Transformers

Recap: Convolution

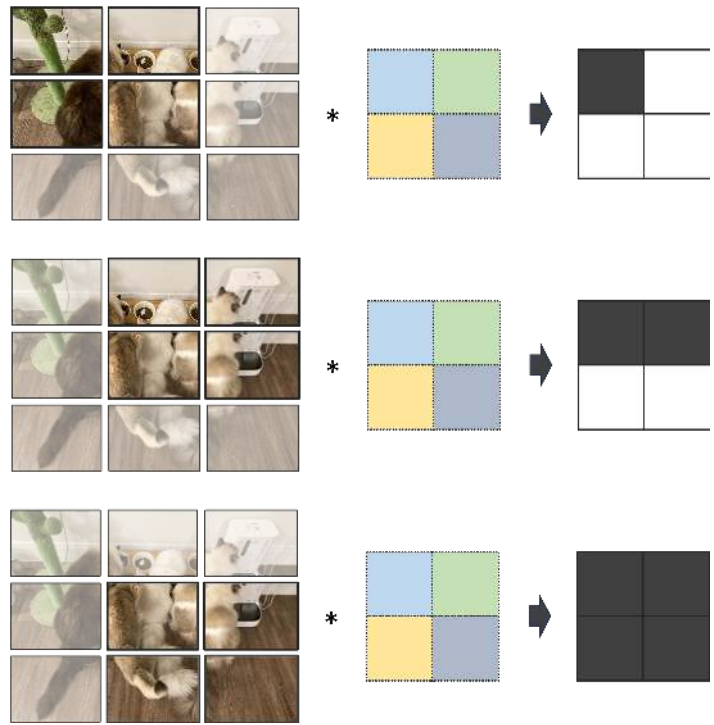
Input: $x \in \mathbb{R}^{C_1 \times H \times W}$

Output: $y \in \mathbb{R}^{C_2 \times (H-h+1) \times (W-w+1)}$

Parameters:

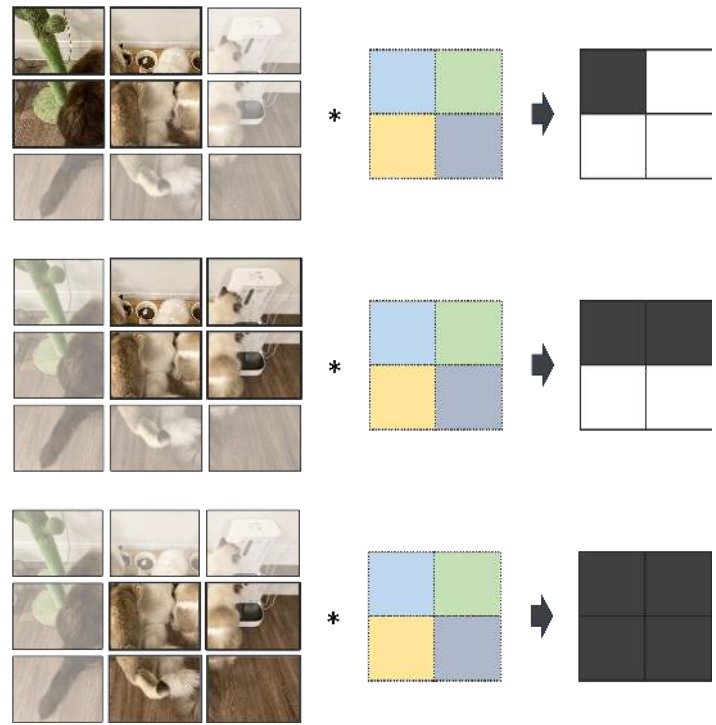
- Kernel: $\omega \in \mathbb{R}^{C_1 \times C_2 \times h \times w}$
- Bias: (optional) $b \in \mathbb{R}^{C_2}$

$$\underbrace{y_{i,j,k}}_{\text{output}} = \underbrace{b_i}_{\text{bias}} + \sum_{l=1}^{C_1} \sum_{m=0}^{h-1} \sum_{n=0}^{w-1} \underbrace{x_{l,j+m,k+n}}_{\text{input}} \cdot \underbrace{\omega_{i,l,m,n}}_{\text{kernel}}$$



Recap: Convolution

Good at parsing relatively **structured** patterns



A New Problem

Sentiment Analysis for Movie Reviews

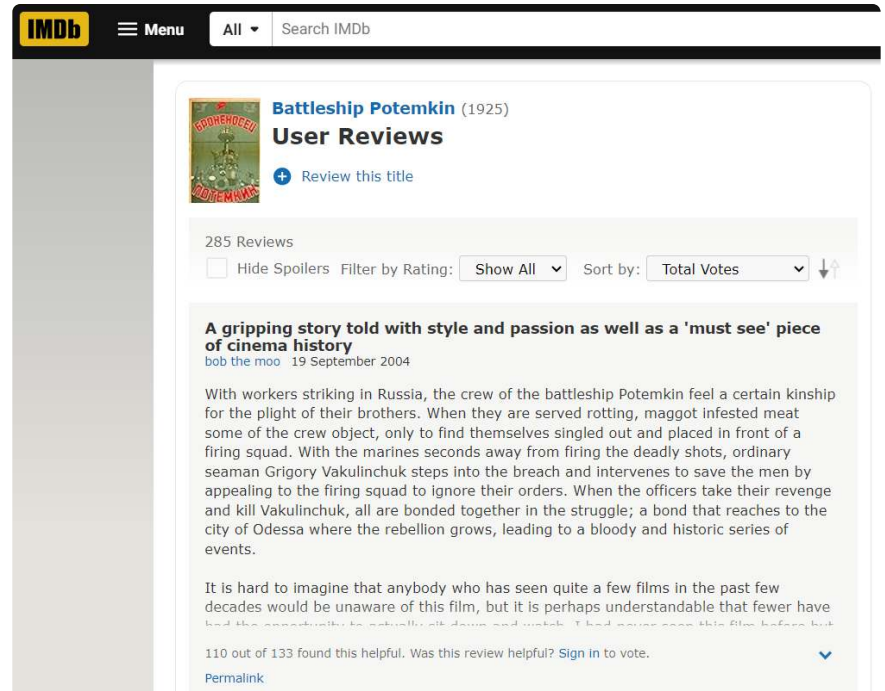
- Predict if review is positive 🙂 or negative 🙁

🙂 `My kid likes this movie`

🙂 `My little kid likes this animated movie`

🙁 `My kid does not like this movie`

Real-world examples: reviews on IMDB[1]:



1. IMDB Reviews ↩

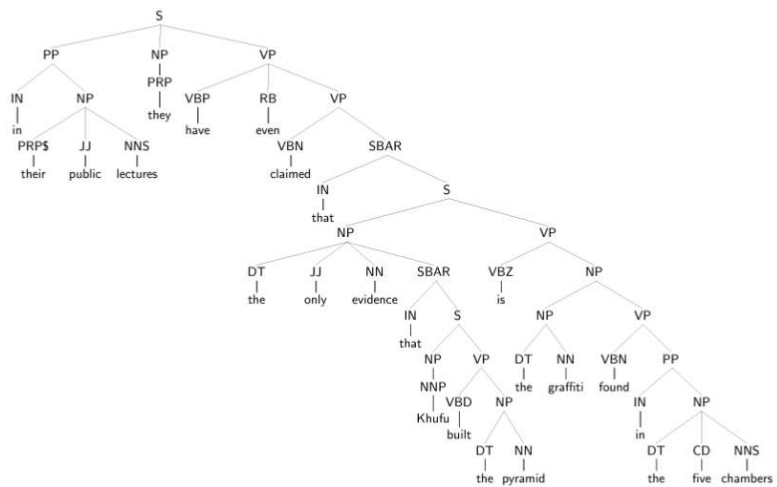
How Is the Task Different From Previous Examples

Images

- Fixed input (resolution)
- Fixed structure

Language

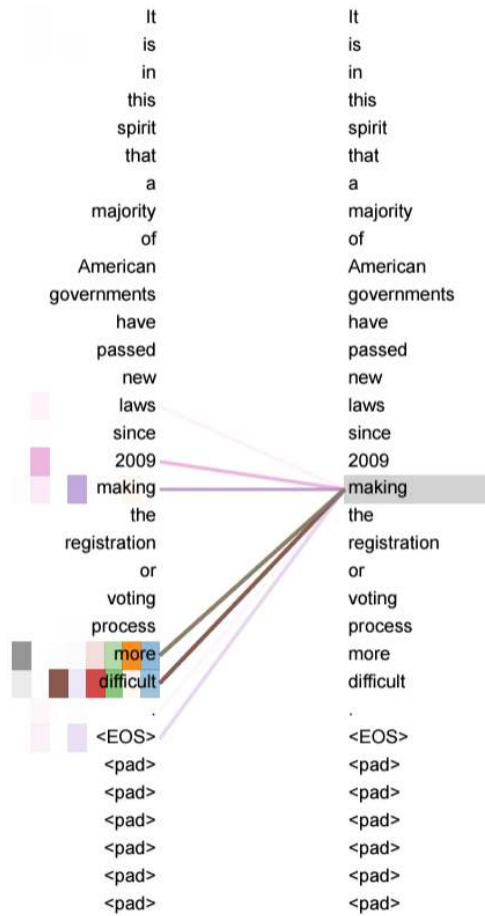
- Variable length
- Diverse structure (tree syntax[1])



1. Tree Syntax of Natural Language ↩

Solution: Transformer

A deep network based on the *attention* mechanism[1]:



1. Visualing attention. ↩