A Simple Reparameterization to Accelerate Training of Deep Neural Networks

November 2, 2017

1 Resúmen

La normalización de los pesos de una red neuronal a través de la reparametrización de vectores pesos nos da una optimización del problema, además acelera la convergencia del descenso de gradiente estocástico. La reparameterización usada en la presente investigación se basa en la normalización por lotes, pero no introduce ninguna dependencia entre los ejemplos en un minilote. Es decir que el presente método también puede aplicarse con éxito a modelos recurrentes como LSTMs y a aplicaciones sensibles al ruido como el aprendizaje de refuerzo profundo o modelos generativos, para los cuales la normalización por lotes es menos adecuada. La sobrecarga computacional del método es pequeña, de esta manera permite más pasos de optimización para ser usados en la misma cantidad de tiempo. Finalmente se presentará ejemplos donde se visualizará la gran utilidad de este método a través de algunas aplicaciones.

2 Introducción

Los éxitos recientes en el aprendizaje profundo han demostrado que las redes neuronales entrenadas por la optimización basada en el gradiente de primer orden son capaces de lograr resultados asombrosos en diversos campos como la visión por computador, el reconocimiento del habla y el modelado del lenguaje. Sin embargo, este método depende en gran medida de la curvatura del objetivo que se optimiza. Si el número de condición de la matriz de Hessiana del objetivo en el óptimo es bajo, entonces el descenso de gradiente de primer orden tendrá dificultad para progresar. La cantidad de curvatura, y la optimización que se plantea, no es invariante a la reparameterización: puede haber múltiples formas equivalentes de parametrizar el mismo modelo. Entonces el principal objetivo sería encontrar la mejor forma de por parametrizar una red neuronal.

Se han desarrollado varios métodos para mejorar el acondicionamiento del gradiente de costos para arquitecturas de redes neuronales generales. Un enfoque consiste en multiplicar explícitamente el gradiente de costos por una inversa aproximada de la matriz de información de Fisher, obteniendo así un gradiente natural aproximadamente blanqueado. Alternativamente, podemos usar descenso de gradiente de primer orden estándar sin precondicionamiento, pero cambiar la parametrización de nuestro modelo para dar gradientes que son más parecidos a los gradientes naturales blanqueados de estos métodos.

Por ejemplo, podemos transformar las salidas de cada neurona para que tengan salida cero y pendiente cero en promedio. De esta manera la transformación diagonaliza aproximadamente la matriz de información de Fisher, blanqueando así el gradiente, conduciendo a un mejor desempeño de optimización. Otro enfoque en esta dirección es la normalización por lotes, un método en

el que la salida de cada neurona (antes de la aplicación de la no linealidad) se normaliza por la media y desviación estándar de las salidas calculadas sobre los ejemplos en el minibatch. Esto reduce el desplazamiento covariable de las salidas de las neuronas y los autores sugieren que también aproxima la matriz de Fisher a la matriz de identidad.

Siguiendo este segundo enfoque para aproximar la optimización del gradiente natural, se presenta un método simple pero general, llamado normalización del peso, para mejorar la optimización de los pesos de los modelos de red neuronal. El método se inspira en la normalización por lotes, pero es un método determinista que no comparte la propiedad de la normalización por lotes de añadir ruido a los gradientes. Además, la sobrecarga impuesta por nuestro es menor: no se requiere memoria adicional y el cálculo adicional es insignificante.

3 Normalización del Peso

Consideramos redes neuronales artificiales estándar donde el cálculo de cada neurona consiste en tomar una suma ponderada de características de entrada:

$$y = \phi(w.x + b)$$

donde w es un vector de peso k-dimensional, b es un término de polarización escalar, x es un vector k-dimensional de características de entrada, $\phi(.)$ denota una no linealidad elemental que denota la salida escalar de la neurona. Después de asociar una función de pérdida a una o más salidas neuronales, dicha red neuronal es comúnmente entrenada por el descenso de gradiente estocástica en los parámetros w, b de cada neurona.

Con la intención de acelerar la convergencia de este procedimiento de optimización, se realizará la reparameterización de cada vector de peso w en términos de un vector de parámetro v y un parámetro escalar g y se calculará el descenso de gradiente estocástico con respecto a esos parámetros. La expresión del vector quedaría expresado de la siguiente forma:

$$w = \frac{g}{||v||}v$$

donde v es un vector k-dimensional, g es un escalar, y $||v||$ denota la norma euclidiana de v. Esta reparameterización tiene el efecto de fijar la norma euclidiana del vector de peso w, siendo ahora $||w|| = g$, independiente de los parámetros v.

Investigaciones anteriores también desarrollaban la idea de normalizar el vector de peso, pero la optimización solo se realizaba mediante la parametrización de w, aplicando solamente la normalización después de cada paso de descenso de gradiente estocástico. Con el presente método se reparameteriza explícitamente el modelo y realizar un descenso de gradiente estocástico en los nuevos parámetros v, g directamente. De esta forma se mejora el acondicionamiento del gradiente y conduce a una convergencia mejorada del procedimiento de optimización. Al desacoplar la norma del vector de peso(g) de la dirección del vector de peso ($v/||v||$), se acelera la convergencia de nuestra optimización de descenso de gradiente estocástico.

3.1 Gradientes

El entrenamiento de una red neuronal mediante la nueva parametrización se realiza utilizando métodos estándar de descenso gradiente estocástico. De esta forma se obtiene el gradiente de una función de pérdida L con respecto a los nuevos parámetros v, g. De la siguiente forma:

$$_gL = \frac{_wL.v}{||v||}, _vL = \frac{g}{||v||}{}_wL - \frac{g_gL}{||v||^2}v$$

donde $_wL$ es el gradiente con respecto a los pesos w que se usan normalmente.

Por lo tanto, la retropropagación mediante la normalización del peso sólo requiere una modificación menor de las ecuaciones habituales y se implementa fácilmente utilizando software de red neural estándar, ya sea especificando directamente la red en términos de los parámetros v, g y dependiendo de la auto-diferenciación o aplicando la ecuación anterior en una etapa posterior al procesamiento. A diferencia de la normalización por lotes, las expresiones anteriores son independientes del tamaño del minilote y, por tanto, causan sólo una sobrecarga computacional mínima.

Una forma alternativa de escribir el gradiente es:

$$_vL = \frac{g}{||v||}M_{ww}L$$

con

$$M_w = I - \frac{ww'}{||w||^2}$$

donde M_w es una matriz de proyección que se proyecta sobre el complemento del vector w. Esto demuestra que la normalización del peso cumple dos cosas: escala el gradiente de peso por $g/||v||$ y proyecta el gradiente lejos del vector de peso actual. Ambos efectos ayudan a aproximar la matriz de covarianza del gradiente a la optimización de la identidad y los beneficios.

Debido a la proyección lejos de w, la norma de v crece monotónicamente con las actualizaciones de los peso cuando una red neuronal aprende con la normalización de peso usando el descenso de gradiente estándar sin momento. Sea $v' = v + v$ denotando la actualización de parámetros, con $v \propto {}_vL$ (ascenso / descenso más pronunciado), entonces v es necesariamente ortogonal al vector de peso actual w, ya que nos proyectamos lejos de él al calcular $_vL$. Puesto que v es proporcional a w, la actualización es también ortogonal a v y aumenta su norma mediante el teorema de Pitágoras. Específicamente, si $||v||/||v|| = c$ el nuevo vector de peso tendrá la norma $||v'|| = \sqrt{||v||^2 + c^2||v||^2} = \sqrt{1 + c^2}||v||||v||$. La tasa de incremento dependerá de la varianza del gradiente de peso. Si nuestros gradientes son ruidosos, c será alto y la norma de v aumentará rápidamente, lo que a su vez reducirá el factor de escala $g/||v||$. Si la norma de los gradientes es pequeña, obtendremos $\sqrt{1 + c^2} \approx 1$, y la norma de v dejará de aumentar. Usando este mecanismo, el gradiente escalado se autoestabiliza su norma.

Empíricamente, la capacidad de crecer de la norma $||v||$ hace que la optimización de redes neuronales con normalización de peso sea muy robusta al valor de la tasa de aprendizaje. Si la tasa de aprendizaje es demasiado grande, la norma de los pesos no normalizados crece rápidamente hasta alcanzar una tasa de aprendizaje efectiva adecuada. Una vez que la norma de los pesos ha crecido grande con respecto a la norma de las actualizaciones, la tasa de aprendizaje eficaz se estabiliza. Por lo tanto, las redes neuronales con normalización de peso funcionan bien con un rango mucho más amplio de tasas de aprendizaje que cuando se usa la parametrización normal. A su vez, las redes neuronales con normalización por lotes también poseen esta propiedad, y pueden ser explicado por este análisis.

Al proyectar el gradiente lejos del vector de peso w, también eliminamos el ruido en esa dirección. Si la matriz de covarianza del gradiente con respecto a w viene dada por C, la matriz de covarianza del gradiente en v viene dada por $D = (g^2/||v||^2)M_wCM_w$. Empíricamente, se determina que w es a menudo un autovector dominante de la matriz de covarianza C, eliminando ese

vector propio da una nueva matriz de covarianza D más cercana a la matriz de identidad, lo que puede acelerar aún más el aprendizaje.

3.2 Relación con la normalización por lotes

Para realizar la reparameterización se usará la normalización por lotes, el cual normaliza las estadísticas de la preactivación t para cada minilote como:

$$t' = \frac{t - \mu[t]}{\sigma[t]}$$

con $[t]$, $[t]$ la media y desviación estándar de las pre-activaciones $t = vx$. Para el caso especial en el que nuestra red sólo tiene una sola capa, y las características de entrada x para esa capa son blanqueadas (distribuidas independientemente con media cero y varianza unitaria), estas estadísticas son dadas por $[t] = 0$ y $[t] = ||v||$. En ese caso, la normalización de las pre-activaciones mediante normalización por lotes es equivalente a la normalización de los pesos mediante la normalización del peso.

Las redes neuronales convolucionales suelen tener mucho menos peso que las pre-activaciones, por lo que normalizar los pesos es a menudo mucho más barato computacionalmente. Además, la norma de v es no estocástica, mientras que la media de minibatch $[t]$ y la varianza $^2[t]$ pueden tener en general una varianza alta para el tamaño de minibatch pequeño. Por lo tanto, la normalización del peso puede verse como una aproximación más barata y menos ruidosa a la normalización por lotes. Aunque la equivalencia exacta no suele mantenerse para arquitecturas más profundas, todavía se encuentra que el método de normalización de peso proporciona gran parte de la aceleración de la normalización por lotes completa.

4 Inicialización de parámetros dependiente de datos

Además de un efecto de reparameterización, la normalización por lotes también tiene el beneficio de fijar la escala de las características generadas por cada capa de la red neuronal. Esto hace que la optimización sea robusta frente a las inicializaciones de parámetros para las cuales estas escalas varían entre capas. Dado que la normalización del peso carece de esta propiedad, es importante inicializar adecuadamente estos parámetros. Entonces se realizará un muestreo de los elementos de v mediante una distribución simple con una escala fija, por ejemplo: distribución normal con media cero y desviación estándar 0,05. Antes de iniciar el entrenamiento, se inicializa los parámetros b y g para fijar las estadísticas de minibatch de todas las pre-activaciones en la red, como en la normalización por lotes, pero solo para un único minibatch de datos y sólo durante la inicialización. Esto se puede realizar de manera eficiente realizando un primer paso de feedforward a través de nuestra red para un único minibatch de datos X, usando el siguiente cálculo en cada neurona:

$$t = \frac{v.x}{||v||}$$

and

$$y = \phi(\frac{t - \mu[t]}{\sigma[t]})$$

donde $[t]$ y $[t]$ son la media y la desviación estándar de la preactivación t sobre los ejemplos en el minibatch. Entonces podemos inicializar la biasa b de la neurona y la escala g como:

$$g \leftarrow \frac{1}{\sigma[t]}, b \leftarrow \frac{-\mu[t]}{\sigma[t]}$$

de manera que $y = \phi(w.x + b)$. Como la normalización por lotes, este método asegura que todas las características inicialmente tienen media cero y varianza unitaria antes de la aplicación de la no linealidad. Con el presente método esto sólo es válido para el minibatch que utilizamos para la inicialización, y los minibutches subsiguientes pueden tener estadísticas ligeramente diferentes, pero experimentalmente se encuentra que este método de inicialización funciona bien. El método también puede aplicarse a redes sin normalización de peso, simplemente haciendo una optimización de gradiente estocástico en los parámetros w directamente, después de la inicialización en términos de v y g. La desventaja de este método de inicialización es que sólo puede aplicarse en casos similares como donde la normalización de lote es aplicable. Para modelos con recursividad, como RNNs y LSTMs, tendremos que recurrir a métodos de inicialización estándar.

5 Normalización de lotes de sólo media

La normalización del peso, hace que la escala de las activaciones neuronales sea aproximadamente independiente de los parámetros v. Sin embargo, a diferencia de la normalización por lotes, las medias de las activaciones neuronales siguen dependiendo de v. con una versión especial de la normalización por lotes, a la que llamamos normalización de lote sólo media. Con este método de normalización, restamos los medias de minibatch como con la normalización de lotes completos, pero no dividimos por las desviaciones estándar de minibatch. Es decir, calculamos las activaciones neuronales utilizando:

$$t = w.x$$
$$t' = t - \mu[t] + b$$
$$y = \phi(t')$$

donde w es el vector de peso, parametrizado usando la normalización del peso, y $[t]$ es la media del minibatch de la preactivación t. Durante el entrenamiento, mantenemos un promedio de la media del minibatch que sustituimos por $[t]$ en el tiempo de prueba.

El gradiente de la pérdida con respecto a la preactivación t se calcula como:

$${}_tL = {}_{t'}L - \mu[{}_{t'}L]$$

donde $[.]$ denota una vez más la operación para tomar la media del minibatch. Por lo tanto, la normalización por lotes de media sólo tiene el efecto de centrar los gradientes que son retropropagados. Esta es una operación comparativamente barata, y la sobrecarga computacional de la normalización de lotes de sólo media es por lo tanto menor que para la normalización por lotes completa. Además, este método produce menos ruido durante el entrenamiento, y el ruido que se produce es más suave, ya que la ley de grandes números asegura que $[t]$ y $[t]$ están aproximadamente distribuidos normalmente.