Supplementary materials for "CDTrans: Cross-domain Transformer for Unsupervised Domain Adaptation"

1 Introduction

From our experiments in the paper, it can be seen that Transformers have a better generalization ability over ConvNets. Here we intend to provide some explanations about why transformers can generalize well from source domain to target domain.

One of the possible reasons is that Transformer is more resilient than CNN to random perturbation made to individual image patches. This is because through the self-attention/cross-attention mechanism, each patch will be combined with all visual similar patches within the same image to form its representation. This combination and weighted averaging process, despite its simplicity, allows us to reduce the impact of noise, like most averaging processes in statistics. To make our argument more rigorous, we provide some theoretical analysis to reveal the power of averaging in self-attention in terms of reducing noise. Our analysis shows that the essential role of self-attention is to distill noises from the input patterns/instances, making the learned model more robust.

2 Problem Definition

Let $x_i \in \mathbb{R}^d, i \in [m]$ be the input instances to self-attention, where $m \gg 1$ is the number of input patterns and $d \gg 1$ is the input dimensionality. We assume that $\{x_i\}_{i=1}^m$ are sampled from $C \ll m$ different Gaussian distributions, denoted by $\mathcal{N}(u_k, \sigma^2/dI), k \in [C]$, where $u_k \in \mathbb{R}^d$ is the center of the kth distribution. We assume that all $u_i, i \in [m]$ are normalized, i.e. $|u_i| = 1, i \in [m]$, and any two center u_i and u_j are well separated, i.e. $r_\ell \leq |u_i - u_j| \leq r_u$, $\forall i, j \in [m]$. We denote by m_k the number of instances that are generated from the kth Gaussian distribution $\mathcal{N}(u_k, \sigma^2/dI)$. For each instance x_i, we denote by k_i the index of Gaussian distribution that x_i is assigned to.

3 Analysis

We will start with a simple case for analysis, followed by a full version analysis of self-attention.

3.1 Analysis I: Simple Case

As a starting point, we consider a simple case for self-attention, where both W_K and W_V are set to be identity matrices. In addition, instead of using softmax to compute the pairwise similarity between input instances, k nearest neighbor is used: for each instance x_i, we identify the first K instances closest x_i, denoted by $x_i^j, j \in [K]$, and calculate the updated pattern x_i' as $x_i' = \sum_{i=1}^K x_i^j/K$. We

assume that K is significantly smaller than $m_i, i \in [C]$. As it will be revealed by Theorem 1, with a high probability,

$$\left|x_i' - u_{k_i}\right| \leq \left|x_i - u_{k_i}\right|, \ i \in [m]$$

implying that the set-attention structure helps reduces the noise in input instances.

Theorem 1. *Assume*

$$d \geq \max\left(8, \frac{2\sigma^2}{(r_\ell - 2\sigma)^2}\right)\log\frac{m}{\delta}, \ K \geq \max\left(C, 9\log\frac{m}{\delta}\right)$$

where C is a universal constant. Then, with a probability $1-3\delta$, for all $i \in [m]$, we have

$$\left|x_i' - u_{k_i}\right| < \left|x_i - u_{k_i}\right|$$

Proof. First, it is easy to verify that $|x_i - u_{k_i}|^2/\sigma^2$ follows a $\frac{1}{d}\chi_d^2$ distribution with d degree of freedom. Using the concentration of χ_d^2 distribution, i.e.

$$\Pr\left(x \geq (1+\delta)d\right) \leq \exp\left(-\frac{d\delta^2}{2}\right), \quad \Pr\left(x \leq (1-\delta)d\right) \leq \exp\left(-\frac{d\delta^2}{2}\right),$$

under the assumption $d \geq 8\log(m/\delta)$, we have, with a probability $1-\delta$, for all $i \in [m]$

$$\sigma\left(1 - \frac{1}{2}\sqrt{\frac{2}{d}\log\frac{m}{\delta}}\right) \leq |x_i - u_{k_i}| \leq \sigma\left(1 + \frac{1}{2}\sqrt{\frac{2}{d}\log\frac{m}{\delta}}\right)$$

As a result, when

$$\sqrt{\frac{2}{d}\log\frac{m}{\delta}} \leq \frac{r_\ell - 2\sigma}{\sigma}$$

or

$$d \geq \frac{2\sigma^2}{(r_\ell - 2\sigma)^2}\log\frac{m}{\delta}$$

with a probability $1-\delta$, for every x_i, its K nearest neighbors $x_i^j, j \in [m]$ are all generated from the k_ith distribution $\mathcal{N}(u_{k_i}, \sigma^2 I)$. Using the vector concentration (Adamczak bound, the unbounded version), we have, for each instance, with a probability $1-\delta$

$$|x_i' - u_{k_i}| \leq C\sigma\left(\frac{1}{K} + \sqrt{\frac{1}{K}\log\frac{m}{\delta}}\right)$$

where C is an universal constant. When K is large enough so that

$$C\sigma\left(\frac{1}{K} + \sqrt{\frac{1}{K}\log\frac{m}{\delta}}\right) \leq \sigma\left(1 - \frac{1}{2}\sqrt{\frac{2}{d}\log\frac{m}{\delta}}\right)$$

we have

$$|x_i' - u_{k_i}| < |x_i - u_{k_i}|, \ i \in [m]$$

Using the assumption $d \geq 8\log(m/\delta)$, we simplify the above expression as

$$K \geq \max\left(\frac{8}{3}C, 9\log\frac{m}{\delta}\right)$$

□

3.2 Analysis II: Full Analysis

We now come to the more challenging case where softmax is used for computing in self-attention. We simplify W_K as λI, where $\lambda \geq 0$ is the parameter to be tuned. For any x_i, after self-attention, it is given as

$$x_i'' = \frac{\sum_{j=1}^m \exp(\lambda \langle x_j, x_i \rangle) W_v x_j}{\sum_{j=1}^m \exp(\lambda \langle x_j, x_i \rangle)} = W_v x_i'$$

where

$$x_i' = \frac{\sum_{j=1}^m \exp(\lambda \langle x_j, x_i \rangle) x_j}{\sum_{j=1}^m \exp(\lambda \langle x_j, x_i \rangle)}$$

We will thus show that with a high probability, $|x_i' - u_{k_i}| < |x_i - u_{k_i}|$.

Before we perform the analysis, the following lemmas provides a few key result that will be used by the analysis.

Lemma 1.

$$|\langle u_i, u_j \rangle| \leq \sqrt{2} \left| 1 - \frac{\gamma_\ell}{\sqrt{2}} \right|$$

Proof. To bound $|\langle u_i, u_j \rangle|$, we have

$$|\langle u_i, u_j \rangle| \leq \sqrt{1 - \gamma_\ell \sqrt{1 - \frac{\gamma^2}{4}}}$$

Define $\delta_\ell = 1 - \gamma_\ell / \sqrt{2}$. We then have

$$|\langle u_i, u_j \rangle| \leq \sqrt{1 - (1 - \delta_\ell)\sqrt{2 - (1 - \delta_\ell)^2}} \leq \sqrt{1 - (1 - \delta_\ell)\left(1 + \delta_\ell - \frac{\delta_\ell^2}{2}\right)} \leq \sqrt{2}\delta_\ell$$

where the last step follows $|\delta_\ell| \leq 1$ □

Lemma 2. *Suppose*

$$\log \frac{m}{\delta} \leq d, \quad \log \frac{m}{2\delta\sigma^2} \leq \frac{d}{2}$$

We have, with a probability $1 - 3\delta$*, for any* i, j

$$|\langle h_i, h_j \rangle| \leq 2\sigma^2 \sqrt{\frac{1}{d} \log \frac{m}{2\delta\sigma^2}},$$

and for i, k

$$|\langle h_i, u_k \rangle| \leq 2\sigma \sqrt{\frac{1}{d} \log \frac{m}{2\delta\sigma}}$$

Proof. Define $h_i = x_i - u_{k_i}$. Since $|h_i|^2$ follow $\frac{1}{d}\chi_d^2$ distribution, we have

$$\Pr\left(|h_i|^2 \geq (1 + t)\right) \leq \exp\left(-\frac{dt^2}{2}\right)$$

There, with a probability $1-\delta$, for any $i \in [m]$, we have

$$|h_i| \le \sigma\left(1+\sqrt{\frac{1}{d}\log\frac{2m}{\delta}}\right) \le 2\sigma$$

For $\langle h_i, h_j\rangle$, by first treating h_i as a constant, we know $\langle h_i, h_j\rangle$ follows $\mathcal{N}(0, \sigma^2|h_i|^2/d)$. Using the concentration property of sub-gaussian distribution, i.e.

$$\Pr\left(|\langle h_i, h_j\rangle| \ge t\right) \le \sqrt{\frac{2}{\pi}}\frac{\exp(-dt^2/[\sigma^2|h_i|^2])}{t},$$

we have, with a probability at least $1-2\delta$, for any i, j,

$$|\langle h_i, h_j\rangle| \le t_0 := \sigma|h_i|\sqrt{\frac{2}{d}\log\frac{m}{t_0\delta}} \le 2\sigma^2\sqrt{\frac{2}{d}\log\frac{m}{t_0\delta}}$$

Using $\log(m/[2\delta\sigma^2]) \le d/2$, we have $t_0 \le 2\sigma^2$ and therefore, with a probability $1-2\delta$,

$$|\langle h_i, h_j\rangle| \le 2\sigma^2\sqrt{\frac{2}{d}\log\frac{m}{2\delta\sigma^2}}$$

For $\langle h_i, u_k\rangle$, it follows $\mathcal{N}(0, \sigma^2/d)$. We derive its bound by following the same method for $\langle h_i, h_j\rangle$ $\square$

Theorem 2. *Suppose $\sigma \le 1$ and*

$$20\lambda\sqrt{\frac{1}{d}\log\frac{2m}{\delta\sigma}} \le \frac{1}{4}, \quad \frac{6m}{m_i}\exp\left(-\frac{\lambda}{2}\right) \le \frac{\sigma}{4}, \quad \left|\sqrt{2}-\gamma_\ell\right| \le \frac{1}{2}, \quad \log\frac{m}{2\delta\sigma^2} \le d$$

Then, with a probability $1-4\delta$, for all $i \in [m]$

$$|x_i' - u_{k_i}| \le \frac{\sigma}{2}$$

Proof. We first analyze the denominator of x_i'. Define $h_i = x_i - u_{k_j}$

$$\begin{aligned}
\sum_{j=1}^m \exp(\lambda\langle x_j, x_i\rangle)x_j &= \sum_{j=1}^m \exp\left\{\lambda\left(\langle u_{k_i}, u_{k_j}\rangle + \langle h_i, h_j\rangle + \langle h_i, u_{k_j}\rangle + \langle h_j, u_{k_i}\rangle\right)\right\}\left(u_{k_j}+h_j\right) \\
&= \underbrace{\sum_{j=1}^m \exp\left(\lambda\langle u_{k_i}, u_{k_j}\rangle\right)u_{k_j}}_{:=a_i} + \underbrace{\sum_{j=1}^m \exp\left\{\lambda\left(\langle u_{k_i}, u_{k_j}\rangle + \langle h_i, h_j\rangle + \langle h_i, u_{k_j}\rangle + \langle h_j, u_{k_i}\rangle\right)\right\}h_j}_{:=b_i} \\
&\quad + \underbrace{\sum_{j=1}^m \left(\exp\left\{\lambda\left(\langle u_{k_i}, u_{k_j}\rangle + \langle h_i, h_j\rangle + \langle h_i, u_{k_j}\rangle + \langle h_j, u_{k_i}\rangle\right)\right\} - \exp\left(\lambda\langle u_{k_i}, u_{k_j}\rangle\right)\right)u_{k_i}}_{:=c_i}
\end{aligned}$$

Below we will bound a_i, b_i and c_i separately.

For a_i, we have

$$\left|a_i - e^\lambda m_i u_{k_i}\right| \le (m-m_i)\max\left\{\exp\left(\lambda\langle u_{k_i}, u_{k_j}\rangle\right) : k_j \ne k_i\right\} \le m\exp\left(\lambda\left|\sqrt{2}-\gamma_\ell\right|\right)$$

where the last step follows from Lemma 1. We then bound b_i. Since each summand in b_i is an independent random vector with zero mean, we have, with a probability $1-\delta$

$$|b_i| \leq mU_i \left(\frac{\max_{j\in[m]} |h_j|}{m} + \sigma\sqrt{\frac{2}{m}\log\frac{2}{\delta}} \right)$$

where

$$U_i = \max_{j\in[m]} \exp\left\{\lambda\left(\langle u_{k_i}, u_{k_j}\rangle + \langle h_i, h_j\rangle + \langle h_i, u_{k_j}\rangle + \langle h_j, u_{k_i}\rangle\right)\right\}$$

Since with a probability $1-3\delta$, for any $i,j\in[m]$ and $k\in[C]$, we have

$$|h_i| \leq 2\sigma,\ |\langle h_i, h_j\rangle| \leq 2\sigma^2\sqrt{\frac{1}{d}\log\frac{m}{2\delta\sigma^2}},\ |\langle h_i, u_k\rangle| \leq 2\sigma\sqrt{\frac{1}{d}\log\frac{m}{2\delta\sigma}}$$

and therefore

$$\begin{aligned} U_i &\leq \exp\left(\lambda\left|\sqrt{2}-\gamma_\ell\right|\right)\exp\left(\lambda\left[2\sigma^2\sqrt{\frac{1}{d}\log\frac{m}{2\delta\sigma^2}} + 4\sigma\sqrt{\frac{1}{d}\log\frac{m}{2\delta\sigma}}\right]\right) \\ &\leq \exp\left(\lambda\left[\left|\sqrt{2}-\gamma_\ell\right| + 6\sigma\sqrt{\frac{1}{d}\log\frac{m}{2\delta\sigma}}\right]\right) \end{aligned}$$

As a result, with a probability $1-4\delta$, for any $i\in[m]$, we have

$$\begin{aligned} |b_i| &\leq \sigma\exp\left(\lambda\left[\left|\sqrt{2}-\gamma_\ell\right| + 6\sigma\sqrt{\frac{1}{d}\log\frac{2m}{\delta\sigma}}\right]\right)\left(2+\sqrt{2m\log\frac{2m}{\delta}}\right) \\ &\leq 3\sigma\sqrt{m\log\frac{2m}{\delta}}\exp\left(\lambda\left[\left|\sqrt{2}-\gamma_\ell\right| + 6\sigma\sqrt{\frac{1}{d}\log\frac{2m}{\delta\sigma}}\right]\right) \end{aligned}$$

To bound c_i, we have

$$\begin{aligned} |c_i| &\leq \sum_{j=1}^m \exp\left(\lambda\langle u_{k_i}, u_{k_j}\rangle\right)\left|\exp\left(\lambda\left[\langle h_i, h_j\rangle + \langle h_i, u_{k_j}\rangle + \langle h_j, u_{k_i}\rangle\right]\right) - 1\right| \\ &\leq e^\lambda \sum_{j:k_j=k_i}\left|\exp\left(\lambda\left[|h_i|^2 + 2\langle h_i, u_{k_i}\rangle\right]\right) - 1\right| \\ &\quad + \sum_{j:k_j\neq k_i}\exp\left(\lambda\langle u_{k_i}, u_{k_j}\rangle\right)\left|\exp\left(\lambda\left[\langle h_i, h_j\rangle + \langle h_i, u_{k_j}\rangle + \langle h_j, u_{k_i}\rangle\right]\right) - 1\right| \end{aligned}$$

Using the bounds for $\langle h_i, h_j\rangle$, $|h_i|^2$, and $\langle h_i, u_k\rangle$, we have, with a probability $1-3\delta$

$$\begin{aligned} |c_i| &\leq m_i e^\lambda\left(\exp\left(2\lambda\left[\sigma^2 + 2\sigma\sqrt{\frac{1}{d}\log\frac{m}{\delta}}\right]\right) - 1\right) \\ &\quad + (m-m_i)\exp\left(\lambda\left|\sqrt{2}-\gamma_\ell\right|\right)\left(\exp\left(\lambda\left[6\sigma\sqrt{\frac{1}{d}\log\frac{2m}{\delta\sigma}}\right]\right) - 1\right) \\ &\leq \lambda\sigma\left(8m_i e^\lambda\sqrt{\frac{1}{d}\log\frac{m}{\delta}} + 12(m-m_i)\sqrt{\frac{1}{d}\log\frac{2m}{\delta\sigma}}\exp\left(\lambda\left|\sqrt{2}-\gamma_\ell\right|\right)\right) \\ &\leq 12m\lambda\sigma e^{\lambda|\sqrt{2}-\gamma_\ell|}\sqrt{\frac{1}{d}\log\frac{2m}{\delta\sigma}} + 8m_i\lambda\sigma e^\lambda\sqrt{\frac{1}{d}\log\frac{m}{\delta}} \end{aligned}$$

Combining the above results, we have, with a probability at least $1-4\delta$,

$$\begin{aligned}
&\left|\sum_{j=1}^{m}\exp\left(\langle x_i,x_j\rangle\right)x_j - e^{\lambda}m_i u_{k_i}\right| \\
&\quad\leq \exp\left(\lambda\left|\sqrt{2}-\gamma_\ell\right|\right)\left(m+6\sigma\sqrt{m\log\frac{2m}{\delta}}+12m\lambda\sigma\sqrt{\frac{1}{d}\log\frac{2m}{\delta\sigma}}\right)+8m_i\lambda\sigma e^{\lambda}\sqrt{\frac{1}{d}\log\frac{2m}{\delta\sigma}} \\
&\quad\leq 5m\exp\left(\lambda\left|\sqrt{2}-\gamma_\ell\right|\right)+8m_i\lambda\sigma e^{\lambda}\sqrt{\frac{1}{d}\log\frac{2m}{\delta\sigma}}
\end{aligned}$$

or

$$\left|\frac{e^{-\lambda}}{m_i}\sum_{j=1}^{m}\exp\left(\langle x_i,x_j\rangle\right)x_j - u_{k_i}\right| \leq \frac{5m}{m_i}\exp\left(-\lambda\left[1-\left|\sqrt{2}-\gamma_\ell\right|\right]\right) \leq \underbrace{\frac{2m}{m_i}\exp\left(-\frac{\lambda}{2}\right)+8\lambda\sigma\sqrt{\frac{1}{d}\log\frac{2m}{\delta\sigma}}}_{:=\nu_i}$$

We finally bound the partition function

$$\sum_{j=1}^{m}\exp(\lambda\langle x_j,x_i\rangle) \;=\; \sum_{j=1}^{m}\exp\left\{\lambda\left(\langle u_{k_i},u_{k_j}\rangle+\langle h_i,h_j\rangle+\langle h_i,u_{k_j}\rangle+\langle h_j,u_{k_i}\rangle\right)\right\}$$

Using the above bounds, we have, with a probability $1-3\delta$

$$\begin{aligned}
&\left|\sum_{j=1}^{m}\exp\left(\lambda\langle x_i,x_j\rangle\right)-m_i e^{\lambda}\right| \\
&\quad= \sum_{j:k_j=k_i} e^{\lambda}\left(\exp\left(\lambda\left[\langle h_i,h_j\rangle+\langle h_i,u_{k_i}\rangle+\langle h_j,u_{k_i}\rangle\right]\right)-1\right) \\
&\qquad+\sum_{j:k_j\neq k_i}\exp\left\{\lambda\left(\langle u_{k_i},u_{k_j}\rangle+\langle h_i,h_j\rangle+\langle h_i,u_{k_j}\rangle+\langle h_j,u_{k_i}\rangle\right)\right\} \\
&\quad\leq m_i e^{\lambda}\left(\exp\left(6\lambda\sigma\sqrt{\frac{1}{d}\log\frac{2m}{\delta\sigma}}\right)-1\right)+(m-m_i)\exp\left(\lambda\left|\sqrt{2}-\gamma_\ell\right|+6\lambda\sigma\sqrt{\frac{1}{d}\log\frac{2m}{\delta\sigma}}\right) \\
&\quad\leq 12m_i e^{\lambda}\lambda\sigma\sqrt{\frac{1}{d}\log\frac{2m}{\delta\sigma}}+m\exp\left(\lambda\left|\sqrt{2}-\gamma_\ell\right|+6\lambda\sigma\sqrt{\frac{1}{d}\log\frac{2m}{\delta\sigma}}\right)
\end{aligned}$$

Hence,

$$\frac{e^{-\lambda}}{m_i}\sum_{j=1}^{m}\exp(\lambda\langle x_i,x_j\rangle)\leq 1+\underbrace{12\lambda\sigma\sqrt{\frac{1}{d}\log\frac{2m}{\delta\sigma}}+\frac{2m}{m_i}\exp\left(-\frac{\lambda}{2}\right)}_{:=\tau_i}$$

Finally, with a high probability, we have

$$\begin{aligned}
&\left| \frac{\sum_{j=1}^m \exp\left(\langle x_i, x_j \rangle\right) x_j}{\sum_{j=1}^m \exp\left(\langle x_i, x_j \rangle\right)} - u_{k_i} \right| \\
&\leq \quad (1+\tau_i) \left| \frac{e^{-\lambda}}{m_i} \sum_{j=1}^m \exp\left(\langle x_i, x_j \rangle\right) x_j - u_{k_i} \right| + \tau_i \leq \tau_i + (1+\tau_i)\nu_i \leq \tau_i + 2\nu_i \\
&= \quad \frac{6m}{m_i} e^{-\lambda/2} + 20\lambda\sigma\sqrt{\frac{1}{d}\log\frac{2m}{\delta\sigma}}
\end{aligned}$$

When

$$20\lambda\sqrt{\frac{1}{d}\log\frac{2m}{\delta\sigma}} \leq \frac{1}{4}, \quad \frac{6m}{m_i}\exp\left(-\frac{\lambda}{2}\right) \leq \frac{\sigma}{4}$$

we have

$$\left| \frac{\sum_{j=1}^m \exp\left(\langle x_i, x_j \rangle\right) x_j}{\sum_{j=1}^m \exp\left(\langle x_i, x_j \rangle\right)} - u_{k_i} \right| \leq \frac{\sigma}{2}$$

□