A Survey of Lie Groups in Machine Learning

Juncheng Wan 120033910148

June 7th, 2021

1 Introduction

As a master student of Computer Science and Engineering, I currently focus on machine learning and data mining[1]. In this area, I use machine learning algorithm to approximately solve non-convex problems, such as approximating complex conditional distribution.

One example is **Machine Translation**, which aiming at translating sentence from one language to another language. Specifically, given source sentence $(x_1, \ldots, x_m)$, the machine needs to translate it into the target sentence $(y_1, \ldots, y_n)$, where $x_i, y_j, (1 \leq i \leq m, 1 \leq j \leq n)$ are all random variables. Thus, this problem is equal to approximating the distribution $p(y_1, \ldots, y_n | x_1, \ldots, x_m)$ with parameterized distribution $q_\theta(y_1, \ldots, y_n | x_1, \ldots, x_m)$, where θ indicates the model parameters. However, this problem is difficult for three reasons:

1. The maximal sentence length could be long. For example, in the common translation benchmarks, such as LDC Chinese-English task and WMT14 English-German task, there are sentences with length larger than 512.
2. The vocabulary size could also be large. The vocabulary size of English is from 50,000 to 10,000. Thus, we need to deal with at least 50000^{512} possibilities.
3. The complex phrase composition,grammatical rules, and syntactic structure.

The above problems is not only for machine translation, but also for other natural language processing problems, such as information retrieval, sequence labeling, and etc.

From my perspective, utilizing various symmetries of language have two advantages:

1. Reducing the size of the space to be modeled. For example, if there are sentences, the only difference between them is *the place of time adverbials*. Then, I think they are almost equal in the meaning, due to the symmetry of syntactic structure.
2. Providing interpretability of the model. Some machine learning algorithms are considered black box models and lack interpretability for parameters of submodule, such as neural networks.

As Lie group is a group of symmetries where the symmetries are continuous, I investigate Lie group in machine learning in this survey out of my interest.

[1]The website of our lab is `http://apex.sjtu.edu.cn/`.

2 Definition

In this section, I clarify basic concepts and definitions.[2]

Feature map vector space Let S be some set, the **feature map vector space** is defined as:

$$V \triangleq \{f | f : S \to \mathbb{R}\} \tag{1}$$

First, I want to introduce the concept of feature map. This is a space of scalar-valued functions on the set S, representing the features of each point in S. For example, in computer vision, S could be $\mathbb{R}^2$ which indicating the 2D coordinates and each $f \in V$ could be gray value function which assigning gray value for each pixel between $[0, 255]$. In natural language processing, S could be $\{1, 2, \ldots, 512\}$ indicating the discrete positions of words in the sentence and range$(f) \in [50000] \subseteq \mathbb{R}$ indicating the words assignment, where [50000] is the coded vocabulary.

G-equivariant Let G be a group, V_1, V_2 be two feature map vector spaces. A map $\Phi : V_1 \to V_2$ is **G-equivariant** with respect to actions ρ_1, ρ_2 of G acting on V_1, V_2 respectively if: $\Phi[\rho_1(g)f] = \rho_2(g)\Phi[f]$ for any $g \in G, f \in V_1$. Then, G-equivariant, a concept from representation theory in undergraduate, is necessary. Because in machine learning I am interested in those operation Φ that is equivariant under the group action. For example, there are rotations and transition in images. I use Φ to map the shallow color features contained in V_1 to useful textural features V_2 and require that they are the same despite the order of rotation or transition of the object.

Group equivariant convolution As a natural generalization of convolutional convolution operation, group equivariant convolution [**?**] enjoys a substantially higher degree of weight sharing than regular convolution layers. The **group equivariant convolution** $\Psi : \mathcal{I}_U \to \mathcal{I}_U$ is defined as :

$$[\Psi f](g) \triangleq \int_G \psi\left(g'^{-1} g\right) f\left(g'\right) dg' \tag{2}$$

where $\psi : G \to \mathbb{R}$ is the convolutional filter and the integral is defined withe respect tot the left Haar mearsure of G.

Define $V = \{f | f : G \to \mathbb{R}\}$ to be the space of scalar-valued functions on the group G, for which we can define a **regular representation** π acting on V as follows:

$$\left[\pi\left(g_\theta\right)(f)\right]\left(g_\phi\right) \triangleq f\left(g_\theta^{-1} g_\phi\right) \tag{3}$$

3 Results

The function composition $f \circ f_K \circ \cdots \circ f_1$ of several equivariant functions $f_k, k \in 1, 2, \ldots, K$ followed by an invariant function f, is an invariant function. Consider group representations $\pi_1, \ldots, \pi_K$ that act on $f_1, \ldots, f_K$ respectively, and representation π_0 that acts on the input space of f_1. If each f_k is equivariant

[2] Though some concepts are fundemental for students of mathematics, they are fresh for me. Thus, I also write them down.

with respect to π_k; π_{k1} such that $f_k \circ \pi_{k1} = \pi_k \circ f_k$, and f is invariant such that $f \circ \pi_k = f$, then we have:

$$\begin{aligned} f \circ f_k \circ \ldots \circ f_1 \circ \pi_0 &= f \circ f_k \circ \ldots \circ \pi_1 \circ f_1 \\ &\vdots \\ &= f \circ \pi_k \circ f_k \circ \ldots \circ f_1 \\ &= f \circ f_k \circ \ldots \circ f_1 \end{aligned} \tag{4}$$

hence $f \circ f_K \circ \cdots \circ f_1$ is invariant.

The group equivariant convolution $\Psi : \mathcal{I}_U \to \mathcal{I}_U$ defined as: $[\Psi f](g) \triangleq \int_G \psi\left(g'^{-1}g\right) f\left(g'\right) dg'$ is equivariant with resepct to the regular representation π of G acting on $\mathcal{I}_U$ as $\left[\pi\left(u\right)\left(f\right)\right]\left(g\right) \triangleq f\left(u^{-1}g\right)$.

Use the invariance of the left Haar measure.

$$\begin{aligned} \Psi[\pi(u)f](g) &= \int_G \psi\left(g'^{-1}g\right)\left[\pi(u)f\right]\left(g'\right) dg' \\ &= \int_{uG} \psi\left(g'^{-1}g\right) f\left(u^{-1}g'\right) dg' \\ &= \int_G \psi\left(g'^{-1}u^{-1}g\right) f\left(g'\right) dg' \\ &= \left[\Psi f\right]\left(u^{-1}g\right) \\ &= \left[\pi(u)[\Psi f]\right](g) \end{aligned} \tag{5}$$

The lifting layer $\mathcal{L}$ is equivariant with respect to the representation π.

Note $\mathcal{L}\left[\pi(u)f_{\mathcal{X}}\right]\left(g\right) = \mathbf{f}_i$ for $g \in s(ux_i)H$ and $\left[\pi(u)\mathcal{L}\left[f_{\mathcal{X}}\right]\right]\left(g\right) = \mathcal{L}[f_{\mathcal{X}}]\left(u^{-1}g\right) = \mathbf{f}_i$ for $g \in us(x_i)H$. Hence $\mathcal{L}\left[\pi(u)f_{\mathcal{X}}\right] = \pi(u)\mathcal{L}\left[f_{\mathcal{X}}\right]$ because the two cosets are equal: $s(ux_i)H = us(x_i)H, \forall u \in G$.

LieSelfAttention is equivariant with respect to the regular representation π.

Let $\mathcal{I}_U = \mathcal{L}(G, \mathbb{R}^D)$ be the space of unconstrained functions $f : G \to \mathbb{R}^D$. We can define the regular representation π of G acting on $\mathcal{I}_U$ as follows:

$$\left[\pi(u)f\right]\left(g\right) = f(u^{-1}g) \tag{6}$$

f is defined on the set $G_f = \bigcup_{i=1}^n s(x_i)H$ (i.e. union of cosets corresponding to each x_i). Note $G_{\pi(u)f} = uG_f$, and G_f does not depend on the choice of section s.

Note that for all provided choices of k_c and k_l, we have:

$$\begin{aligned} k_c\left(\left[\pi(u)f\right](g), \left[\pi(u)f\right]\left(g'\right)\right) &= k_c\left(f\left(u^{-1}g\right), f\left(u^{-1}g'\right)\right) \\ k_l\left(g^{-1}g'\right) &= k_l\left(\left(u^{-1}g\right)^{-1}\left(u^{-1}g'\right)\right) \end{aligned} \tag{7}$$

Hence for all choices of F, we have that

$$\begin{aligned} \alpha_{\pi(u)f}\left(g, g'\right) &= F\left(k_c\left(\left[\pi(u)f\right](g), \left[\pi(u)f\right]\left(g'\right)\right), k_l\left(g^{-1}g'\right)\right) \\ &= F\left(k_c\left(f\left(u^{-1}g\right), f\left(u^{-1}g'\right)\right), k_l\left(\left(u^{-1}g\right)^{-1}u^{-1}g'\right)\right) \\ &= \alpha_f\left(u^{-1}g, u^{-1}g'\right) \end{aligned} \tag{8}$$

We thus prove equivariance for the below choice of LieSelfAttention $\Phi : \mathcal{I}_U \to \mathcal{I}_U$ that uses softmax normalisation, but a similar proof holds for constant normalisation. Let $A_f(g, g') \triangleq \exp(\alpha_f(g, g'))$, hence Equation (10) also holds for A_f:

$$\begin{aligned}(g) &= \int_{G_f} w_f(g, g') f(g') \, dg' \\ &= \int_{G_f} \frac{A_f(g, g')}{\int_{G_f} A_f(g, g'') \, dg''} f(g') \, dg'\end{aligned} \tag{9}$$

Hence:

$$\begin{aligned}w_{\pi(u)f}(g, g') &= \frac{A_{\pi(u)f}(g, g')}{\int_{G_{\pi(u)f}} A_{\pi(u)f}(g, g'') \, dg''} \\ &= \frac{A_f\left(u^{-1}g, u^{-1}g'\right)}{\int_{uG_f} A_f(u^{-1}g, u^{-1}g'') \, dg''} \\ &= \frac{A_f\left(u^{-1}g, u^{-1}g'\right)}{\int_{G_f} A_f(u^{-1}g, g'') \, dg''} \\ &= w_f\left(u^{-1}g, u^{-1}g'\right)\end{aligned} \tag{10}$$

Then we can show that Φ is quivariant with respect to the representation π as follows:

$$\begin{aligned}\Phi[\pi(u)f](g) &= \int_{G_{\pi(u)f}} w_{\pi(u)f}(g, g') \, [\pi(u)f](g') \, dg' \\ &= \int_{uG_f} w_f\left(u^{-1}g, u^{-1}g'\right) f\left(u^{-1}g'\right) dg' \\ &= \int_{G_f} w_f\left(u^{-1}g, g'\right) f(g') \, dg' \\ &= [\Phi f]\left(u^{-1}g\right) \\ &= [\pi(u)[\Phi f]](g)\end{aligned} \tag{11}$$

Let operator $\mathcal{K} : \mathbb{L}_2(X) \to \mathbb{L}_2(Y)$ be linear and bounded, let X, Y be homogeneous spaces on which Lie group G act transitively, and $d\mu_X$ a Radon measure on X, then:

1. K is a kernel operator, i.e., $\exists_{\tilde{k} \in \mathbb{L}_1(Y \times X)} : (\mathcal{K}f)(y) = \int_X \tilde{k}(y, x) f(x) \mathrm{d}\mu_X$
2. under the G-equivariance constraint of Eq. (3) the map is defined by a one-argument kernel:

$$\tilde{k}(y, x) = \frac{\mathrm{d}\mu_X\left(g_y^{-1} \odot x\right)}{\mathrm{d}\mu_X(x)} k\left(g_y^{-1} \odot x\right) \tag{12}$$

   for any $g_y \in G$ such that $y = g_y \odot y_0$ for some fixed origin $y_0 \in Y$
3. if $Y \equiv G/H$ is the quotient of G with $H = \mathrm{Stab}_G(y_0) = \{g \in G \mid g \odot y_0 = y_0\}$ then the kernel is constrained via:

$$\forall_{h \in H}, \forall_{x \in X} : \quad k(x) = \frac{\mathrm{d}\mu_X\left(g_y^{-1} \odot x\right)}{\mathrm{d}\mu_X(x)} k\left(h^{-1} \odot x\right) \tag{13}$$

1. It follows from Dunford-Pettis Theorem, that if $\mathcal{K}$ is linear and bounded it is an integral operator.

2. The left-equivariance constraint then imposes bi-left-invariance of the kernel $\tilde{k}$ as follows, where $\forall_{g\in G}$ and $\forall_{f\in\mathbb{L}_2(X)}$:

$$\begin{aligned}
&\left(\mathcal{K}\circ\mathcal{L}_g^{G\to\mathbb{L}_2(X)}\right)(f) = \left(\mathcal{L}_g^{G\to\mathbb{L}_2(Y)\circ\mathcal{K}}\right)(f) \quad \Leftrightarrow \\
&\int_X \tilde{k}(y,x) f\left(g^{-1}x\right)\mathrm{d}x = \int_X \tilde{k}\left(g^{-1}y,x\right) f(x)\mathrm{d}x \overset{\text{in r.h.s. integral}}{\Leftrightarrow} x \leftarrow g^{-1}x \\
&\int_X \tilde{k}(y,x) f\left(g^{-1}x\right)\mathrm{d}x = \int_X \tilde{k}\left(g^{-1}y,g^{-1}x\right) f\left(g^{-1}x\right)\mathrm{d}\left(g^{-1}x\right) \Leftrightarrow \\
&\int_X \tilde{k}(y,x) f\left(g^{-1}x\right)\mathrm{d}x = \int_X \tilde{k}\left(g^{-1}y,g^{-1}x\right) f\left(g^{-1}x\right)\frac{1}{|\det g|}\mathrm{d}x
\end{aligned} \tag{14}$$

   Since the final equation holds for all $f\in\mathbb{L}_2(X)$ we obtain:

$$\forall_{g\in G}: \quad \tilde{k}(y,x) = \frac{1}{|\det g|}\tilde{k}\left(g^{-1}y,g^{-1}x\right) \tag{15}$$

   Furthermore, since G acts transitively on Y we have that $\forall_{y,y_0\in Y}\exists_{g_y\in G}$ such that $y=g_y y_0$ and thus

$$\tilde{k}(y,x) = \tilde{k}\left(g_y y_0,x\right) = \frac{1}{|\det g_y|}\tilde{k}\left(y_0,g_y^{-1}x\right) =: \frac{1}{|\det g_y|}k\left(g_y^{-1}x\right) \tag{16}$$

   for every $g_y\in G$ such that $y=g_y y_0$ with arbitrary fixed origin $y_0\in Y$.

3. Every homogeneous space Y of G can be identified with a quotient group $G=H$. Choose an origin $y_0\in Y$ s.t. $\forall_{h\in H}: hy_0=y_0$, i.e., $H=\mathrm{Stab}_G\, y_0$, then

$$\tilde{k}\left(y_0,x\right) = \tilde{k}\left(hy_0,x\right) \Leftrightarrow k(x) = \frac{1}{|\det h|}k\left(h^{-1}x\right) \tag{17}$$

4 Conclusion