Bridging the Gap : Pretraining for Multi-Modal Settings

How Pretraining is changing the world

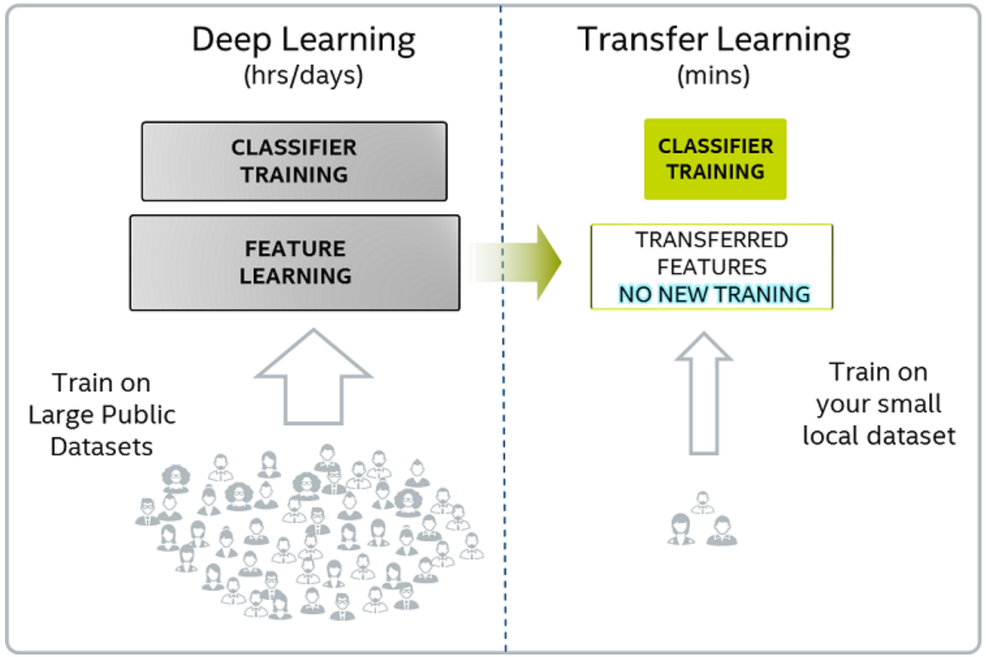
By Hrituraj Singh

#AdobeRemix

Jon Noorlander

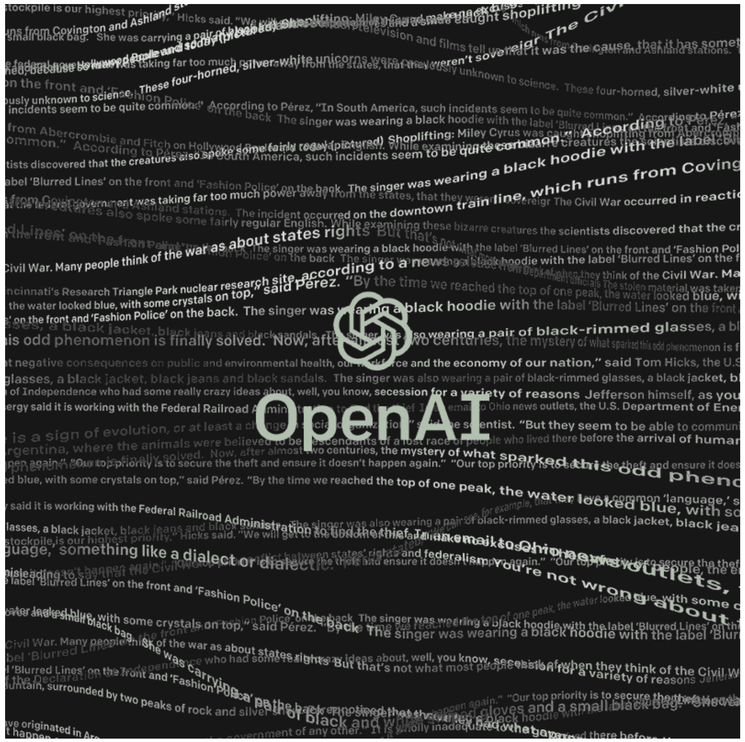
Introduction

ImageNet challenge introduced us to models which after being trained on ImageNet Dataset could be used for a number of pretraining tasks. This came to be known as *Transfer learning*.



Introduction

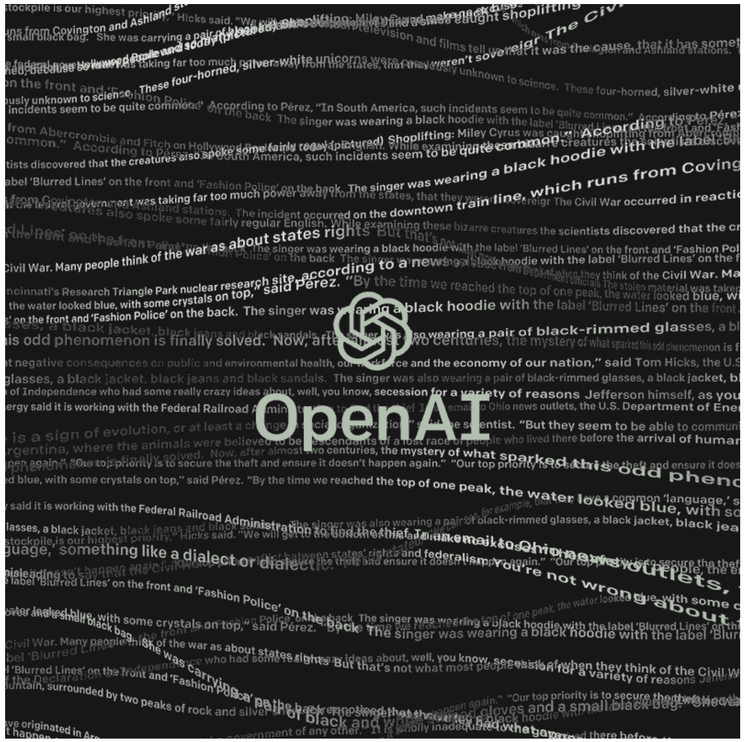
This was however earlier limited to Vision up until two years ago when this happened :



Introduction

This was however earlier limited to Vision up until two years ago when this happened :

Improving Language Understanding with Unsupervised Learning – Radford et al.



Introduction

Sea of pretrained models

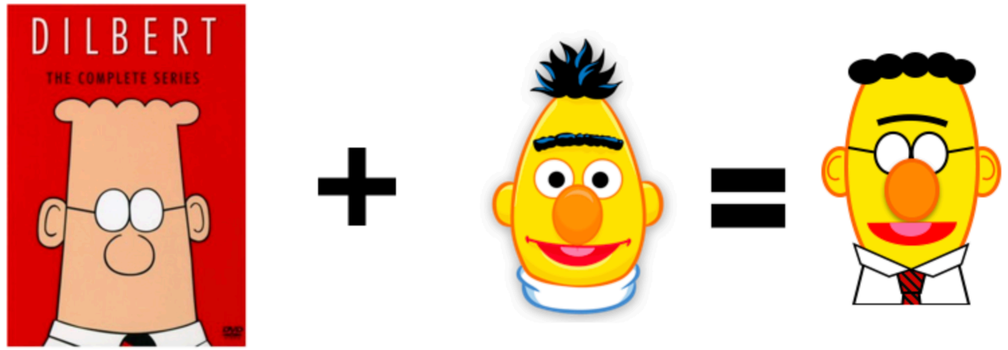
BERT

GPT-2

ELMo

Introduction

What this is about: Vision + Language

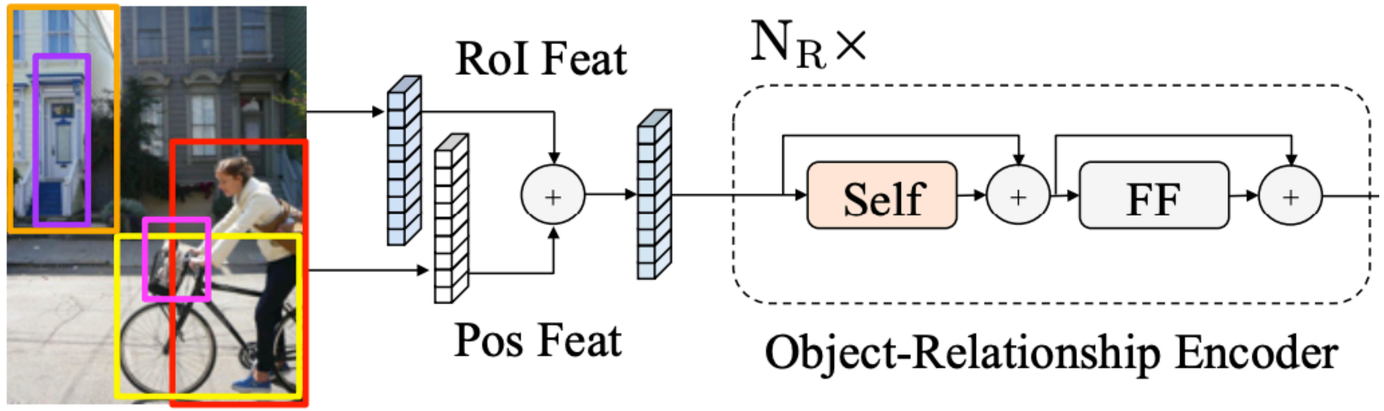


EMNLP 2019 – LXMERT

Encoders : Object Relationship Encoder

1. Detect *m* objects
2. Get their ROI features and bounding box positions
3. Get the object embedding as :

$$\hat{f}_j = \text{LayerNorm}\left(W_\text{F} f_j + b_\text{F}\right)$$
$$\hat{p}_j = \text{LayerNorm}\left(W_\text{P} p_j + b_\text{P}\right)$$
$$v_j = \left(\hat{f}_j + \hat{p}_j\right)/2$$

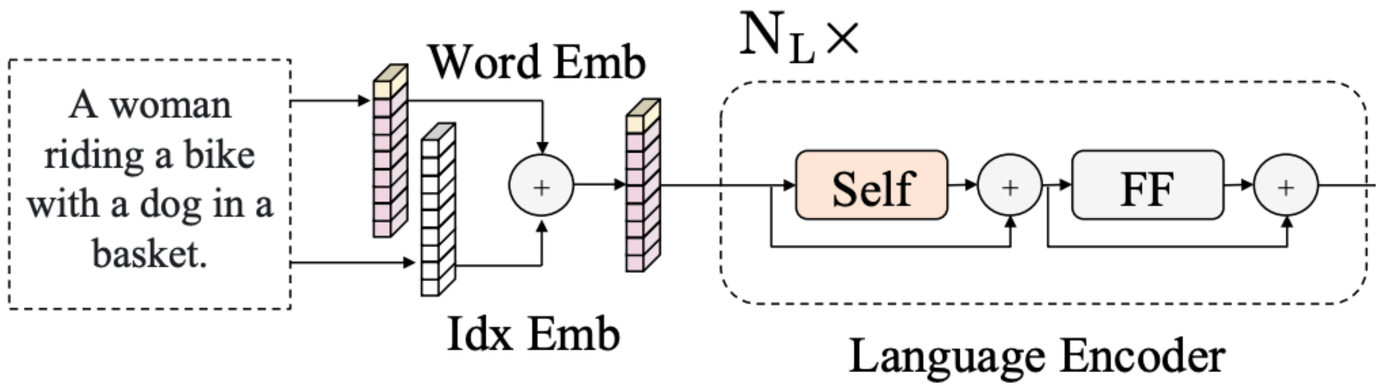


EMNLP 2019 – LXMERT

Encoders : Language Encoder : Basically BERT

1. Get tokens from the sentence/text input
2. Get each word *Wi* and its index *i*
3. Get the word embedding as :

$$\hat{w}_i = \text{WordEmbed}\left(w_i\right)$$
$$\hat{u}_i = \text{IdxEmbed}\left(i\right)$$
$$h_i = \text{LayerNorm}\left(\hat{w}_i + \hat{u}_i\right)$$

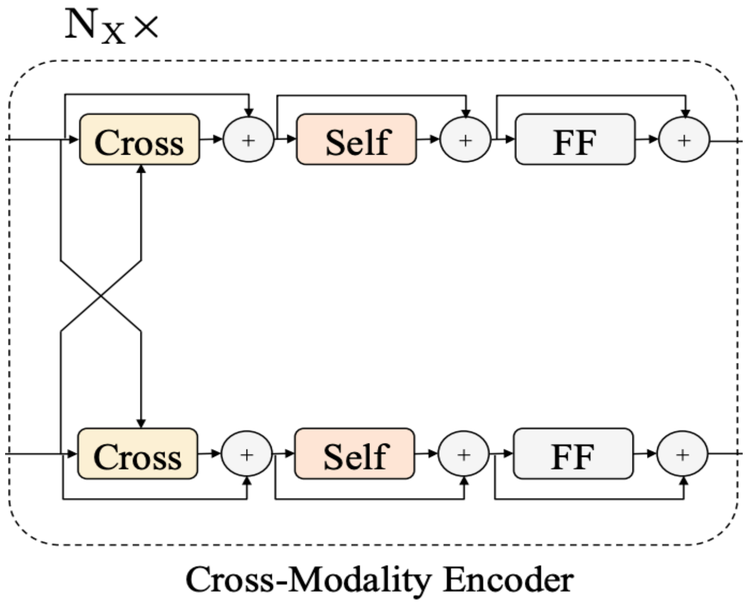


EMNLP 2019 – LXMERT

- Encoders : Cross Modality Encoder : Basically Enc-Dec Attention

1. Get contextual embeddings for *m* objects
2. Get contextual embeddings for *n* tokens
3. Make them interact as:

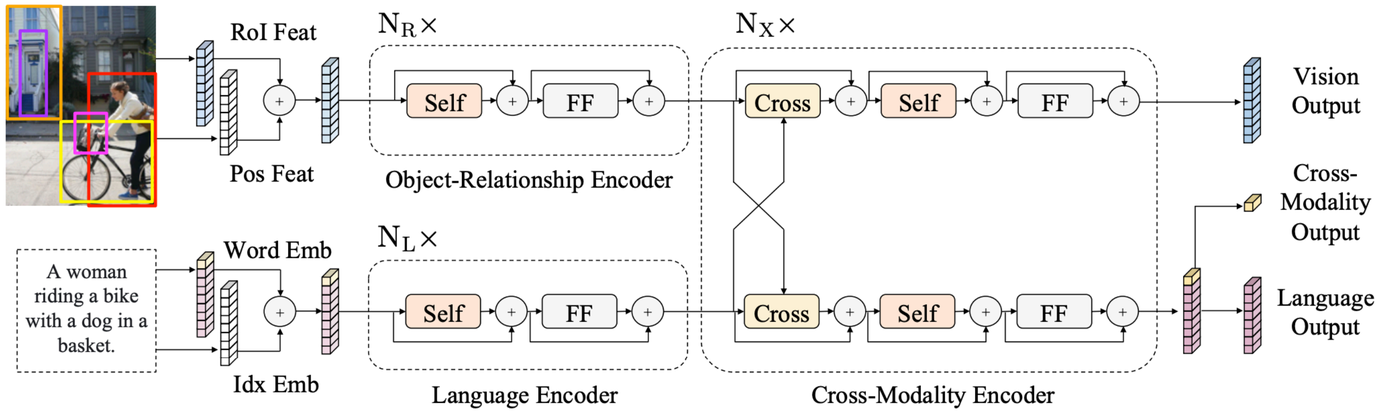
$$\hat{h}_i^k = \text{CrossAtt}_{\text{L}\rightarrow\text{R}}\left(h_i^{k-1}, \{v_1^{k-1}, \ldots, v_m^{k-1}\}\right)$$
$$\hat{v}_j^k = \text{CrossAtt}_{\text{R}\rightarrow\text{L}}\left(v_j^{k-1}, \{h_1^{k-1}, \ldots, h_n^{k-1}\}\right)$$



Cross-Modality Encoder

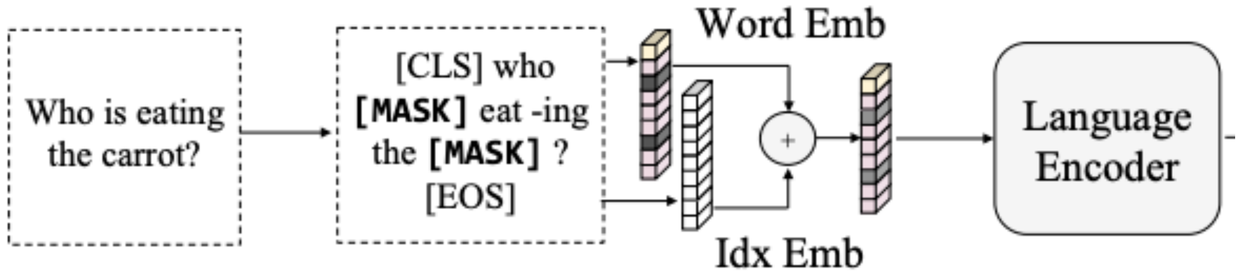
EMNLP 2019 – LXMERT

Encoders : Overall Pipeline



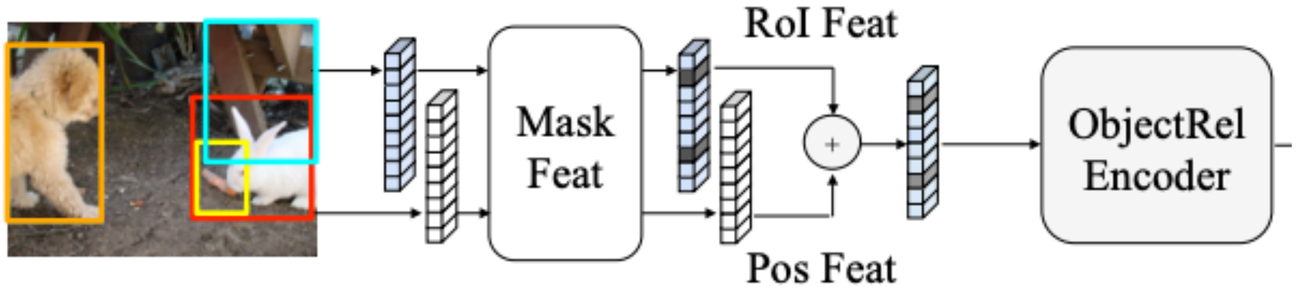
EMNLP 2019 – LXMERT

- Pretraining Tasks : Language Task : Masked Language Model



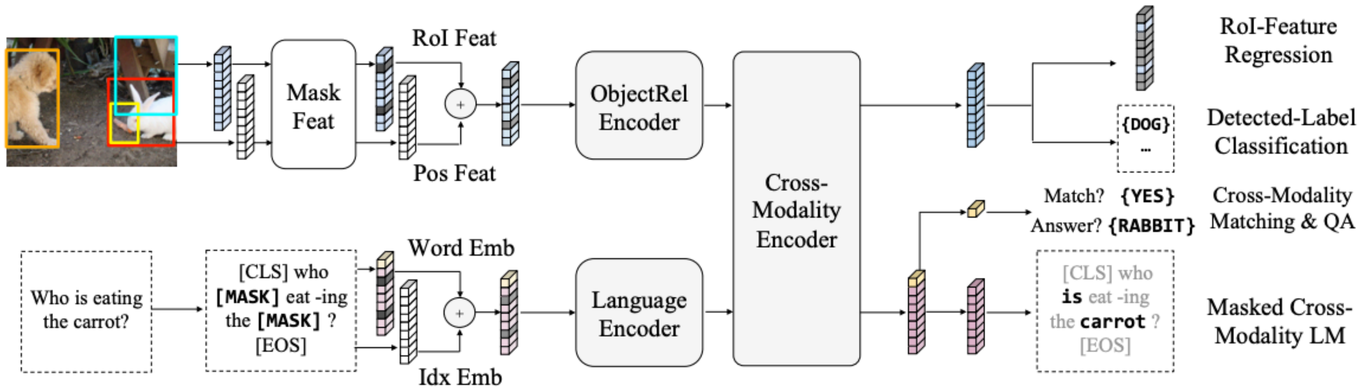
EMNLP 2019 – LXMERT

Pretraining Tasks : Vision Task : Masked Object Prediction



EMNLP 2019 – LXMERT

Pretraining Tasks : Cross Modal Task : Cross Modality Matching and VQA



EMNLP 2019 – LXMERT

Results on Downstream Task

Method	VQA				GQA			NLVR^2	
	Binary	Number	Other	**Accu**	Binary	Open	**Accu**	Cons	**Accu**
Human	-	-	-	-	91.2	87.4	89.3	-	96.3
Image Only	-	-	-	-	36.1	1.74	17.8	7.40	51.9
Language Only	66.8	31.8	27.6	44.3	61.9	22.7	41.1	4.20	51.1
State-of-the-Art	85.8	53.7	60.7	70.4	76.0	40.4	57.1	12.0	53.5
LXMERT	**88.2**	**54.2**	**63.1**	**72.5**	**77.8**	**45.0**	**60.3**	**42.1**	**76.2**

Takeaway

"... spend the summer linking a camera to a computer and getting the computer to describe what it saw."

Marvin Minsky on the goal of a 1966 undergraduate summer research project [1]

[1] Margaret A. Boden. Mind as Machine: A History of Cognitive Science. Oxford University Press, 2008

Transformer