Visio-Linguistic Brain Encoding

August 13, 2022

COLING 2022

Subba Reddy Oota[1,2], Jashn Arora[2], Vijay Rowtula[2], Manish Gupta[2,3], Bapi Raju Surampudi[2]

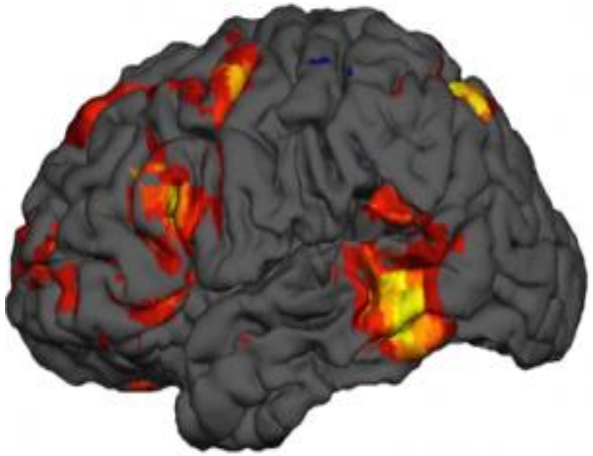
[1]Inria Bordeaux France, [2]IIIT-Hyderabad, [3]Microsoft India

What is fMRI?

(Bird)

Concept + Picture

A vision-language task in the scanner



fMRI Brain Activity

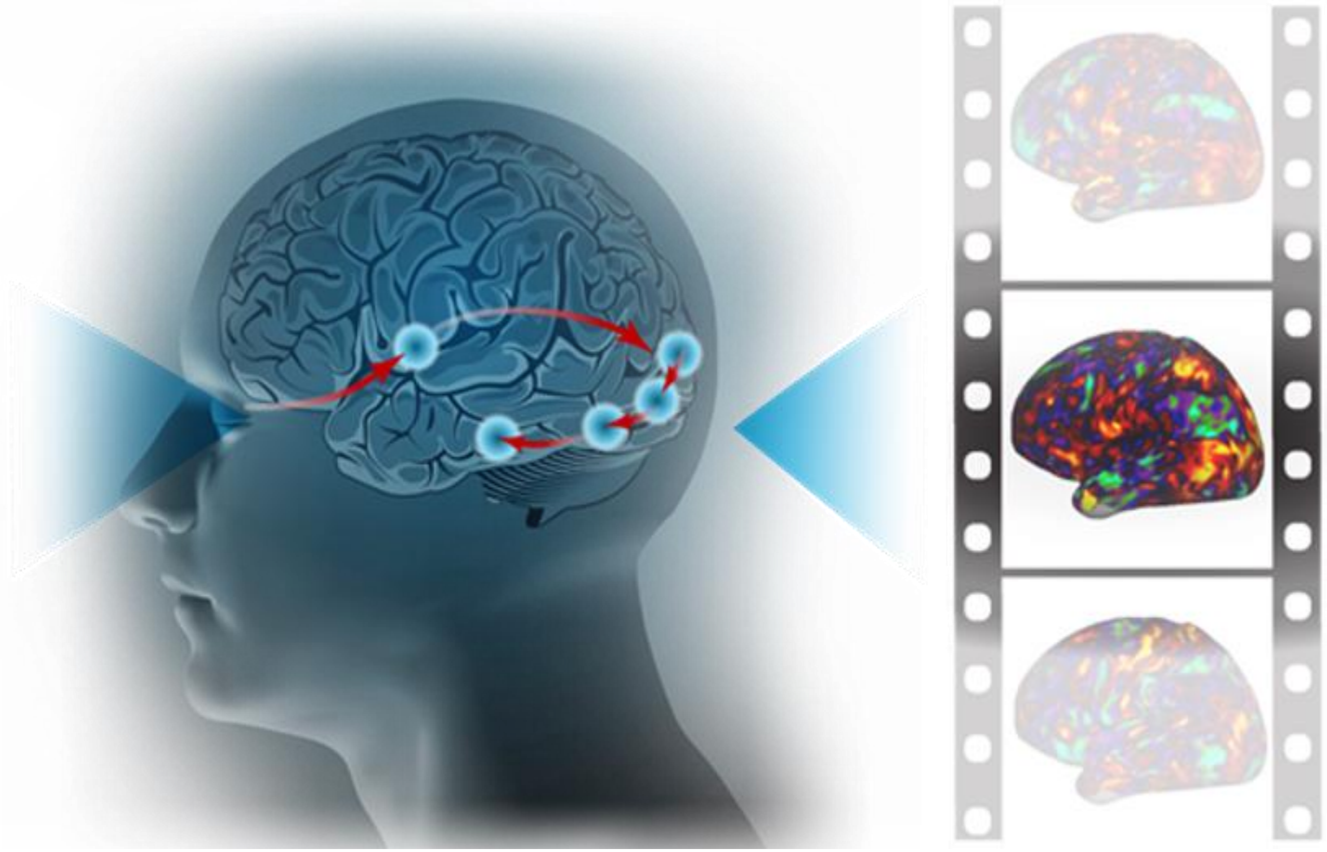
https://www.biopac.com/events/fmri-psych/

Brain Encoding vs Decoding

Stimulus Representation → Encoding → fMRI

(Bird)

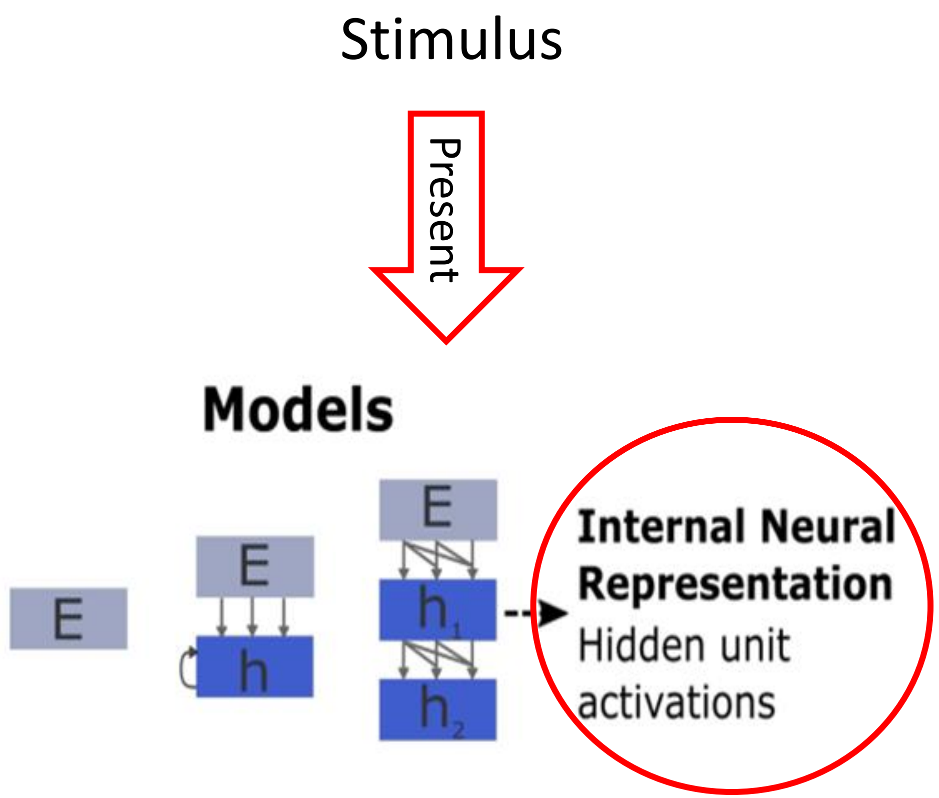
Concept + Picture



Stimulus Representation ← Decoding ← fMRI

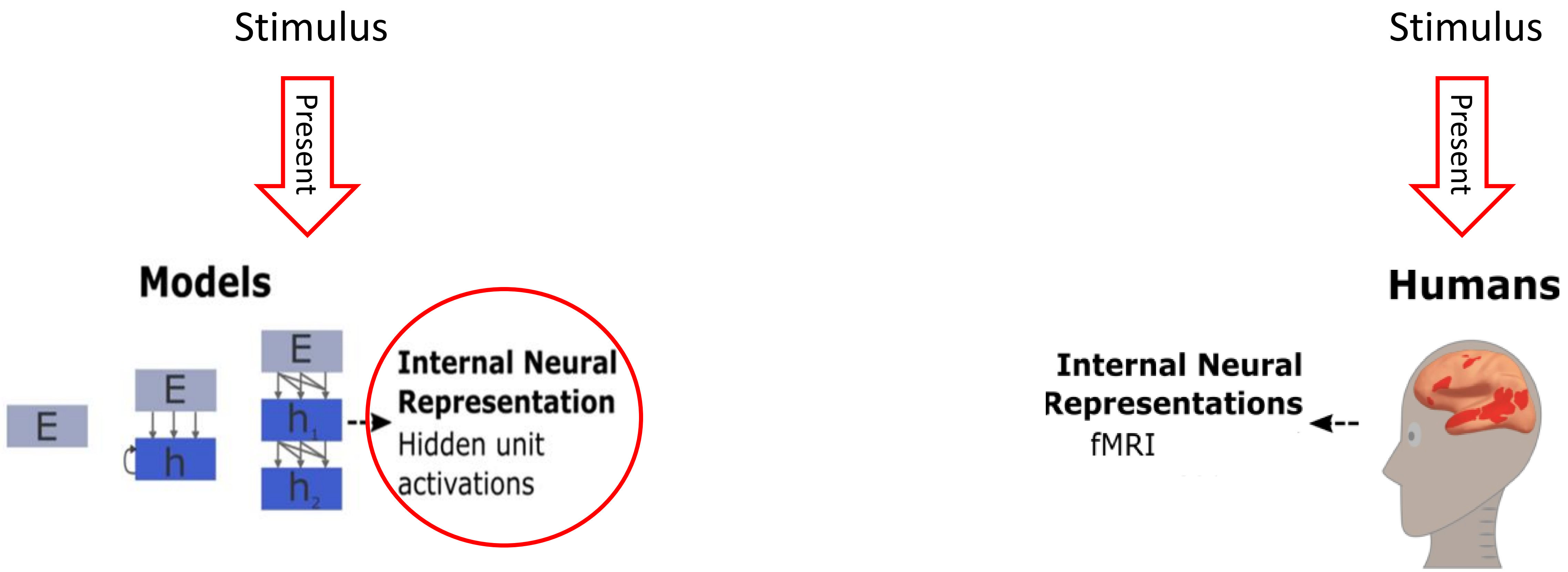
Haiguang Wen et al, 2017

What is Brain Encoding?



Schrimpf et al. 2021 fMRI

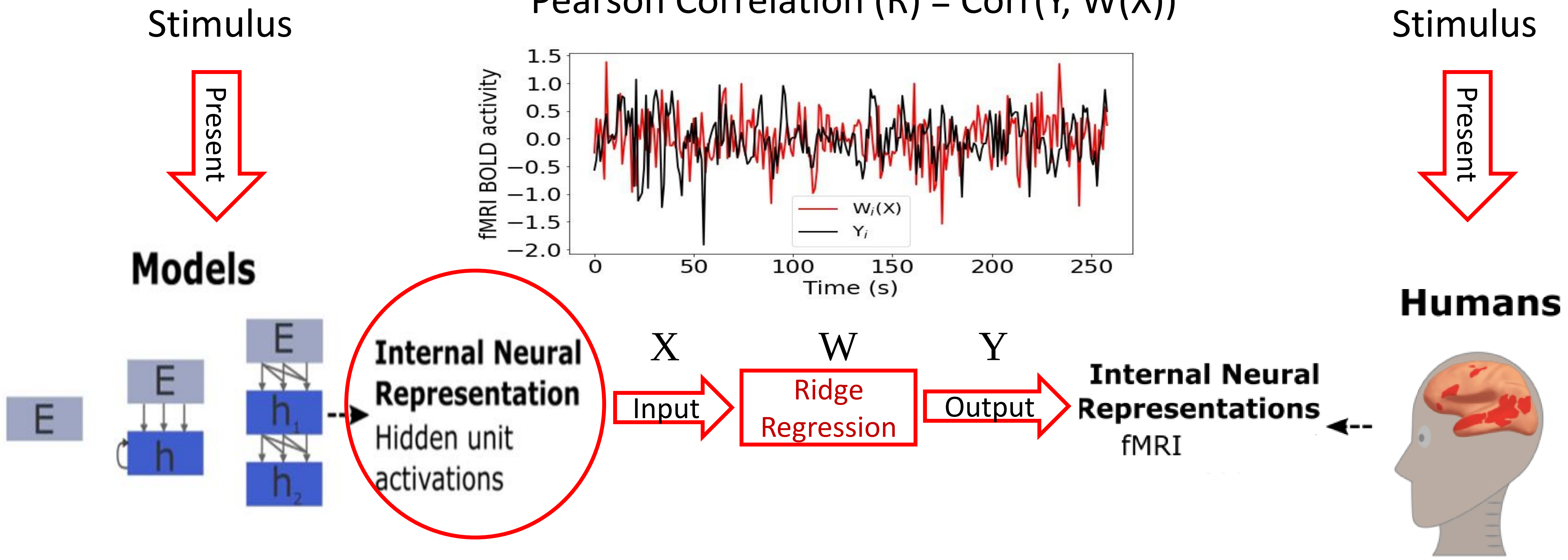
What is Brain Encoding?



Schrimpf et al. 2021 fMRI

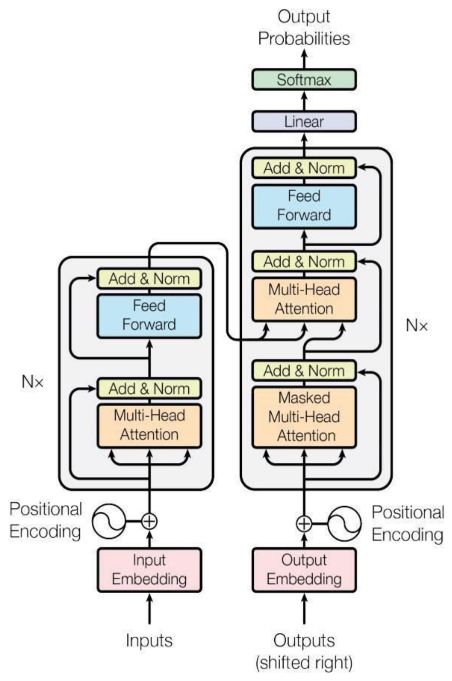
What is Brain Encoding?

Pearson Correlation (R) = Corr(Y, W(X))

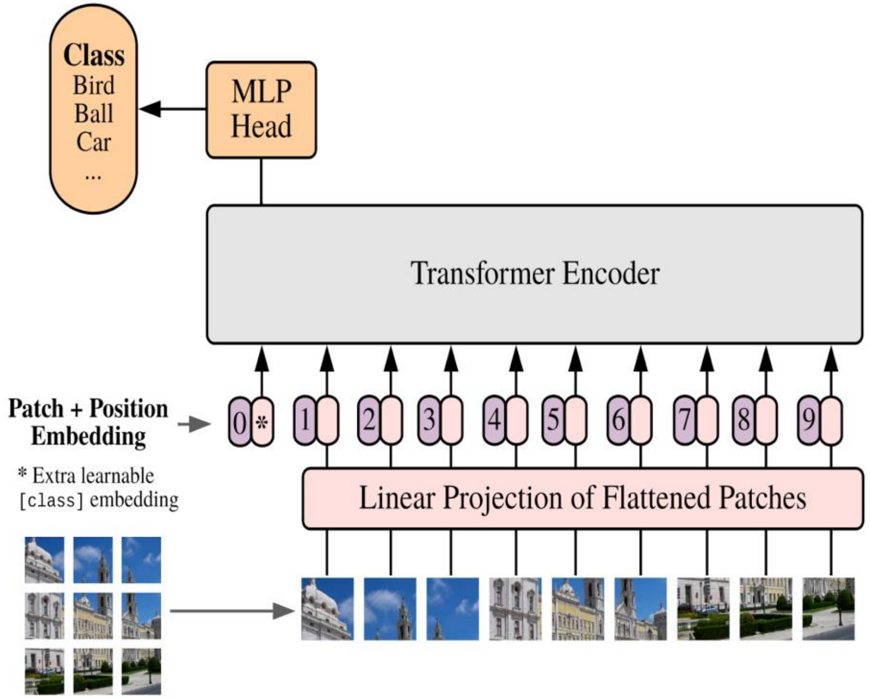


Schrimpf et al. 2021

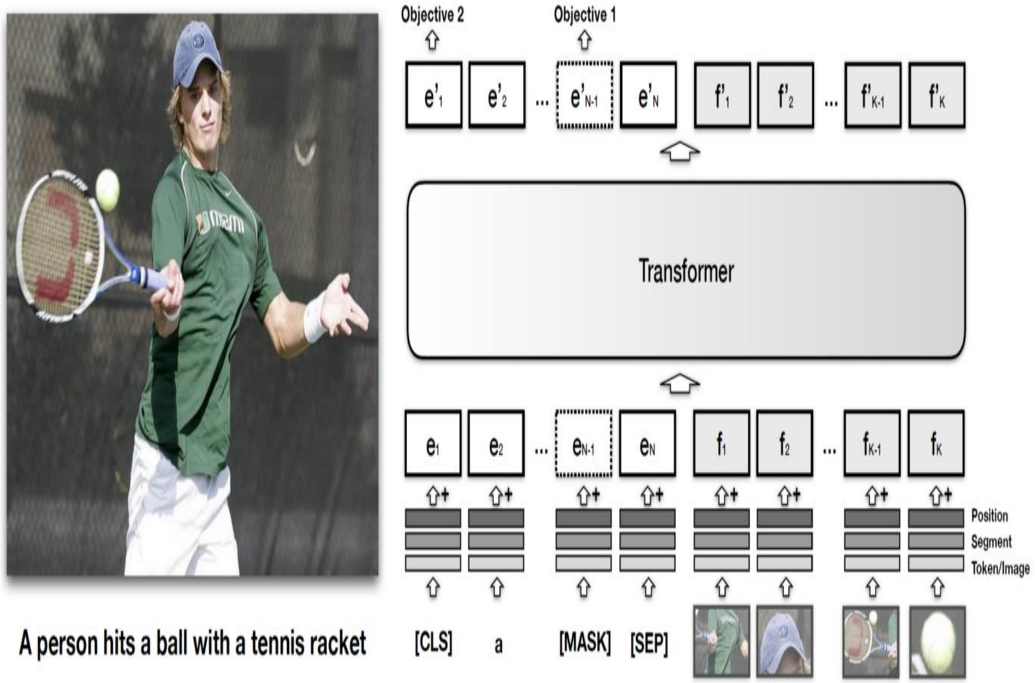
Most popular models are Transformers



Transformer language models



Vision Transformer (ViT)



Multi-modal Transformer

Vaswani et al. 2017, Dosovitskiy et al. 2021, Harold Li et al. 2019

Brain encoding for single-mode stimuli: Vision

Brain-Score: Which Artificial Neural Network for Object Recognition is most Brain-Like?

Martin Schrimpf[*,1,2], Jonas Kubilius[*,3,4], Ha Hong[5], Najib J. Majaj[6], Rishi Rajalingham[1], Elias B. Issa[7], Kohitij Kar[1,3], Pouya Bashivan[1,3], Jonathan Prescott-Roy[1], Kailyn Schmidt[1], Daniel L. K. Yamins[8,9], and James J. DiCarlo[1,2,3]

Integrative Benchmarking to Advance Neurally Mechanistic Models of Human Intelligence

Martin Schrimpf,[1,2,3] Jonas Kubilius,[2,4,5] Michael J. Lee,[1,2] N. Apurva Ratan Murty,[1,2,3] Robert Ajemian,[1,2]
and James J. D
[1]Department of B
[2]McGovern Instit
[3]Center for Brain
[4]Brain and Cogni
[5]Three Thirds, Vil
*Correspondence
https://doi.org/10

SUMMARY

A potentially
intelligence a
experimental
the next step
considered t

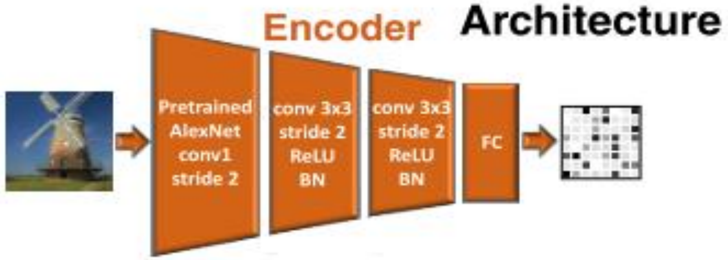
The inte
works (A
ternal n
primate
evolve, a
are most

Neural Taskonomy: Inferring the Similarity of Task-Derived Representations from Brain Activity

Aria Y. Wang, Carnegie Mellon University, ariawang@cmu.edu
Michael J. Tarr, Carnegie Mellon University, michaeltarr@cmu.edu
Leila Wehbe, Carnegie Mellon University, lwehbe@cmu.edu

Abstract

Convolutional neural networks (CNNs) trained for object classification have been widely used to account for visually-driven neural responses in both human and primate brains. However, because of the generality and complexity of object classification, despite the effectiveness of CNNs in predicting brain activity, it is



Shrimpf et al. 2019, Shrimpf et al. 2019, Wang et al. 2019

Brain encoding for single-mode stimuli: Text

The neural architecture of language: Integrative modeling converges on predictive processing

Linking artificial and human neural representations of language

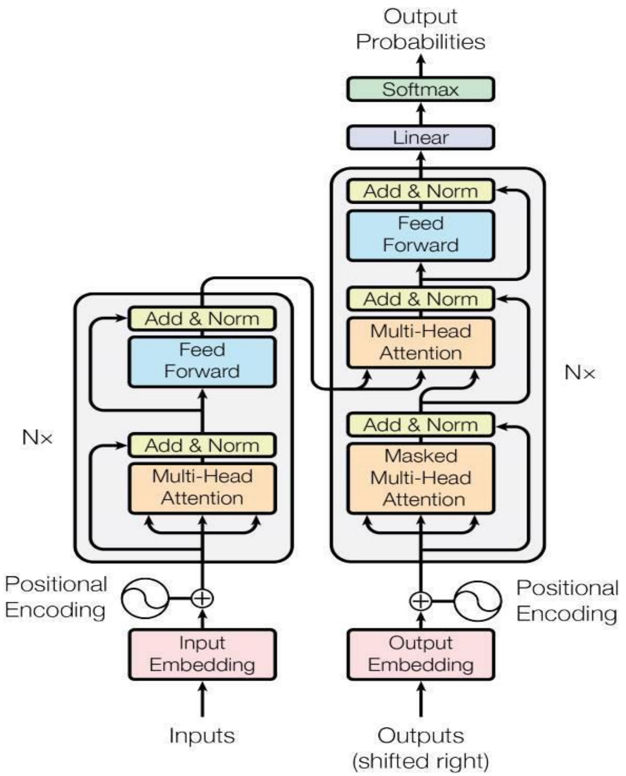
Jon Gauthier and **Roger P. Levy**
Massachusetts Institute of Technology
Department of Brain and Cognitive Sciences
jon@gauthiers.net, rplevy@mit.edu

Abstract

What information from an act of sentence understanding is robustly represented in the human brain? We investigate this question by comparing sentence encoding models on a brain decoding task, where the sentence that an

theories of language understanding, many are specified at too high a level of analysis to plausibly map onto neural structures without serious further revision (Poeppel, 2012).

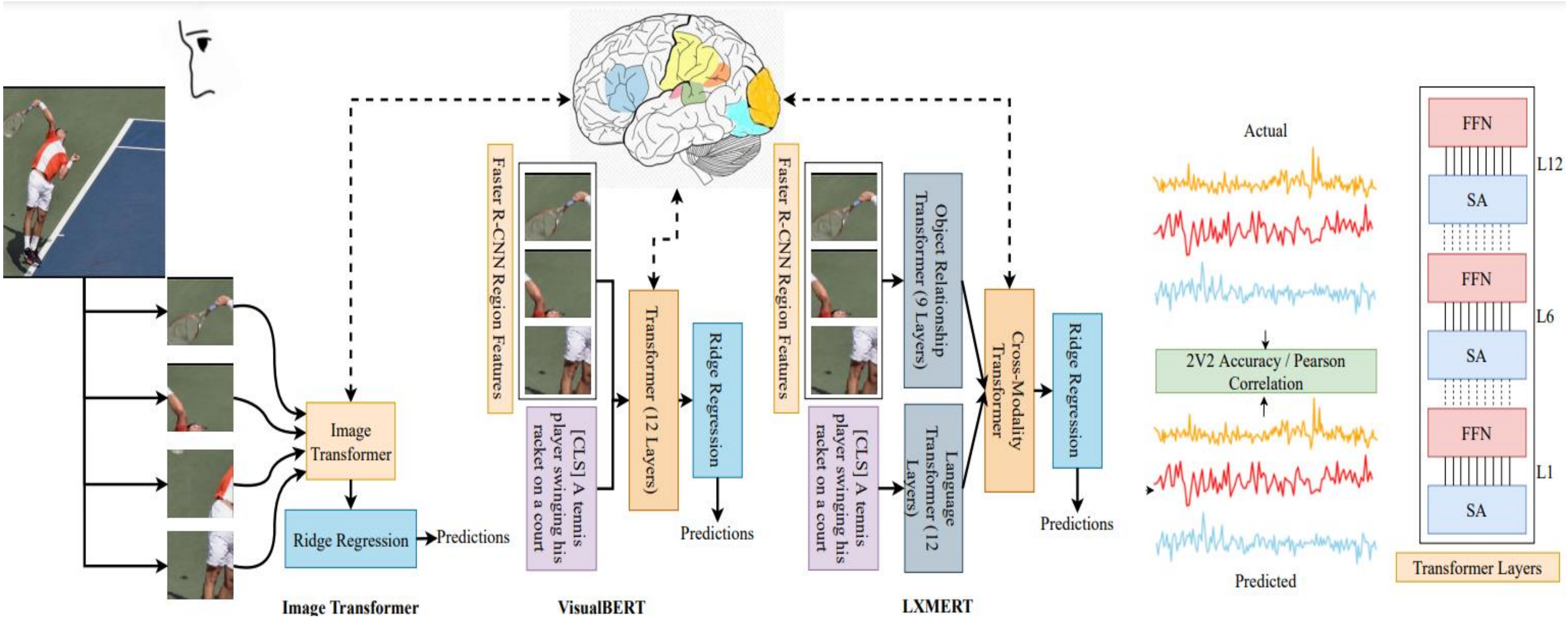
Studies which draw on these high-level representations must therefore also assume some link



Transformer language models (BERT, XLM, GPT,…)

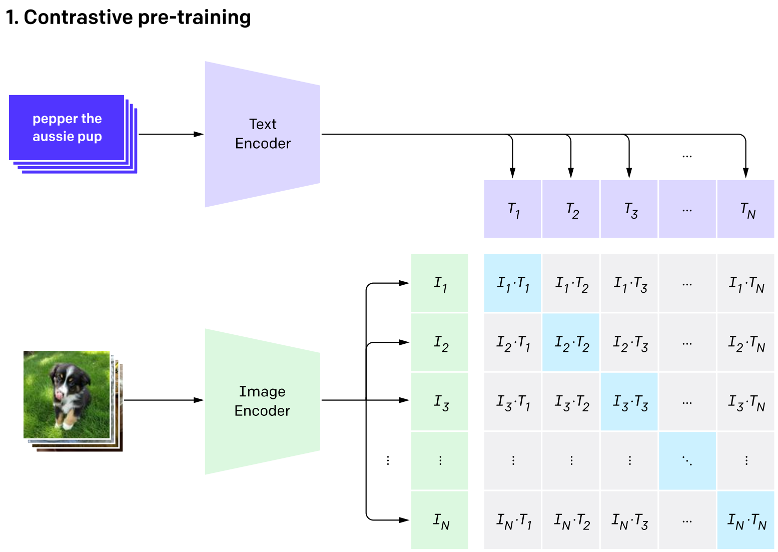
Vaswani et al. 2017, Gauthier et al. 2019, Schrimpf et al. 2021

Can image-based and multi-model Transformers accurately perform fMRI encoding?

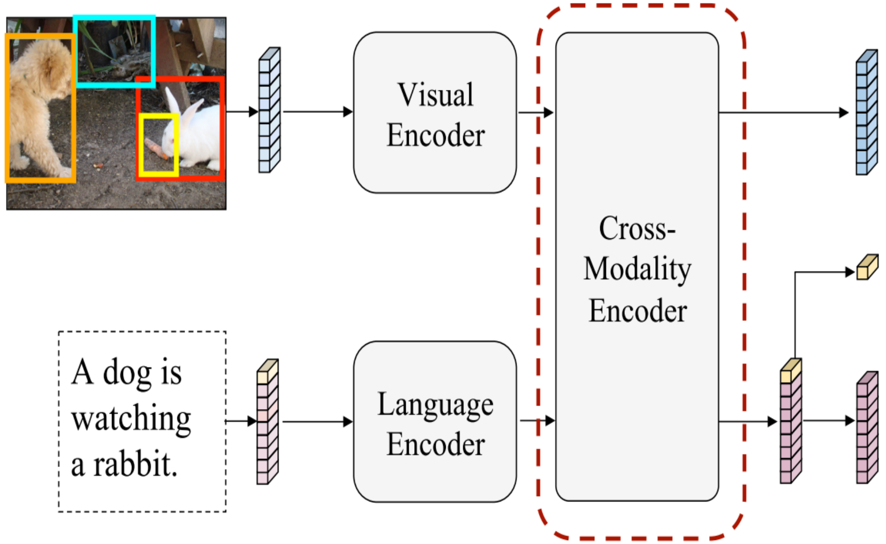


Dosovitskiy et al. 2021, Tan et al. 2019, Harold Li et al. 2019

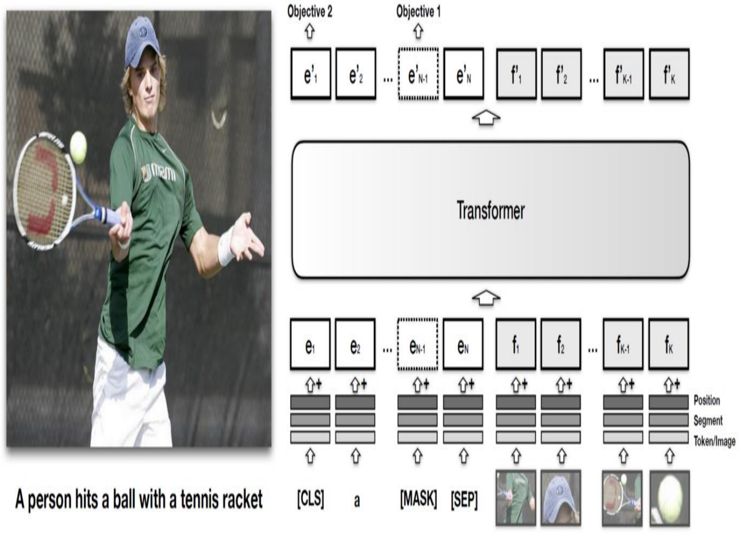
Models used: Multi-Modal Transformers



CLIP



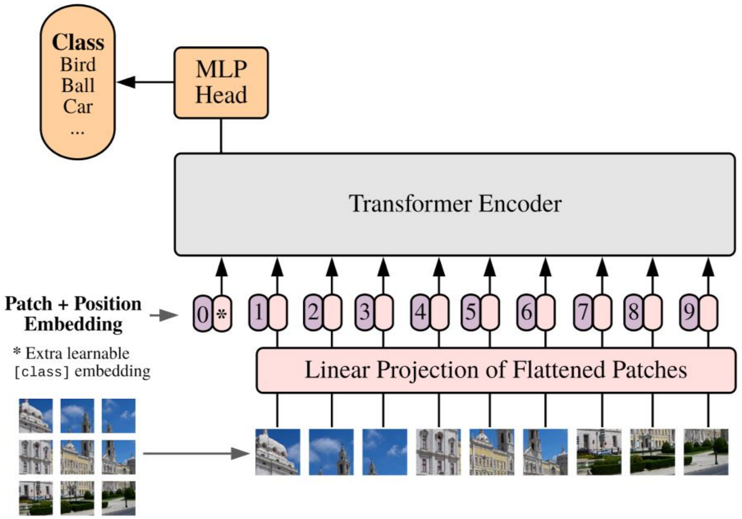
LXMERT



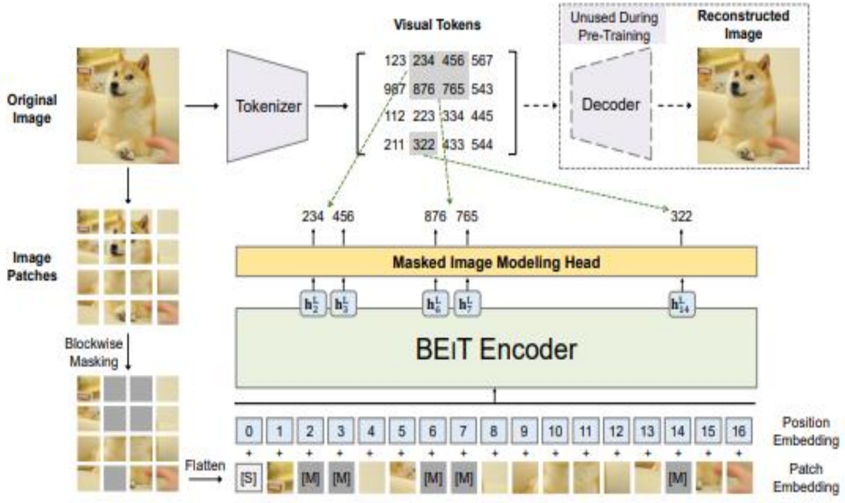
VisualBERT

Radford et al. 2021, Tan et al. 2019, Harold Li et al. 2019

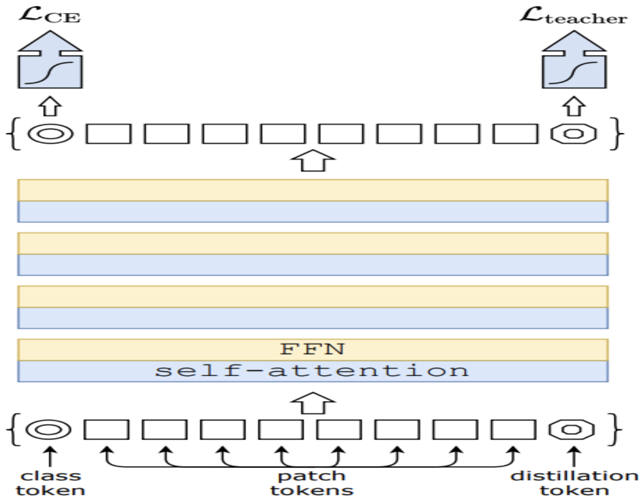
Models used: Image Transformers



ViT



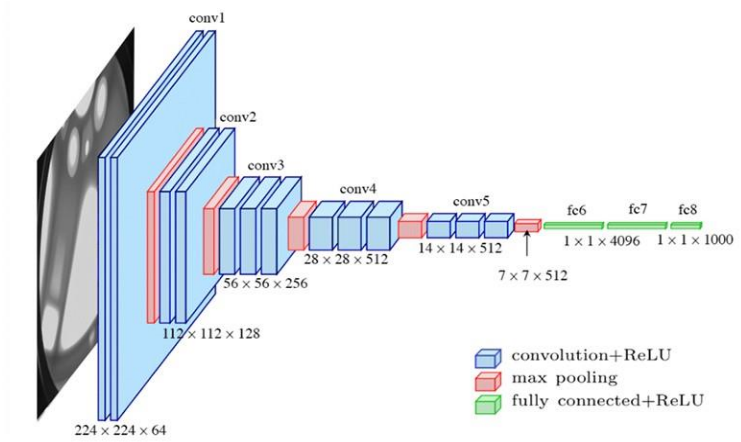
BEiT



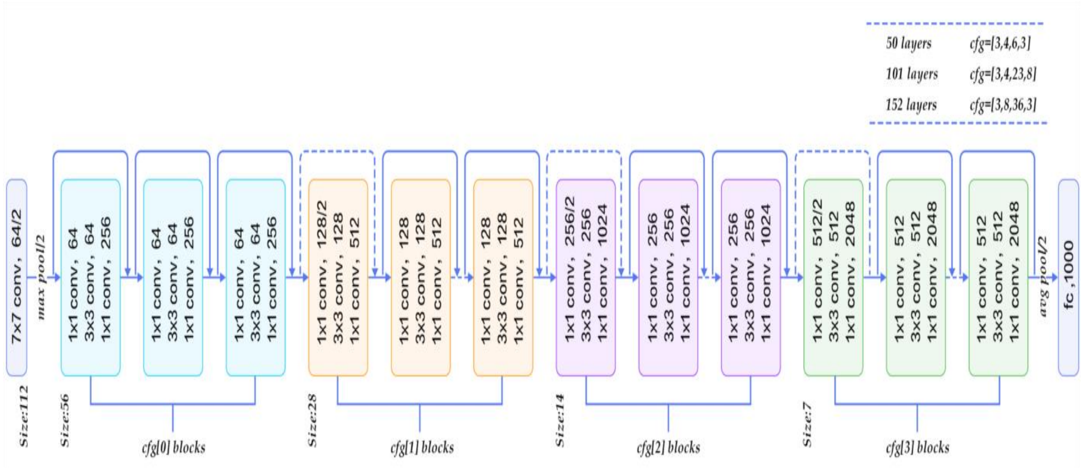
DEiT

Dosovitskiy et al. 2021, Hangbo et al. 2021, Touvron et al. 2021

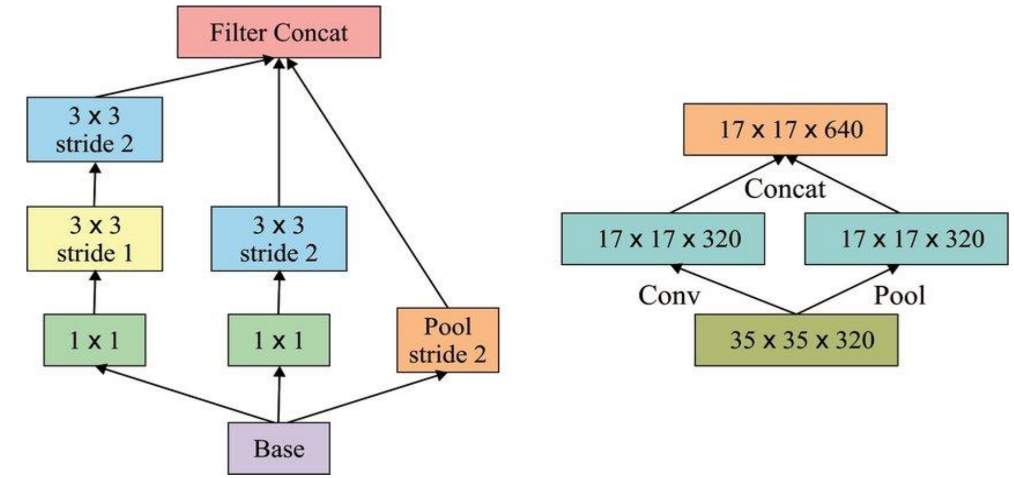
Models used: CNNs



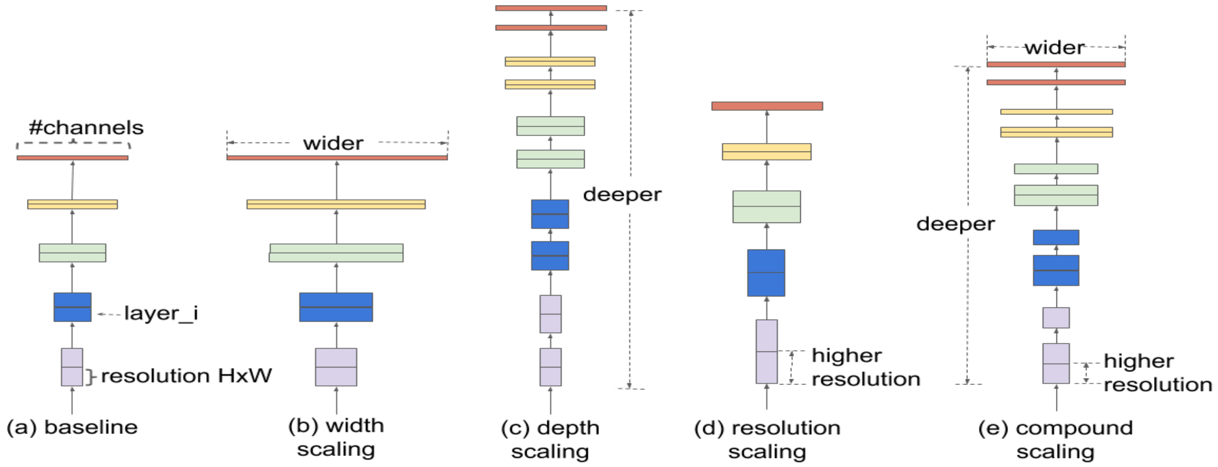
VGGNET



RESNET50



InceptionV2



EfficientNET

Simonyan et al. 2014, He et al. 2016, Szegedy et al. 2017, Tan et 2019

Dataset Details

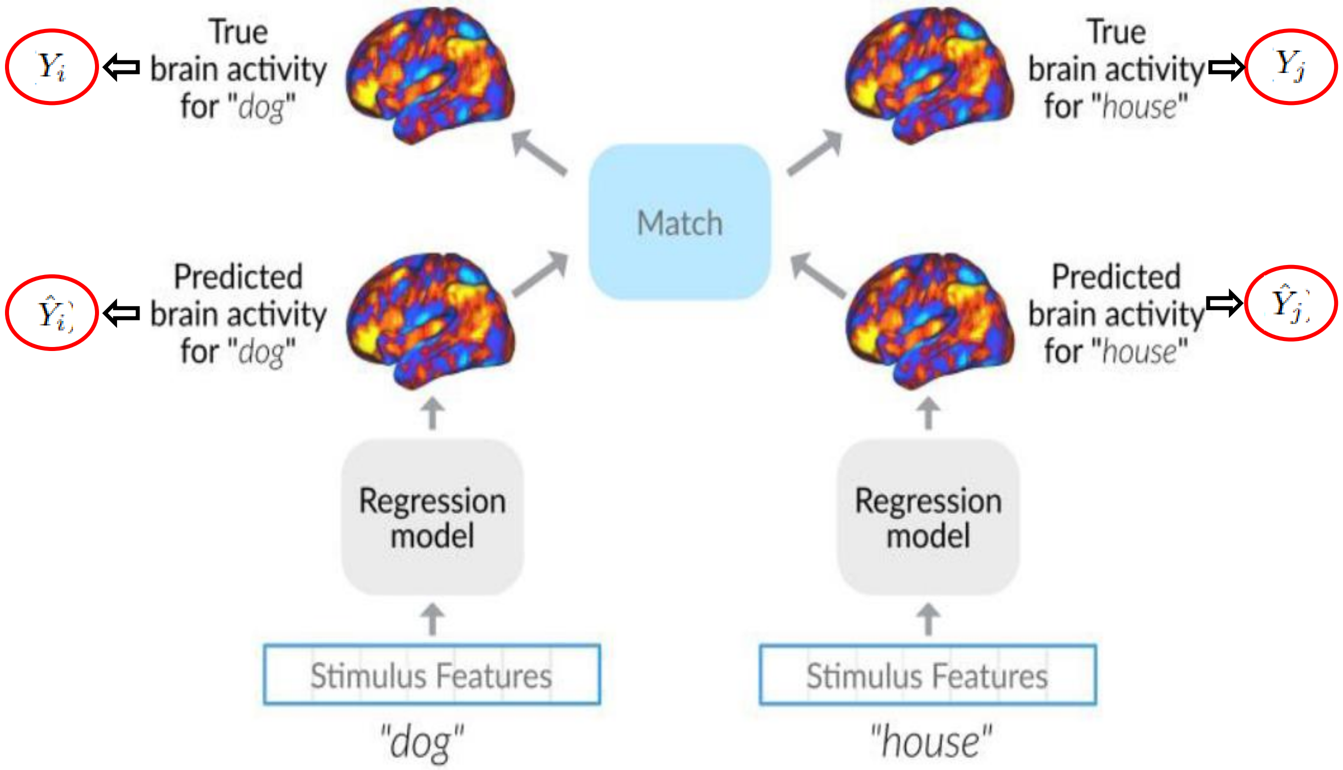
Concept+Picture
(Bird)

Periera

BOLD5000

Periera et al. 2018, Nadine et al. 2019

Evaluation Metrics: 2V2 and Pearson



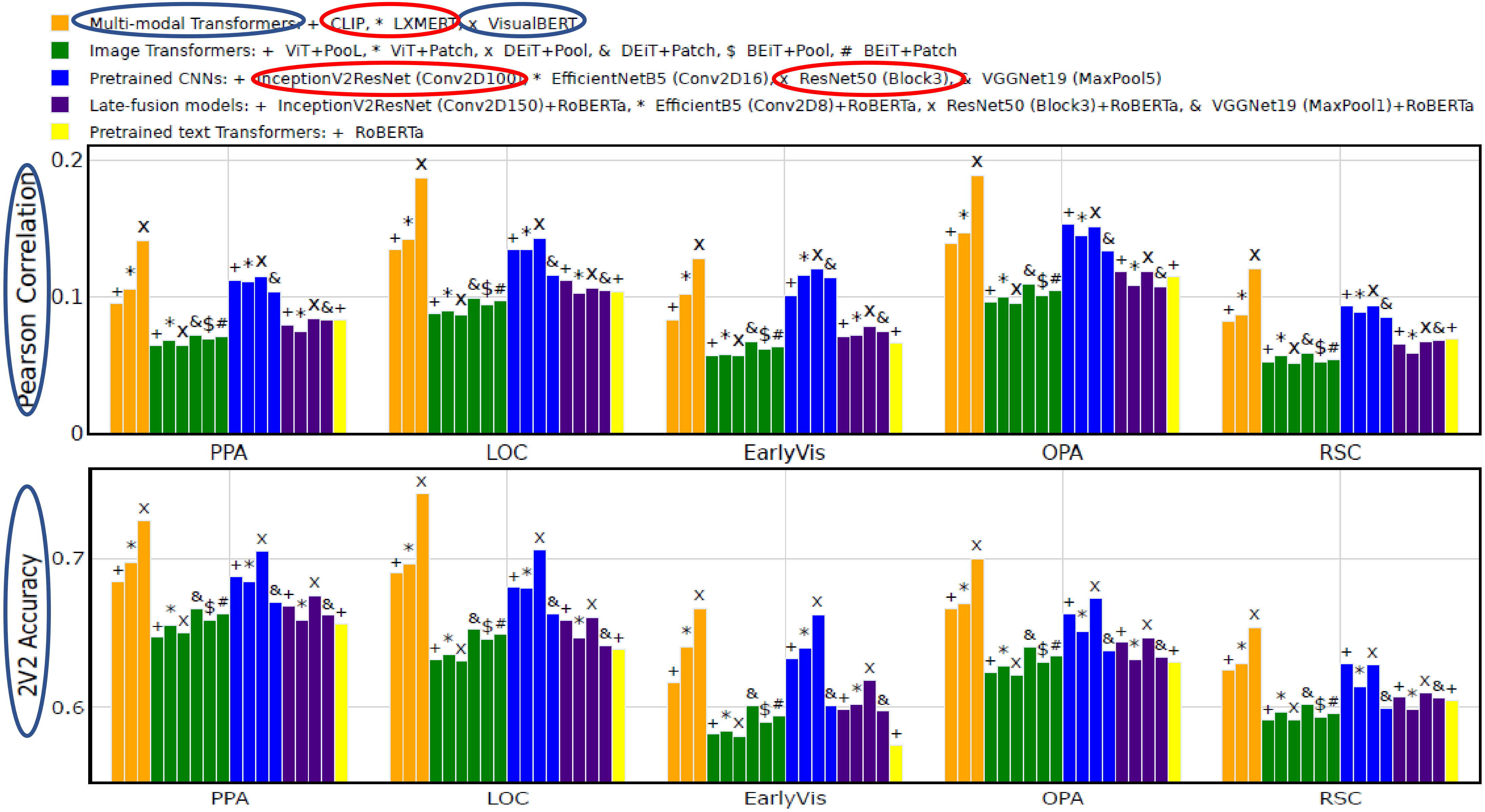
2V2 Accuracy

$$\text{2V2 Accuracy} = \frac{1}{N_{C_2}} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} I[\{cosD(Y_i, \hat{Y}_i) + cosD(Y_j, \hat{Y}_j)\} < \{cosD(Y_i, \hat{Y}_j) + cosD(Y_j, \hat{Y}_i)\}]$$

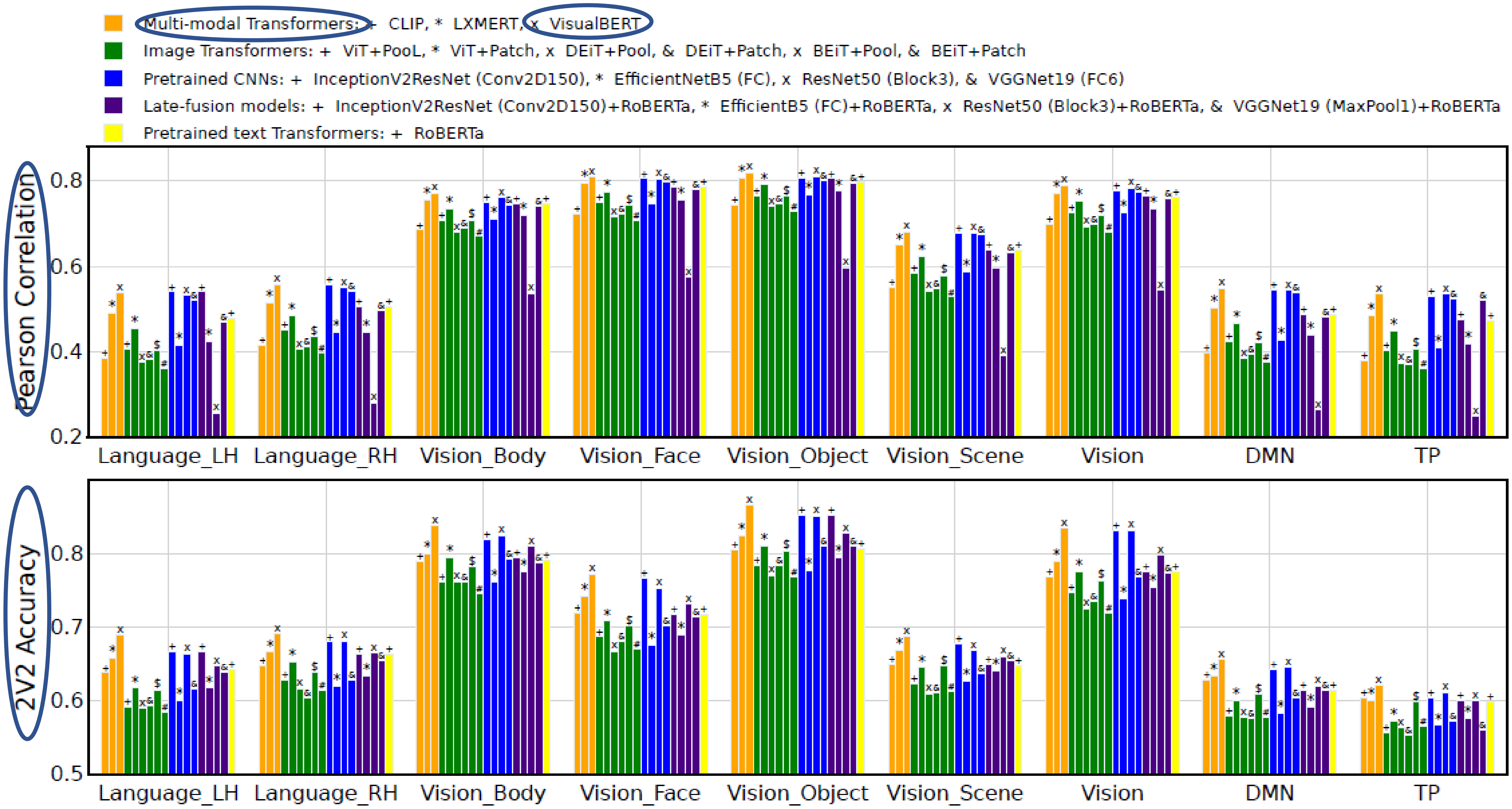
Cosine distance

Toneva et al. 2020

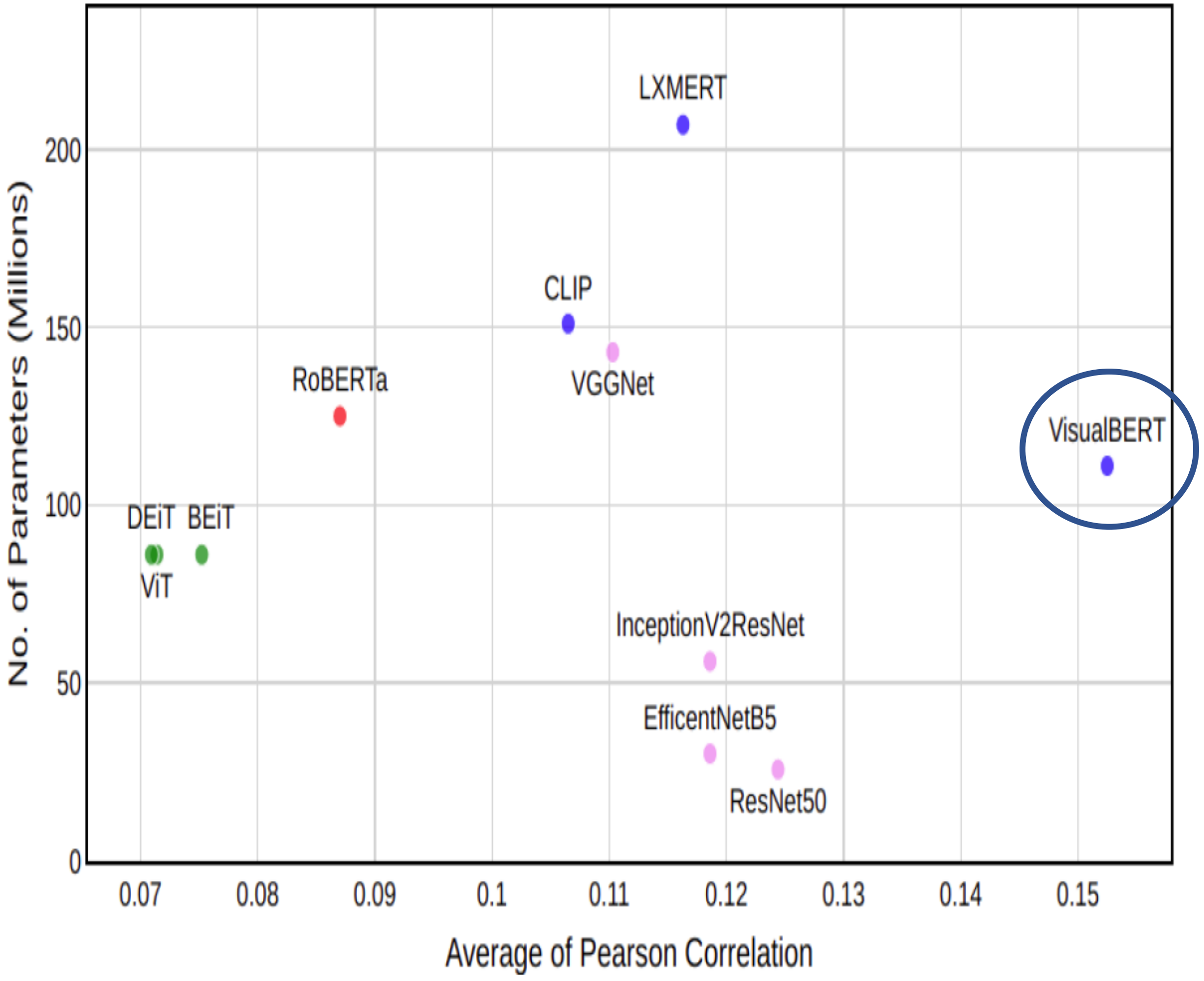
Encoding performance (BOLD5000)



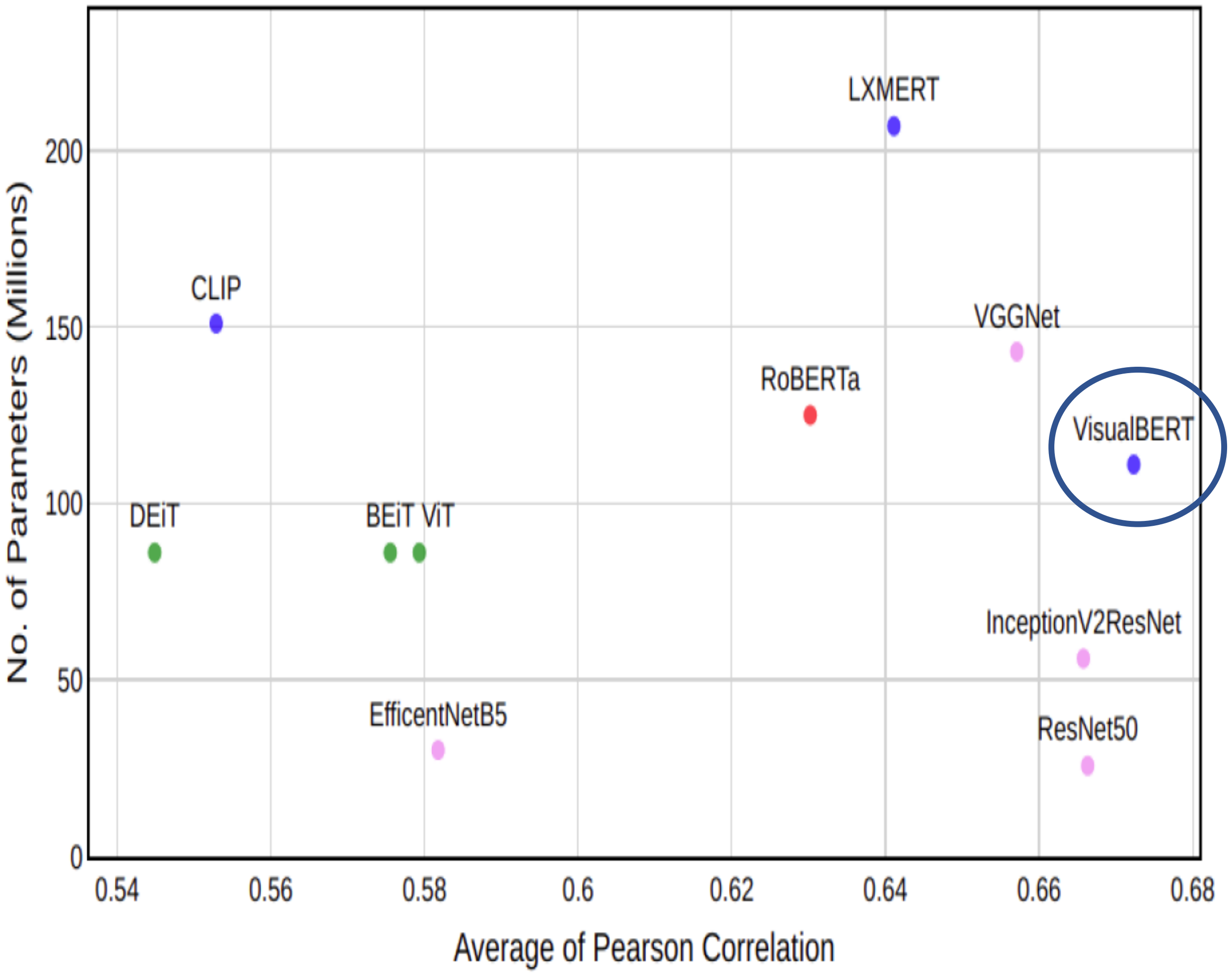
Encoding performance (Periera)



Model size vs Efficacy



BOLD5000



Periera

Single Stream vs Dual Stream

Dual Stream

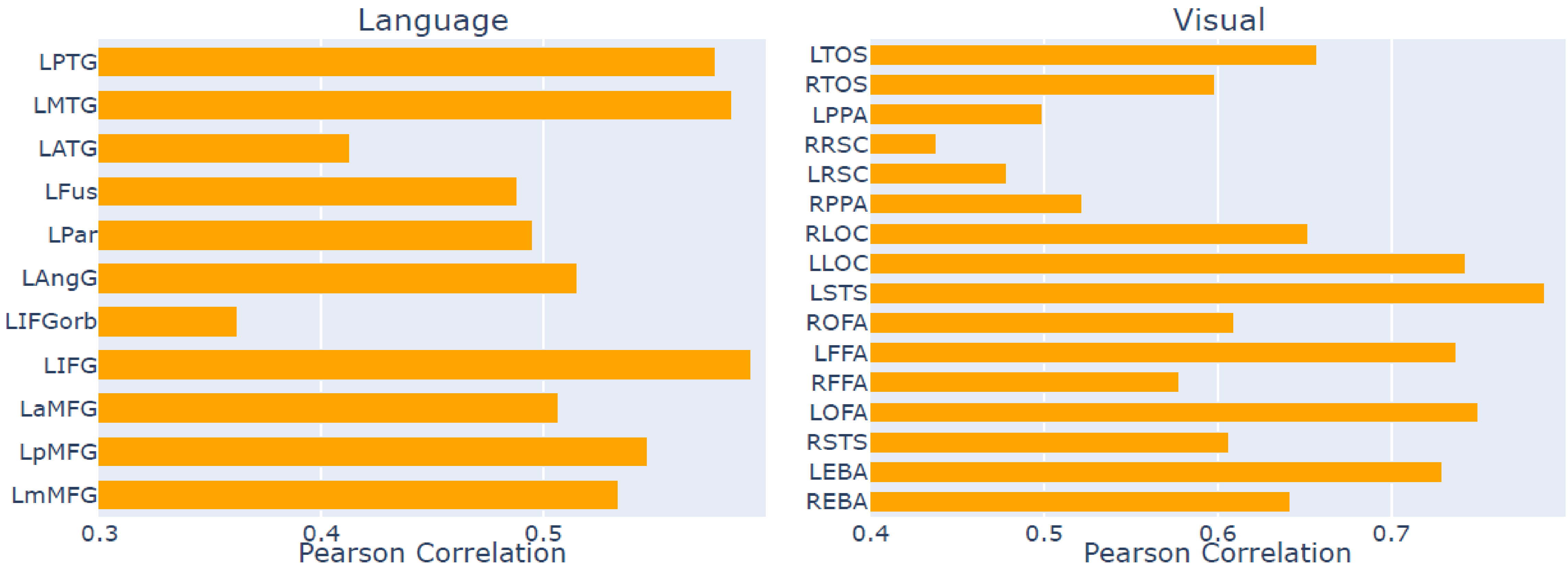
Models compared	PPA	LOC	EarlyVis	OPA	RSC
CLIP	0.095	0.134	0.083	0.139	0.082
LXMERT	0.106	0.142	0.102	0.146	0.087
VisualBERT	**0.141**	**0.187**	**0.128**	**0.188**	**0.12**
ViLBERT	0.057	0.078	0.052	0.087	0.045

Single Stream

Is Linguistic Information Important in Multi-Modal Transformers?

	Models compared	PPA	LOC	EarlyVis	OPA	RSC
Correct Image-Text pairs	CLIP	0.095	0.134	0.083	0.139	0.082
	LXMERT	0.106	0.142	0.102	0.146	0.087
	VisualBERT	**0.141**	**0.187**	**0.128**	**0.188**	**0.12**
	ViLBERT	0.057	0.078	0.052	0.087	0.045
Randomize Image-Text pairs	CLIP-Random	0.020	0.024	0.033	0.031	0.002
	LXMERT-Random	0.035	0.041	0.035	0.049	0.029
	VisualBERT-Random	0.072	0.102	0.062	0.109	0.060
	ViLBERT-Random	0.018	0.011	0.013	0.017	0.017

Does Language Influence Vision?

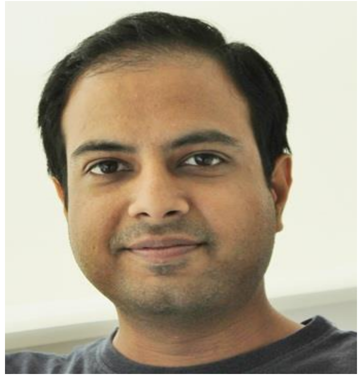


Collaborators

Subba Reddy Oota

Jashn Arora

Vijay Rowtula



Manish Gupta

Bapi Raju Surampudi