Multimodality extension to Universal Multilingual BPE Text Tokenizer

This abstract is proposing **Multimodality extension** to the paper One Tokenizer To Rule Them All..[1]. The referenced paper [1] mainly uses bucketed weighting scheme on unseen/expanded languages by script e.g. (Devnagari, Hindi) or (Latin, Polish) pair and trains Byte-Pair Encoding (BPE) model on diverse **text corpus** across ~69 languages (combination of languages used in pretraining and many others that are only intended for tokenizer coverage). It also provides byte-fallback for edge cases outside training data. This abstract is about **concrete enhancements to the above paper's** [1] bucketed weighting scheme and explicitly accounts for **multiple modalities like Images, Speech, OCR, Text etc**. The below enhancements are aimed at achieving >= 0.95 CMS score to consider the resultant tokenizer as **Multimodal tokenizer.** 1.Modality-aware buckets[2] Extend buckets to [script, modality] pairs e.g., (Devanagari, OCR), (Arabic, ASR) and assign higher sampling weights to underrepresented pairs and monitor per-bucket coverage in tokens/word and bytes/token. Keeping total vocabulary same; train BPE on weighted samples. Measure improvement in OCR/ASR-related tasks for same scripts; per-bucket tokens/word. **Success:** ≥ small positive lift (0.5–2 pts) on OCR-heavy tasks for those scripts vs baseline.2.Confidence-weighted sampling Use OCR/ASR confidence scores to downweight low-confidence examples or to *preferentially* sample medium-confidence ones for tokenizer training. Integrate OCR/ASR confidence scores and sample with prob $\propto (\alpha + \text{conf}^{\beta})$. Precompute confidences; tune α (e.g., 0.05) and β (e.g., 1.0→2.0). Keep some low-confidence included via α. **Measure:** Token noise (tokens seen only in low-confidence data), downstream VQA/DocVQA on OCR; stability of merges. **Success:** Cleaner merges (fewer spurious tokens) and small downstream improvement; reduction in tokens primarily seen in noisy buckets.3. Adaptive Reweighting with Feedback[3] During tokenizer training, periodically evaluate downstream proxy tasks (small VQA/ASR validation slices). Reweight buckets that show poor downstream performance. Every N steps, compute per-bucket validation loss; increase sampling weight for buckets with high loss (up to a cap). **Measure:** Convergence speed on proxies; stability of vocab. **Success:** Faster improvements on held-out proxies; stable vocabulary.4. Cross-Modal Coverage Balancing[4] Upweight text segments that are aligned to images/speech (e.g., OCR region + image) so merges capture visually-grounded tokens by marking multimodal aligned text and multiply sample weight by γ (1.5–3.0) and **measure** improvement in grounded retrieval/DocVQA EM and fewer mis-OCR tokens. **Success**: Noticeable lift on grounding tasks (>= 1-3 pts) 5. Curriculum-Based Bucket Scheduling[5] Systematic phases in the training as in Phase A, clean high-confidence multimodal pairs. Phase B: gradually add noisy/augmented examples for N steps. **Measure:** Merge stability (fewer reversions), downstream robustness to noisy OCR. **Success:** Better OCR robustness and fewer low-quality tokens. 6. Multimodal-Aware Validation Metrics[6][7][8][9] We want metrics that reflect multimodal performance, not just perplexity or compression. Composite Multimodal Score (CMS) should be >= 0.95 AND no single Primary Metric falls below target.

Metric (Mi)	Dataset(s)	Expected Value / Success Target	Priority
Text Perplexity	WikiText, Flores, mC4 (clean text)	Within 5% of baseline tokenizer	Primary
Avg. Tokens per Word (Text CR)	WikiText, Flores	≤ 1.1x baseline	Primary
OCR Compactness Ratio (OCR-CR)	FUNSD, SROIE, IIT-CDIP, SynthText	≤ 1.3x text CR	Primary
OCR Robustness Score (Edit distance overlap)	FUNSD, SROIE	≥ 90% overlap with GT tokens	Secondary
ASR Token Error Rate (TER)	LibriSpeech (clean/other),	Should ≈ WER (gap ≤ 3%)	Primary
WER-to-TER Gap	LibriSpeech, MLS	≤ +3% gap	Secondary
DocVQA Exact Match (EM)	DocVQA, TextVQA	≥ 95% of baseline EM	Primary
DocVQA F1 Score	DocVQA	≥ 95% of baseline F1	Secondary
Caption Compactness (tokens per caption)	COCO Captions, Flickr30k	Within 10% of baseline tokenizer	Primary
Alignment Overlap (IOU@token)	COCO Captions + CLIP labels	≥ baseline tokenizer IOU	Secondary
Composite Multimodal Score (weighted across modalities)	All above datasets	Overall ≥ 0.95 relative to baseline	Primary

$$CMS = \sum w_i \cdot M_i/T_i$$

Mi = measured metric value
(Refer Metric column from side table)
Ti = target threshold (so ratio ≥1 is "good")
wi = weight assigned to modality
(wi values = [Core text = 20%, OCR = 25%, ASR = 25%, DocVQA = 15%, Captions = 15%])

References -
[1] https://arxiv.org/pdf/2506.10766 [2] https://www.rohan-paul.com/p/multilingual-and-multimodal-llms [3] https://aclanthology.org/D17-1158/ [4] https://arxiv.org/abs/1909.11740 [5] https://dl.acm.org/doi/10.1145/1553374. 1553380[6] https://arxiv.org/abs/2007. 00398
[7]https://arxiv.org/abs/1904.08920[8] https://www.openslr.org/12[9] https://commonvoice.mozilla.org/en

Tools used – ChatGPT for validating the approaches and finding the references to the proposed approaches in this abstract for citation purpose.