Dual Supervised Learning (Supplementary Document)

Yingce Xia [1] **Tao Qin** [2] **Wei Chen** [2] **Jiang Bian** [2] **Nenghai Yu** [1] **Tie-Yan Liu** [2]

A. Theoretical Analysis

As we know, the final goal of the dual learning is to give correct predictions for the unseen test data. That is to say, we want to minimize the (expected) risk of the dual models, which is defined as follows[1]:

$$R(f,g)=\mathbb{E}\left[\frac{\ell_1(f(x),y)+\ell_2(g(y),x)}{2}\right],\forall f\in\mathcal{F},g\in\mathcal{G},$$

where $\mathcal{F}=\{f(x;\theta_{xy});\theta_{xy}\in\Theta_{xy}\},\mathcal{G}=\{g(x;\theta_{yx});\theta_{yx}\in\Theta_{yx}\}$, Θ_{xy} and Θ_{yx} are parameter spaces, and the $\mathbb{E}$ is taken over the underlying distribution P. Besides, let $\mathcal{D}$ denote the product space of the two models satisfying probabilistic duality, i.e., the constraint in Eqn.(4). For ease of reference, define $\mathcal{H}_{\text{dual}}$ as $(\mathcal{F}\times\mathcal{G})\cap\mathcal{D}$.

Define the empirical risk on the n sample as follows: for any $f\in\mathcal{F},g\in\mathcal{G}$,

$$R_n(f,g)=\frac{1}{n}\sum\nolimits_{i=1}^{n}\frac{\ell_1(f(x_i),y_i)+\ell_2(g(y_i),x_i)}{2}.$$

Following (Bartlett & Mendelson, 2002), we introduce Rademacher complexity for dual supervised learning, a measure for the complexity of the hypothesis.

Definition 1. *Define the Rademacher complexity of DSL, $\mathfrak{R}_n^{DSL}$, as follows:*

$$\mathfrak{R}_n^{DSL}=\mathop{\mathbb{E}}_{\boldsymbol{z},\sigma}\Big[\sup_{(f,g)\in\mathcal{H}_{dual}}\big|\frac{1}{n}\sum_{i=1}^{n}\sigma_i\big(\ell_1(f(x_i),y_i)+\ell_2(g(y_i),x_i)\big)\big|\Big],$$

where $\boldsymbol{z}=\{z_1,z_2,\cdots,z_n\}\sim P^n$, $z_i=(x_i,y_i)$ in which $x_i\in\mathcal{X}$ and $y_i\in\mathcal{Y}$, $\boldsymbol{\sigma}=\{\sigma_1,\cdots,\sigma_m\}$ are i.i.d sampled with $P(\sigma_i=1)=P(\sigma_i=-1)=0.5$.

Based on $\mathfrak{R}_n^{\text{DSL}}$, we have the following theorem for dual supervised learning:

Theorem 1 ((Mohri et al., 2012))**.** *Let $\frac{1}{2}\ell_1(f(x),y)+\frac{1}{2}\ell_2(g(y),x)$ be a mapping from $\mathcal{X}\times\mathcal{Y}$ to $[0,1]$. Then, for any $\delta\in(0,1)$, with probability at least $1-\delta$, the following inequality holds for any $(f,g)\in\mathcal{H}_{dual}$,*

$$R(f,g)\le R_n(f,g)+2\mathfrak{R}_n^{DSL}+\sqrt{\frac{1}{2n}\ln(\frac{1}{\delta})}. \quad (1)$$

Similarly, we define the Rademacher complexity for the standard supervised learning $\mathfrak{R}_n^{\text{SL}}$ under our framework by replacing the $\mathcal{H}_{\text{dual}}$ in Definition 1 by $\mathcal{F}\times\mathcal{G}$. With probability at least $1-\delta$, the generation error bound of supervised learning is smaller than $2\mathfrak{R}_n^{\text{SL}}+\sqrt{\frac{1}{2n}\ln(\frac{1}{\delta})}$.

Since $\mathcal{H}_{\text{dual}}\in\mathcal{F}\times\mathcal{G}$, by the definition of Rademacher complexity, we have $\mathfrak{R}_n^{\text{DSL}}\le\mathfrak{R}_n^{\text{SL}}$. Therefore, DSL enjoys a smaller generation error bound than supervised learning.

The approximation of dual supervised learning is defined as

$$R(f_{\mathcal{F}}^*,g_{\mathcal{F}}^*)-R^* \quad (2)$$

in which

$$R(f_{\mathcal{F}}^*,g_{\mathcal{F}}^*)=\inf R(f,g),\ s.t.\ (f,g)\in\mathcal{H}_{\text{dual}};$$
$$R^*=\inf R(f,g).$$

The approximation error for supervised learning is similarly defined.

Define $\mathcal{P}_{y|x}=\{P(y|x;\theta_{xy})|\theta_{xy}\in\Theta_{xy}\}$, $\mathcal{P}_{x|y}=\{P(x|y;\theta_{yx})|\theta_{yx}\in\Theta_{yx}\}$. Let $P_{y|x}^*$ and $P_{x|y}^*$ denote the two conditional probabilities derived from P. We have the following theorem:

Theorem 2. *If $P_{y|x}^*\in\mathcal{P}_{y|x}$ and $P_{x|y}^*\in\mathcal{P}_{x|y}$, then supervised learning and DSL has the same approximation error.*

Proof. By definition, we can verify both of the two approximation errors are zero. □

B. Details about the Language Models for Marginal Distributions

We use the LSTM language models (Sundermeyer et al., 2012; Mikolov et al., 2010) to characterize the marginal distribution of a sentence x, defined as $\prod_{i=1}^{T_x}P(x_i|x_{<i})$,

[1]School of Information Science and Technology, University of Science and Technology of China, Hefei, Anhui, China [2]Microsoft Research, Beijing, China. Correspondence to: Tao Qin <taoqin@microsoft.com>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017.

[1]The parameters θ_{xy} and θ_{yx} in the dual models will be omitted when the context is clear.

where x_i is the i-th word in x, T_x denotes the number of words in x, and the index $< i$ indicates $\{1, 2, \cdots, i-1\}$. The embedding dimension and hidden node are both 1024. We apply 0.5 dropout to the input embedding and the last hidden layer before softmax. The validation perplexities of the language models are shown in Table 1, where the validation sets are the same as those for machine translation tasks.

Table 1. Validation Perplexities of Language Models

En↔Fr		En↔De		En↔Zh	
En	Fr	En	De	En	Zh
88.72	58.90	101.44	90.54	70.11	113.43

As shown in Table 1, the perplexities of different language models vary a lot, but out DSL can make improvements on all the translation tasks (Please refer to Table 1 of the main text). This shows that DSL is not very sensitive to the qualities of the two marginal distributions.

For the marginal distributions for sentences of sentiment classification, we choose the LSTM language model again like those for machine translation applications. The two differences are: (i) the vocabulary size is 10000; (ii) the word embedding dimension is 500. The perplexity of this language model is 58.74.

References

Bartlett, Peter L and Mendelson, Shahar. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov): 463–482, 2002.

Mikolov, Tomas, Karafiát, Martin, Burget, Lukas, Cernockỳ, Jan, and Khudanpur, Sanjeev. Recurrent neural network based language model. In *Interspeech*, volume 2, pp. 3, 2010.

Mohri, Mehryar, Rostamizadeh, Afshin, and Talwalkar, Ameet. *Foundations of machine learning*. MIT press, 2012.

Sundermeyer, Martin, Schlüter, Ralf, and Ney, Hermann. Lstm neural networks for language modeling. In *Interspeech*, pp. 194–197, 2012.