Notas para a Apresentação

SLIDE 4

Não saturada

Na parte positiva, do gráfico **a função continua a crescer**, o que significa que a sua **derivada se mantém constante e não nula**.

Suave e Contínua

Esta função, tal como a Swish, é suave **(não tem pontos angulosos)** e contínua **(não tem pontos de descontinuidade)**.

Diferenciável em todo o seu Domínio

Fruto da propriedade anterior, a Mish é sempre **derivável em qualquer ponto** (como sabemos o algoritmo de descida do gradiente usa informação das derivadas de, entre outras, a função de ativação).

Ativação para inputs negativos

Ao contrário da ReLU, **a Mish mantém ativações pequenas para valores negativos pequenos** (esta informação pode ser útil para **manter a expressividade do modelo**), tentendo para 0 de seguida, de modo a **desencorajar valores negativos demasiado pequenos** (grandes em valor absoluto).

Regularização Intrínseca

Valores demasiado pequenos são então "esquecidos", o que se torna especialmente importante no **início do treino quando ativações demasiadamente negativas são mais comuns**.

SLIDE 5

Na figura podemos ver o *output* de uma rede com **2 entradas** (coordenadas (x,y)), **1 saída** (pintada a azul ou vermelho) e **5 camadas (com pesos inicializados aleatoriamente)**.

No caso da **ReLU**, vemos **variações bruscas de valor** (dado ao ponto anguloso que ela apresenta e falta de suavidade em geral). **A Mish e a Swish partilham a suavidade desejada**. Vemos transições bem uniformes entre azul e vermelho.

SLIDE 9

Nesta tabela podemos ver várias informações:

- Temos a coluna com os **resultados originais do *paper***;
- A ***accuracy* na época final e na melhor época**;
- Temos também os **tempos de treino para cada configuração**;

De forma geral, **conseguimos replicar os resultados originais**.

Contudo, pelos nossos testes **não houve uma clara vantagem de nenhuma função em relação às demais**, em termos de *accuracy*.

SLIDE 11

Por fim, temos a implementação híbrida que tentámos fazer. **Partindo do código do SWA juntámos a função Mish**.

A 2 primeiras linhas têm as **taxas de aprendizagem por defeito (para o *paper* do SWA)**. Vemos que a **ReLU surge na frente, com uma margem bem diminuta**.

As 2 linhas seguintes mostram 2 **taxas de aprendizagem mais baixas**. Isto porque no *paper* da Mish é feita a alusão à **ideia de que esta função beneficia de taxas mais baixas**.

Depois fizemos mais 1 teste com **taxas de aprendizagem intermédias**. Neste caso, a **Mish ganhou vantagem para a ReLU, novamente por uma margem pequena**.