The Hebrew University Of Jerusalem

Ben-Gurion University of the Negev

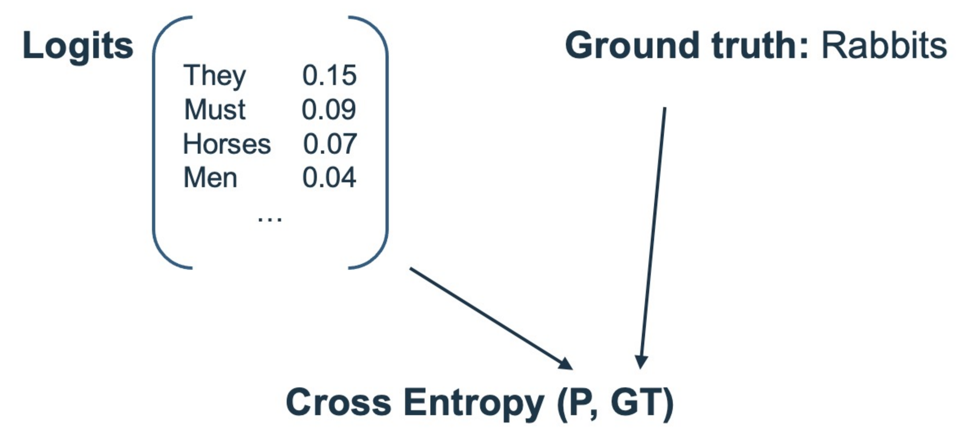
Data Efficient Masked Language Modeling for Vision and Language

Yonatan Bitton, Gabriel Stanovsky, Michael Elhadad, Roy Schwartz

Overview

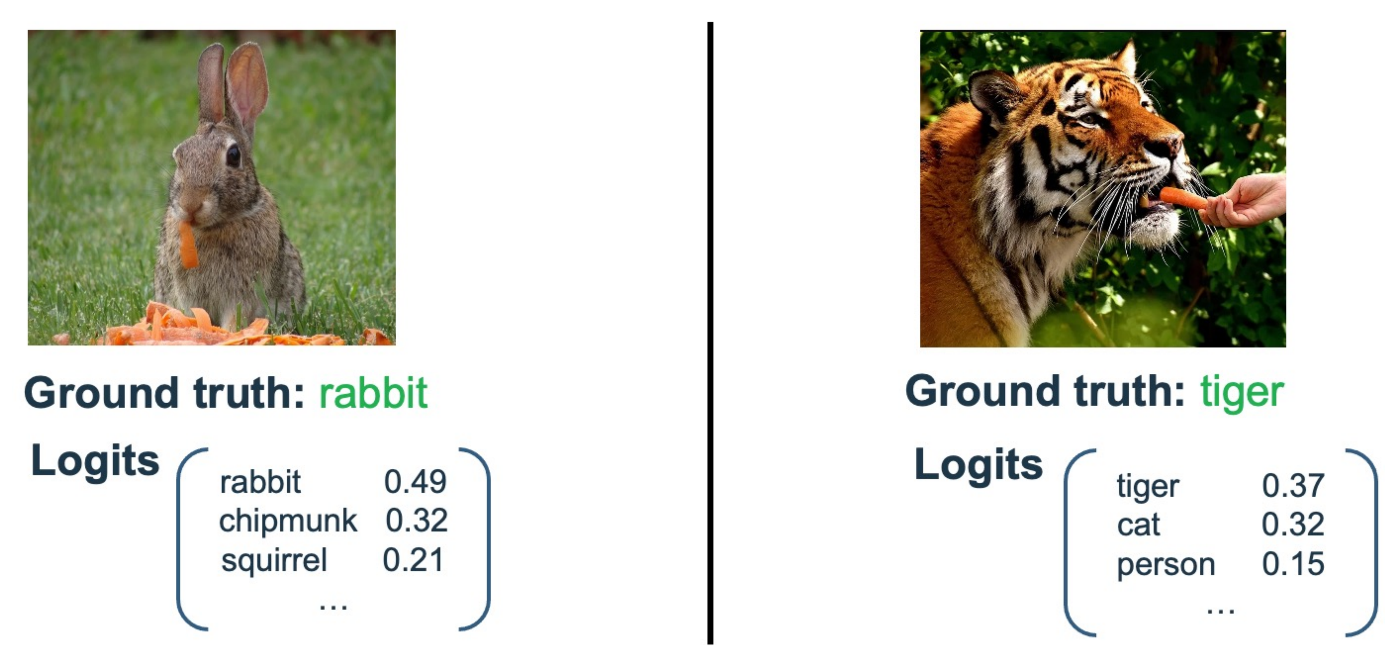
Masked language modeling (MLM) is a key pre-training objective in text transformers.

[MASK] have muscled hind legs that allow for maximum force, maneuverability, and acceleration

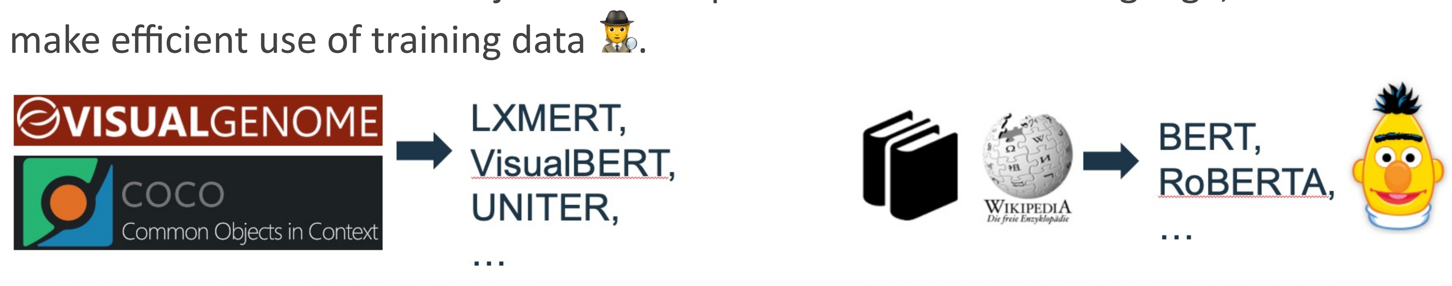


The difference in the cross-modal setting, is that the model takes into account both the textual context and the image 👓.

A **[MASK]** is eating the carrot



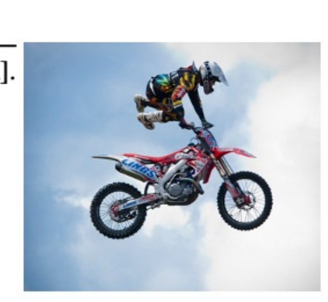
We find the current MLM objective sub-optimal for vision and language, as it does not make efficient use of training data 🕵.



- In many cases, no token is masked
- ~50% of tokens in pre-train data are stop-words or punctuation marks

Focusing on stop-words is leading to under-utilization of the image 🖼.

Sentence	A person performs a stunt jump on a [MASK].
Masked token	motorcycle
Top 5 predictions	motorcycle, bike, ramp, bicycle, cycle
Top 5 predictions w/o image	building, wall, beach, field, street
Loss	0.25
Loss w/o image	3.96
Δ *image loss*	3.71

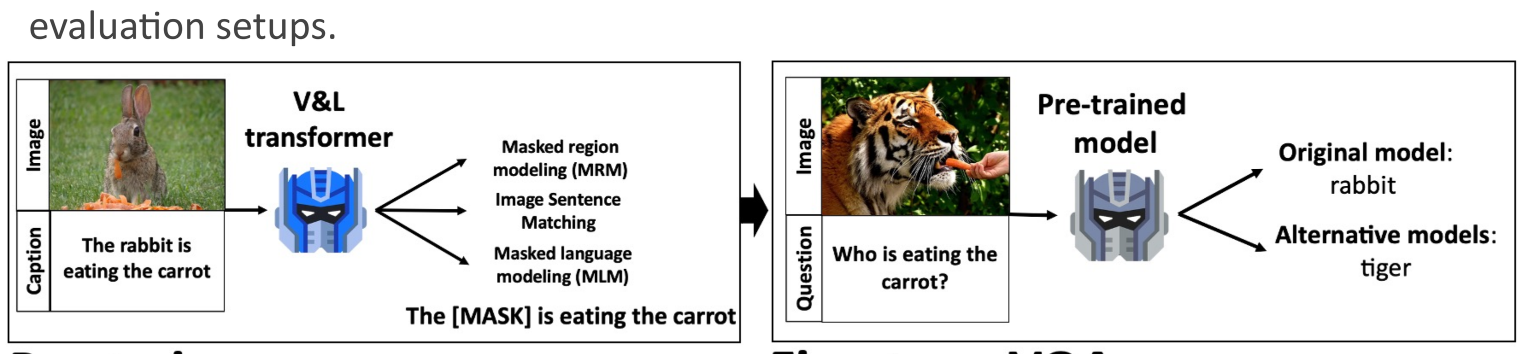


Masking strategy	With Image		Without Image		Image Necessity	
Metric	image loss (exp)	Accuracy @ 5	image loss (exp)	Accuracy @ 5	Δ *image loss* (exp)	Accuracy @ 5
Baseline MLM	3.2	89%	8.9	78%	5.7	10%
Stop-words & punctuation, 15%	1.5	98%	2.9	96%	1.4	2%
Content words, 15%	9.4	76%	38.7	56%	29.3	20%

We suggest alternative masking strategies, specific to the cross-modal settings, addressing these shortcomings 📈.

- Content words
- Objects
- Top concrete

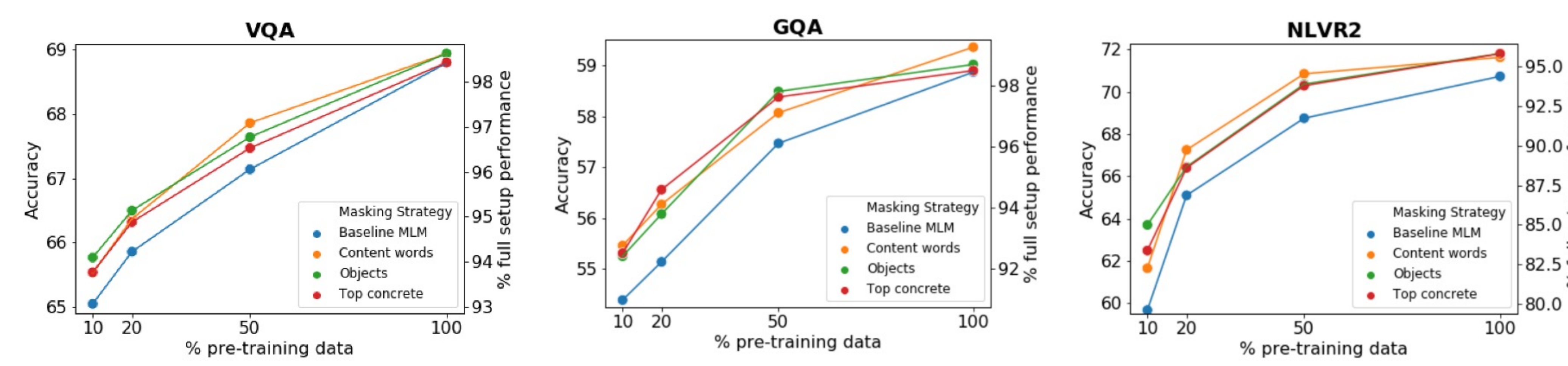
Our method masks words that require the image in order to be predicted.
Our pre-train masking strategy consistently improves over the baseline strategy in two evaluation setups.



Experiments

Downstream tasks

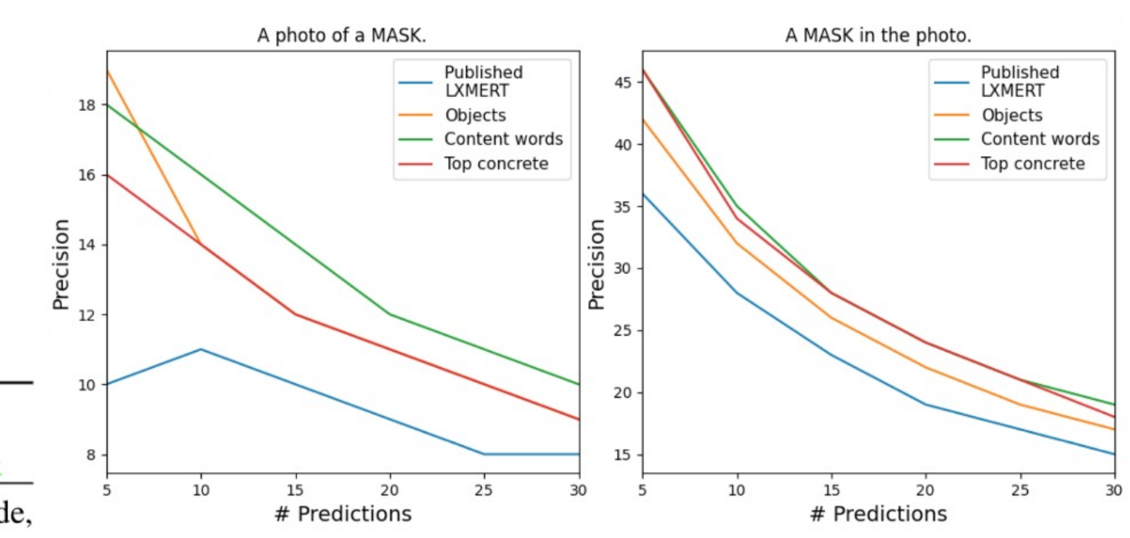
Our alternative masking strategies consistently outperform the baseline MLM strategy, especially in low resource settings



Prompt based object detection

Our alternative models are more responsive to the image contents

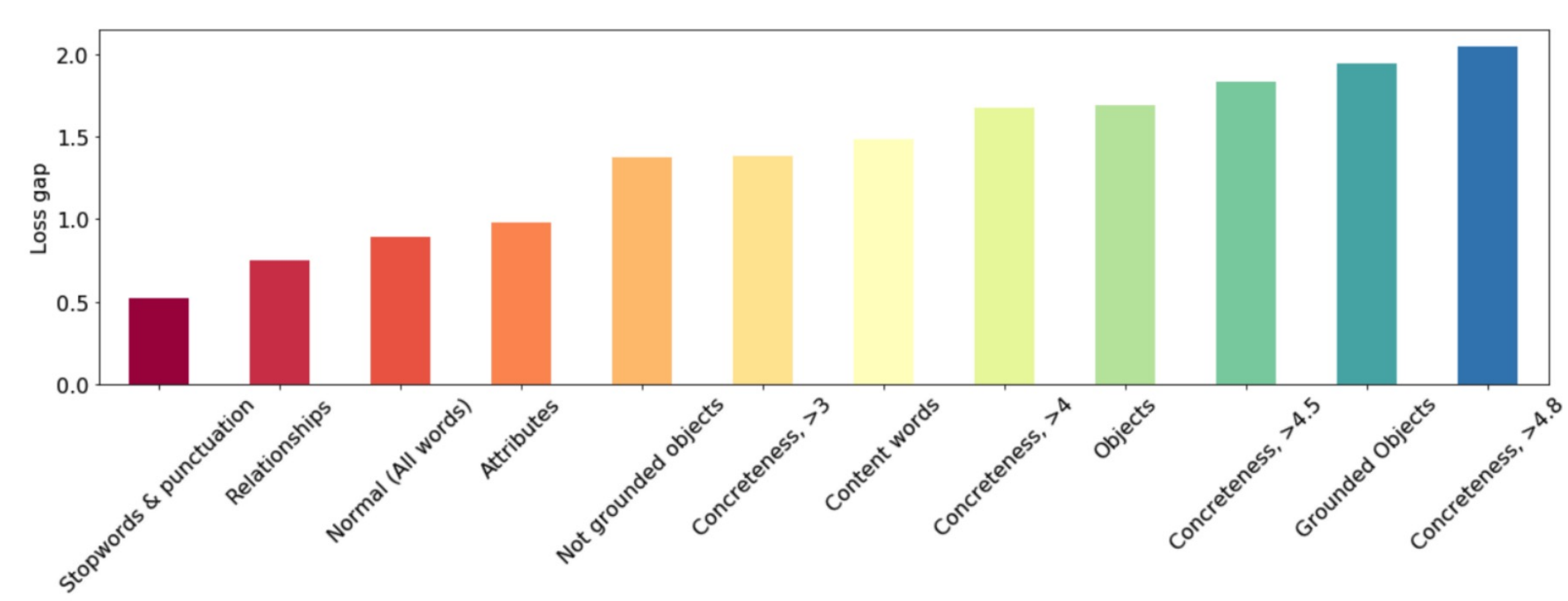
Published LXMERT	bathroom, beach, city, kitchen, woman
Objects	motorcycle, bathroom, parade, man, crowd
Ground truth objects	glasses, gang, motorcycle, shirt, man, parade, woman, road, ducati, cellphone, light, hat, ...



Analysis

Hierarchy of Masked Semantic Classes

Which kind of tokens will make the model to actively rely on the image?



MLM Performance across Word Classes

Does a model trained with Objects strategy learn to complete words from other classes?

Model Masking Strategy	Baseline MLM	Objects	Content words	Top concrete
Baseline MLM	87%	27%	70%	36%
Stop-words & punctuation, 15%	98%	4%	80%	13%
Content words, 15%	74%	57%	62%	62%
Objects	76%	85%	82%	83%
Attributes	70%	22%	59%	50%
Relationships	89%	15%	75%	25%