Supplementary Section: Video Event Understanding using Natural Language Descriptions

Vignesh Ramanathan* Percy Liang† Li Fei-Fei†
*Department of Electrical Engineering, Stanford University
†Computer Science Department, Stanford University
{vigneshr, pliang, feifeili}@cs.stanford.edu

A. Optimization for Posterior Regularization

The optimization problem solved to learn our model with Posterior Regularization (PR) is shown again in Eq. 1. To recount, $\mathbf{a}, \mathbf{r}$ are the action and role assignments to all the human tracklets across all videos and $Q(\mathbf{a}, \mathbf{r})$ is the corresponding probability of the assignments. The CRF log-likelihood of the assignments is given by $L(\mathbf{a}, \mathbf{r}; w)$, where w are the weights to be learnt.

$$\begin{aligned}\min_{\substack{w,Q,\\ \delta\geq 0,\eta\geq 0}} \quad & \frac{\|w\|^2}{2} - \mathbb{E}_Q[L(\mathbf{a},\mathbf{r};w)] + \sum_i \left\{ \sum_a \delta_i^a + \sum_r \eta_i^r \right\} - H_Q \\ \text{subject to} \quad & \mathbb{E}_Q[N_i(a)] \geq 1 - \delta_i^a, \quad \forall y_i^a = 1 \\ & \mathbb{E}_Q[N_i(a)] \leq \delta_i^a, \quad \forall y_i^a = -1 \\ & \mathbb{E}_Q[M_i(r)] \geq 1 - \eta_i^r, \quad \forall z_i^r = 1 \\ & \mathbb{E}_Q[M_i(r)] \leq \eta_i^r, \quad \forall z_i^r = -1,\end{aligned} \tag{1}$$

where H_Q is the entropy of Q, $N_i(a) = \sum_h \mathbf{1}(a_h^i = a)$ and $M_i(r) = \sum_h \mathbf{1}(r_h^i = r)$.

In the absence of the slack constraints, Eq. 1 is traditionally learnt through an Expectation Maximization (EM) procedure. Similarly, we will discuss an EM algorithm with a modified expectation step to account for the constraints. The modified expectation step (E'-step) and the maximization step (M-step) at an iteration t are explained below.

A.1. E'-step

The E'-step at an iteration t refers to the optimization of Eq. 1, while keeping the model weights fixed from the previous iteration $w^{(t-1)}$. The minimizing Q provides $Q^{(t)}$. The E'-step at iteration t is shown in Eq. 2.

$$\begin{aligned}\max_{\substack{Q\\ \delta\geq 0,\eta\geq 0}} \quad & \mathbb{E}_Q[L(\mathbf{a},\mathbf{r};w^{(t-1)})] - \sum_i \left(\sum_a \delta_i^a + \sum_r \eta_i^r \right) + H_Q \\ \text{subject to} \quad & \mathbb{E}_Q[N_i(a)] \geq 1 - \delta_i^a, \quad \forall y_i^a = 1 \\ & \mathbb{E}_Q[N_i(a)] \leq \delta_i^a, \quad \forall y_i^a = -1 \\ & \mathbb{E}_Q[M_i(r)] \geq 1 - \eta_i^r, \quad \forall z_i^r = 1 \\ & \mathbb{E}_Q[M_i(r)] \leq \eta_i^r, \quad \forall z_i^r = -1,\end{aligned} \tag{2}$$

We notice that the above problem is convex due to the linear nature of constraints arising from our model. The above equation can be solved exactly by minimizing the dual. In order to solve the dual, we first define a modified CRF log-likelihood function $\tilde{L}$ with dual variables λ, γ as shown in Eq. 3.

$$\begin{aligned}\tilde{L}(\mathbf{a},\mathbf{r};w,\lambda,\gamma) \quad = \quad & \sum_{i=1}^n \sum_h \Big(w_g(a_i^h, r_i^h) \cdot f_g^{x_i} + \\ & w_{ac.}(a_i^h) \cdot f_{ac.}^h + w_{ro.}(r_i^h) \cdot f_{ro.}^h \\ & + w_{in.}(a_i^h, r_i^h) + \sum_a y_i^a \lambda_i^a \cdot \mathbf{1}(a_i^h = a) + \\ & \sum_r z_i^r \gamma_i^r \cdot \mathbf{1}(r_i^h = r) \Big) - \log \tilde{Z}_i(w, \lambda_i, \gamma_i),\end{aligned} \tag{3}$$

where $\tilde{Z}_i$ is the partition function for video x_i corresponding to the modified CRF potential.

We obtain the optimal dual variables λ^*, γ^* as shown in Eq. 4.

$$\begin{aligned}\lambda_i^*, \gamma_i^* \quad = \quad & \underset{\substack{0\leq\lambda_i\leq 1\\ 0\leq\gamma_i\leq 1}}{\operatorname{argmin}} \Big\{ \log \tilde{Z}_i(w^{(t-1)}, \lambda_i, \gamma_i) - \\ & \sum_a \lambda_i^a \cdot \mathbf{1}(y_i^a = 1) - \sum_r \gamma_i^r \cdot \mathbf{1}(z_i^r = 1) \Big\}\end{aligned} \tag{4}$$

Now, $Q^{(t)}(\mathbf{a}, \mathbf{r})$ is obtained by running inference on a CRF whose log-likelihood is given by

$\tilde{L}(\mathbf{a}, \mathbf{r}; w^{(t-1)}, \lambda^*, \gamma^*)$. Thus the problem is tractable since we solve Eq. 4 and run inference separately for each video.

A.2. M-step

The M-step is shown in Eq. 5.

$$w^{(t)} = \underset{w}{\operatorname{argmin}} \frac{\|w\|^2}{2} - \mathbb{E}_{Q^{(t)}}[L(\mathbf{a}, \mathbf{r}; w)] \tag{5}$$

As seen, the M-step remains the same as the traditional EM procedure and can be solved through coordinate descent.

B. Self paced learning to handle outliers

We wish to use the textual features to handle outliers and ensure that only good examples are chosen based on the natural language descriptions. Hence, we define a new potential $\phi_d(x, h, t, a, r)$ corresponding to the assignment of action a and role r to the human track h in video x with textual description t as shown in Eq. 6, which includes an additional global potential corresponding to the textual features f_d^t with corresponding weights given by w_d.

$$\begin{aligned} \Phi_d(x, h, t, a, r) &= w_g(a, r) \cdot f_g^x + w_{in.}(a, r) \\ &\quad + w_{ac.}(a) \cdot f_{ac.}^h + w_{ro.}(r) \cdot f_{ro.}^h \\ &\quad w_d(a, r) \cdot f_d^t \end{aligned} \tag{6}$$

The new log-likelihood L_d corresponding to this modified potential is shown below.

$$\begin{aligned} p_d(a_i, r_i; w, w_d) &= \frac{1}{Z_i^d} \exp\left(\sum_{h \in \mathcal{H}_i} \Phi_d(x_i, h, t_i, a_i^h, r_i^h) \right) \\ L_d(\mathbf{a}, \mathbf{r}; w, w_d) &= \sum_i \log p_d(a_i, r_i; w, w_d), \end{aligned}$$

where Z_i^d is the modified partition function for the video x_i.

The model learning at an iteration of the self-paced scheme is then defined as shown in Eq. 7.

$$\begin{aligned} \min_{\substack{w, w_d, Q, \mathcal{V} \\ \delta \geq 0, \eta \geq 0}} \; & \frac{\|w\|^2}{2} + \frac{\|w_d\|^2}{2\sigma_d^2} - \mathbb{E}_Q[L_d(\mathbf{a}, \mathbf{r}; w, w_d)] - H_Q \\ & + \sum_a \left\{ \sum_{x_i \in \mathcal{V}^a} \delta_i^a - \frac{|\mathcal{V}^a|}{K} \right\} + \sum_r \left\{ \sum_{x_j \in \mathcal{V}^r} \eta_j^r - \frac{|\mathcal{V}^r|}{K} \right\} \\ \text{subject to } & \mathbb{E}_Q[N_i(a)] \geq 1 - \delta_i^a, \quad \forall y_i^a = 1 \\ & \mathbb{E}_Q[N_i(a)] \leq \delta_i^a, \quad \forall y_i^a = -1 \\ & \mathbb{E}_Q[M_i(r)] \geq 1 - \eta_i^r, \quad \forall z_i^r = 1 \\ & \mathbb{E}_Q[M_i(r)] \leq \eta_i^r, \quad \forall z_i^r = -1, \end{aligned} \tag{7}$$

where, $\mathcal{V}^a$ and $\mathcal{V}^r$ are the sets of training examples corresponding to action a and role r respectively, whose corresponding action or role labels are used at an iteration of the self paced procedure to enforce PR constraints. In our experiments, we initialize $\mathcal{V}$ to be the same as the examples considered for baseline models not using the natural language descriptions.

We run the model for three iterations for the sake of tractable learning and anneal the parameter σ_d from 1 to 0 in the second iteration. This ensures that the textual features are considered initially, but are neglected midway when the model is more confident. The optimization in Eq. 7 can be solved approximately through an alternate search strategy, which alternates between choosing the optimal set of examples $\mathcal{V}$ with fixed model weights and optimizing the model weights keeping $\mathcal{V}$ fixed. We refer the reader to [1] for further details on the optimization. K is annealed linearly from 10 to 2, so that at the end, the model includes only those examples which are not extreme outliers.

References

[1] M. P. Kumar, B. Packer, and D. Koller. Self-paced learning for latent variable models. In *NIPS*, 2010. 2