Weakly-Supervised Alignment of Video With Text

Piotr Bojanowski, Rémi Lagugie, Edouard Grave, Ivan Laptev, Jean Ponce and Cordelia Schmid
INRIA / ENS / CNRS

Goal

Decoupling Data and Alignment

We propose to decompose the assignment variable as $Z = \Psi Y$

Clearly separate data from optimization variables:

$$\min_{Z \in \mathcal{Z}, Y} \|Z - \Psi \Phi\|_F^2 \quad \min_{\Psi \in \mathcal{\Psi}} \|\Psi - W\Phi\|_F^2$$

Advantages

- Easier to add constraints and priors
- Clearly separate variables and data
- Continuous representations for $\Psi$

Hot Topic

Continuous sentence representations similar to word embedding (sent2vec...)

Prior

Model prone to degenerate solutions

All time intervals assigned to same event

Gets worse when $T \to \infty$

If we use a bias:

$$\|\Psi Y - W\Phi - b_1 \Psi + b_2\|_F^2 = 0$$

Constant row solution has objective 0

$$W^* = 0 \quad b^*_1 = \Psi e_1^T \quad Y^* = e_1^T$$

Duration prior

Penalize event duration when expected duration $\mu_2$ is available

$$\frac{1}{2\sigma^2} \|Y - \mu_2\|^2$$

Dualize constraint

Add a linear term

$$\alpha \text{Tr}(Y^T Y)$$

Semi Supervised Setting

How can we add supervision to our model?

Action annotation in videos:

- Start of "open door"
- End of "open door"
- End of "sit down"

Potential inconsistency with the weakly-supervised predictions

Instead of fixing the assignment to GT, we constrain it

$Y$ is forced to select at least one relevant frame per event

Experiments

Experiments on the dataset from [1]

Direct comparison to [1]: same data, and build $\Psi$ from the sequence of labels

Compare different sorts of rounding schemes

Best working rounding:

$$\min_{\Psi} \|\Psi(Y - Y^*)\|_F^2$$

Using priors helps a bit

Using supervision does not give the performance drop observed before

Tacos dataset

127 videos of cooking tasks

Multiple annotators per video (Mturks)

Replicate video per annotator

Simple text descriptions

Video features: 500 dimensional BOW (HOG + HOF + MBH)

Pooling over windows of 150 frames, stride of 50

Evaluating the priors

Use priors for various sets of parameters

We observe a tradeoff for both proposed methods

Comparing sentence representations

We evaluate several kinds of sentence representations

ROOT: indicator of the root

ROOT + DOB: root and direct object

VNA: bag of verb + nouns + adjectives

W2V: word2vec on root + dobj

Qualitative results

We observe several forms of the root and the object