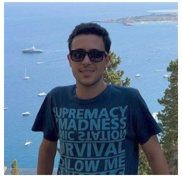
UNIVERSITY OF COPENHAGEN

Language Modelling with Pixels

Phillip Rust[1] Jonas F. Lotz[1] Emanuele Bugliarello[1] Elizabeth Salesky[3] Miryam de Lhoneux[1,2,4] Desmond Elliott[1]






[1]University of Copenhagen [2]KU Leuven [3]Johns Hopkins University [4]Uppsala University

Summary

We train a **pixel-based encoder of language (PIXEL)**,
a **language model** trained solely on **images** of **rendered text.**

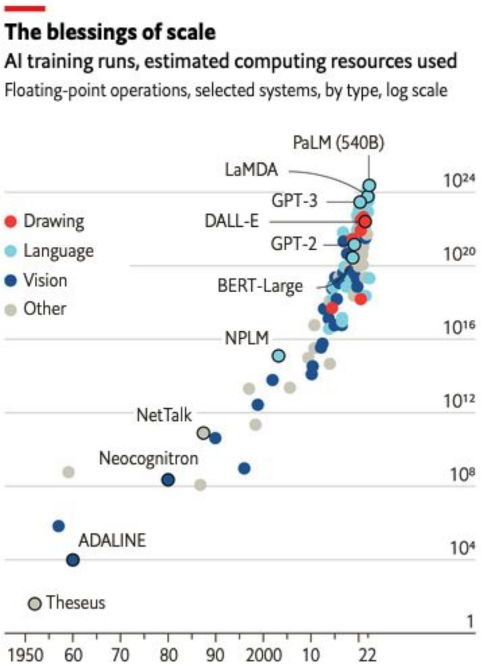
Some of PIXEL's strengths are

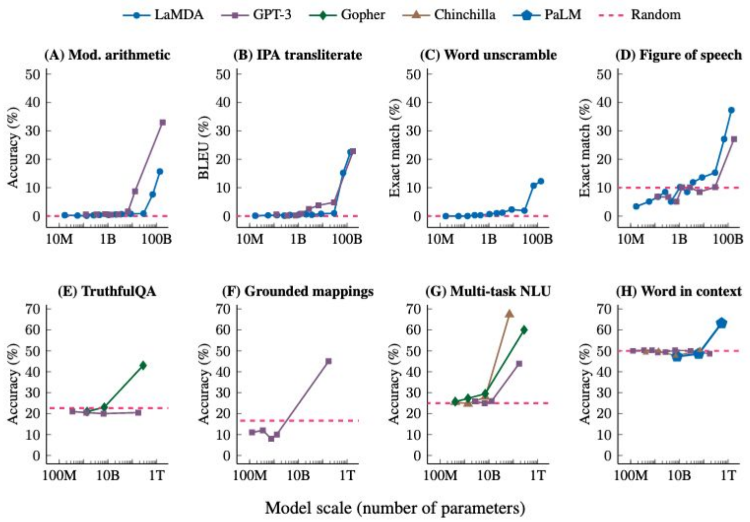
Out-of-the-box **transfer** to **unseen languages** and **scripts**

Robustness to orthographic attacks & code-switching*

* See our paper for the code-switching results

NLP in the Era of Scale





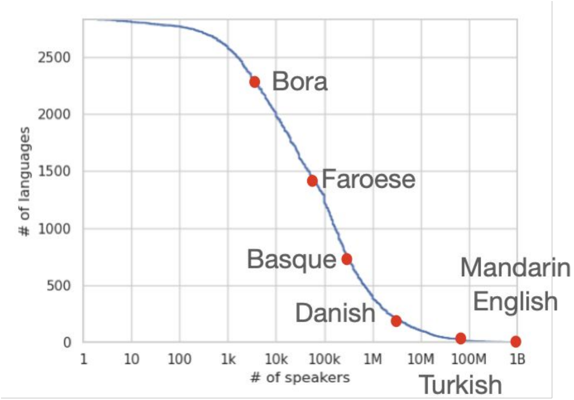
Emergent Abilities of Large Language Models
(Wei+ TMLR'22)

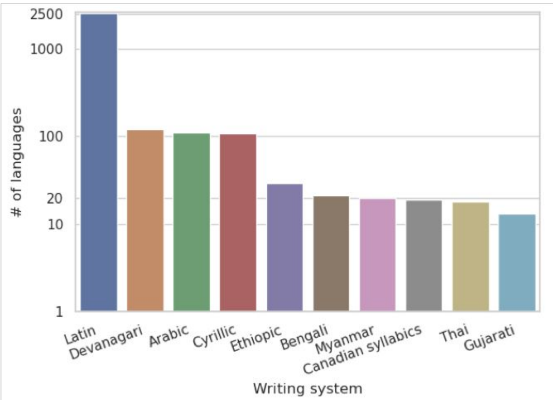
Source: www.economist.com/interactive/briefing/2022/06/11/huge-foundation-models-are-turbo-charging-ai-progress

NLP for all written languages?

There are **~7000 spoken languages**, of which **~3000** are **written** and at least **400** have **>1M speakers**

Most NLP only covers **100 languages** *(van Esch+ LREC'22)*
→ **Lack** of **technological inclusion** for **billions of people**



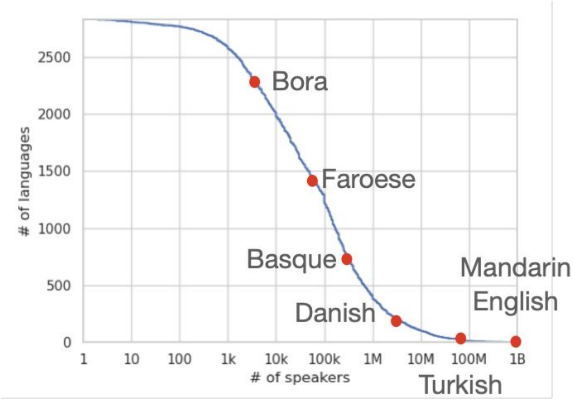


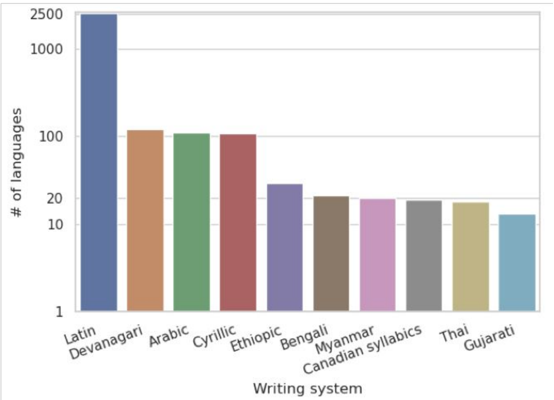
Slide credit: Sebastian Ruder

What’s left? NLP for all written languages

There are **~7000 spoken languages**, of which **~3000** are **written** and at least **400** have **>1M speakers**

Most NLP only covers **100 languages** *(van Esch+ LREC’22)*
→ **Lack** of **technological inclusion** for **billions of people**





Slide credit: Sebastian Ruder

Question: What's stopping us?

NLP is an **open vocabulary problem**.

A language model's ability to **process unseen words** is **determined by its vocabulary**:

1. "Trained" over a corpus: Byte-Pair Encoding *(Sennrich+ ACL'16)*
→ Unseen tokens not in the vocabulary (unless w/ byte-level fallback)

2. Corpus-independent: characters *(Clark+ TACL'22)* / bytes *(Xue+ ACL'22)*
→ Need to deal with longer sequence lengths

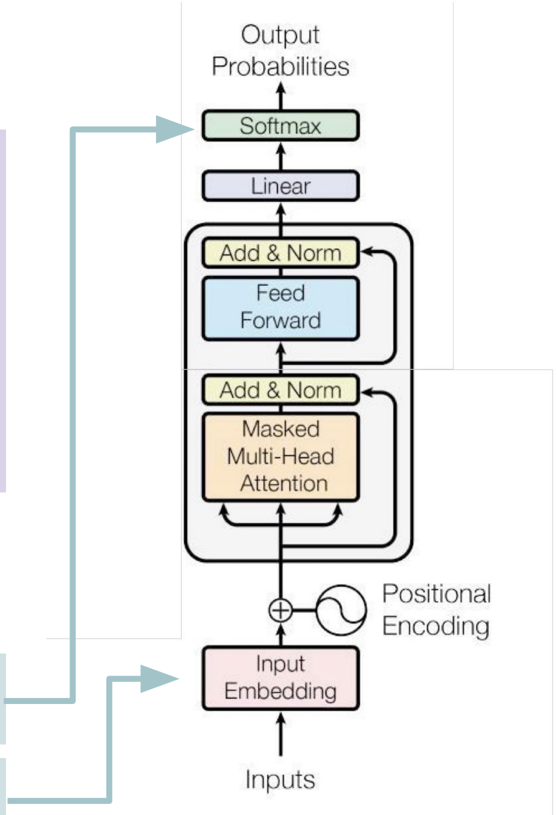
Answer: The *Vocabulary Bottleneck*

Language models have **discrete** input and output **vocabularies** expressed over a **finite inventory** of tokens, characters, words, sub-words, etc.

→ **This creates a bottleneck in two places**

Computational bottleneck in the output layer

Representational bottleneck in the embedding layer



Vaswani+ NeurIPS'17

TL;DR of our paper

We attempt to crack the *vocabulary bottleneck* with pixels.

But what does that mean?

The NLP pipeline simplified

The NLP pipeline simplified

My cat, Dr. Beans II., sleeps 22h a day.

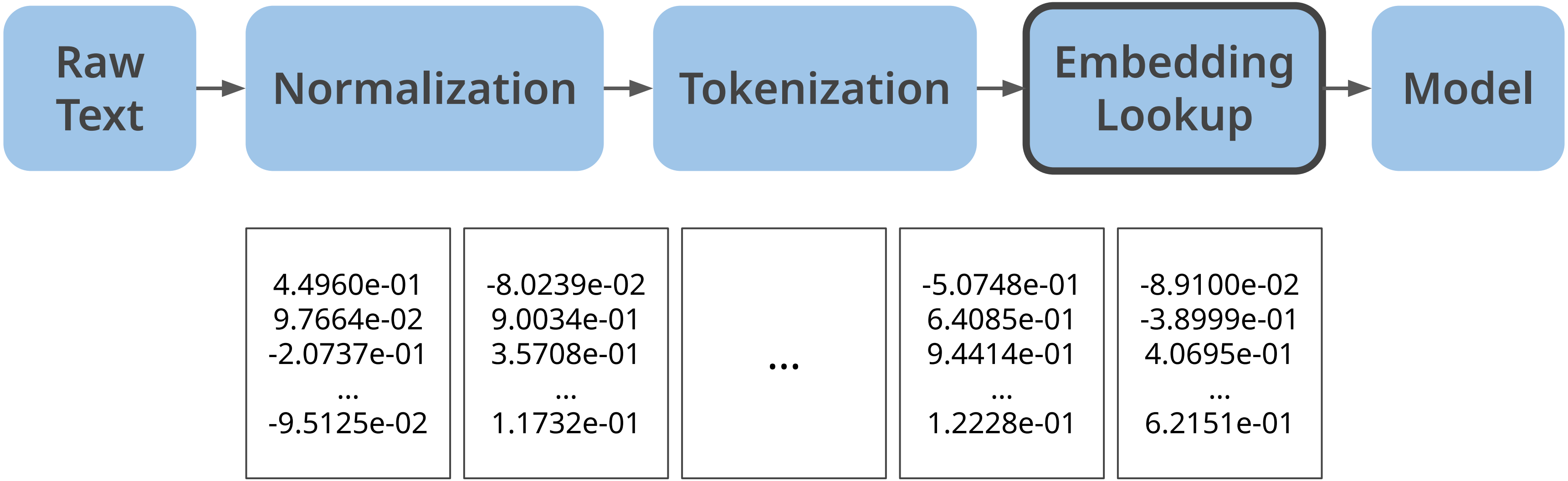
The NLP pipeline simplified

My cat , Dr . Beans II . , sleeps 22h a day .

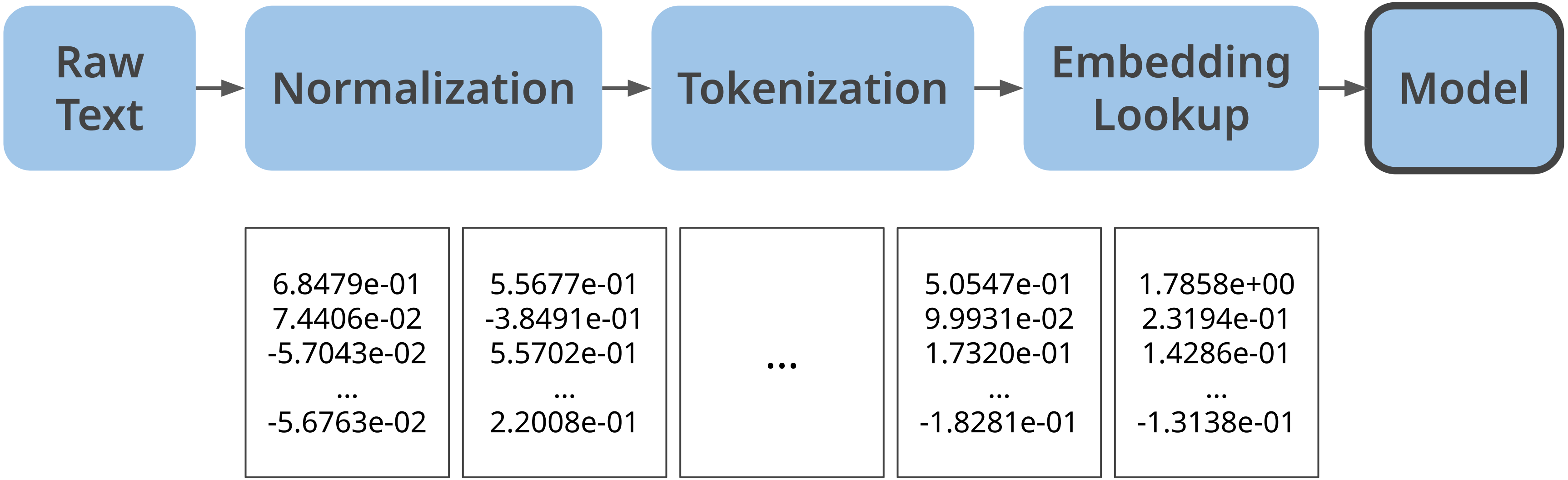
The NLP pipeline simplified

[CLS]^{101} *My*^{1422} *cat*^{5855} *,*^{117} *Dr*^{1987} *.*^{119} *Bean*^{21561} *##s*^{1116} *II*^{1563} *.*^{119} *,*^{117}
sleep^{2946} *##s*^{1116} *22*^{1659} *##h*^{1324} *a*^{170} *day*^{1285} *.*^{119} *[SEP]*^{102}

The NLP pipeline simplified



The NLP pipeline simplified



Cracking the *vocabulary bottleneck* with pixels

Treat **language processing** as **visual processing**

Cracking the *vocabulary bottleneck* with pixels

Treat **language processing** as **visual processing**

Cracking the *vocabulary bottleneck* with pixels

Treat **language processing** as **visual processing**

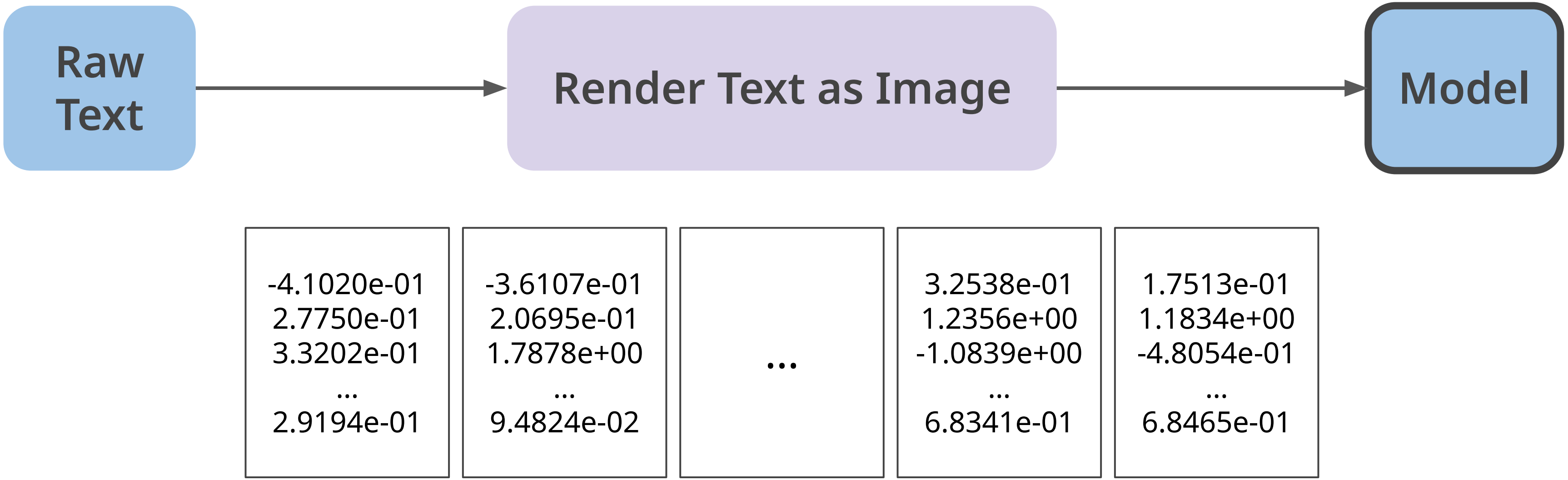
My cat, Dr. Beans II., sleeps 22h a day.

Cracking the *vocabulary bottleneck* with pixels

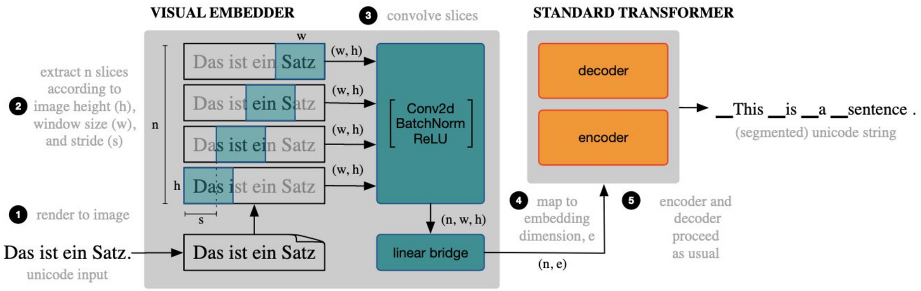
Treat **language processing** as **visual processing**

Cracking the *vocabulary bottleneck* with pixels

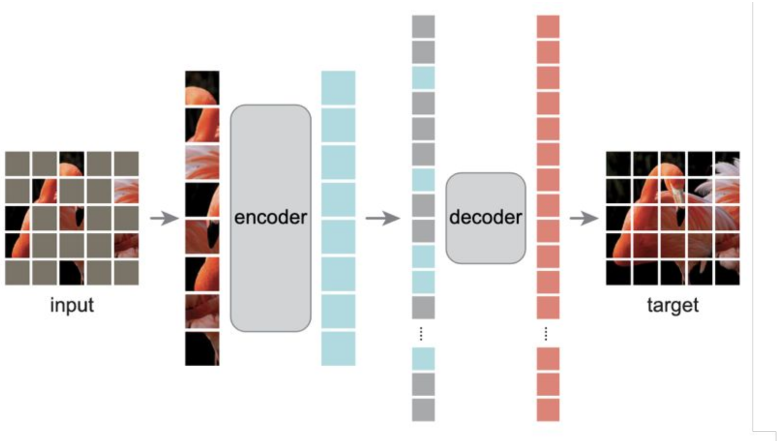
Treat **language processing** as **visual processing**



Inspiration



Robust Open-Vocabulary Translation from Visual Text Representations
(Salesky+ EMNLP'21)



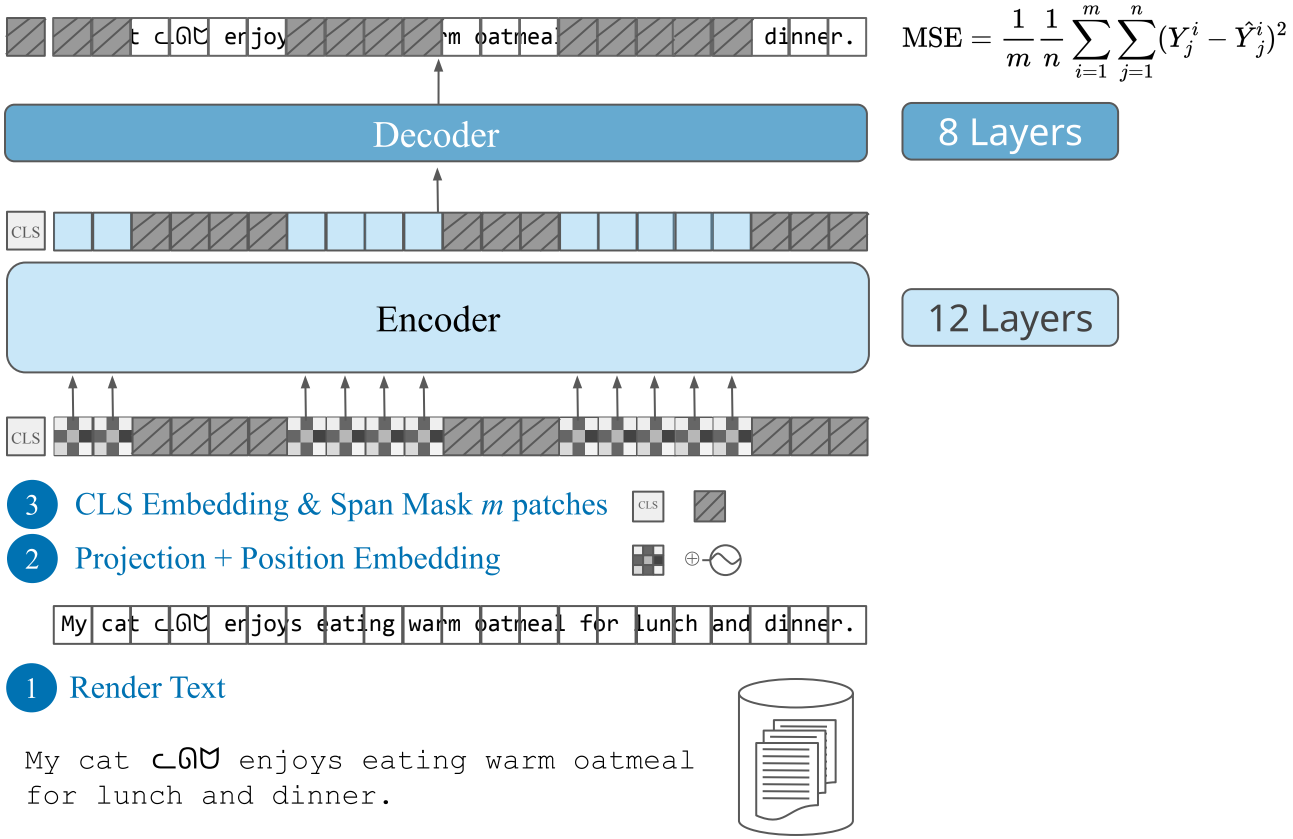
Masked Autoencoders are Scalable Visual Learners
(He+ CVPR'22)

Pixel-based Encoder of Language (PIXEL)

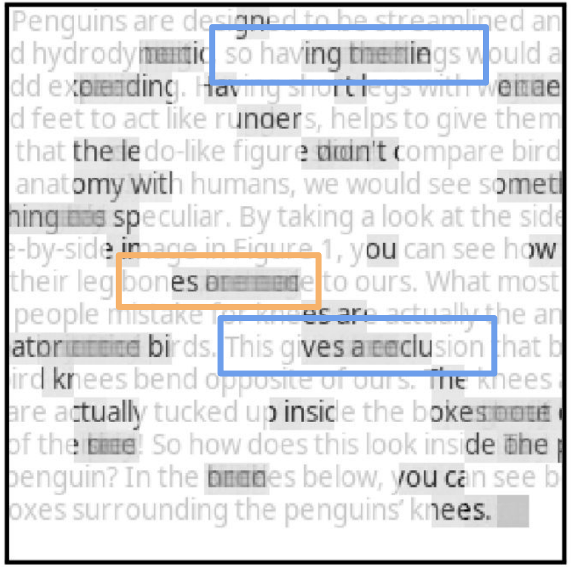
16x16 patch resolution

Google Noto Fonts

PyGame / PangoCairo



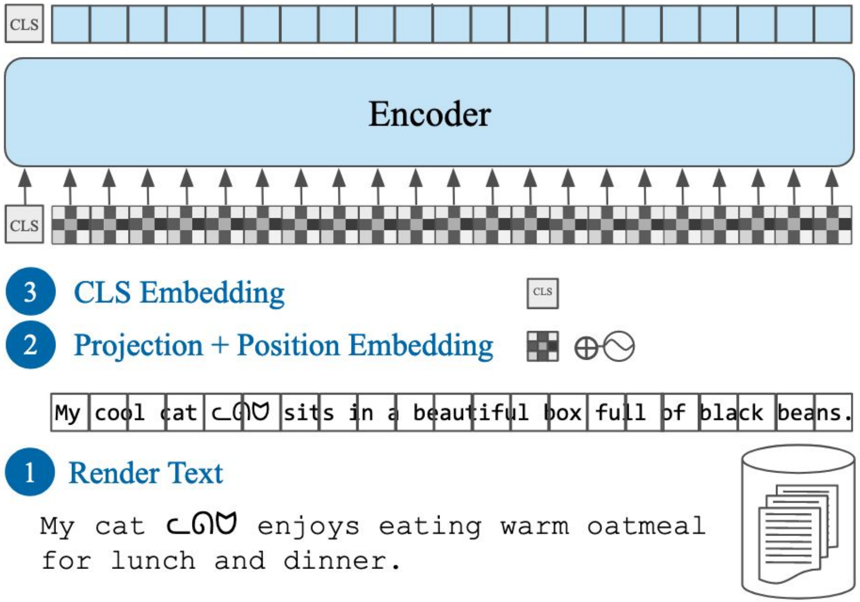
PIXEL learns to reconstruct text



100k steps

🤗Demo: https://huggingface.co/Team-PIXEL/pixel-base

Downstream Task Fine-Tuning



Flexible Text Renderer

Color Emoji

My cat 🐈 loves pancakes 🥞 and grapes 🍇.

Left-to-right, right-to-left, and logosyllabic writing systems

一隻貓正在吃碗中的猫粮 قط جاثم على غصن شجرة

Word-level rendering

አቢሲኒያውያን በጣም ጠንካራ ድመቶች ናቸው ።

The Benefits of Pixels

PIXEL can **process anything that can be rendered**
→ **Open vocabulary** which is easily extensible to **unseen text**
→ Support all written languages

Complete parameter sharing from the input representation
(unlike separate-but-related subwords in an embedding matrix)

Nothing language-specific in the **input / output**
→ **Greater flexibility** to process written language in **different forms**
(PDFs, scanned newspapers, etc.)

Experiments

Pretraining

🤗https://huggingface.co/Team-PIXEL/pixel-base

Dataset	
Masking	
Max. Seq. Length	
Compute	
Parameters	

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: *Wànlǐ Chángchéng*)

Finetuning Experiments

🤗https://huggingface.co/Team-PIXEL

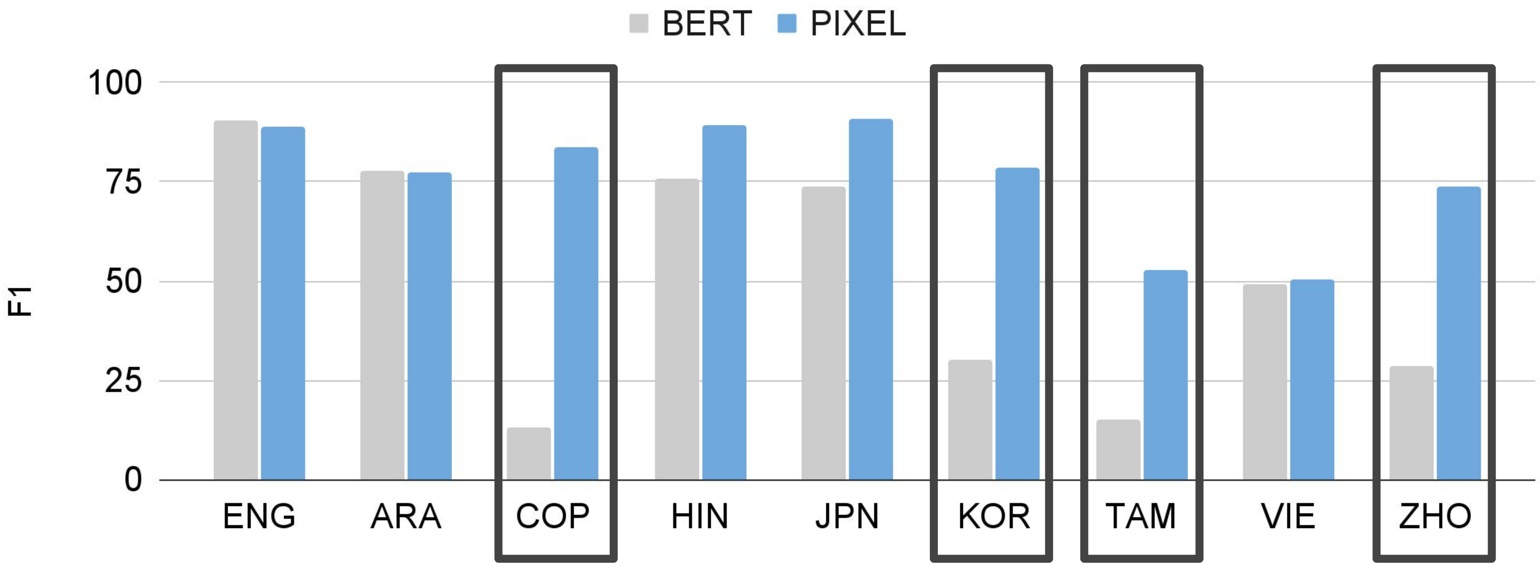
Datasets	.

Finetuning Experiments

🤗https://huggingface.co/Team-PIXEL

Datasets	.

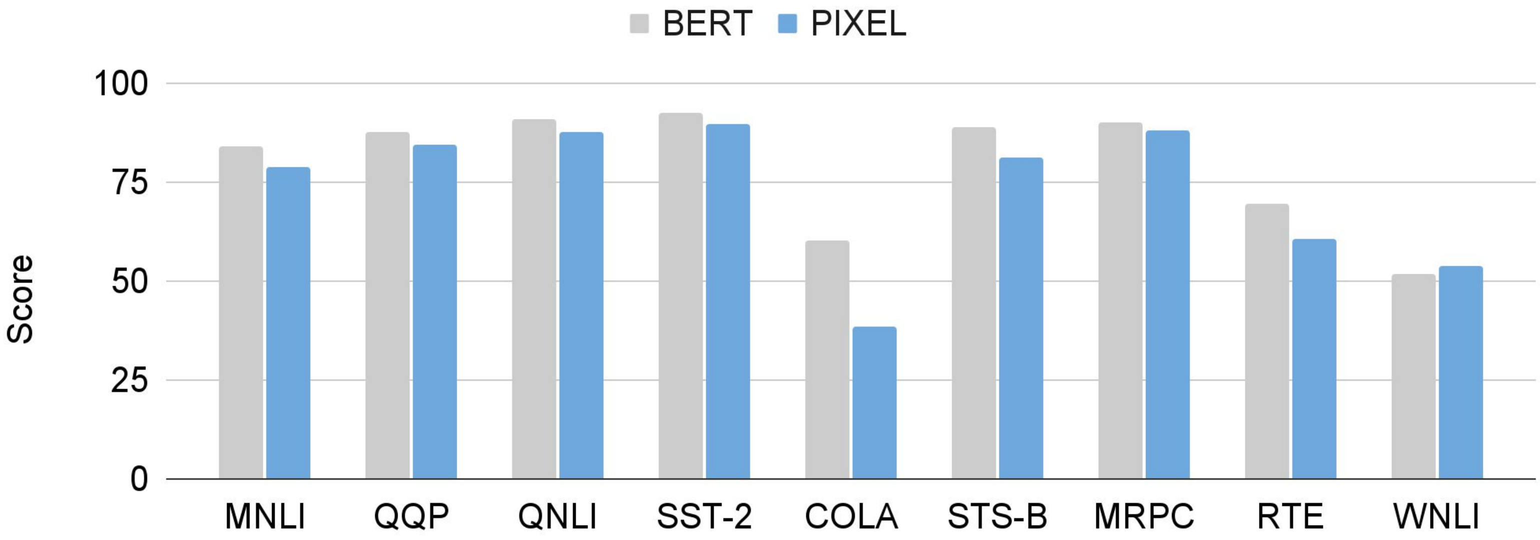
Dependency Parsing Results



BERT UNK	0%	1%	94%	33%	46%	85%	82%	5%	73%

PIXEL (vastly) outperforms BERT on unseen scripts

GLUE Results



BERT outperforms PIXEL on English sentence-level tasks

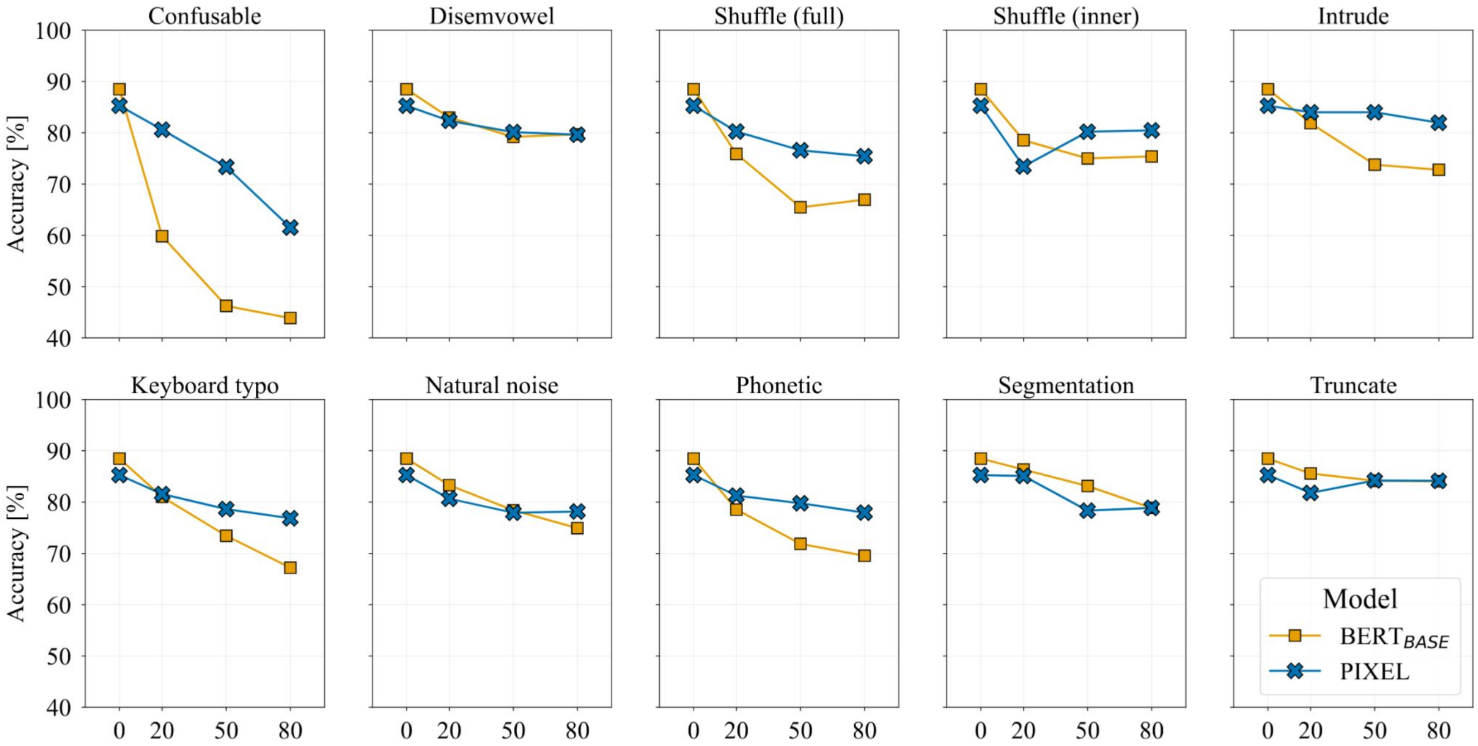
Robustness against orthographic attacks (Zeroé)

Attack	Sentence
NONE	Penguins are designed to be streamlined

Robustness against orthographic attacks (Zeroé)

Attack	Sentence
None	Penguins are designed to be streamlined

PIXEL is more robust than BERT



Conclusions

PIXEL is a **new type of language model** that renders **text as images** instead of splitting text into a finite set of tokens.

Rendered text makes it possible to achieve **high-quality transfer** to **unseen scripts** in syntactic and semantic tasks.

Pixel-based learning could be a **promising research direction** to **make NLP technology accessible** to more people.

PIXEL Resources

https://github.com/xplip/pixel

https://huggingface.co/Team-PIXEL