Supplementary Material: Visual Semantic Search: Retrieving Videos via Complex Textual Queries

Dahua Lin[1] **Sanja Fidler**[1,2] **Chen Kong**[3] **Raquel Urtasun**[1,2]
TTI Chicago[1] University of Toronto[2] Tsinghua University[3]
dhlin@ttic.edu, kc10@mails.tsinghua.edu.cn, {fidler,urtasun}@cs.toronto.edu

Abstract

This document provides some technical details related to the learning problem presented in the paper [1]. In particular, we review the concept of conciseness, and provide the proof to the Proposition 1 in [1], which establishes the fact that our learning problem is concise, and finally give the detailed derivation of the simplified optimization problem given in Eq.(7).

1 The Learning Problem

The problem of learning the optimal combination weights of scores was formulated in section 4.5.1 of the paper. For self-containedness, we briefly revisit the problem below.

$$\begin{aligned} &\text{minimize} \ \ \frac{1}{2}\|\mathbf{w}\|^2 + C \sum_i \xi_i \\ &\text{s.t.} \quad \mathbf{w}^T \boldsymbol{\phi}_i(\mathbf{y}^{(i)}) \geq \mathbf{w}^T \boldsymbol{\phi}_i(\mathbf{y}) + \Delta(\mathbf{y}, \mathbf{y}^{(i)}) - \xi_i, \ \ \forall \mathbf{y}^{(i)} \in \mathcal{Y}^{(i)}. \\ &\qquad\ \ \xi_i \geq 0, \ \ \forall i = 1, \dots, N. \end{aligned} \tag{1}$$

Here, $\mathbf{y}^{(i)}$ is the ground-truth matching for the i-th instance, $\boldsymbol{\phi}_i(\mathbf{y})$ is a vector of matching scores for $\mathbf{y}$, and $\Delta(\mathbf{y}, \mathbf{y}^{(i)})$ the loss function. In particular, $\boldsymbol{\phi}_i(\mathbf{y})$ can be expressed as

$$\boldsymbol{\phi}_i(\mathbf{y}) = [\phi_i^{(1)}(\mathbf{y}), \dots, \phi_i^{(K)}(\mathbf{y})], \quad \text{with } \phi_i^{(k)}(\mathbf{y}) = \sum_{uv} f_{uv}^{(ik)} y_{uv}. \tag{2}$$

We use the Hamming loss, as

$$l(\mathbf{y}; \mathbf{y}^{(i)}) = \sum_{uv} \mathbb{1}(y_{uv} \neq y_{uv}^{(i)}) = a^{(i)} - \sum_{uv} y_{uv} y_{uv}^{(i)}, \tag{3}$$

where $a^{(i)} = \sum_u s_u^{(i)}$ is the total number of matching edges, which is a constant.

The domain $\mathcal{Y}^{(i)}$ depends on particular instance, and can be written as

$$\mathcal{Y}^{(i)} = \left\{ \mathbf{y} : \ \sum_v y_{uv} = s_u^{(i)}, \ \sum_u y_{uv} \leq t_v^{(i)}, \ 0 \leq y_{uv} \leq c_{uv}^{(i)} \right\}. \tag{4}$$

2 The Notion of Conciseness

The learning problem given by Eq.(1) can be re-written as

$$\begin{aligned} &\text{minimize} \ \ \frac{1}{2}\|\mathbf{w}\|^2 + C \sum_i \xi_i \\ &\text{s.t.} \quad \mathbf{w}^T \boldsymbol{\phi}_i(\mathbf{y}^{(i)}) \geq \max_{\mathbf{y} \in \mathcal{Y}^{(i)}} \left(\mathbf{w}^T \boldsymbol{\phi}_i(\mathbf{y}) + \Delta(\mathbf{y}; \mathbf{y}^{(i)}) \right) - \xi_i, \ \ \xi_i \geq 0, \quad \forall i = 1, \dots, N. \end{aligned} \tag{5}$$

This model is called ***concise*** if there exists a function $\tilde{f}_i$ that is concave in $\boldsymbol{\mu}$ and a convex set $\mathcal{U}^{(i)}$ for each i such that

$$\max_{\mathbf{y}\in\mathcal{Y}^{(i)}} \left(\mathbf{w}^T\boldsymbol{\phi}_i(\mathbf{y}) + \Delta(\mathbf{y};\mathbf{y}^{(i)})\right) = \max_{\boldsymbol{\mu}\in\mathcal{U}^{(i)}} \tilde{f}_i(\mathbf{w},\boldsymbol{\mu}). \tag{6}$$

Next, we review how *conciseness* can be exploited to simplify the learning problem. Without losing generality, we express $\boldsymbol{\mu}\in\mathcal{U}^{(i)}$ using a convex function $\mathbf{g}_i$ as

$$\mathbf{g}_i(\boldsymbol{\mu}) \leq 0. \tag{7}$$

Then the Lagrangian for $\tilde{f}_i(\mathbf{w},\boldsymbol{\mu})$ is

$$L_i(\boldsymbol{\mu},\boldsymbol{\lambda};\mathbf{w}) = \tilde{f}_i(\mathbf{w},\boldsymbol{\mu}) - \boldsymbol{\lambda}^T\mathbf{g}_i(\boldsymbol{\mu}) \quad \text{with } \boldsymbol{\lambda}\geq 0. \tag{8}$$

This provides an upper bound for $\tilde{f}_i(\mathbf{w},\boldsymbol{\mu})$ within $\mathcal{U}^{(i)}$. By strong duality (which can be easily verified), we have:

$$\begin{aligned}\max_{\boldsymbol{\mu}\in\mathcal{U}^{(i)}} \tilde{f}_i(\mathbf{w},\boldsymbol{\mu}) &= \max_{\boldsymbol{\mu}\in\mathcal{U}^{(i)}}\min_{\boldsymbol{\lambda}\geq 0} L_i(\boldsymbol{\mu},\boldsymbol{\lambda};\mathbf{w}),\\ &= \min_{\boldsymbol{\lambda}\geq 0}\max_{\boldsymbol{\mu}\in\mathcal{U}^{(i)}} L_i(\boldsymbol{\mu},\boldsymbol{\lambda};\mathbf{w}).\end{aligned} \tag{9}$$

Suppose $\max_{\boldsymbol{\mu}\in\mathcal{U}^{(i)}} L_i(\boldsymbol{\mu},\boldsymbol{\lambda},\boldsymbol{\nu};\mathbf{w})$ has a Lagrangian dual given by

$$\rho_i(\boldsymbol{\lambda};\mathbf{w}) \ \text{ s.t. } \ \boldsymbol{\eta}_i(\boldsymbol{\lambda};\mathbf{w}) \leq 0. \tag{10}$$

Then, we have

$$\max_{\boldsymbol{\mu}\in\mathcal{U}^{(i)}} \tilde{f}_i(\mathbf{w},\boldsymbol{\mu}) = \min_{\boldsymbol{\eta}^{(i)}(\boldsymbol{\lambda};\mathbf{w})\leq 0} \rho_i(\boldsymbol{\lambda};\mathbf{w}) \tag{11}$$

For conciseness, the condition $\boldsymbol{\lambda}\geq 0$ is merged into $\boldsymbol{\eta}_i(\boldsymbol{\lambda};\mathbf{w})\leq 0$. Incorporating this into Eq.(5) results in

$$\begin{aligned}&\text{minimize} \ \ \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i \xi_i \\ &\text{s.t.} \quad \mathbf{w}^T\boldsymbol{\phi}_i(\mathbf{y}^{(i)}) \geq \min_{\boldsymbol{\eta}^{(i)}(\boldsymbol{\lambda};\mathbf{w})\leq 0} \rho_i(\boldsymbol{\lambda};\mathbf{w}) - \xi_i, \ \ \xi_i \geq 0, \quad \forall i = 1,\ldots,N.\end{aligned} \tag{12}$$

Combining the optimization over $\mathbf{w}$ and that over $\boldsymbol{\lambda}$, we finally gets the following problem:

$$\begin{aligned}&\text{minimize} \ \ \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i \xi_i \\ &\text{s.t.} \quad \mathbf{w}^T\boldsymbol{\phi}_i(\mathbf{y}^{(i)}) \geq \rho_i(\boldsymbol{\lambda},\boldsymbol{\nu};\mathbf{w}) - \xi_i, \quad \forall i = 1,\ldots,N, \\ &\qquad \boldsymbol{\eta}_i(\boldsymbol{\lambda},\boldsymbol{\nu};\mathbf{w}) \leq 0, \ \ \xi_i \geq 0, \quad \forall i = 1,\ldots,N.\end{aligned} \tag{13}$$

3 Proof of Proposition 1

Proposition 1 in the paper establishes the fact that our learning problem is *concise*. Below, we prove this proposition.

With Eq.(2) and Eq.(3), we have

$$\begin{aligned}\mathbf{w}^T\boldsymbol{\phi}_i(\mathbf{y}) + \Delta(\mathbf{y};\mathbf{y}^{(i)}) &= \sum_{k=1}^{K} w_k \sum_{uv} f_{uv}^{(ik)} y_{uv} + \left(a^{(i)} - \sum_{uv} y_{uv} y_{uv}^{(i)}\right) \\ &= a^{(i)} + \sum_{uv}\left(\sum_{k=1}^{K} w_k f_{uv}^{(ik)} - y_{uv}^{(i)}\right) y_{uv} \\ &= a^{(i)} + \left(\mathbf{F}^{(i)}\mathbf{w} - \mathbf{y}^{(i)}\right)^T \mathbf{y}.\end{aligned} \tag{14}$$

Here, each $\mathbf{F}^{(i)}$ is an mn-by-K matrix, where each row corresponding to a particular matching pair (u, v) and each column corresponds to a score channel. According to Eq.(6), we can conclude that this model is *concise*, with

$$\begin{aligned} \tilde{f}_i(\mathbf{w}, \boldsymbol{\mu}) &= a^{(i)} + \left(\mathbf{F}^{(i)}\mathbf{w} - \mathbf{y}^{(i)}\right)^T \boldsymbol{\mu} \\ &= a^{(i)} + \sum_{uv} \left(\mathbf{w}^T \mathbf{f}_{uv}^{(i)} - y_{uv}^{(i)}\right) \mu_{uv}. \end{aligned} \tag{15}$$

Here, $\mathbf{f}_{uv}^{(i)}$ is the uv-th row of $\mathbf{F}^{(i)}$, which is a K-dimensional vector. In addition, the constraint $\boldsymbol{\mu} \in \mathcal{U}^{(i)}$ can be written explicitly as

$$\sum_v \mu_{uv} = s_u^{(i)} \ \ \forall u, \quad \sum_u \mu_{uv} \leq t_v^{(i)} \ \ \forall v, \quad 0 \leq \mu_{uv} \leq c_{uv}^{(i)} \ \ \forall u, v. \tag{16}$$

The proof is completed.

4 Simplified Optimization Problem

Then, we can derive the Lagrangian dual as follows

$$\rho^{(i)}(\boldsymbol{\lambda}, \boldsymbol{\eta}, \boldsymbol{\nu}, \mathbf{w}) = a^{(i)} + \sum_u \lambda_u s_u^{(i)} + \sum_v \eta_v t_v^{(i)} + \sum_{uv} \nu_{uv} c_{uv}^{(i)}, \tag{17}$$

with

$$\mathbf{w}^T \mathbf{f}_{uv}^{(i)} \leq y_{uv}^{(i)} + \lambda_u + \eta_v + \nu_{uv}, \ \ \eta_v \geq 0, \ \ \nu_{uv} \geq 0 \ \ \forall u, v. \tag{18}$$

Finally, according to Eq.(13), the learning problem can be written as

$$\begin{aligned} &\text{minimize} \ \ \frac{1}{2}\|\mathbf{w}\|^2 + C \sum_i \xi_i \qquad (19) \\ &\text{s.t.} \quad \mathbf{w}^T \mathbf{z}^{(i)} \geq \rho^{(i)}(\boldsymbol{\lambda}, \boldsymbol{\eta}, \boldsymbol{\nu}, \mathbf{w}) - \xi_i, \quad \forall i = 1, \ldots, N, \\ &\qquad\ \ \mathbf{w}^T \mathbf{f}_{uv}^{(i)} \leq y_{uv}^{(i)} + \lambda_u^{(i)} + \eta_v^{(i)} + \nu_{uv}^{(i)}, \quad \forall u, v, i \\ &\qquad\ \ \eta_v^{(i)} \geq 0, \ \ \nu_{uv}^{(i)} \geq 0, \ \ \xi^{(i)} \geq 0, \quad \forall u, v, i \end{aligned}$$

Here, $\mathbf{z}^{(i)} = [z_1^{(i)}, \ldots, z_K^{(i)}]$ with $z_k^{(i)} = \sum_{uv} f_{uv}^{(ik)} y_{uv}^{(i)}$.

References

[1] D. Lin, S. Fidler, C. Kong, and R. Urtasun. Visual semantic search: Retrieving videos via complex textual queries. In *CVPR*, 2014.