

Daniel Ricardo Rodriguez Olarte

Guión proyecto machine Learning

Está organizado como # de diapositiva al lado el tiempo que se demora el diálogo y abajo el diálogo:

1 /9 segundos:

Buenas tardes mi nombre es Daniel Ricardo Rodriguez Olarte y hoy voy a presentar acerca de la generalización en Deep Learning

2/11 segundos

Problema general

¿Qué prácticas promueven y no la generalización? ¿Qué mide y qué no mide la generalización?

3/8 segundos

Problema específico

El objetivo es comprender la capacidad de modelo eficaz de las redes neuronales de avance.

¿La regularización es una razón fundamental para explicar la generalización?

4/16 segundos

¿Por qué es importante explicar la generalización en aprendizaje de máquina?

La generalización se refiere a la capacidad de un modelo para adaptarse correctamente a datos nuevos, nunca antes vistos, por lo que esta situación plantea un desafío conceptual a la teoría del aprendizaje estadístico, ya que las medidas tradicionales de complejidad del modelo luchan por explicar la capacidad de generalización de las grandes redes neuronales artificiales

5/18 segundos

¿Entonces cómo se busca la generalización en un modelo?

En un primer paso de optimización, llamado *entrenamiento*, ajustamos un modelo a un conjunto de datos.

En un segundo paso, llamado *prueba*, juzgamos el modelo por su rendimiento con los datos recién generados del mismo proceso

6/11 segundos

Arquitectura CNN

En Deep Learning, una red neuronal convolucional (CNN) es una clase de redes neuronales profundas, que se aplica más comúnmente al análisis de imágenes visuales. Las redes neuronales convolucionales son modelos de última generación para clasificación de imágenes, segmentación, detección de objetos y muchas otras tareas de procesamiento de imágenes.

7/24 segundos

Experimentos

Daniel Ricardo Rodriguez Olarte

Para estos experimentos se usa las arquitecturas CNN: Inception, Alexnet y MLP

Se usa un conjunto de datos llamados CIFAR-10 e ImageNet

CIFAR-10 consta de 60000 imágenes en color de 32x32 en 10 clases, con 6000 imágenes por clase. Hay 50000 imágenes de entrenamiento y 10000 imágenes de prueba.

8/21 segundos

Para estos experimentos Se tomó una arquitectura candidata y se entrenó tanto con los datos verdaderos como con una copia de los datos en la que las etiquetas verdaderas fueron reemplazadas por etiquetas aleatorias.

. Y Para obtener más información sobre este fenómeno..... Se experimentó con diferentes niveles de aleatorización explorando el continuo entre el ruido sin etiqueta y las etiquetas completamente corruptas

9/15segundos

Se realizo los experimentos con las siguientes modificaciones de las etiquetas y las imágenes de entrada

Etiquetas verdaderas, Etiquetas parcialmente dañadas, Etiquetas aleatorias, Píxeles mezclados:, Píxeles aleatorios, Gaussiana:

10 /35 segundos

¿ qUÉ SE OBTUVO EN los EXPERIMENTOS?

Esta figura muestra las curvas de aprendizaje del modelo Inception en el conjunto de datos CIFAR10 en varias configuraciones , Sorprendentemente, aunque las etiquetas aleatorias destruyen por completo la relación entre imágenes y etiquetas. El descenso de gradiente estocástico con configuraciones de hiperparámetros sin cambios puede optimizar los pesos para que se ajusten perfectamente a etiquetas aleatorias

Y aunque se rompió aún más la estructura de las imágenes mezclando los píxeles de la imagen e incluso volviendo a muestrear por completo los píxeles aleatorios de una distribución gaussiana la red sigue encajando después de pasar por el conjunto de entrenamiento varias veces

11/19 segundos

Etiquetas parcialmente dañadas

Entonces en las etiquetas parcialmente dañadas Se inspeccionó el comportamiento del entrenamiento de redes neuronales con un nivel variable de corrupción de etiquetas de 0 (sin corrupción) a 1 (etiquetas aleatorias completas) en el conjunto de datos CIFAR10.

, sin embargo como ya habia mencionado.... Las redes se ajustan perfectamente al conjunto de entrenamiento corrupto para todos los casos.

12/28 segundos

Comparación de arquitecturas

Bueno comparando las 3 arquitecturas antes mencionadas entonces en la figura b se muestra la desaceleración del tiempo de convergencia con el aumento del nivel de ruido de las etiquetas

y en la figura c se muestra los errores de prueba después de la convergencia.

Dado que los errores de entrenamiento son siempre cero, los errores de prueba son los mismos que los errores de generalización. A medida que el nivel de ruido se acerca a 1, los errores de generalización convergen al 90%

13/18 segundos

Conclusiones acerca de los experimentos serian

No necesitamos cambiar el ritmo de aprendizaje del programa

Una vez que comienza la adaptación, converge rápidamente

Converge para adaptarse perfectamente al conjunto de entrenamiento.

Las redes neuronales convolucionales pueden encajar ruido aleatorio.

14/10 segundos

Implicaciones

A la luz de los experimentos de aleatorización se discutió cómo los hallazgos plantean un desafío para varios enfoques tradicionales para el razonamiento sobre la generalización

15/31 segundos

La complejidad de Rademacher se usa comúnmente y es una medida de complejidad flexible de una clase de hipótesis.

La complejidad empírica de Rademacher de una función de clase F en un conjunto de datos $\{z_1, \dots, z_n\}$ se define como:

donde $\sigma_1, \dots, \sigma_n \in \{\pm 1\}$ son independientes e igualmente distribuidas variables aleatorias uniformes.

Específicamente, $R_n(H)$ mide la capacidad de H para ajustarse a asignaciones de etiquetas binarias al azar ± 1 . Si bien consideramos problemas multiclase, es sencillo considerar problemas de clasificación binaria relacionados para los que se mantienen las mismas observaciones experimentales.

16/42 segundos

Bueno entonces la definición anterior se parece mucho a nuestra prueba de aleatorización

Dado que nuestras Dado que nuestras pruebas de aleatorización sugieren que muchas redes neuronales se ajustan perfectamente al conjunto de entrenamiento con etiquetas aleatorias, esperamos que $R_n(H) \approx 1$ para la clase de modelo correspondiente H .

Este es, por supuesto, un límite superior trivial en la complejidad de Rademacher que no conduce a límites de generalización útiles en entornos realistas, como la dimensión vc tiene un razonamiento muy

similar al de Rademacher entonces concluimos que tampoco sirve para explicar el rendimiento de generalización de redes neuronales de última generación.

17/16 segundos

El papel de la regularización

La mayoría de las pruebas de aleatorización se realizan con la regularización explícita desactivada.

Los regularizadores son la herramienta estándar en teoría y práctica para mitigar el sobreajuste en el régimen cuando hay más parámetros que puntos de datos.

La idea básica es que aunque la hipótesis original es demasiado grande para generalizar bien, los regularizadores ayudan a confinar el aprendizaje a un subconjunto del espacio de hipótesis con complejidad manejable.

18/26 segundos

AIGUNOS REGULARIZADORES

En lugar de hacer una encuesta completa de todo tipo de técnicas de regularización introducidas para el aprendizaje profundo, simplemente tomamos varias arquitecturas de red de uso común y comparamos el comportamiento al apagar los regularizadores equipados. Se cubren los siguientes regularizadores:.....

Aumento de datos: se aumenta el conjunto de entrenamiento mediante transformaciones específicas de dominio. Para los datos de imagen, las transformaciones de uso común incluyen recorte aleatorio, perturbación aleatoria de brillo, saturación, tono y contraste.

Deterioro de peso: equivalente a una restricción estricta de los pesos a una bola euclidiana, con el radio decidido por la cantidad de caída del peso.

Abandono :enmascara cada elemento de la salida de una capa al azar con una probabilidad de expulsión determinada.

19/17 segundos

La Tabla muestra los resultados de Inception, Alexnet y MLP en CIFAR10, alternando el uso de aumento de datos y disminución de peso.

Ambas técnicas de regularización ayudan a mejorar el rendimiento de la generalización

.....pero incluso con todos los regularizadores desactivados, todos los modelos siguen generalizándose muy bien. Es difícil decir que los regularizadores cuentan como un cambio de fase fundamental en la capacidad de generalización de redes profundas.

20/15 segundos

La normalización por lotes es un método que se utiliza para hacer que las redes neuronales artificiales sean más rápidas y estables mediante la normalización de las entradas de las capas al volver a centrar y escalar.

Aunque no está diseñado explícitamente para la regularización, la normalización por lotes suele mejorar el rendimiento de la generalización

21/38 segundos

Con respecto a la regularización implícita en estas figura podemos ver los efectos de los regularizadores implícitos sobre el desempeño de la generalización

Donde aug es aumento de datos; wd es la disminución del peso; BN es normalización por lotes.

Entonces en la figura a se muestra la precisión del entrenamiento y las pruebas en ImageNet. El área sombreada indica la mejor precisión acumulada de la prueba, como referencia de la ganancia potencial de rendimiento para una parada anticipada.

Y en la figura b compara las curvas de aprendizaje de las dos variantes de Inception en CIFAR10, con todos los regularizadores explícitos desactivados.

Sin embargo se puede ver que en el conjunto de datos CIFAR10, no observamos ningún beneficio potencial de la interrupción temprana.

22/9 segundos

El descenso de gradiente estocástico Es un método iterativo para optimizar una función objetivo con propiedades de suavidad adecuadas Además es un algoritmo popular para entrenar una amplia gama de modelos en machine learning

23/28 segundos

Regularización implícita

En las redes neuronales, casi siempre elegimos nuestro modelo como la salida de la ejecución del descenso de gradiente estocástico.

Apelando a los modelos lineales, analizamos cómo SGD actúa como un regularizador implícito.

Para modelos lineales, SGD siempre converge a una solución con una pequeña norma.

Por lo tanto, el algoritmo en sí mismo está regularizando implícitamente la solución.

24/26 segundos

¿ Se puede explicar la generalización con la regularización ?

Las observaciones sobre los regularizadores explícitos e implícitos están sugiriendo consistentemente que los regularizadores, cuando se ajustan correctamente, podrían ayudar a mejorar el desempeño de la generalización. Sin embargo, es poco probable que los regularizadores sean la razón fundamental para la generalización, ya que las redes continúan funcionando bien después de eliminar todos los regularizadores.

RECURSOS :UNDERSTANDING DEEP LEARNING REQUIRES RETHINKING GENERALIZATION

<https://arxiv.org/pdf/1611.03530.pdf>

https://en.wikipedia.org/wiki/Batch_normalization

https://en.wikipedia.org/wiki/Vapnik%E2%80%93Chervonenkis_dimension

https://en.wikipedia.org/wiki/Rademacher_complexity

https://en.wikipedia.org/wiki/Stochastic_gradient_descent

<https://towardsdatascience.com/this-thing-called-weight-decay-a7cd4bcfccab>