EMNLP 2024

Autoregressive Pre-Training on Pixels and Texts

EMNLP 2024

Yekun Chai, Qingyi Liu, Jingwu Xiao, Shuohuan Wang, Yu Sun, Hua Wu

Baidu Inc.

Bai du 百度

Code: https://github.com/ernie-research/pixelgpt

Model: https://huggingface.co/baidu/PixelGPT

Background

Pixel-based training show its potential to leverage the image modality of texts; previous works are mainly:

(1) encoder-based, such as PIXEL [1];

(2) encoder-decoder based, [2].

Challenges：

(1) The feasibility of tokenization-free autoregressive pre-training ;

(2) The synergistic benefits of multimodal pre-training between the duality of pixels and texts.

Autoregressive Pre-Training on Pixels and Texts

The integration of visual and textual information represents a promising direction in the advancement of language models. In this paper, we explore the dual modality of language—both visual and textual—within an autoregressive framework, pre-trained on both document images and texts. Our method employs a multimodal training strategy, utilizing visual data through next patch prediction with a regression head and/or textual data through next token prediction with a classification head. …

Text

Autoregressive Pre-Training on Pixels and Texts

Yekun Chai♠ Qingyi Liu*♡ Jingwu Xiao*◇
Shuohuan Wang♠ Yu Sun♠ Hua Wu♠
♠Baidu Inc. ♡Sun Yat-sen University ◇Peking University
{chaiyekun,wangshuohuan}@baidu.com
{liuqy95}@mail2.sysu.edu.cn

Abstract

The integration of visual and textual information represents a promising direction in the advancement of language models. In this paper, we explore the dual modality of language—both visual and textual—within an autoregressive framework, pre-trained on both document images and texts. Our method employs a multimodal training strategy, utilizing visual data through next patch prediction with a regression head and/or textual data through next token prediction with a classification head. We focus on understanding the interaction between these two modalities and their combined impact on model performance. Our extensive evaluation across a wide range of benchmarks shows that incorporating both visual and textual data significantly improves the performance of pixel-based language models. Remarkably, we find that a unidirectional pixel-based model trained solely on visual data can achieve comparable results to state-of-the-art bidirectional models on several language understanding tasks. This work uncovers the untapped potential of integrating visual and textual modalities for more effective language modeling. We release our code, data, and model checkpoints at https://github.com/ernie-research/pixelgpt.

1 Introduction

Recent advancements in large language models (LLMs) have pushed the boundaries of their capabilities in diverse applications, including language assistant (Touvron et al., 2023a), code generation (Lozhkov et al., 2024; Chai et al., 2023), and multimodal comprehension (OpenAI, 2023; Anil et al., 2023). LLMs typically tokenize input text into sequences of discrete subword units, allowing for a wide array of applications. However, tokenization-based approaches struggle with visually complex textual content, such as PDFs, where converting visual data into plain text often results in significant information loss. Traditional solutions rely on optical character recognition (OCR) models for extracting text from images, but these methods are inherently limited by the accuracy of text extraction and the fidelity of the original document structure.

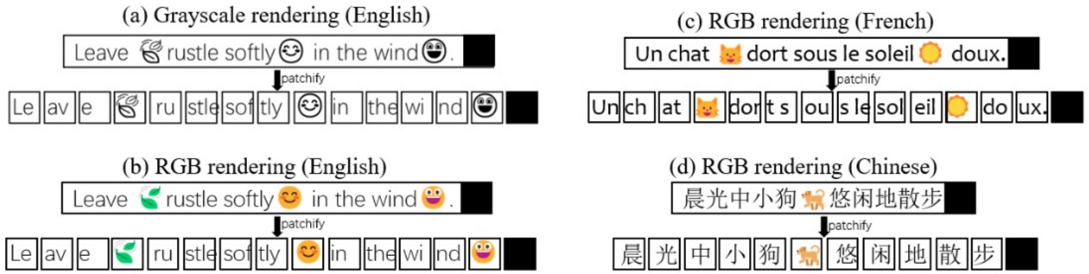
To address these challenges, recent work has introduced a new paradigm: pixel-based language modeling. This approach learns directly from the visual representation of text (as images) rather than relying solely on tokenized text. Models such as PIXEL (Rust et al., 2023) exemplify this shift, offering solutions that circumvent the limitations of traditional tokenization by treating text as image data. Pixel-based modeling also addresses the *vocabulary bottleneck*—a trade-off between input encoding granularity and the computational costs associated with vocabulary estimation in conventional language models (Rust et al., 2023).

In the previous literature, the development of pixel-based language models has been bifurcated into encoder-based (Rust et al., 2023; Tschannen et al., 2023) or encoder-decoder architectures (Salesky et al., 2023), encompassing models that either employ bidirectional mechanisms akin to MAE (He et al., 2022) or utilize an encoder-decoder framework, where a pixel-based model serves as the encoder, paired with a unidirectional language decoder. Despite these advancements, the exploration of pixel-based models employing a decoder-centric approach remains in its infancy.

Moreover, current research often processes visual text as 8-bit grayscale (Rust et al., 2023) or 2-bit binary images (Tai et al., 2024). This approach constrain the richness of the visual input, especially when processing content with color information, such as emojis or highlighted text. This limitation suggests that processing real-valued RGB images could offer a more detailed representation of visual text. However, the potential of pre-training

*Work done during QL and JX's internship at Baidu.

Visual Document



Text rendering.

[1] Language modelling with pixels. ICLR 2023.
[2] Multilingual pixel representations for translation and effective cross-lingual transfer. EMNLP 2023.

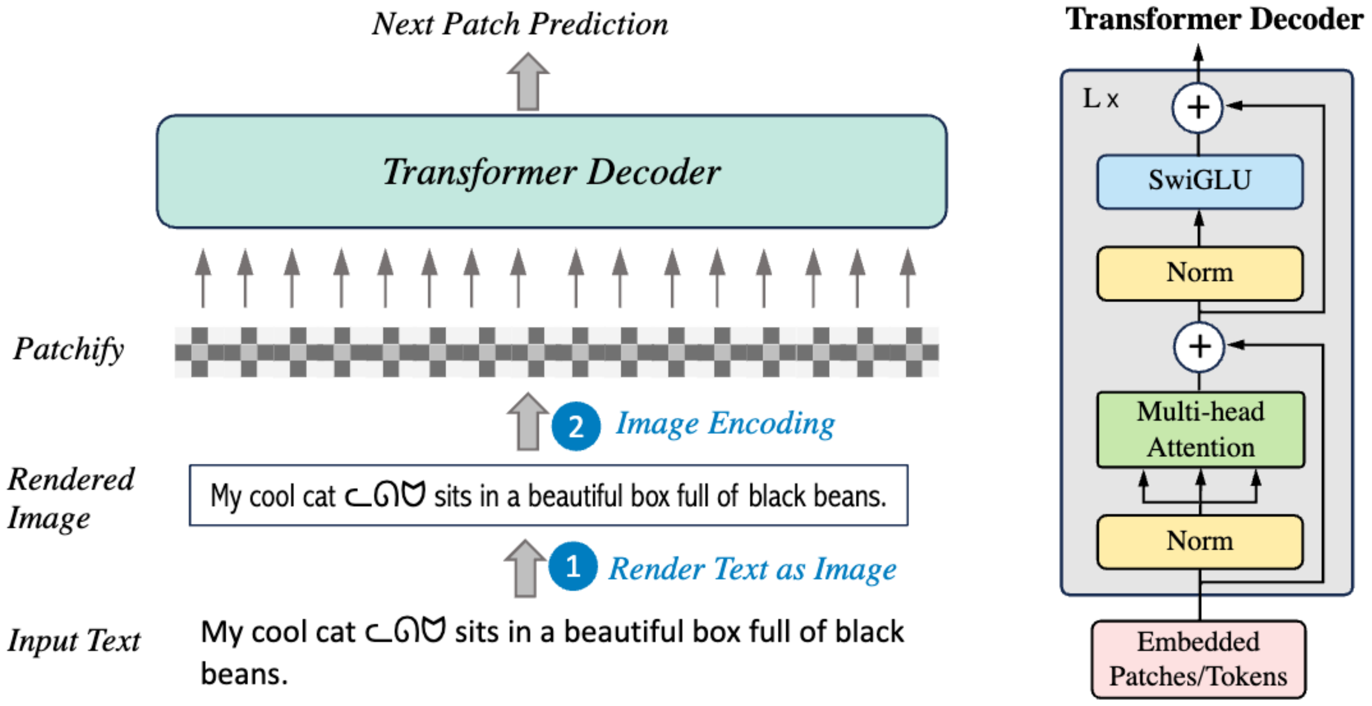
Autoregressive Pre-Training on Pixels and Texts

Pixel Input Preprocessing

① **Text rendering.** Utilize text renderer by converting texts into a visually-rich RGB images.

② **Image encoding**. Split rendered images into patches as in vision transformers.

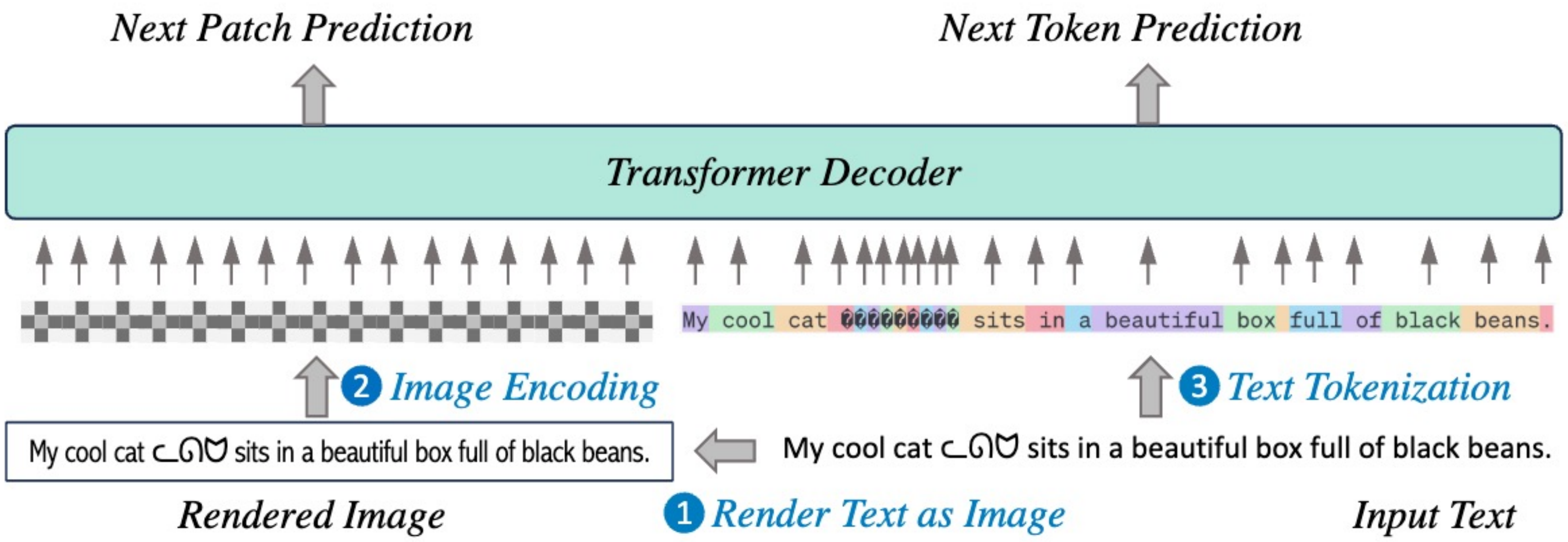
③ **Autoregressive Training.** Predict next patch based on its historical patches.



(a) Visual text image pre-training (*PixelGPT*).

(b) Model architecture.

Autoregressive Pre-Training on Pixels and Texts

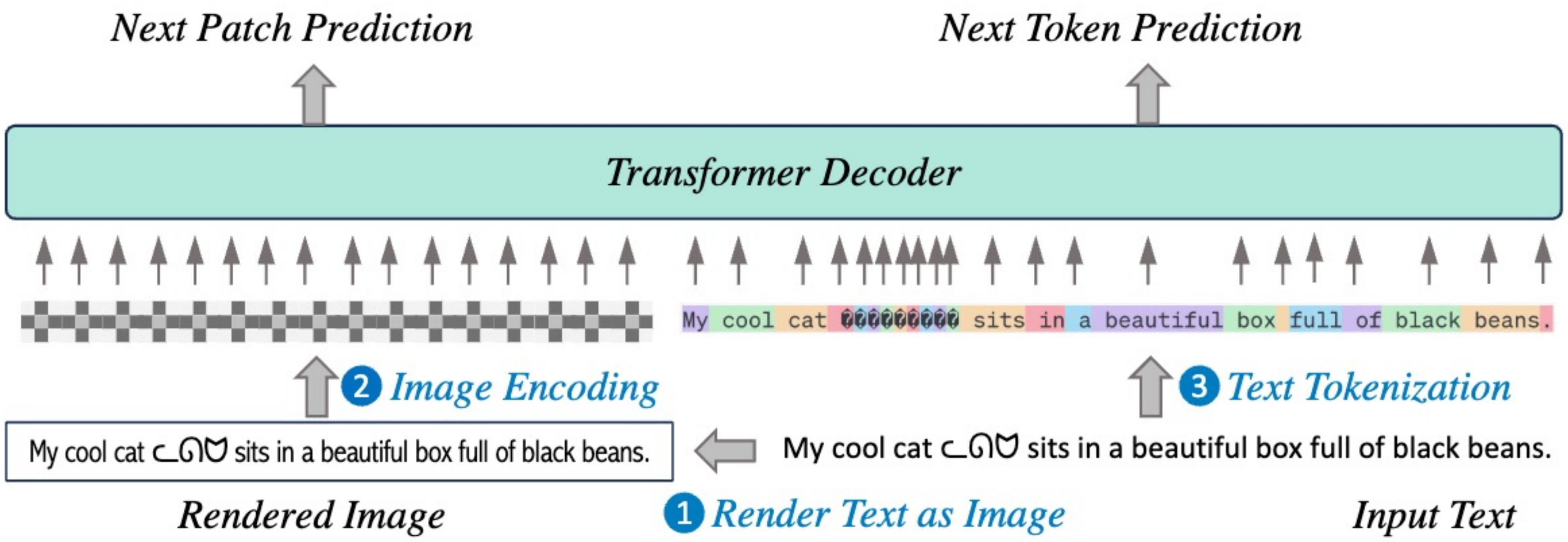


Pretraining Objectives

- **Image**: *Next patch prediction*. Given a sequence of N visual patches $x\_p = (x_p^1, x_p^2, \dots, x_p^N)$ where each visual patch x_t^p is a flattened patch embedding. We use a normalized mean squared error (MSE) loss quantifies the pixel reconstruction accuracy:

- **Text**: *Next token prediction*. We optimize a cross-entropy loss that evaluates the fidelity of predicted token sequences generated via teacher-forcing.

Autoregressive Pre-Training on Pixels and Texts



Pretraining Recipe

- **PixelGPT:** Trained solely on rendered image using MSE loss.
- **MonoGPT:** Trained on separate streams of rendered image and text data without any intermodal pairing.
- **DualGPT**: Trained on unpaired image and text input, and on paired image-text data (dual-modality).

Experiments

➢ Language Understanding

Model	#Param	Input Modality		MNLI-m/mm	QQP	QNLI	SST-2	CoLA	STS-B	MRPC	RTE	WNLI	Avg.
		Text	Pixel	Acc	F1	Acc	Acc	MCC	Spear.	F1	Acc	Acc	
BERT	110M	✓	✗	84.0/84.2	87.6	91.0	92.6	60.3	88.8	90.2	69.5	51.8	80.0
GPT-2	126M	✓	✗	81.0	89.4	87.7	92.5	77.0	74.9	71.5	52.0	54.9	75.6
DONUT	143M	✗	✓	64.0	77.8	69.7	82.1	13.9	14.4	81.7	54.9	57.7	57.2
CLIPPO	93M	✗	✓	77.7/77.2	85.3	83.1	**90.9**	28.2	**83.4**	84.5	59.2	-	-
PIXAR	85M	✗	✓	78.4/78.6	85.6	85.7	89.0	**39.9**	81.7	83.3	58.5	**59.2**	74.0
PIXEL	86M	✗	✓	78.1/**78.9**	84.5	**87.8**	89.6	38.4	81.1	**88.2**	60.5	53.8	74.1
PixelGPT	317M	✗	✓	**79.0**/78.2	**86.0**	85.6	90.1	35.3	80.3	84.6	**63.9**	**59.2**	**74.2**

- **Autoregressive Pixel-based Pre-training Rivals PIXEL**. PixelGPT outperforms PIXEL on QQP (+1.5), RTE (+3.4), and WNLI (+5.4).

Experiments

➢ Multilingual Evaluation

Model	#lg	#Param	Input Modality		ENG	ARA	BUL	DEU	ELL	FRA	HIN	RUS	SPA	SWA	THA	TUR	URD	VIE	ZHO	Avg.
			Text	Pixel																
Fine-tune model on all training sets (Translate-train-all)																				
mBERT	104	179M	✓	✗	83.3	73.2	77.9	78.1	75.8	78.5	70.1	76.5	79.7	67.2	67.7	73.3	66.1	77.2	77.7	74.8
XLM-R base	100	270M	✓	✗	85.4	77.3	81.3	80.3	80.4	81.4	76.1	79.7	82.2	73.1	77.9	78.6	73.0	79.7	80.2	79.1
BERT	1	110M	✓	✗	83.7	64.8	69.1	70.4	67.7	72.4	59.2	66.4	72.4	62.2	35.7	66.3	54.5	67.6	46.2	63.9
PIXEL	1	86M	✗	✓	77.2	**58.9**	66.5	68.0	64.9	69.4	57.8	63.4	70.3	60.8	**50.2**	64.0	54.1	64.8	**52.0**	62.8
PixelGPT	1	317M	✗	✓	**77.7**	55.4	**66.7**	**69.0**	**67.4**	**71.2**	**59.1**	**65.6**	**71.4**	**61.7**	47.0	**65.2**	**54.4**	**66.1**	50.5	**63.2**

- **PixelGPT matches the performance of BERT,** and consistently surpasses the in average accuracy across multilingual XNLI dataset.

Experiments

Model	Input Modality		MNLI-m/mm	QQP	QNLI	SST-2	CoLA	STS-B	MRPC	RTE	WNLI	Avg.
	Text	Pixel	Acc	F1	Acc	Acc	MCC	Spear.	F1	Acc	Acc	
TextGPT (text only)	✓	✗	79.9/80.0	86.1	86.1	91.5	47.3	**85.8**	86.3	63.5	56.3	76.3
MonoGPT (text+pixel)	✓	✗	80.0/**80.5**	85.9	**87.3**	90.1	40.2	83.8	87.0	62.8	56.3	75.4
	✗	✓	64.7/65.9	78.9	77.3	74.8	11.6	73.2	83.5	59.9	57.7	64.8
DualGPT (text+pixel+pair)	✓	✗	**80.1**/80.4	**86.5**	86.8	**91.6**	**49.0**	85.4	**87.6**	65.7	56.3	**76.9**
	✗	✓	71.5/71.7	82.8	81.6	83.4	17.2	80.2	84.1	**66.4**	**59.2**	69.4

Model	Input Modality		ENG	ARA	BUL	DEU	ELL	FRA	HIN	RUS	SPA	SWA	THA	TUR	URD	VIE	ZHO	Avg.
	Text	Pixel																
Fine-tune model on all training sets (Translate-train-all)																		
TextGPT (text only)	✓	✗	72.4	60.4	62.8	64.8	63.3	65.0	58.5	61.5	65.2	57.7	**59.9**	61.2	54.9	63.6	**63.1**	62.3
MonoGPT (text+pixel)	✓	✗	**72.9**	60.8	63.2	63.5	63.5	63.6	57.9	60.7	64.4	58.8	59.4	60.6	55.2	63.2	60.7	61.9
	✗	✓	66.8	47.1	61.2	61.8	63.4	64.5	56.7	59.2	64.9	56.8	48.7	61.8	52.1	61.0	50.7	58.4
DualGPT (text+pixel+pair)	✓	✗	72.7	**61.6**	63.8	64.7	63.9	65.1	58.8	61.6	65.4	59.0	59.8	62.2	**55.8**	63.4	62.1	**62.7**
	✗	✓	71.7	55.0	**67.6**	**66.5**	**66.8**	**68.4**	**59.0**	**64.4**	**68.9**	**61.3**	48.7	**64.3**	54.7	**65.8**	54.4	62.5

Table 5: Ablation results of model performance on XNLI under *Translate-Train-All* settings.

- Paired dual-modality data improves the language understanding tasks.

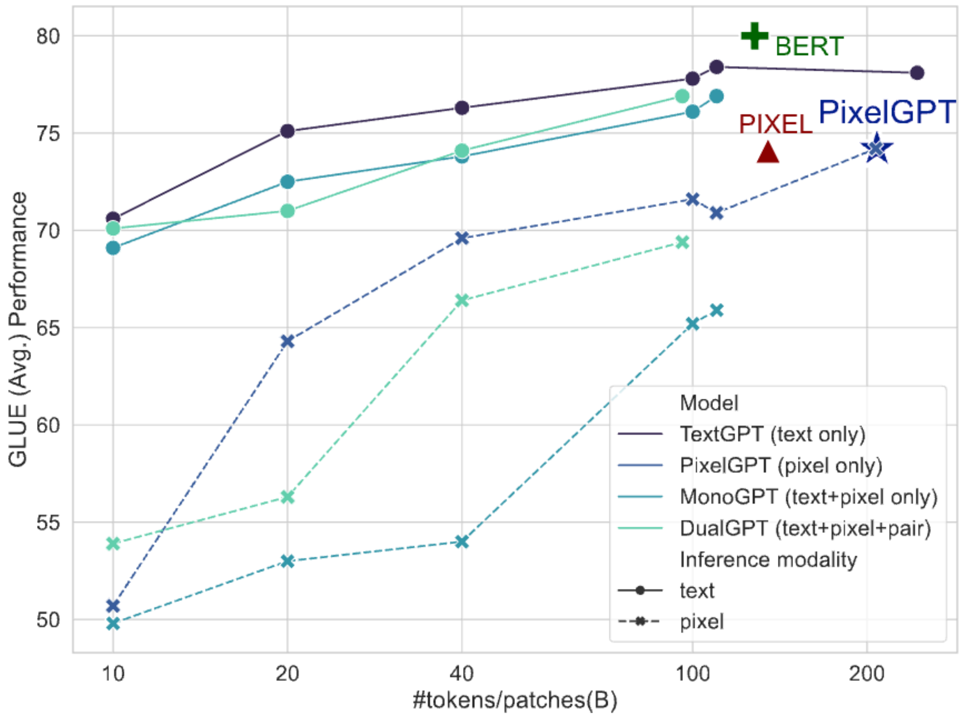


Figure 3: Training tokens/patches versus overall performance on GLUE benchmark.

Scaling Training Tokens vs. GLUE Performance

① **Pixel-based training exhibit an increased data demand.**

② **Utilizing paired dual-modality data improves multimodal learning, particularly for pixel-based input.**

Experiments

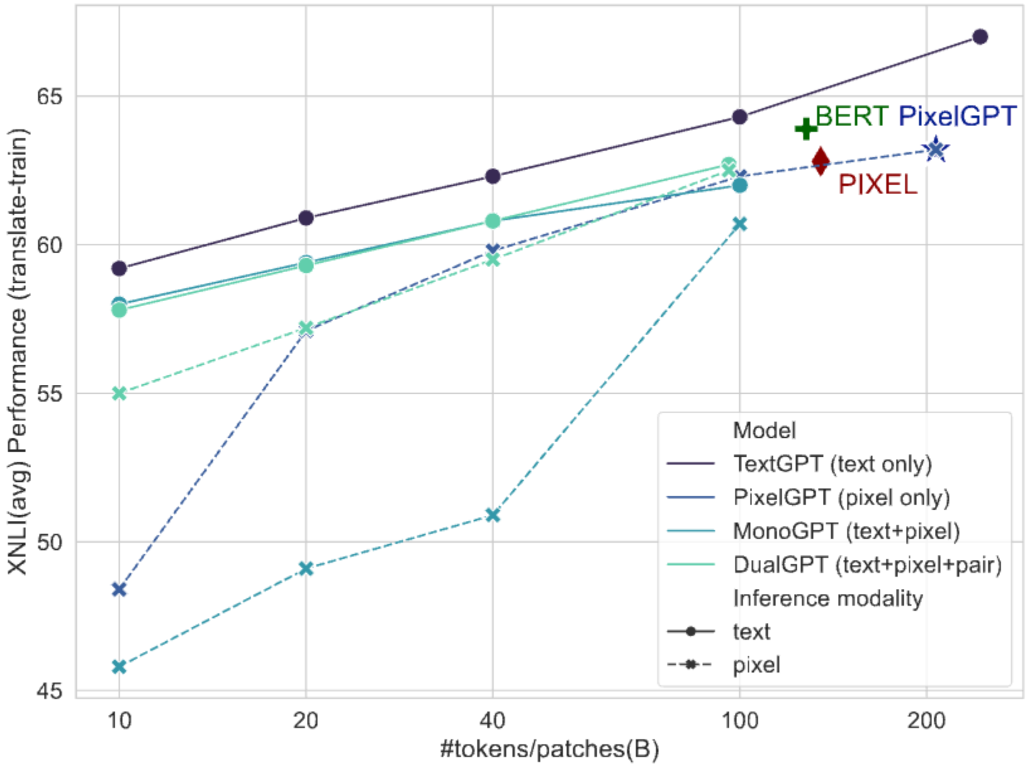


Figure 4: Training tokens/patches versus overall performance on XNLI benchmark.

Scaling Training Tokens vs. XNLI (Translate-Train-All) Performance

① **Pixel-based training exhibit an increased data demand for multilingual tasks.**

② **Utilizing paired dual-modality data at the early stages improves the pixel-based models.**

③ **Our text baseline (TextGPT) outperforms BERT.**

Experiments

➢ Analysis

A larger batch size improves stable training.

Robust to generalize across varied font representations.

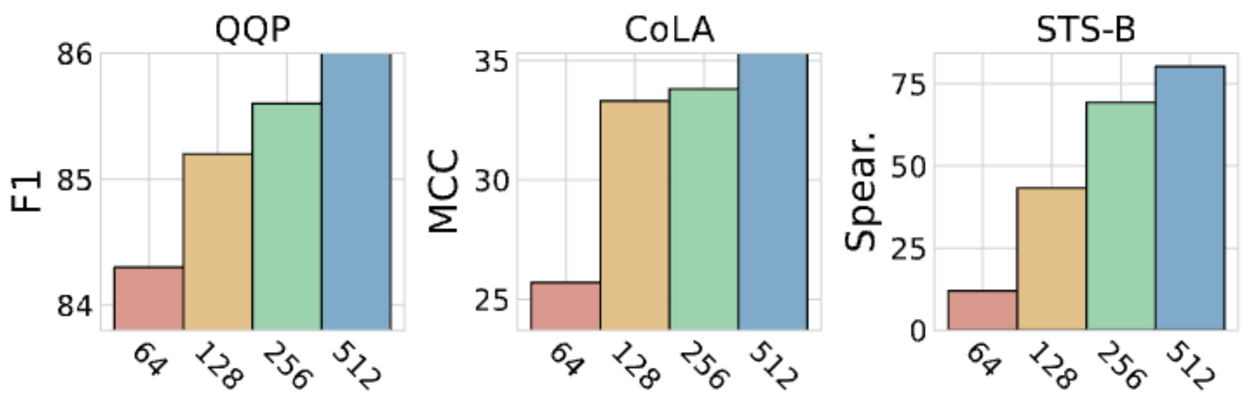


Figure 5: Analysis of escalating the global batch size.

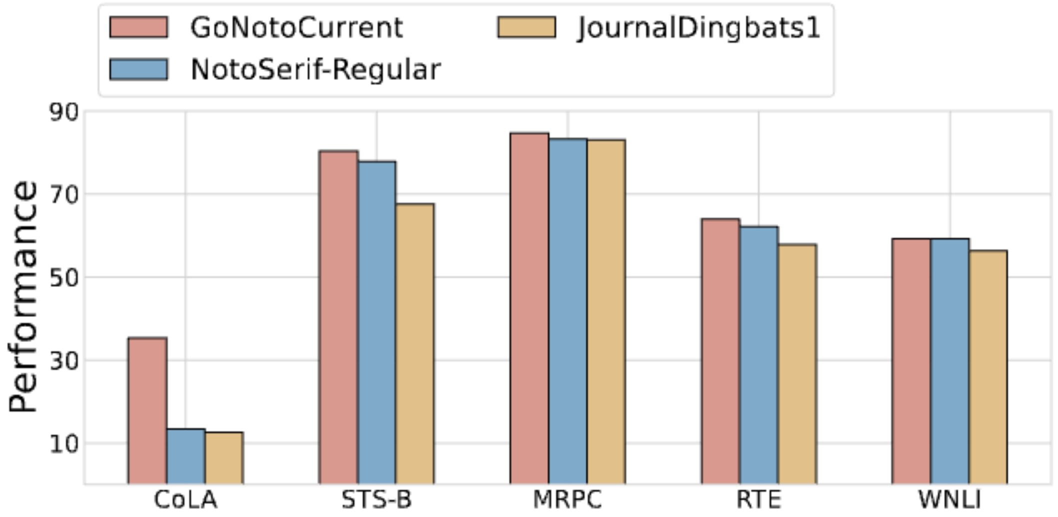


Figure 6: Analysis of fine-tuning on different fonts.

➢ Impact Analysis of Color Retention

Render Mode	Font	Acc	△
Grayscale	Apple Emoji	58.7	-
RGB		61.4	+2.7

Table 6: Comparison performance on HatemojiBuild dataset with grayscale and RGB rendering.

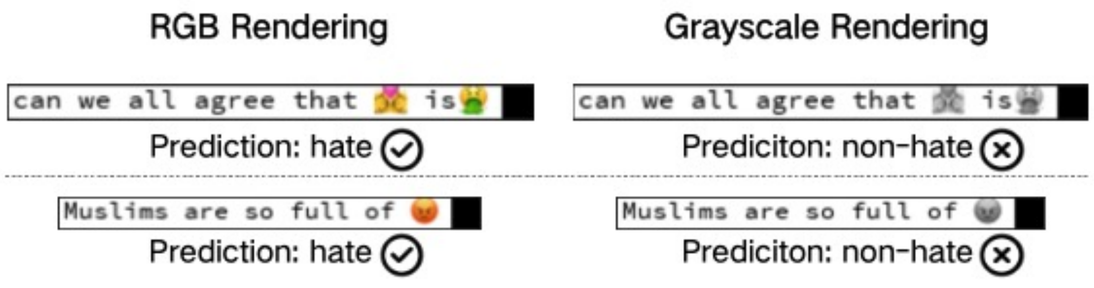


Figure 7: Example cases of **HatemojiBuild** predictions. ✓ and ✗ indicate the correct and incorrect predictions.

- **Color information matters.** RGB-rendered data finetuning outperforms its grayscale counterpart on HatemojiBuild dataset.

Code & Datasets & Checkpoints

README MIT license

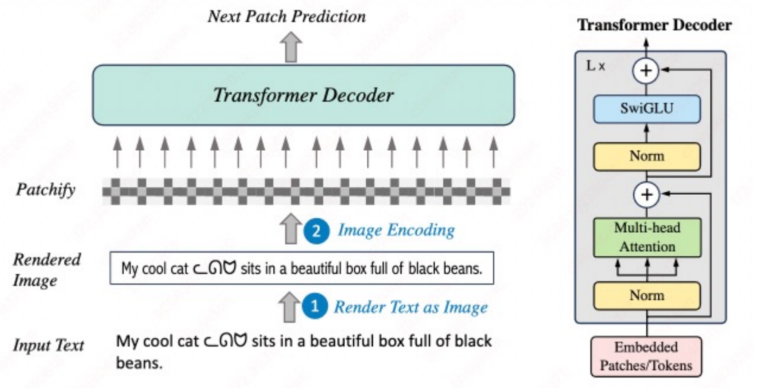
EMNLP'24 | Autoregressive Pre-Training on Pixels and Texts

Models | Data(rendered GLUE) | Data(rendered XNLI) | Paper | Proceedings EMNLP2024

The official repository which contains the code and model checkpoints for our paper Autoregressive Pre-Training on Pixels and Texts (EMNLP 2024).

🔥 News

- **21 September, 2024:** 🎉 Our work has been accepted to EMNLP 2024! 🎉
- **1 May, 2024:** 🎉 We release the official codebase and model weights of `PixelGPT`, `MonoGPT`, and `DualGPT`. Stay tuned!🔥



(a) Visual text image pre-training (*PixelGPT*). (b) Model architecture.

Code: https://github.com/ernie-research/pixelgpt

Model: https://huggingface.co/baidu/PixelGPT

Thank You!