when they do not match with the ground truth annotations (e.g., "put board" in the first example and "put plate" in the last example). VNMCE overall allows to obtain better predictions thanks to the extra emphasis which is put on *verbs* and *nouns*. An interesting example is given in the third row of Fig. 3, where the "put towel" action is correctly anticipated even if it appears only 19 times in the dataset, whereas "put" appears 251 times and "glass" appears 513 times. The predictions of VNMCE+T3 are similar to the ones of VNMCE, but VNMCE+T3 often ranks the ground truth action higher than the other methods.

Finally, Table 3 reports the results of action recognition experiments. In particular, we compare the use of the proposed VNMCE loss with respect to the separate classification of *verbs* and *nouns* (VN-CE) and standard cross entropy on actions (A-CE). Following [6], we use TOP-K accuracy as class agnostic measures and average class precision and recall for class-aware measures. Also in

Table 3. Action recognition results of the investigated methods according to different evaluation measures. Best per-column results are reported in bold.

	TOP-1 Accuracy%			TOP-3 Accuracy%			TOP-5 Accuracy%			Avg. Class Precision%			Avg. Class Recall%		
METHOD	VERB	NOUN	ACTION	VERB	NOUN	ACTION	VERB	NOUN	ACTION	VERB	NOUN	ACTION	VERB	NOUN	ACTION
VN-CE [6]	51.67	**36.34**	23.28	**78.00**	**53.65**	36.39	**86.56**	**61.37**	42.67	**50.32**	**38.89**	**12.31**	23.03	**31.69**	**10.27**
A-CE	51.72	35.15	26.48	77.25	51.69	40.23	84.04	58.34	46.71	43.03	38.48	06.92	27.85	28.80	04.83
VNMCE	**53.02**	**36.34**	**27.15**	77.97	52.65	**41.33**	84.64	59.29	**47.72**	47.00	36.05	09.04	**29.05**	30.28	05.98

this case, we use the provided many shot *nouns*, *verbs* and *actions* to compute the average precision and recall values. Similarly to what observed in the case of action anticipation, the independent prediction of *verbs* and *nouns* generally leads to suboptimal action recognition results (compare the action recognition scores of VN-CE with those obtained by the other methods). This happens for all measures except average class precision. Modeling actions directly (A-CE) allows to generally obtain better results (e.g., A-CE achieves a TOP-1 Accuracy of 26.48% vs 23.28% of VN-CE). Interestingly, VNMCE allows to systematically improve action recognition performances according to all class-agnostic measures (e.g., 27.15% vs 26.48% in the case of TOP-1 accuracy and 47.72% vs 46.71% in the case of TOP-5 accuracy). Moreover, VNMCE always obtains higher *verb* and *noun* accuracies with respect to A-CE for class-agnostic measures (e.g., 53.02% vs 51.72 TOP-1 verb accuracy).

7 Conclusion

We have studied the role of uncertainty of egocentric action anticipation in the definition of suitable evaluation measures and loss functions. We first showed that action anticipation can be seen as a multi-label learning problem in the presence of missing label. Under this perspective, we highlighted that TOP-K criteria should be preferred when evaluating action anticipation methods. We further extended the analysis showing how the uncertainty of egocentric action anticipation can be leveraged to design loss functions capable of diversifying the predictions and improve anticipation results. Specifically, we introduced a novel Verb-Noun Marginal Cross Entropy Loss (VNMCE) which encourages the model to focus on *verbs* and *nouns* in addition to *actions* and explored the potential of TOP-K losses for action anticipation. Experiments and qualitative results have shown that TOP-K losses allow to obtain promising action anticipation results. Finally, the proposed VNMCE loss is shown to improve egocentric action recognition results.

Acknowledgment

This research has been supported by Piano della Ricerca 2016-2018 linea di Intervento 2 of DMI of the University of Catania.

References

1. Abu Farha, Y., Richard, A., Gall, J.: When will you do what?-anticipating temporal occurrences of activities. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5343–5352 (2018)
2. Aliakbarian, M.S., Saleh, F.S., Salzmann, M., Fernando, B., Petersson, L., Andersson, L.: Encouraging lstms to anticipate actions very early. In: IEEE International Conference on Computer Vision (ICCV). vol. 1 (2017)
3. Berrada, L., Zisserman, A., Kumar, M.P.: Smooth loss functions for deep top-k classification. International Conference on Learning Representations (2018)
4. Bokhari, S.Z., Kitani, K.M.: Long-term activity forecasting using first-person vision. In: Asian Conference on Computer Vision. pp. 346–360. Springer (2016)
5. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4724–4733 (2017)
6. Damen, D., Doughty, H., Farinella, G.M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., Wray, M.: Scaling egocentric vision: The epic-kitchens dataset. In: European Conference on Computer Vision (ECCV) (2018)
7. Fan, C., Lee, J., Ryoo, M.S.: Forecasting hand and object locations in future frames. CoRR **abs/1705.07328** (2017), http://arxiv.org/abs/1705.07328
8. Fathi, A., Farhadi, A., Rehg, J.M.: Understanding egocentric activities. In: International Conference on Computer Vision. pp. 407–414 (2011)
9. Fathi, A., Li, Y., Rehg, J.: Learning to recognize daily actions using gaze. In: European Conference on Computer Vision. pp. 314–327 (2012)
10. Feichtenhofer, C., Pinz, A., Zisserman, A.: Convolutional two-stream network fusion for video action recognition. In: Computer Vision and Pattern Recognition. pp. 1933–1941 (2016)
11. Furnari, A., Battiato, S., Grauman, K., Farinella, G.M.: Next-active-object prediction from egocentric videos. Journal of Visual Communication and Image Representation **49**, 401–411 (2017). https://doi.org/10.1016/j.jvcir.2017.10.004
12. Gao, J., Yang, Z., Nevatia, R.: Red: Reinforced encoder-decoder networks for action anticipation. In: British Machine Vision Conference (2017)
13. Huang, D.A., Kitani, K.M.: Action-reaction: Forecasting the dynamics of human interaction. In: European Conference on Computer Vision. pp. 489–504. Springer (2014)
14. Jain, A., Koppula, H.S., Raghavan, B., Soh, S., Saxena, A.: Car that knows before you do: Anticipating maneuvers via learning temporal driving models. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3182–3190 (2015)
15. Kanade, T., Hebert, M.: First-person vision. Proceedings of the IEEE **100**(8), 2442–2453 (aug 2012). https://doi.org/10.1109/JPROC.2012.2200554
16. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: Computer Vision and Pattern Recognition. pp. 1725–1732 (2014)
17. Koppula, H.S., Saxena, A.: Anticipating human activities using object affordances for reactive robotic response. IEEE Transactions on Pattern Analysis and Machine Intelligence **38**(1), 14–29 (Jan 2016). https://doi.org/10.1109/TPAMI.2015.2430335

18. Lan, T., Chen, T.C., Savarese, S.: A hierarchical representation for future action prediction. In: European Conference on Computer Vision. pp. 689–704. Springer (2014)
19. Lapin, M., Hein, M., Schiele, B.: Analysis and optimization of loss functions for multiclass, top-k, and multilabel classification. IEEE transactions on pattern analysis and machine intelligence (2017)
20. Li, Y., Ye, Z., Rehg, J.M.: Delving into egocentric actions. In: Computer Vision and Pattern Recognition. pp. 287–295 (2015)
21. Ma, M., Fan, H., Kitani, K.M.: Going deeper into first-person activity recognition. In: Computer Vision and Pattern Recognition. pp. 1894–1903 (2016)
22. Ma, S., Sigal, L., Sclaroff, S.: Learning activity progression in lstms for activity detection and early detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1942–1950 (2016)
23. Mahmud, T., Hasan, M., Roy-Chowdhury, A.K.: Joint prediction of activity labels and starting times in untrimmed videos. In: Computer Vision (ICCV), 2017 IEEE International Conference on. pp. 5784–5793 (2017)
24. Park, H.S., Hwang, J.J., Niu, Y., Shi, J.: Egocentric future localization. In: Cvpr 2016. pp. 4697–4705 (2016). https://doi.org/10.1109/CVPR.2016.508
25. Rhinehart, N., Kitani, K.M.: First-person activity forecasting with online inverse reinforcement learning. In: ICCV (2017)
26. Ryoo, M.S., Fuchs, T.J., Xia, L., Aggarwal, J.K., Matthies, L.: Robot-centric activity prediction from first-person videos: What will they do to me? In: IEEE International Conference on Human-Robot Interaction. pp. 295—-302 (2015). https://doi.org/10.1145/2696454.2696462
27. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: Advances in Neural Information Processing Systems. pp. 568–576 (2014)
28. Singh, K.K., Fatahalian, K., Efros, A.A.: Krishnacam: Using a longitudinal, single-person, egocentric dataset for scene understanding tasks. In: IEEE Winter Conference on Applications of Computer Vision (2016). https://doi.org/10.1109/WACV.2016.7477717
29. Soran, B., Farhadi, A., Shapiro, L.: Generating notifications for missing actions: Don't forget to turn the lights off! In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4669–4677 (2016). https://doi.org/10.1109/ICCV.2015.530
30. Spriggs, E.H., De La Torre, F., Hebert, M.: Temporal segmentation and activity classification from first-person sensing. In: Computer Vision and Pattern Recognition Workshops. pp. 17–24 (2009)
31. Vondrick, C., Pirsiavash, H., Torralba, A.: Anticipating visual representations from unlabeled video. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 98–106 (2016)
32. Walker, J., Doersch, C., Gupta, A., Hebert, M.: An uncertain future: Forecasting from static images using variational autoencoders. In: European Conference on Computer Vision. pp. 835–851. Springer (2016)
33. Wang, H., Kläser, A., Schmid, C., Liu, C.L.: Dense trajectories and motion boundary descriptors for action recognition. International Journal of Computer Vision **103**(1), 60–79 (2013)
34. Wang, H., Schmid, C.: Action recognition with improved trajectories. In: International Conference on Computer Vision. pp. 3551–3558 (2013)

35. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L.: Temporal segment networks: towards good practices for deep action recognition. In: European Conference on Computer Vision. pp. 20–36 (2016)
36. Yu, H.F., Jain, P., Kar, P., Dhillon, I.: Large-scale multi-label learning with missing labels. In: International conference on machine learning. pp. 593–601 (2014)
37. Zhang, M., Ma, K.T., Lim, J.H., Zhao, Q., Feng, J.: Deep future gaze: Gaze anticipation on egocentric videos using adversarial networks. In: Conference on Computer Vision and Pattern Recognition. pp. 4372–4381 (2017)
38. Zhang, M.L., Zhou, Z.H.: A review on multi-label learning algorithms. IEEE transactions on knowledge and data engineering **26**(8), 1819–1837 (2014)
39. Zhou, Y., Berg, T.L.: Temporal perception and prediction in ego-centric video. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4498–4506 (2016). https://doi.org/10.1109/ICCV.2015.511