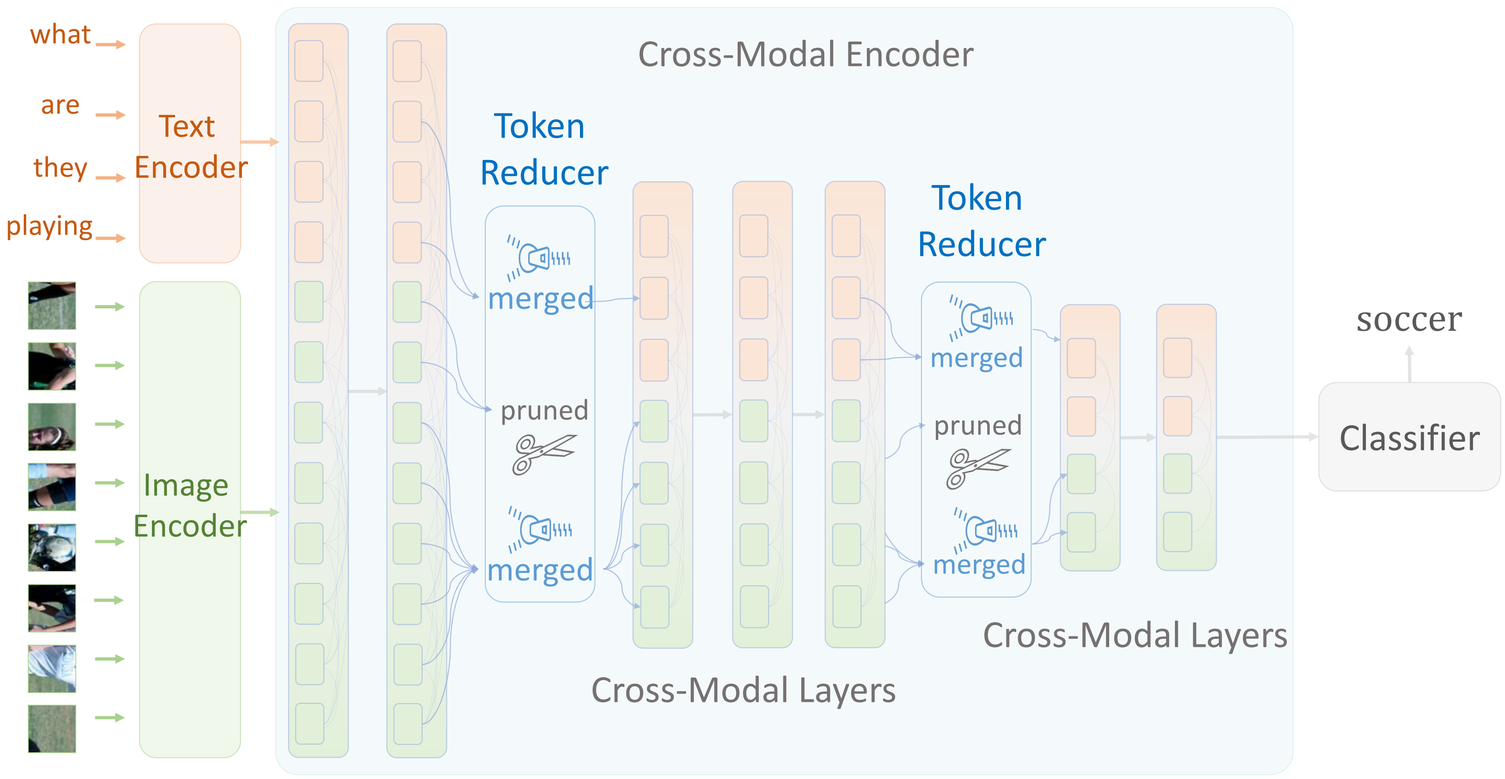
PuMer: Pruning and Merging Tokens for Efficient Vision Language Models

UWNLP

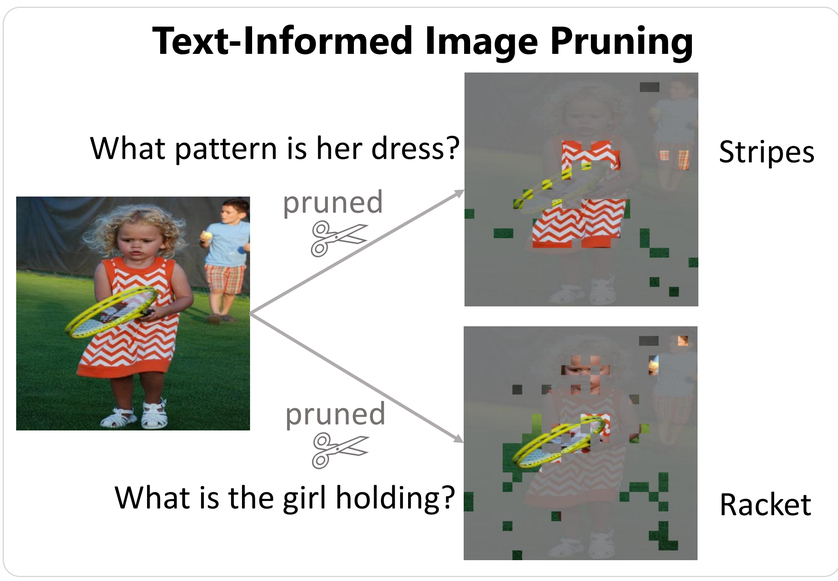
Qingqing Cao, Bhargavi Paranjape and Hanna Hajishirzi

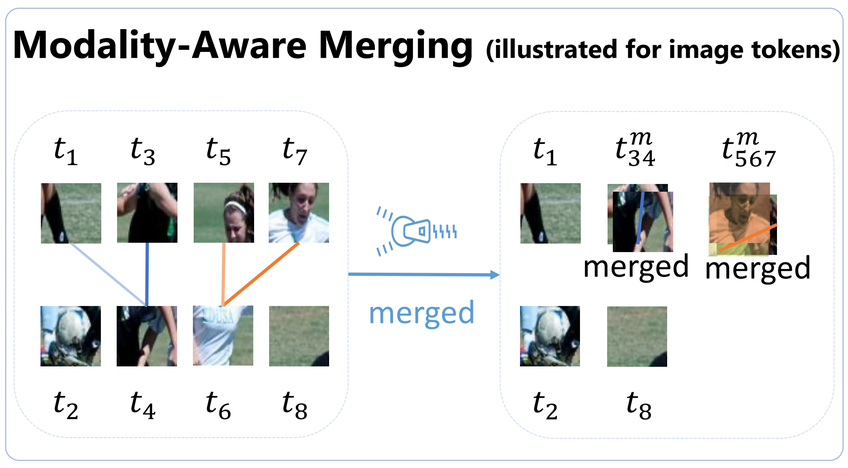
UNIVERSITY of WASHINGTON

ACL 2023

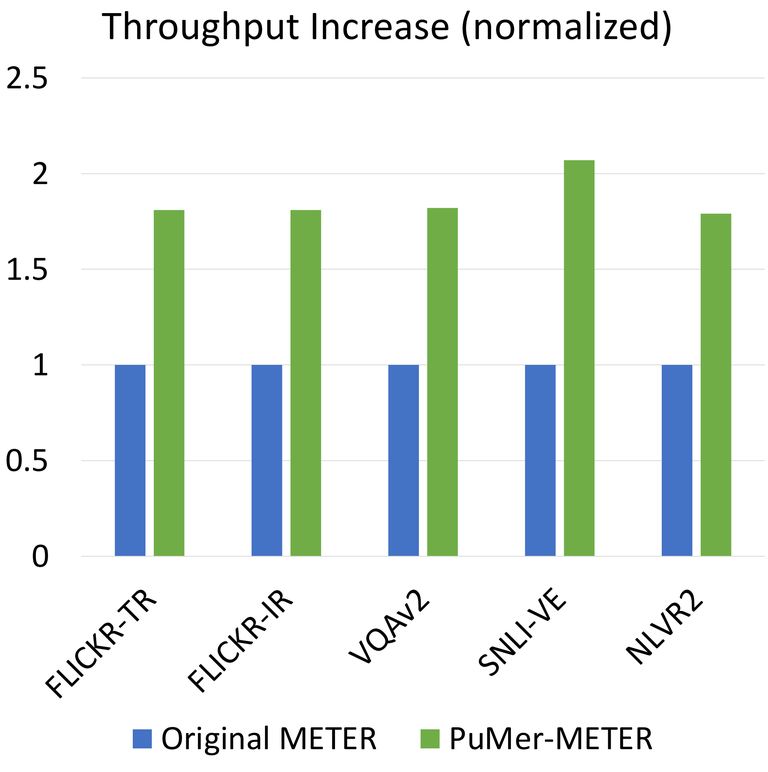


PuMer applies **text-informed image pruning** and **modality-aware merging** to the computational-intensive cross-modal layers of a VL model via lightweight **token reducers**.

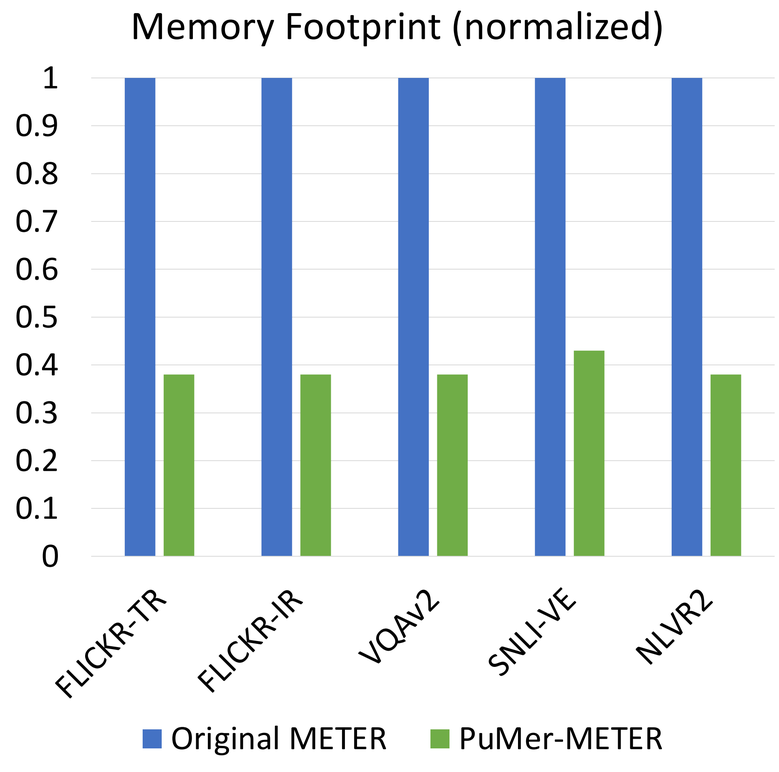




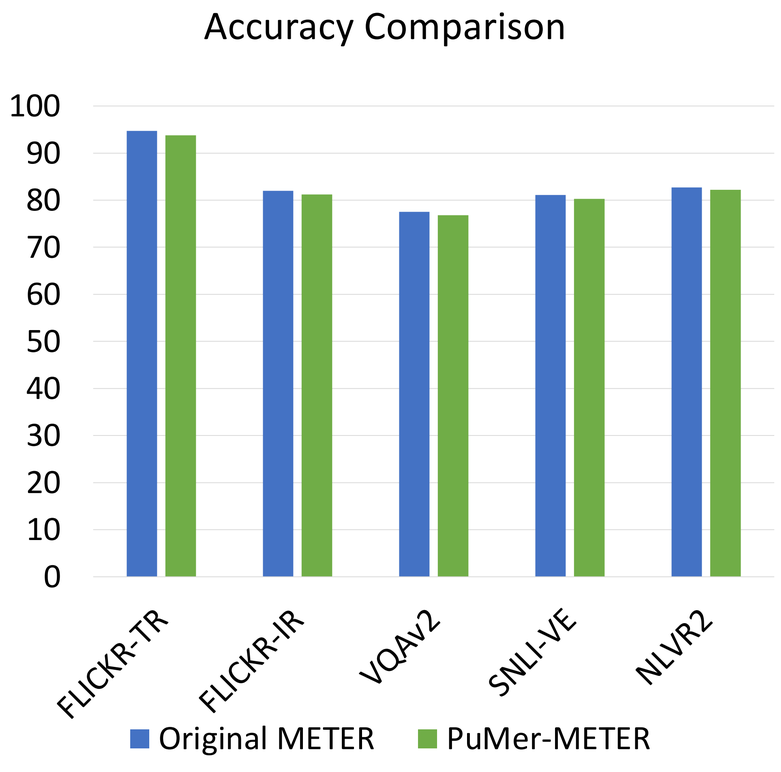
PuMer is a token reduction framework that progressively removes and combines the input text and image tokens via text-informed image pruning and modality-aware token merging, increasing 2x inference throughput and reducing 50% memory footprint with <1% accuracy drop.



PuMer brings **1.8 ~ 2x** inference throughput increase for SoTA VL models



PuMer reduces **38% ~ 43%** inference memory footprint for SoTA VL models



PuMer causes **<1%** accuracy drop for SoTA VL models over all studied VL tasks