Comment est entraîné un modèle d'IA ?

Les duels de l'IA
par compar:IA

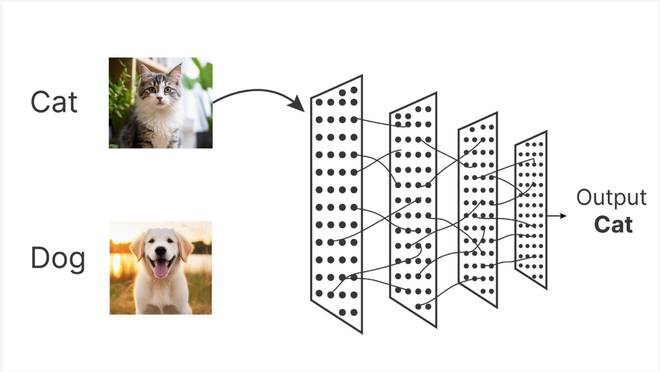
Des **données** pour l'entraînement

+

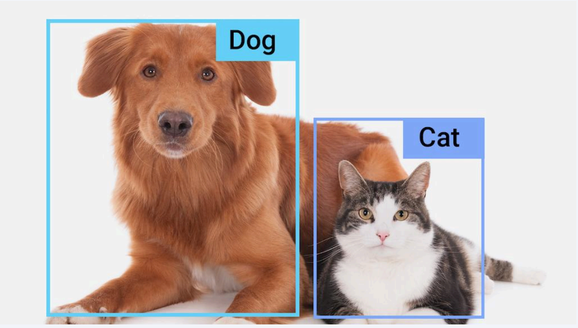
Beaucoup de **calculs** pour "apprendre"

=

Un modèle d'IA



Prédit sur de nouvelles données



Mai 2025, licence etalab-2.0

Comment est entraîné un modèle d'IA conversationnelle ?

Les duels de l'IA
par compar:IA

Des **données** pour l'entraînement

+

Beaucoup de **calculs** pour "apprendre"

=

Un modèle d'IA

- Quasiment tout Internet public
- Peu de filtrage des contenus sous droits d'auteur

Un jeu de données de taille moyenne tel que FineWeb pèse 51TB de mémoire.

$$\begin{aligned}
\mathbf{Layer}(x) &= \mathrm{LayerNorm}\left[x + \mathrm{FFN}\left(\mathrm{Concat}\left(\mathrm{Attention}(Q,K,V)\right)W^O\right)\right], \\
\text{where} \quad Q &= (x+PE)W_Q, \\
K &= (x+PE)W_K, \\
V &= (x+PE)W_V, \\
PE &= \mathrm{PositionalEncoding}(\mathrm{positions}), \\
\mathrm{Attention}(Q,K,V) &= \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V, \\
\mathrm{FFN}(z) &= \mathrm{LayerNorm}\left[z + \mathrm{ReLU}(zW_1+b_1)W_2 + b_2\right],
\end{aligned}$$

Un ensemble de chiffres qui permettent de prédire la réponse la plus probable à la requête.

Mai 2025, licence etalab-2.0