Language Modelling with Pixels

Explorations of pixel-based encoding of language

Jonas F. Lotz
Department of Computer Science, University of Copenhagen
Large Language Models?

\[ p_\theta(x_i|x_{<i}) = \frac{e^{z_i}}{\sum_j e^{z_j}} \]

j-dimensional vector \( z \) All words in vocabulary

Deep Neural Network

Billions of learned features

The woman drew many famous
Large Language Models

- Spelling and grammar checking
- Machine translation
- Web search
- Text prediction
- Topic modelling
- Chatbots
In the Era of Scale


Emergent Abilities of Large Language Models. (Wei+ TMLR)
What’s left?

天地玄黄 《I h โทร》
ABCD अइ०
NLP for All Written Languages

- There are 7,000 spoken languages, of which 3,000 are written
  - There is at least 400 languages with >1M speakers
- But NLP only covers 100 languages (van Esch+ LREC22)
  - Lack of technological inclusion for billions of people

Slide credit: Sebastian Ruder
Today: Pixel-based Language Modelling

- Key insight: treat language processing as visual processing

Søren Kierkegaard (d. 1855) was a Golden Age philosopher.
A new type of generative model

Penguins are designed to be streamlined and hydrodynamic, so having long legs would add extra drag. Having short legs with webbed feet to act like runners, helps to give them the penguin-like figure. We compare bird anatomy with humans, we would see something very peculiar. By taking a look at the side-by-side image in Figure 1, you can see how their leg bones line up to ours. What most people mistake for knees are actually the anatomies of birds. This gives the illusion that bird knees bend opposite of ours. The knees are actually tucked up inside the penguin's body. So how does this look inside the penguin? In the images below, you can see boxes surrounding the penguins' knees.

100K steps
500K steps
1M steps

Demo: https://huggingface.co/Team-PIXEL/pixel-base
Results: Dependency Parsing

PIXEL vastly outperforms BERT on unseen scripts
Exploring Text Rendering Strategies

Bigram text rendering produces much better models
Pretraining

- **English Dataset**: English Wikipedia and Books Corpus
- **Masking**: 25% Span Masking
- **Maximum sequence length**: 529 patches (16 x 8464 pixels)
- **Compute**: 8 x 40GB A100 GPUs for 8 days
- **Parameters**: 86M encoder + 26M decoder

There is only 0.05% non-English text in our pretraining data (estimated by Blevins and Zettlemoyer 2022)

The Great Wall of China *(traditional Chinese: 萬里長城; simplified Chinese: 万里长城; pinyin: Wàn lǐ Chángchéng)*

Pretrained model: [https://huggingface.co/Team-Pixel/pixel-base](https://huggingface.co/Team-Pixel/pixel-base)
Ongoing Work

• Improve sentence level reasoning tasks
  • Contrastive objective or multi-scale modelling

• Scale to multilingual pretraining
  • Better representations from reconstructing multiple scripts?

• Understanding cross-script transfer
  • What causes the PIXEL model transfer well to other scripts?