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

Dahua Lin\textsuperscript{1}  Sanja Fidler\textsuperscript{1,2}  Chen Kong\textsuperscript{3}  Raquel Urtasun\textsuperscript{1,2}
TTI Chicago\textsuperscript{1}  University of Toronto\textsuperscript{2}  Tsinghua University\textsuperscript{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.

\[
\text{minimize} \quad \frac{1}{2} \|w\|^2 + C \sum_i \xi_i \\
\text{s.t.} \quad w^T \phi_i(y^{(i)}) \geq w^T \phi_i(y) + \Delta(y, y^{(i)}) - \xi_i, \quad \forall y^{(i)} \in \mathcal{Y}^{(i)}. \\
\xi_i \geq 0, \quad \forall i = 1, \ldots, N.
\]

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

\[
\phi_i(y) = [\phi_1^{(i)}(y), \ldots, \phi_K^{(i)}(y)], \quad \text{with} \quad \phi_k^{(i)}(y) = \sum_{u,v} f_{uv}^{(i,k)} y_{uv}.
\]

We use the Hamming loss, as

\[
l(y; y^{(i)}) = \sum_{u,v} 1(y_{uv} \neq y_{uv}^{(i)}) = a^{(i)} - \sum_{u,v} y_{uv} y_{uv}^{(i)}, \quad \text{with} \quad a^{(i)} = \sum_{u,v} s_u^{(i)}, \text{and} \quad \mathcal{Y}^{(i)}\text{ depends on particular instance, and can be written as}
\]

\[
\mathcal{Y}^{(i)} = \left\{ y : \sum_v y_{uv} = s_u^{(i)}, \sum_u y_{uv} \leq t_v^{(i)}, \ \text{and} \ \mathcal{Y}^{(i)} \right\}.
\]

2 The Notion of Conciseness

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

\[
\text{minimize} \quad \frac{1}{2} \|w\|^2 + C \sum_i \xi_i \\
\text{s.t.} \quad w^T \phi_i(y^{(i)}) \geq \max_{y \in \mathcal{Y}^{(i)}} \left( w^T \phi_i(y) + \Delta(y, y^{(i)}) \right) - \xi_i, \quad \xi_i \geq 0, \quad \forall i = 1, \ldots, N.
\]
This model is called **concise** if there exists a function \( \tilde{f}_i \) that is concave in \( \mu \) and a convex set \( \mathcal{U}(i) \) for each \( i \) such that

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

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

\[
g_i(\mu) \leq 0. \tag{7}
\]

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

\[
L_i(\mu, \lambda; w) = \tilde{f}_i(w, \mu) - \lambda^T g_i(\mu) \quad \text{with} \quad \lambda \geq 0. \tag{8}
\]

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

\[
\max_{\mu \in \mathcal{U}(i)} \tilde{f}_i(w, \mu) = \max_{\mu \in \mathcal{U}(i)} \min_{\lambda \geq 0} L_i(\mu, \lambda; w),
\]

\[
= \min_{\lambda \geq 0} \max_{\mu \in \mathcal{U}(i)} L_i(\mu, \lambda; w). \tag{9}
\]

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

\[
\eta_i(\lambda; w) \quad \text{s.t.} \quad \eta_i(\lambda; w) \leq 0. \tag{10}
\]

Then, we have

\[
\max_{\mu \in \mathcal{U}(i)} \tilde{f}_i(w, \mu) = \min_{\eta_i(\lambda; w) \leq 0} \rho_i(\lambda; w). \tag{11}
\]

For conciseness, the condition \( \lambda \geq 0 \) is merged into \( \eta_i(\lambda; w) \) \( \leq 0 \). Incorporating this into Eq.\( \text{[5]} \) results in

\[
\text{minimize} \quad \frac{1}{2} \|w\|^2 + C \sum_{i} \xi_i \quad \text{s.t.} \quad w^T \phi_i(y^{(i)}) \geq \min_{\eta_i(\lambda; w) \leq 0} \rho_i(\lambda; w) - \xi_i, \quad \xi_i \geq 0, \quad \forall i = 1, \ldots, N. \tag{12}
\]

Combining the optimization over \( w \) and that over \( \lambda \), we finally get the following problem:

\[
\text{minimize} \quad \frac{1}{2} \|w\|^2 + C \sum_{i} \xi_i \quad \text{s.t.} \quad w^T \phi_i(y^{(i)}) \geq \rho_i(\lambda, \nu; w) - \xi_i, \quad \forall i = 1, \ldots, N,
\]

\[
\quad \eta_i(\lambda, \nu; w) \leq 0, \quad \xi_i \geq 0, \quad \forall i = 1, \ldots, N. \tag{13}
\]

### 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.\( \text{[2]} \) and Eq.\( \text{[3]} \), we have

\[
w^T \phi_i(y) + \Delta(y; y^{(i)}) = \sum_{k=1}^{K} w_k \sum_{uv} f^{(ik)}_{uv} y_{uv} + \left( a^{(i)} - \sum_{uv} y_{uv} y^{(i)}_{uv} \right)
\]

\[
= a^{(i)} + \sum_{uv} \left( \sum_{k=1}^{K} w_k f^{(ik)}_{uv} - y^{(i)}_{uv} \right) y_{uv}
\]

\[
= a^{(i)} + \left( \mathbf{F}^{(i)} w - y^{(i)} \right)^T \mathbf{y}. \tag{14}
\]
Here, each $F^{(i)}$ 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

$$
\tilde{f}_i(w, \mu) = a^{(i)} + \left( F^{(i)} w - y^{(i)} \right)^T \mu
$$

$$
= a^{(i)} + \sum_{uv} \left( w^T f^{(i)}_{uv} - y^{(i)}_{uv} \right) \mu_{uv}.
$$

(15)

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

$$
\sum_v \mu_{uv} = s^{(i)}_u \forall u, \quad \sum_u \mu_{uv} \leq t^{(i)}_v \forall v, \quad 0 \leq \mu_{uv} \leq c^{(i)}_{uv} \forall u,v.
$$

(16)

The proof is completed.

4 Simplified Optimization Problem

Then, we can derive the Lagrangian dual as follows

$$
\rho^{(i)}(\lambda, \eta, \nu, w) = a^{(i)} + \sum_u \lambda_u s^{(i)}_u + \sum_v \eta_v t^{(i)}_v + \sum_{uv} \nu_{uv} c^{(i)}_{uv},
$$

(17)

with

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

(18)

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

$$
\text{minimize} \quad \frac{1}{2} \|w\|^2 + C \sum_i \xi_i
$$

$$
\text{s.t.} \quad w^T z^{(i)} \geq \rho^{(i)}(\lambda, \eta, \nu, w) - \xi_i, \quad \forall i = 1, \ldots, N,
$$

$$
w^T f^{(i)}_{uv} \leq y^{(i)}_{uv} + \lambda_u + \eta_v + \nu_{uv}, \quad \forall u, v, i
$$

$$
\eta^{(i)}_v \geq 0, \quad \nu^{(i)}_{uv} \geq 0, \quad \xi^{(i)} \geq 0, \quad \forall u, v, i
$$

Here, $z^{(i)} = [z^{(i)}_1, \ldots, z^{(i)}_K]$ with $z^{(i)}_k = \sum_{uv} f^{(i)}_{uv} y^{(i)}_{uv}$.

References