From Pixels to Tokens: Byte-Pair Encoding on Quantized Visual Modalities

ICLR 2025

Wanpeng Zhang[1], Zilong Xie[2], Yicheng Feng[1], Yijiang Li[3], Xingrun Xing[4,5], Sipeng Zheng[4], Zongqing Lu[1,6]

[1]PKU [2]CUHK [3]UCSD [4]BAAI [5]CASIA [6]BeingBeyond

Background

- MLLMs struggle with effectively aligning visual and textual modalities
- Current approaches:
  - Late-fusion with specialized encoders: Complex alignment challenges
  - Early-fusion token-based: Lacks explicit structural information
- Text-only LLMs benefit from BPE tokenization but visual modalities lack equivalent

Theoretical Foundation

Proposition 1. *For data generating processes described in either Scenario 1 or Scenario 2, as $m \to \infty$, the optimal cross-entropy loss among unigram model family $\mathcal{Q}_{\text{1-gram}}$ satisfies*

$$\liminf_{m \to \infty} \min_{Q \in \mathcal{Q}_{\text{1-gram}}} \mathcal{L}_m(Q) \geq H(\pi) = \sum_{a \in \mathcal{C}} \pi(a) \log(\pi(a)). \tag{3}$$

In contrast, the optimal unconstrained cross entropy loss satisfies

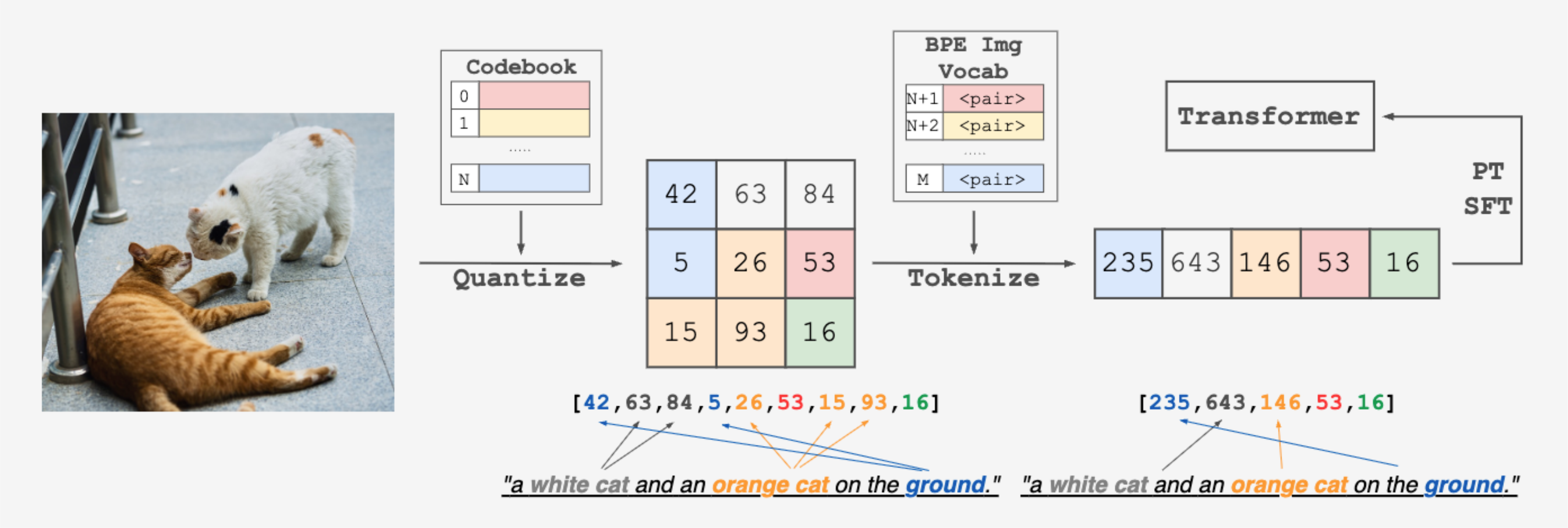
$$\lim_{m \to \infty} \min_{Q} \mathcal{L}_m(Q) = H_\infty \triangleq - \sum_{a \in \mathcal{C}} \sum_{a' \in \mathcal{C}} \pi(a) P(a' \mid a) \log \left(P(a' \mid a) \right). \tag{4}$$

Proposition 2. *For data generating processes described in either Scenario 1 or Scenario 2, assume that $\delta \triangleq \min_{a,a' \in \mathcal{C}} P(a'|a) > 0$. Then there exists a tokenizer with a dictionary containing at most D tokens, along with an encoding function $\mathrm{enc}(\cdot)$ applied to $\boldsymbol{X}$, such that*

$$\limsup_{m \to \infty} \min_{Q \in \mathcal{Q}_{\text{1-gram}}} \mathcal{L}_m(Q \circ \mathrm{enc}(\cdot)) \leq \frac{1}{1 - \varepsilon} H_\infty, \tag{5}$$

where $\varepsilon = \log(1/\delta)/(0.99 \log(D))$ and $D \in \mathbb{N}$ is an arbitrary constant that is sufficiently large.

Our Method



- **BPE Image Tokenizer**:
  - Step 1: Quantize image into initial token IDs using VQ-GAN (codebook size 8192)
  - Step 2: Learn merged tokens based on frequency patterns (like text BPE)
  - Step 3: Combine tokens into semantically meaningful units with structural information

Our Method

- Information loss theoretical upper bound:

$$L_{bpe} \leq (|D_{bpe}| - |D_{vq}|) \times (-p_{\min} \log(p_{\min})).$$

- For typical configuration: ~0.35% information loss, acceptable trade-off.

Our Method

Algorithm 4.1 BPE Image Tokenizer training procedure.

Input v_0, m, D. ▷ v_0: initial vocab size, m: new vocab size, D: training data
$v \leftarrow v_0$ ▷ v: current vocab size
$A \leftarrow \text{zeros}(v \times v)$ ▷ A: adjacency matrix
$V \leftarrow \emptyset$ ▷ V: extended vocabulary
for $i \leftarrow 1$ to m **do**
  $A \leftarrow \text{UpdateMatrix}(D)$
  $(p, f) \leftarrow \text{MaxFreqPair}(A)$ ▷ p: best pair, f: frequency
  if $f = 0$ **then break**
  end if
  $V \leftarrow V \cup \{(p, v)\}$
  $D' \leftarrow \emptyset$
  for each $d \in D$ **do**
    $d' \leftarrow$ Replace p with v in d
    $D' \leftarrow D' \cup \{d'\}$
  end for
  $D \leftarrow D'$
  $v \leftarrow v + 1$ ▷ set next id for new token
end for
return V

Training Pipeline

- **Base Model**: Llama-3.1-8B
- **Two-Stage Training Process**:
  - **PT (Pretraining)**:
    - Freeze original text embeddings
    - Train only visual embeddings
    - 595K images (CC-3M) + 558K (LCS)
  - **SFT (Supervised Fine-Tuning)**:
    - Unfreeze all weights
    - 1.27M entries from LLaVA-OneVision Dataset
- **Key Difference**: Direct fusion of image modality without separate encoders

Experiment Results

	Training type	VQAv2	MMBench	MME^p	MME^c	POPE	VizWiz
LLM+VQ	SFT	51.1	35.9	972.3	231.8	73.8	43.1
	PT(full)+SFT	53.7	37.0	1037.2	261.4	75.3	44.2
	PT(freeze)+SFT	55.4	37.6	1054.5	277.0	76.0	45.3
LLM+VQ+BPE (Being-VL-0)	SFT	52.2	35.4	1029.7	269.6	76.3	45.3
	PT(full)+SFT	56.5	38.6	1144.6	284.3	77.3	45.8
	PT(freeze)+SFT	57.1	40.9	1223.5	307.1	79.0	46.0
Additional scaling (PT)	+RefCOCO(50.6K)	58.6	42.3	1257.4	314.3	79.8	47.1
	+AOKVQA (66.2K)	59.6	43.1	1288.1	321.4	80.4	47.5
Additional scaling (SFT)	+ShareGPT4o (57.3K)	60.2	43.7	1304.5	327.7	80.9	47.8
	+ALLaVA Inst (70K)	60.6	44.0	1316.2	331.0	81.3	48.2

- BPE Image Tokenizer consistently outperforms direct VQ approach
- Performance improves with vocabulary size up to 8K (balanced utilization)
- Model shows strong category-specific improvements (Existence: 145.0 vs 113.3)

Summary & Future Work

- First explicit tokenization of multimodal data like text-only LLMs
- Theoretical analysis validates benefits of tokenization for 2D data
- Significant performance gains with limited training data (0.1% of typical CLIP training)
- Future: Scale up data, apply to video, explore more sophisticated tokenization

Thanks!