# Transformer (Architektura AI)

**Transformer** je architektura hlubokého učení (Deep Learning), kterou v roce 2017 představili vědci z Google Brain v přelomovém článku *„Attention Is All You Need“*. Na rozdíl od předchozích modelů (jako RNN nebo LSTM) nezpracovává data popořadě, ale všechna najednou, což umožňuje masivní paralelizaci a lepší pochopení souvislostí v textu.

Dnes tvoří základ téměř všech velkých jazykových modelů a nachází uplatnění i v počítačovém vidění (Vision Transformers).

## Klíčový koncept: Self-Attention (Sebepozornost)

Hlavní inovací Transformeru je mechanismus **Self-Attention**. Ten umožňuje modelu při zpracování určitého slova (nebo tokenu) „dívat se“ na všechna ostatní slova ve větě a určit, která z nich jsou pro pochopení významu nejdůležitější.

- **Příklad:** Ve větě *„Zvíře nepřešlo silnici, protože bylo příliš **unavené**“*, mechanismus pozornosti propojí slovo „unavené“ se slovem „zvíře“.
- Pokud větu změníme na *„...protože byla příliš **široká**“*, model automaticky zaměří pozornost na slovo „silnice“.

## Architektura: Encoder a Decoder

Původní Transformer se skládá ze dvou hlavních částí:

### 1. Encoder (Kodér)

Analyzuje vstupní sekvenci a vytváří její bohatou číselnou reprezentaci (vektory). Modely založené pouze na encoderu (např. **BERT**) jsou vynikající pro pochopení textu, klasifikaci nebo analýzu sentimentu.

### 2. Decoder (Dekodér)

Bere reprezentaci z encoderu a generuje výstupní sekvenci (slovo po slově). Modely založené pouze na decoderu (např. **GPT**) jsou optimalizovány pro generování textu.

## Proč Transformer změnil svět?

- **Paralelizace:** Starší modely musely číst text slovo po slově (zleva doprava). Transformery vidí celou větu (nebo odstavec) najednou, což umožnilo trénovat modely na obrovských grafických kartách (GPU).
- **Dlouhá paměť:** Mechanismy pozornosti netrpí „ztrátou paměti“ u dlouhých textů, což byl

hlavní problém starších architektur.
- **Přenositelnost (Transfer Learning):** Model se může naučit základy jazyka na obrovském množství dat a poté být snadno „doladěn“ (fine-tuned) pro konkrétní úkol (např. lékařskou diagnostiku).

## Klíčové vrstvy Transformeru

| Vrstva | Funkce |
|---|---|
| **Positional Encoding** | Protože model vidí všechna slova najednou, tato vrstva mu dodává informaci o tom, v jakém pořadí slova ve větě jsou. |
| **Multi-Head Attention** | Umožňuje modelu sledovat několik různých typů vztahů mezi slovy současně. |
| **Feed-Forward Network** | Standardní neuronová síť, která dále zpracovává informace získané z pozornosti. |
| **Layer Normalization** | Zajišťuje stabilitu tréninku a zrychluje konvergenci modelu. |

> **Zajímavost:** Původní motivací pro vytvoření Transformeru byl strojový překlad. Ukázalo se však, že stejný princip funguje skvěle i pro generování zdrojového kódu, skládání hudby nebo predikci struktury bílkovin (AlphaFold).

Zpět na AI rozcestník