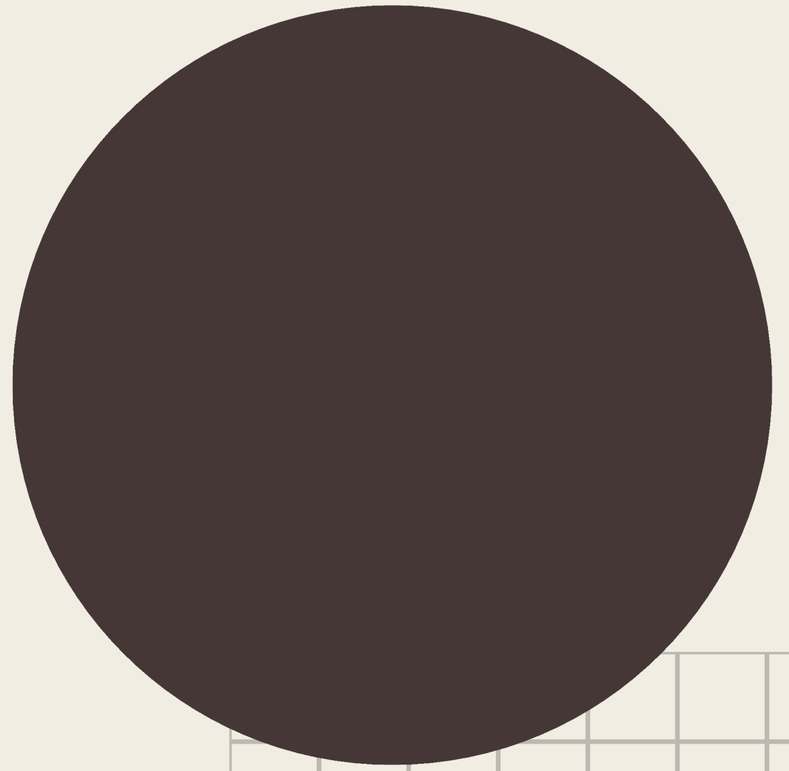
Metodos números 2

GRADIENTE ESTOCÁSTICO, ADAM
PROYECTO 2

Problematica!

Las redes neuronales profundas tienen millones de parámetros que deben ajustarse durante el entrenamiento para minimizar una función de costo (o pérdida). Esta función de costo cuantifica la diferencia entre las predicciones del modelo y los valores reales



Desafios:

- Paisaje de la función de costos
- Escalabilidad
- Velocidad de convergencia
- Gradientes desaparecientes y explosivos

Gradiente estocástico

El SGD es un método de optimización que actualiza los pesos del modelo utilizando una muestra aleatoria en cada iteración

$$\theta = \theta - \eta \nabla J(\theta; x^i, y^i)$$

donde η es la tasa de aprendizaje, θ son los parámetros del modelo y J es el gradiente evaluado en una muestra de datos.

Ventajas y desventajas

Ventajas

- Rápida actualización
- Escape de mínimos locales

Desventajas

- Convergencia ruidosa
- Dependencia de la taza de aprendizaje

Class Break!

¿Alguna duda?

Optimizador ADAM

1

Adam (Adaptive Moment Estimation) es un método de optimización que combina las ventajas de Momentum y RMSprop para ofrecer actualizaciones de gradientes estables y adaptativas.

2

Es ampliamente utilizado en el entrenamiento de redes neuronales profundas debido a su capacidad de adaptarse automáticamente a los parámetros individuales de los modelos.

3

Aborda varios desafíos comunes de la optimización en aprendizaje profundo, como la convergencia inestable de SGD y la elección fija de la tasa de aprendizaje.

Funcionamiento basico

- Calcula promedios móviles de primer orden
- Calcula promedios móviles de segundo orden
- corrección de sesgo para evitar valores iniciales subestimados.
- Actualizaciones de los parámetros

$$\theta = \theta - \frac{\eta}{\sqrt{\hat{v}_t} + \epsilon}\hat{m}_t$$

Formulas importantes

- Promedios moviles

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1)\nabla J(\theta)$$
$$v_t = \beta_2 v_{t-1} + (1 - \beta_2)(\nabla J(\theta))^2$$

- Corrección de sesgo

$$\hat{m}_t = \frac{m_t}{1 - \beta_1^t}, \quad \hat{v}_t = \frac{v_t}{1 - \beta_2^t}$$

Ventajas y desventajas

Ventajas

- Convergencia rapida
- Adapta tasas de aprendizaje individualmente

Desventajas

- Posible sobreajuste
- Mayor costo computacional

Class Break!

¿Alguna duda?

Presentracion del codigo

Análisis de resultados

Comparaciones de las convergencias de ambos modelos

El efecto que trae cada tasa de aprendizaje en el desarrollo de cada modelo

Importancia de las buenas practicas en la seleccion de los hiperparametros y la importancia de los mismos

¿Alguna pregunta?

Métodos numéricos 2

Gracias