



Intermediate Deep Learning for Engineers

Spring 2025, Deep Learning for Engineers
April 15, 2025, 10th Session

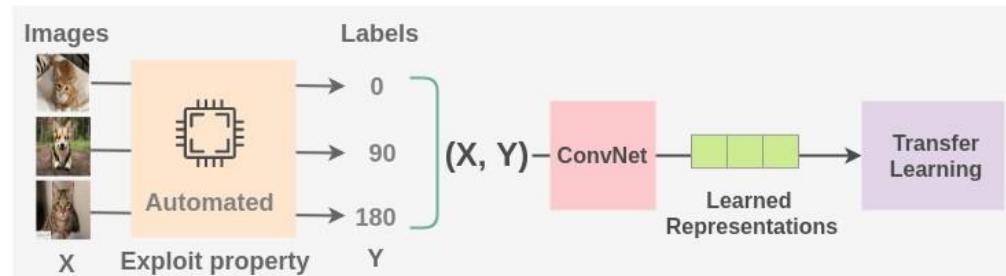
Amir Barati Farimani

*Associate Professor of Mechanical Engineering and Bio-Engineering
Carnegie Mellon University*

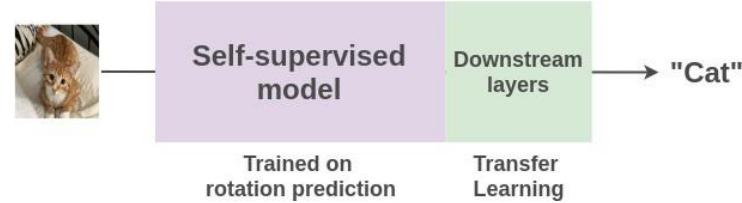
Self-Supervised Learning (Recap)

Self-supervised learning example

Pretext task: train a model to predict the rotation degree of rotated images with cats and dogs (we can collect million of images from internet, labeling is not required)

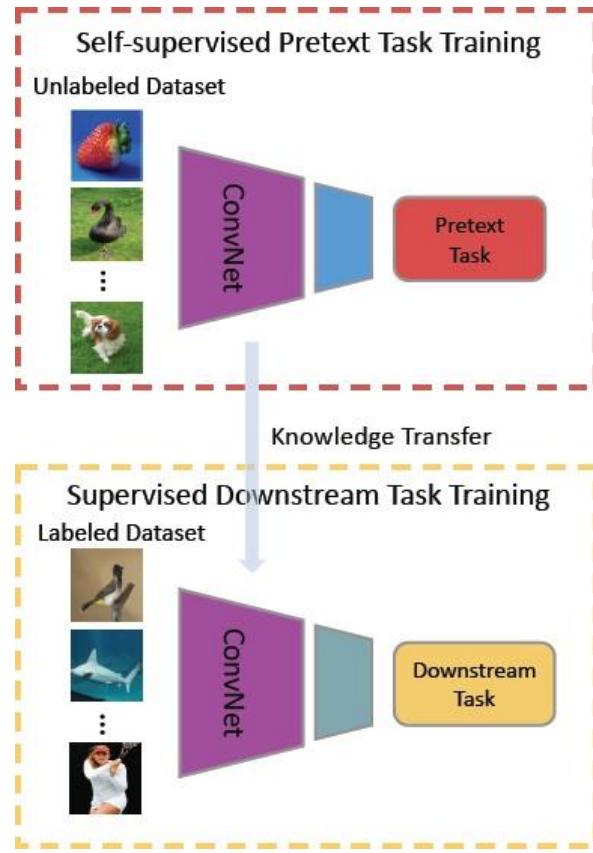


Downstream task: use transfer learning and fine-tune the learned model from the pretext task for **classification** of cats vs dogs with very few labeled examples



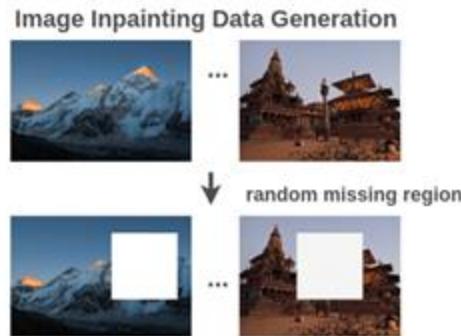
Self-Supervised Learning (Recap)

One more depiction of the general pipeline for self-supervised learning is shown in the figure

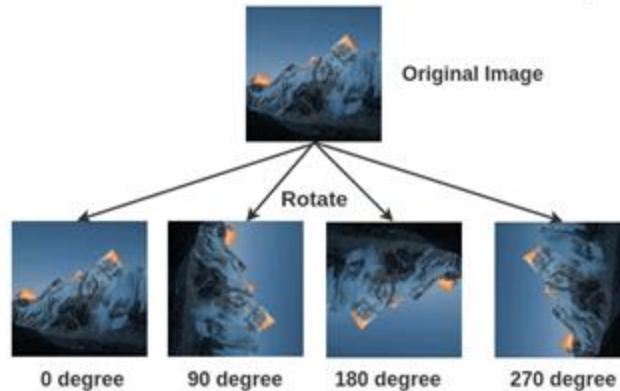


For the downstream task, reuse the trained ConvNet base model, and fine-tune the top layers on a small labeled dataset

SSL for Vision



Data Generation for Geometric Transformation Recognition



SimCLR

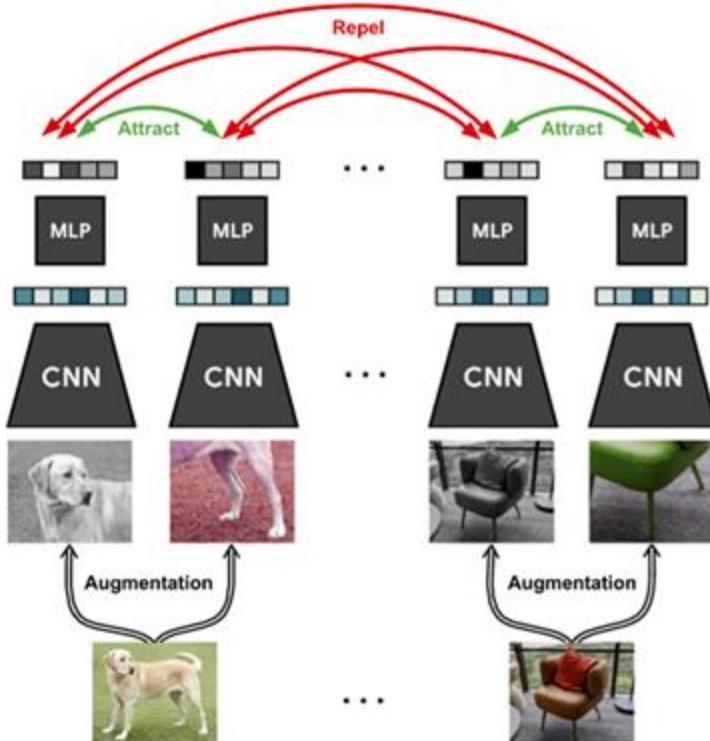


Figure taken from <https://sh-tsang.medium.com/review-simclr-a-simple-framework-for-contrastive-learning-of-visual-representations-5de42ba0bc66>

SimCLR (cont'd)

Augmented Images in Batch



Pair 1

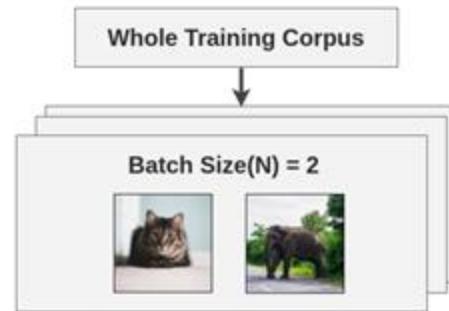


Pair 2

$$L_{NCE} = - \sum_{i=1}^n \log \frac{e^{\theta(t_1^i, t_2^i)}}{\frac{1}{b} \sum_{j=1}^b e^{\theta(t_1^i, t_2^j)}}$$

.

SimCLR (cont'd)



Augmented Images in Batch



$$L_{NCE} = - \sum_{i=1}^n \log \frac{e^{\theta(t_1^i, t_2^i)}}{\frac{1}{b} \sum_{j=1}^b e^{\theta(t_1^i, t_2^j)}}$$

$$I(\begin{array}{|c|c|}\hline \text{cat} & \text{elephant} \\\hline\end{array}, \begin{array}{|c|c|}\hline \text{cat} & \text{elephant} \\\hline\end{array}) = -\log \left(\frac{e^{\text{similarity}(\begin{array}{|c|c|}\hline \text{cat} & \text{cat} \\\hline\end{array})}}{\text{similarity}(\begin{array}{|c|c|}\hline \text{cat} & \text{cat} \\\hline\end{array}) + \frac{e^{\text{similarity}(\begin{array}{|c|c|}\hline \text{cat} & \text{elephant} \\\hline\end{array})}}{e^{\text{similarity}(\begin{array}{|c|c|}\hline \text{cat} & \text{elephant} \\\hline\end{array})}} + \text{similarity}(\begin{array}{|c|c|}\hline \text{elephant} & \text{elephant} \\\hline\end{array})} \right)$$

SimCLR (cont'd)

Augmented Images in Batch



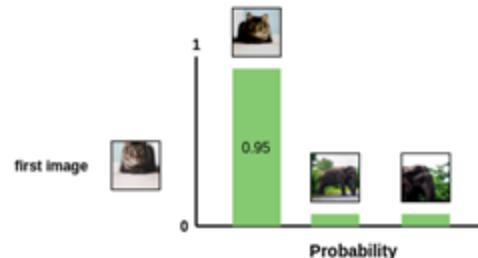
Pair 1



Pair 2

$$L_{NCE} = - \sum_{i=1}^n \log \frac{e^{\theta(t_1^i, t_2^i)}}{\frac{1}{b} \sum_{j=1}^b e^{\theta(t_1^i, t_2^j)}}$$

$$I(\text{[tiger]}, \text{[tiger]}) = -\log \left(\frac{e^{\text{similarity}(\text{[tiger]}, \text{[tiger]})}}{e^{\text{similarity}(\text{[tiger]}, \text{[elephant]})} + e^{\text{similarity}(\text{[tiger]}, \text{[elephant]})} + e^{\text{similarity}(\text{[tiger]}, \text{[elephant]})}} \right)$$



Figures taken from
<https://amitness.com/2020/03/illustrated-simclr/>

The importance of Augmentations in SimCLR

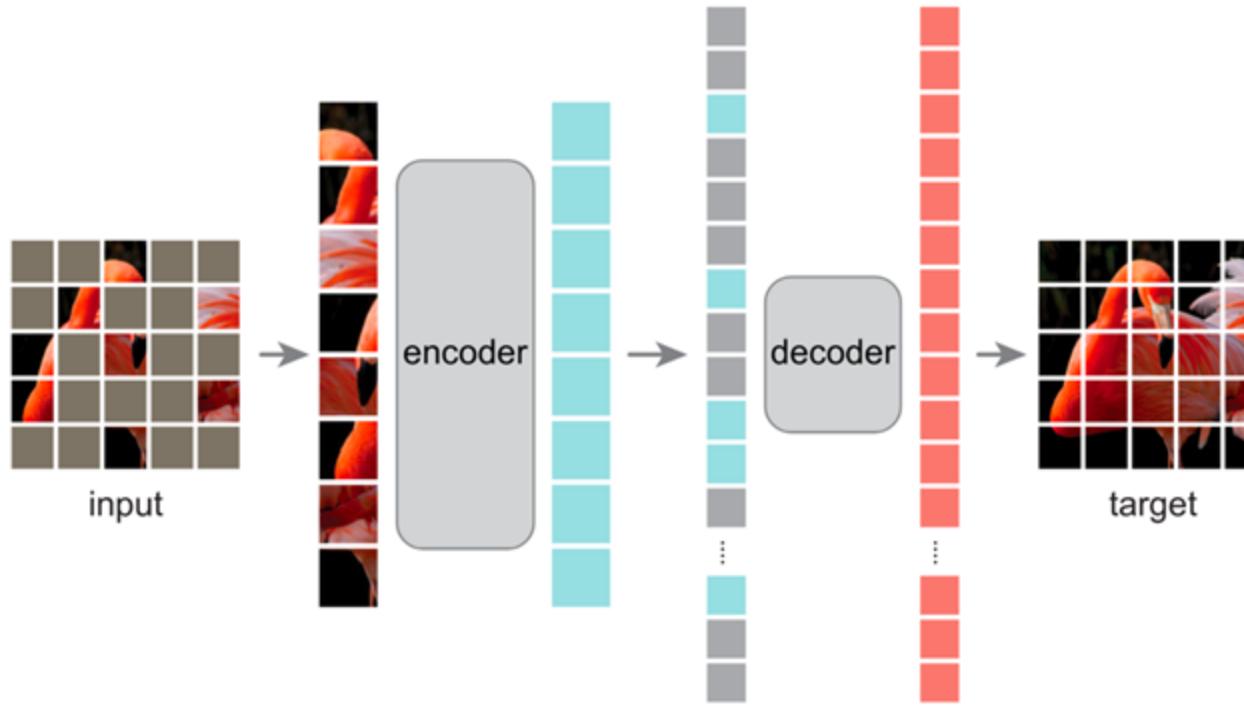
WHAT AUGMENTATIONS TO USE DEPENDS ON THE DOWNSTREAM TASK, AND THE DATASET ITSELF.

LET'S SAY YOU WANT TO CLASSIFY A DATA OF APPLES OF DIFFERENT COLORS.

WOULD YOU USE COLOR AUGMENTATION IN THIS TASK?



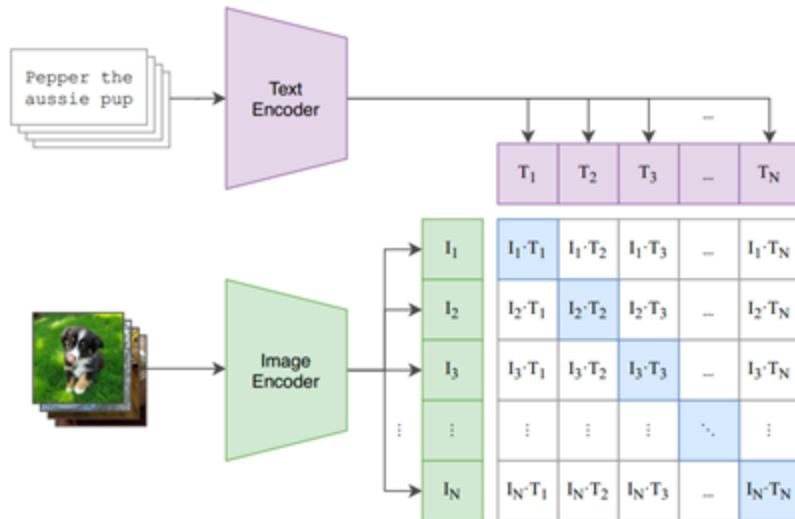
Masked Autoencoders



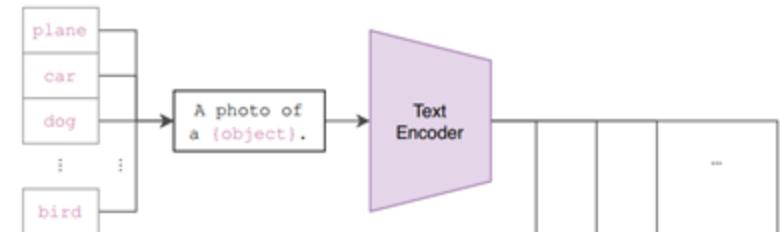
He, Kaiming, et al. "Masked autoencoders are scalable vision learners." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022.

CLIP

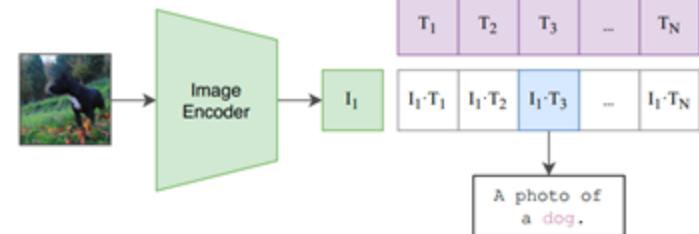
(1) Contrastive pre-training



(2) Create dataset classifier from label text

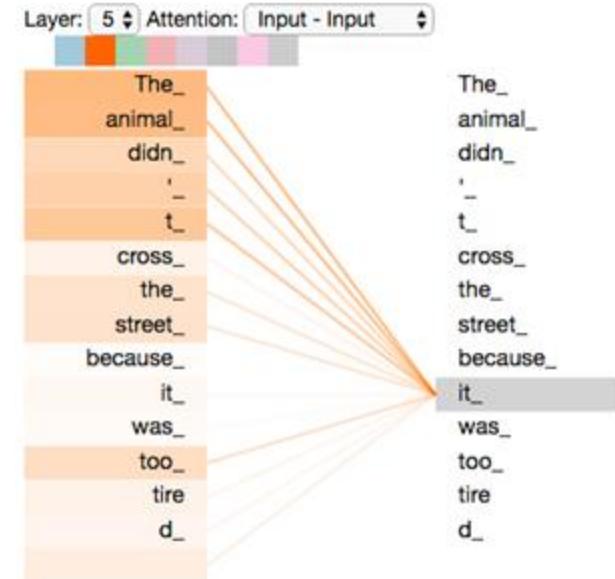
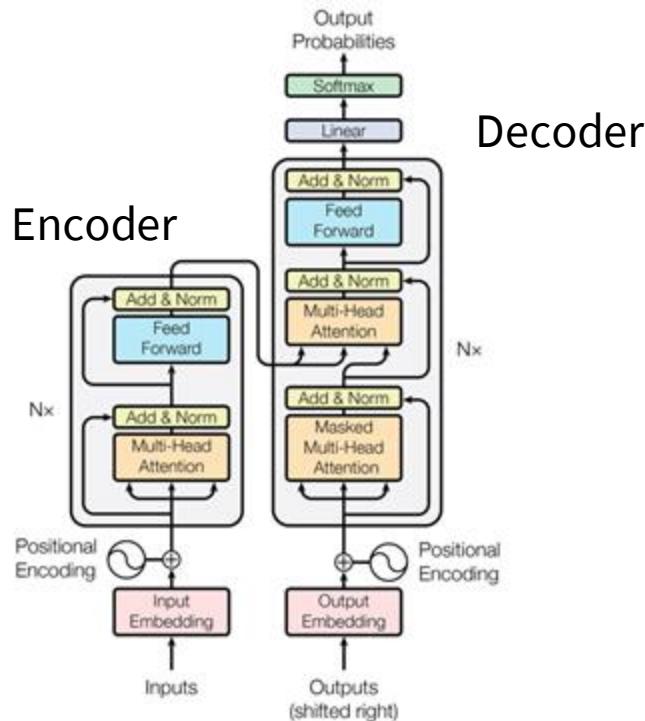


(3) Use for zero-shot prediction

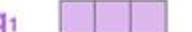
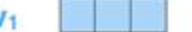


Radford, Alec, et al. "Learning transferable visual models from natural language supervision." *International conference on machine learning*. PMLR, 2021.

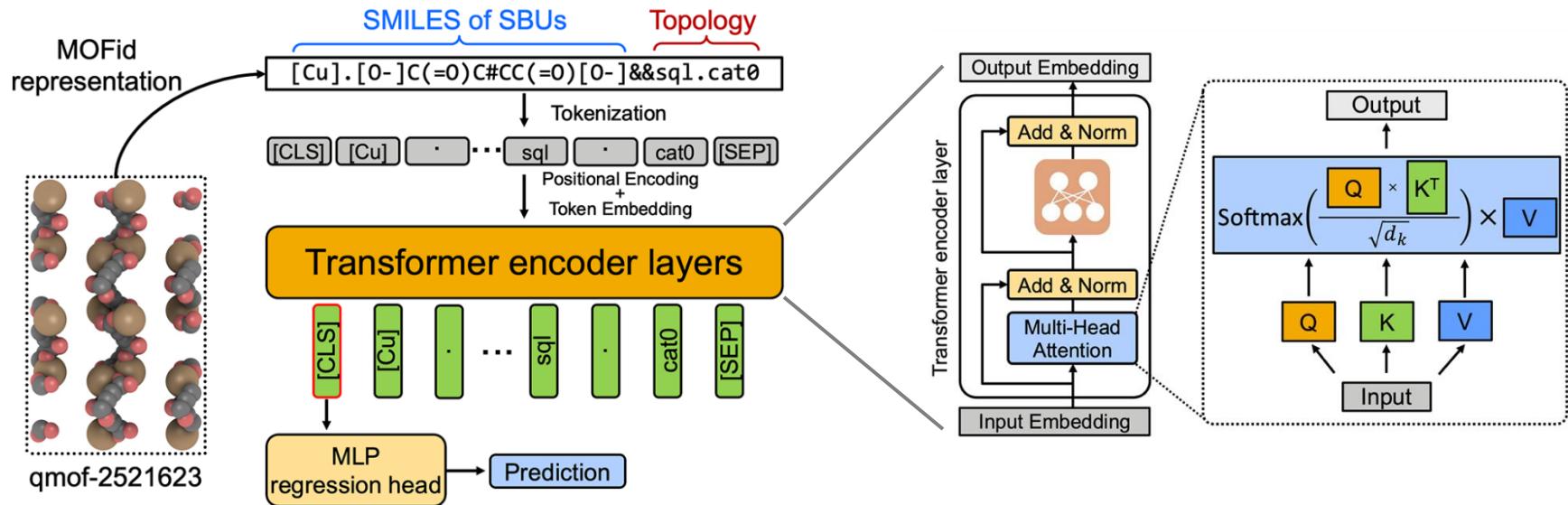
Transformer (Recap)



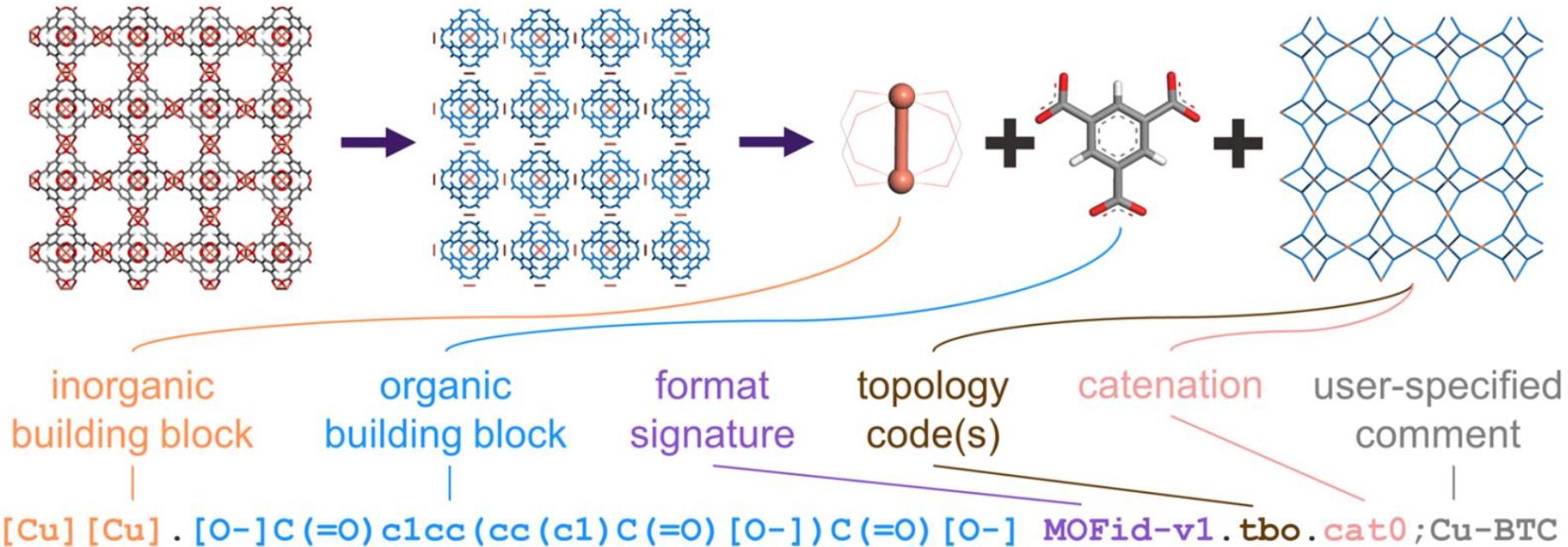
Transformer (Recap)

Input		
Embedding	x_1	
Queries	q_1	
Keys	k_1	
Values	v_1	
Score	$q_1 \cdot k_1 = 112$	$q_1 \cdot k_2 = 96$
Divide by 8 ($\sqrt{d_k}$)	14	12
Softmax	0.88	0.12

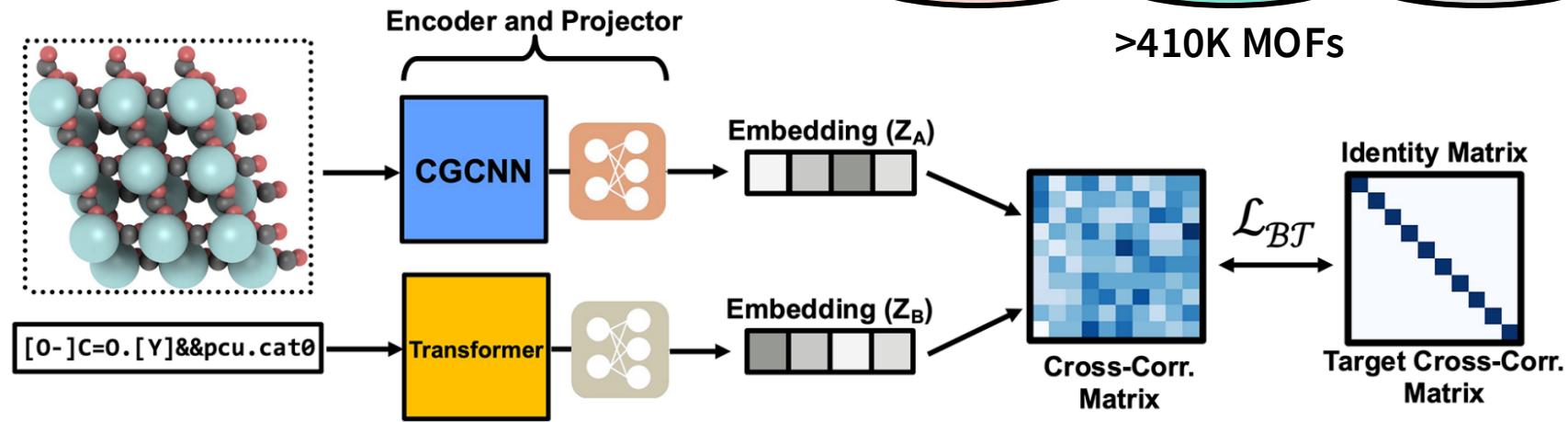
MOFormer



MOFid



Multimodal pre-training

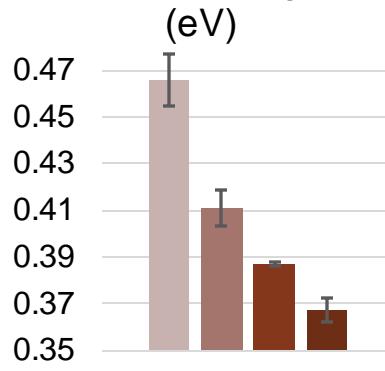


$$L_{\mathcal{BT}} \stackrel{\Delta}{=} \sum_i (1 - C_{ii})^2 + \lambda \sum_i \sum_{j \neq i} C_{ij}^2$$

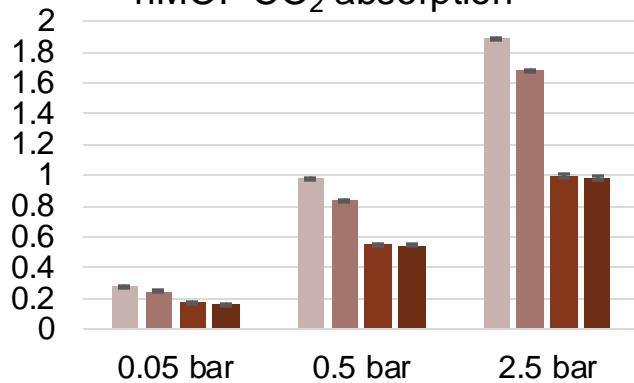
$$C_{ij} \stackrel{\Delta}{=} \frac{\sum_b Z_{b,i}^A Z_{b,j}^B}{\sqrt{\sum_b (Z_{b,i}^A)^2} \sqrt{\sum_b (Z_{b,j}^B)^2}}$$

Band Gap & Gas Adsorption Prediction

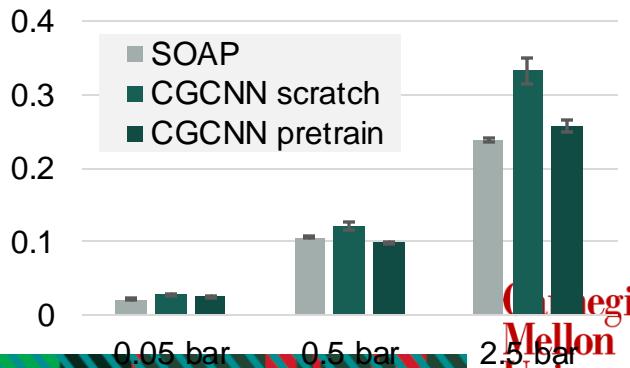
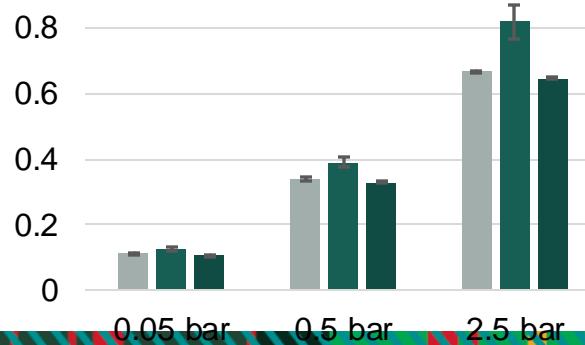
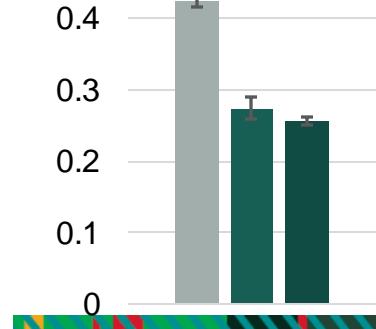
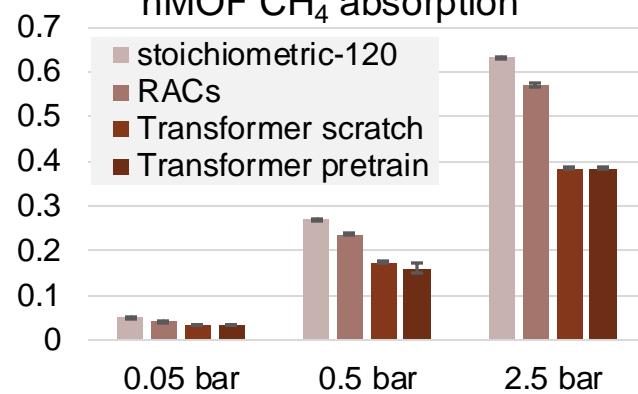
QMOF band gap (eV)



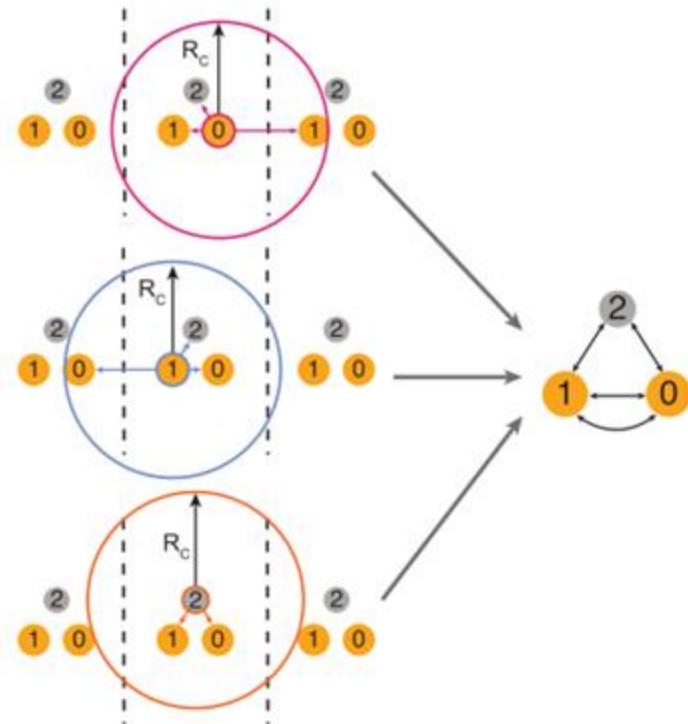
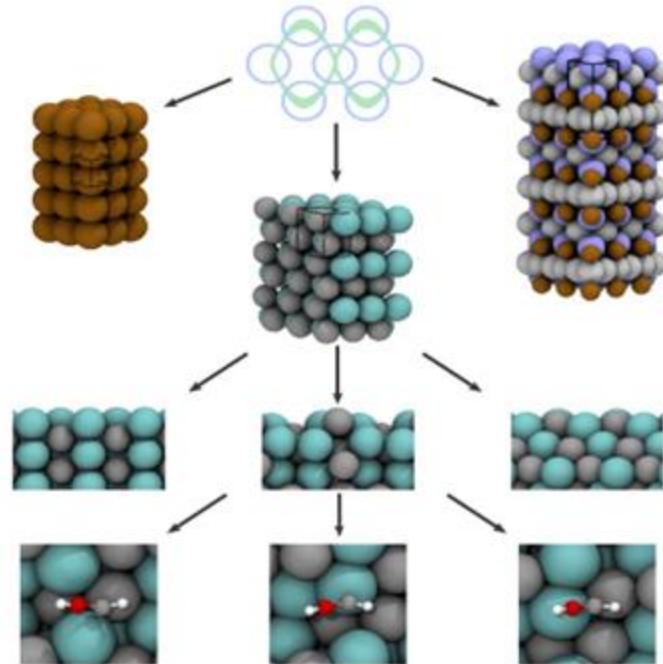
hMOF CO₂ absorption



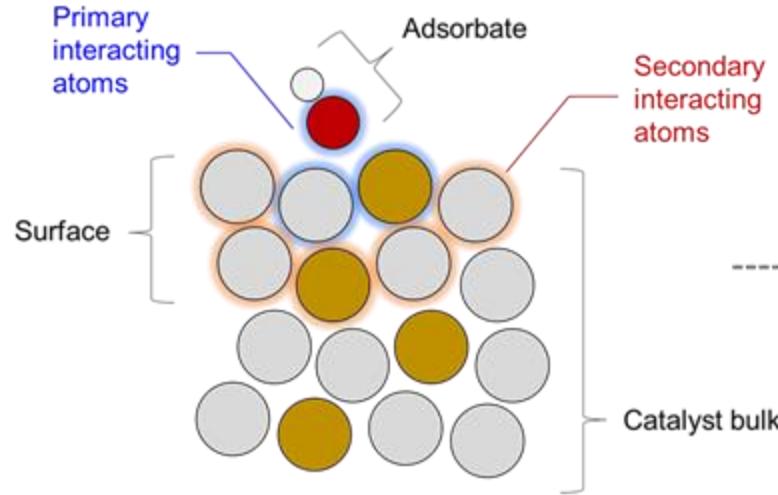
hMOF CH₄ absorption



Graph representation requires a precise understanding of atomic coordinates



Seamless Incorporation of Observable Features through Language Format



Textual String

NO Ti6Pd2 (1 0 0) N Ti Pd bridge [Ti Ti Ti Ti Ti Ti Pd Pd Pd N] ...

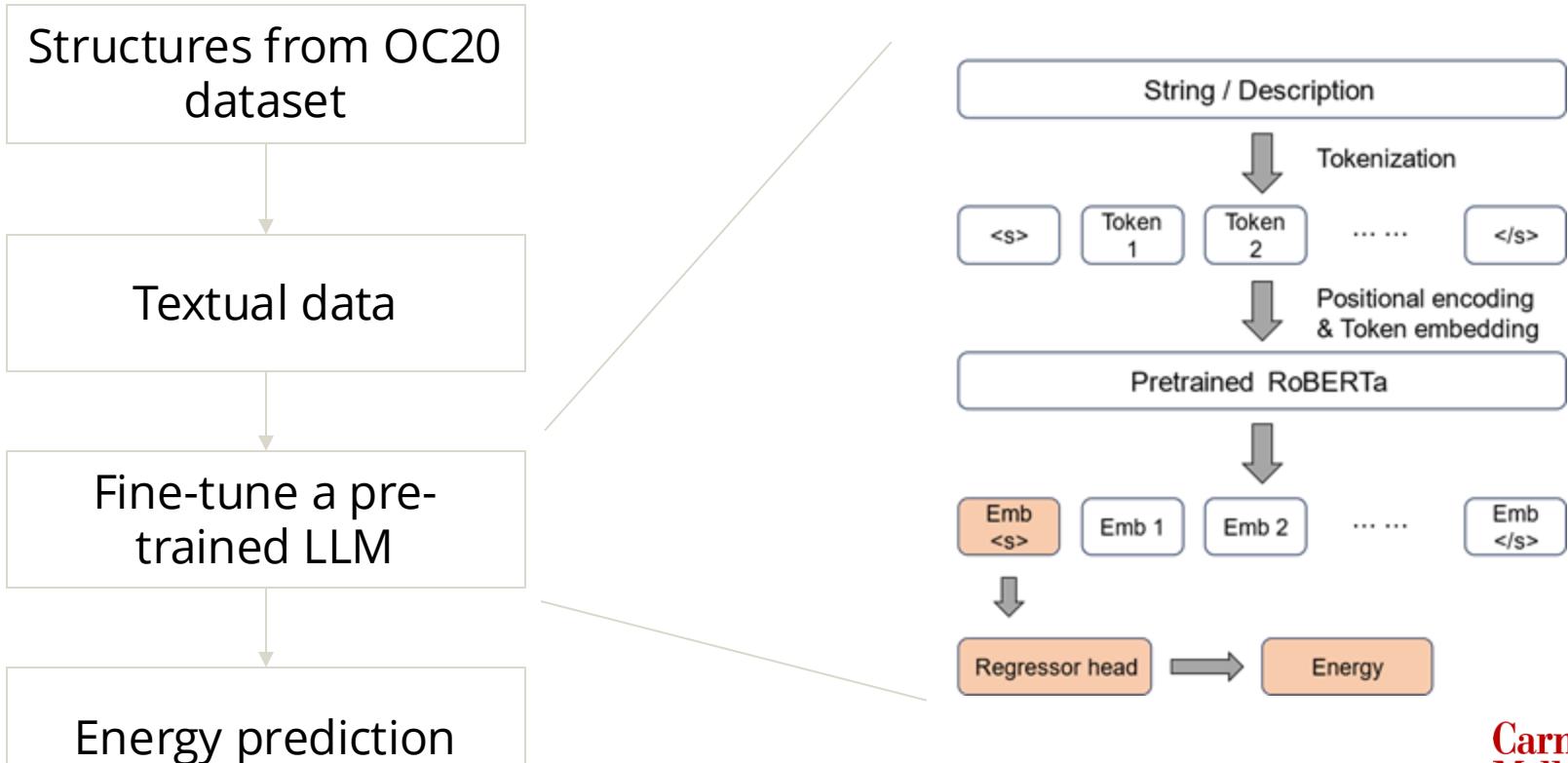
Textual Description

User NO molecule characteristics?
The molecule has a linear geometry with a bond angle of 180 degrees ... The bonding type in NO is covalent ...

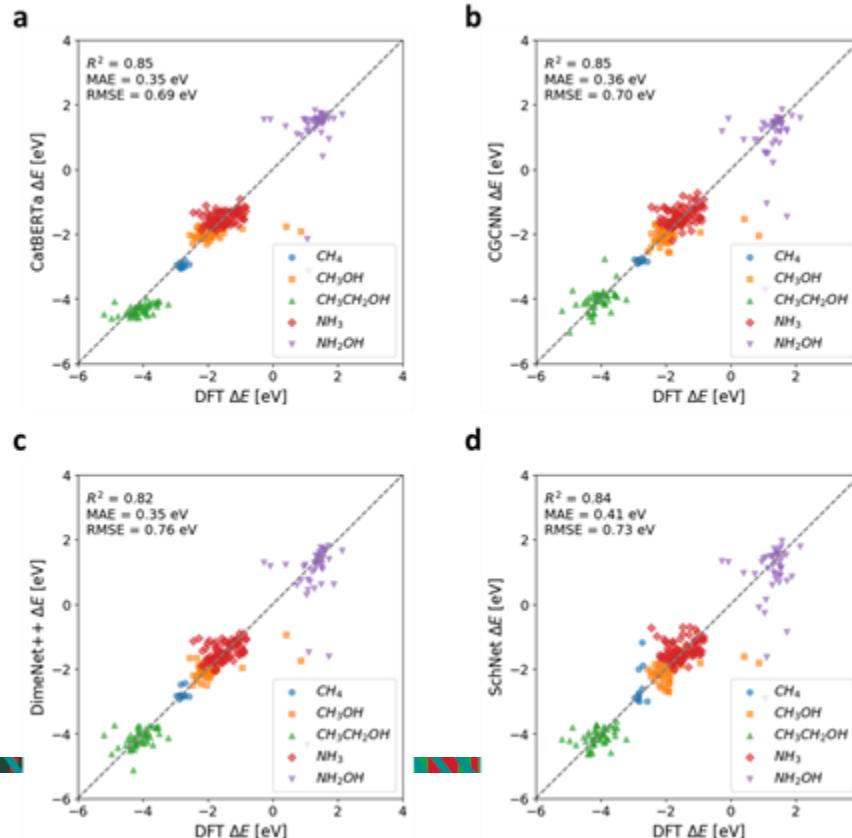
User Surface type of Ti6Pd2 (1,0,0) surface? Valence electron number and electronegativity?
The type of surface is BCC. The coordination number is 8... Surface atoms of Ti: Valence electron number: 4 Pauling electronegativity: 1.54 ...

User Structure of Ti6Pd2?
Ti₃Pd crystallizes in the cubic Pm̄3n space group. Ti is bonded in a 6-coordinate geometry to two equivalent Ti and four equivalent Pd atoms ...

Framework

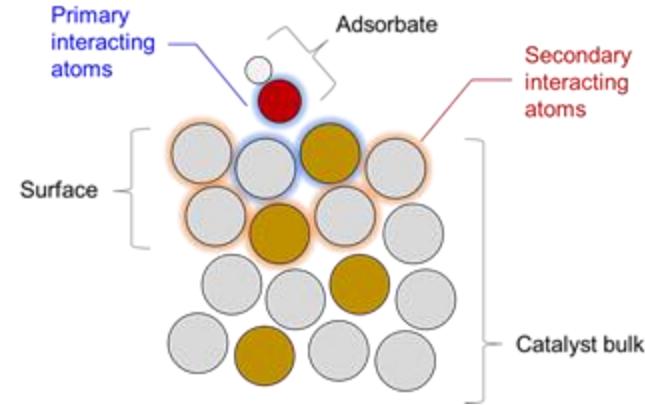


Identifying High-Accuracy Subsets with CatBERTa and GNNs



Exploring Features through Textual Format Input (1)

- Interacting atoms serve as impactful features
- Atomic properties and bonding distance do not universally enhance accuracy



No.	Textual String Format	MAE [eV]	
		Subset	Total
String 1	< s > adsorbate SMILES < /s > bulk composition (Miller index) < /s >	0.44	0.85
String 2	String 1 + [primary interacting atoms, site type]	0.37	0.79
String 3	String 1 + [primary interacting atoms, site type] [atomic properties]	0.47	0.79
String 4	String 1 + [primary interacting atoms, site type] [secondary interacting atoms]	0.35	0.75
String 5	String 1 + [primary interacting atoms, site type, bond distance] [secondary interacting atoms]	0.34	0.77

Exploring Features through Textual Format Input (2)

- Description 1 contains general information and does not derive from adsorption modeling studies
- Description 2 includes properties utilized in adsorption modeling
- Accuracy: description 2 > description 1

No.	Entity	Input Feature	Tool	MAE [eV]	
				Subset	Total
Description 1	System	Adsorbate SMILES, bulk composition, Miller index, primary interacting atoms, adsorption site type	Pymatgen	0.50	0.84
	Adsorbate	Bonding type, angle, length, molecule size, dipole moment, orbital characteristics	ChatGPT		
	Catalyst	Composition, space group, bonding geometry, length, atom arrangement	RoboCrystallographer		
Description 2	System	Adsorbate SMILES, bulk composition, Miller index, primary interacting atoms, adsorption site type	Pymatgen	0.43	0.79
	Adsorbate	Central atom, coordination number	ChatGPT		
	Catalyst	Surface type, coordination number, valence electron number, electronegativity	ChatGPT		



Comparison with GNNs

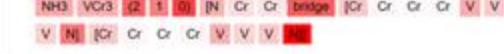
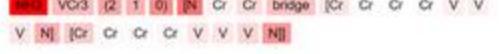
- CatBERTa exhibits enhanced effectiveness when trained on a dataset of limited size
- Subset demonstrating good accuracy in GNNs also yields consistent results in CatBERTa
- CatBERTa does not perform well when encountering unseen adsorbates

Data Size	Model	MAE [eV] (lower is better)					
		Subset	Total	ID	OOD _{ads}	OOD _{cat}	OOD _{both}
10k	CGCNN	0.66	0.83	0.88	0.85	0.80	0.79
	SchNet	0.73	0.90	0.89	0.93	0.87	0.90
	DimeNet++	0.51	0.76	0.77	0.80	0.72	0.75
	CatBERTa	0.46±0.02	0.82±0.02	0.75±0.02	0.95±0.02	0.71±0.01	0.88±0.02
100k	CGCNN	0.36	0.63	0.58	0.72	0.55	0.66
	SchNet	0.41	0.61	0.61	0.61	0.64	0.58
	DimeNet++	0.35	0.57	0.54	0.63	0.52	0.60
	CatBERTa	0.35±0.01	0.75±0.01	0.65±0.01	0.90±0.01	0.61±0.01	0.86±0.02

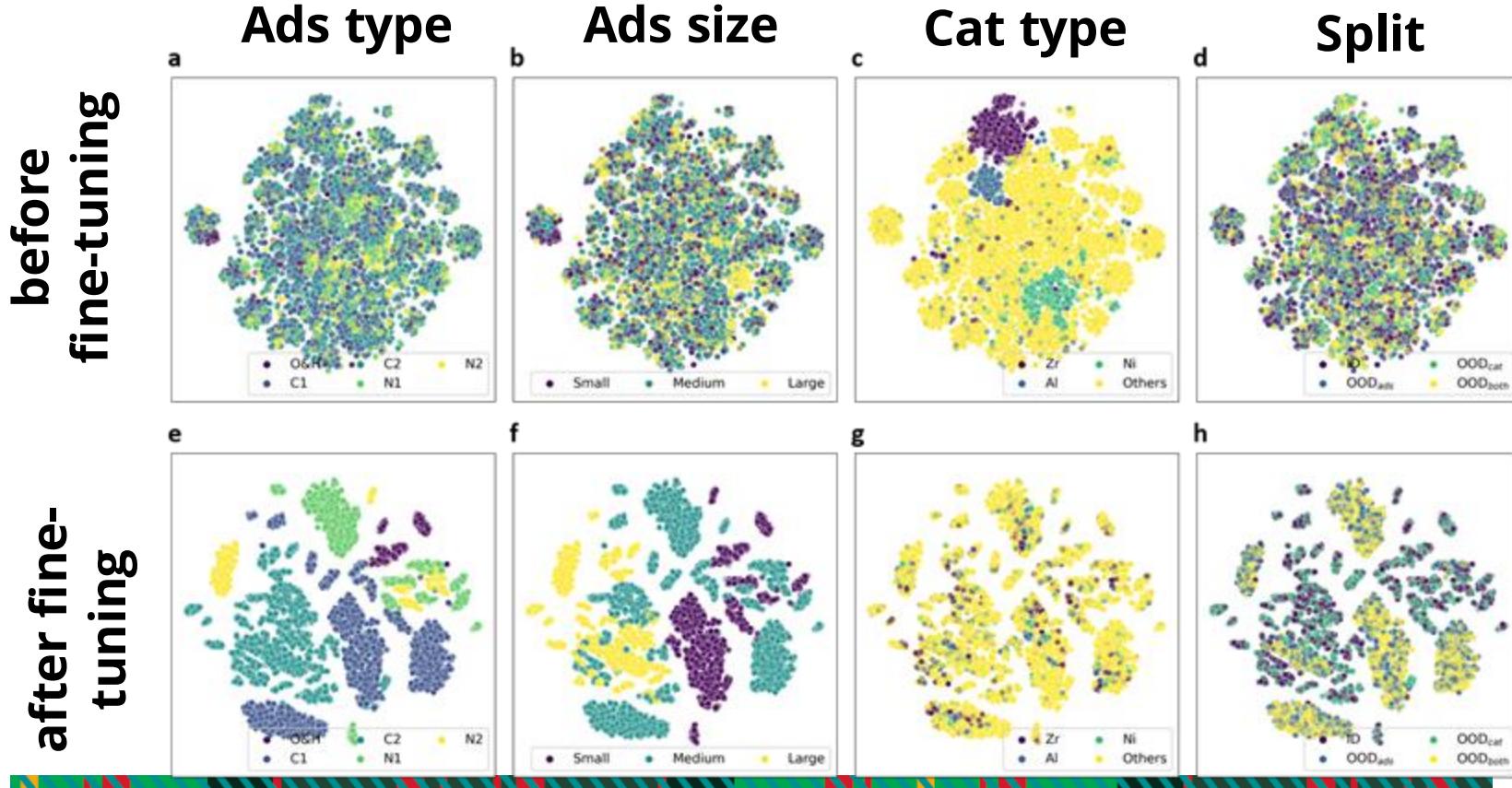
In-domain (ID): Both adsorbate and catalyst are seen during training

Out-of-domain (OOD): A specific entity is not seen during training

The attention score can provide interpretability

First hidden layer	Last hidden layer
 <p>Adsorbate NH₃ is adsorbed on the catalytic surface V₂O₅ with a Molar Index of 0.1. The N atom of the adsorbate is placed on the bridge site and is bonding to two oxygen atoms. O₂ or the surface atom in the V₂O₅ molecule is bonded to the nitrogen atom. The O₂ molecule is a triatomic molecule consisting of one nitrogen atom and three hydrogen atoms. It has a tetrahedral pyramidal geometry with a bond angle of 107.5 degrees and a bond length of 1.10 Å. The nitrogen atom is not hybridized, while the hydrogen atoms have a sp^3 hybridization. The bonding in NH₃ is partially covalent with the nitrogen atom donating one of its valence electrons to each of the three atoms. This remaining pair of electrons on the nitrogen atom makes it a Lewis base capable of forming hydrogen bonds with other molecules. The dipole moment of NH₃ is 1.47 D and arises from the unequal sharing of electrons between nitrogen and hydrogen atoms resulting in a partial negative charge on nitrogen and partial positive charges on hydrogen atoms. Overall, NH₃ is polar because due to its asymmetric molecular geometry and its lone pair interactions with metal catalysts in various catalytic reactions. V₂O₅ is a polymeric structure and crystallizes in the FCC space group. V is bonded in a distorted body-centered cubic geometry to four O atoms. There are two inequivalent O sites in the first V site. O is bonded in a distorted body-centered cubic geometry to four equivalent V and four equivalent O atoms. All O-O bond lengths are 2.48 Å. In the second V site, O is bonded in a distorted body-centered cubic geometry to four equivalent V and four equivalent O atoms.</p>	 <p>Adsorbate NH₃ is adsorbed on the catalytic surface V₂O₅ with a Molar Index of 0.1. The N atom of the adsorbate is placed on the bridge site and is bonding to two oxygen atoms. O₂ or the central atom in the V₂O₅ molecule is bonded to the nitrogen atom. The O₂ molecule is a triatomic molecule consisting of one nitrogen atom and three hydrogen atoms. It has a tetrahedral pyramidal geometry with a bond angle of 107.5 degrees and a bond length of 1.10 Å. The nitrogen atom is not hybridized, while the hydrogen atoms have a sp^3 hybridization. The bonding in NH₃ is partially covalent with the nitrogen atom donating one of its valence electrons to each of the three hydrogen atoms. The remaining pair of electrons on the nitrogen atom makes it a Lewis base capable of forming hydrogen bonds with other molecules. The dipole moment of NH₃ is 1.47 D and arises from the unequal sharing of electrons between nitrogen and hydrogen atoms resulting in a partial negative charge on nitrogen and partial positive charges on hydrogen atoms. NH₃ is a polar molecule due to its asymmetric molecular geometry and its lone pair interactions with metal catalysts in various catalytic reactions. V₂O₅ is a polymeric structure and crystallizes in the FCC space group. V is bonded in a distorted body-centered cubic geometry to four O atoms. There are two inequivalent O sites in the first V site. O is bonded in a distorted body-centered cubic geometry to four equivalent V and four equivalent O atoms. All O-O bond lengths are 2.48 Å. In the second V site, O is bonded in a distorted body-centered cubic geometry to four equivalent V and four equivalent O atoms.</p>
 <p>Adsorbate NH₃ is adsorbed on the catalytic surface V₂O₅ with a Molar Index of 0.1. The N atom of the adsorbate is placed on the bridge site and is bonding to two oxygen atoms. O₂ or the central atom in the V₂O₅ molecule is bonded to the nitrogen atom. The O₂ molecule is a triatomic molecule consisting of one nitrogen atom and three hydrogen atoms. It has a tetrahedral pyramidal geometry with a bond angle of 107.5 degrees and a bond length of 1.10 Å. The nitrogen atom is not hybridized, while the hydrogen atoms have a sp^3 hybridization. Therefore, the difference between the maximum number of bonds the central atom can have and the actual number of bonds it has is 2. The V₂O₅ crystal has a Miller index of 0.1. It has a lattice of FCC (Face-Centered Cubic) type. The coordination number for this surface is 12. To determine the valence electron number and Pauling electronegativity of the surface atoms in V₂O₅, we first need to find the atomic number and the electron configuration of the elements involved. The atomic number of Vanadium (V) is 23. Its electron configuration is [Ar] 3d³ 4s². The valence electron number is 5. [Ar] 3d³ 4s² The atomic number of Chromium (Cr) is 24. Its electron configuration is [Ar] 3d⁵ 4s¹. The valence electron number is 6. Cr and V both have a valence electron number of 6. The Pauling electronegativity of Vanadium is 1.6. The Pauling electronegativity of Chromium is 1.66. Therefore, the valence electron number and Pauling electronegativity of the surface atoms in V₂O₅ are both 6 and 1.66.</p>	 <p>Adsorbate NH₃ is adsorbed on the catalytic surface V₂O₅ with a Molar Index of 0.1. The N atom of the adsorbate is placed on the bridge site and is bonding to two oxygen atoms. O₂ or the central atom in the V₂O₅ molecule is bonded to the nitrogen atom. The O₂ molecule is a triatomic molecule consisting of one nitrogen atom and three hydrogen atoms. It has a tetrahedral pyramidal geometry with a bond angle of 107.5 degrees and a bond length of 1.10 Å. The nitrogen atom is not hybridized, while the hydrogen atoms have a sp^3 hybridization. Therefore, the difference between the maximum number of bonds the central atom can have and the actual number of bonds it has is 2. The V₂O₅ crystal has a Miller index of 0.1. It has a lattice of FCC (Face-Centered Cubic) type. The coordination number for this surface is 12. To determine the valence electron number and the Pauling electronegativity of the surface atoms in V₂O₅, we first need to find the atomic number and the electron configuration of the elements involved. The atomic number of Vanadium (V) is 23. Its electron configuration is [Ar] 3d³ 4s². The valence electron number is 5. [Ar] 3d³ 4s² The atomic number of Chromium (Cr) is 24. Its electron configuration is [Ar] 3d⁵ 4s¹. The valence electron number is 6. Cr and V both have a valence electron number of 6. The Pauling electronegativity of Vanadium is 1.6. The Pauling electronegativity of Chromium is 1.66. Therefore, the valence electron number and Pauling electronegativity of the surface atoms in V₂O₅ are both 6 and 1.66.</p>

The capability of latent space capturing features

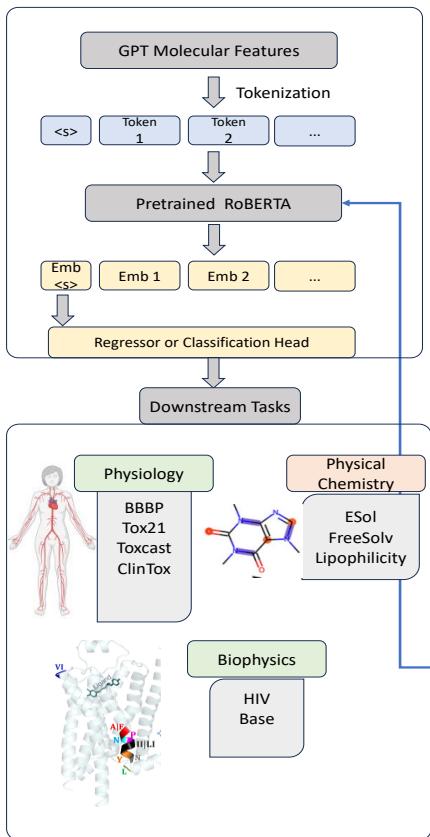


• MoleculeNet

OCC(C)(C)C(C)C=O

ChatGPT

GPT-MolBerta



You are able to generate important and verifiable features about molecular SMILES

Certainly! I can provide important and verifiable features about a molecular SMILES notation. If you have a specific SMILES notation in mind or need information about a particular molecule, please provide the SMILES notation or the molecule's name, and I'll generate relevant information for you.

Generate a description about the following SMILES molecule: OCC(C)(C)C(C)C=O

The SMILES notation OCC(C)(C)C(C)C=O corresponds to the organic compound diethyl ether. Diethyl ether, also known as ethyl ether, is a colorless, highly volatile liquid with a characteristic sweet, ethereal odor. Here is a description of this molecule:

Chemical Structure:

- Functional Groups:** Diethyl ether consists of several functional groups. The central part of the molecule contains an oxygen atom bonded to two carbon atoms, forming the ether functional group (COC). Additionally, there is a carbonyl group (C=O) at the end of the molecule, which imparts some reactivity.

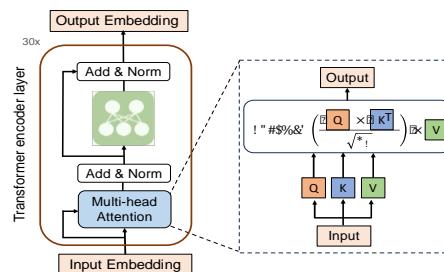
Physical Properties:

- State:** Diethyl ether is typically a colorless liquid at room temperature but can easily vaporize, forming a highly flammable and volatile gas.
- Odor:** It has a distinct, sweet, and somewhat pungent odor, which makes it easily recognizable.
- Density:** Diethyl ether is less dense than water, which means it will float on the surface of water.

Chemical Properties and Uses:

Task (Metric)	Dataset	# Molecules
Regression(RMSE)	ESOL	1128
	FreeSolv	642
	Lipophilicity	4200
Classification (ROC-AUC)	HIV	41127
	BACE	1513
	BBBP	2039
	Tox21	7831
	SIDER	1427
	ClinTox	1478

Balaji, S., Magar, R., Yayati, J. Barati Farimani, A. GPT-MoBERTa: GPT Molecular Features Language Model for molecular property prediction, arxiv, 2023



Prompt Engineering

S1: Example of Molecular Description

The given SMILES molecule, OC12C3CC(NC1=O)C23, represents a cyclic compound with a complex structure. Here is a description of its important features:

- Ring Structure:** The molecule consists of a fused ring system with three rings. The first ring is a six-membered ring (C1-C6), the second ring is a five-membered ring (C2-C6), and the third ring is a four-membered ring (C2-C5).
- Oxygen Atom:** There is an oxygen atom (O) attached to the first carbon atom (C1) of the six-membered ring. This oxygen atom contributes to the overall polarity of the molecule.
- Nitrogen Atom:** A nitrogen atom (N) is present in the five-membered ring (C2-C6). It is connected to the carbon atom (C3) and forms a bond with the oxygen atom (O) in the first ring (C1).
- Carbonyl Group:** The nitrogen atom (N) in the five-membered ring is also connected to a carbonyl group (C=O), represented by the SMILES notation NC1=O. This carbonyl group adds a functional group to the molecule and can participate in various chemical reactions.
- Substituents:** The molecule has additional substituents attached to the carbon atoms in the rings. The specific nature of these substituents is not provided in the SMILES notation.

Overall, the given SMILES molecule represents a complex cyclic compound with oxygen and nitrogen atoms, a carbonyl group, and additional substituents. The specific arrangement and nature of the substituents will likely determine the molecule's properties and potential biological activities.

Iteration	Query Prompt
1	<pre>"role": "system", "content": "-" "role": "user", "content": f"Give a description of the following molecule value"</pre>
2	<pre>"role": "system", "content": "You are an expert in everything about molecules" "role": "user", "content": f"Generate verifiable information about the following SMILES value"</pre>
3	<pre>"role": "system", "content": 'You are able to generate important and verifiable features about molecular SMILES' "role": "user", "content": f"Generate a description about the following SMILES molecule value"</pre>

GPT-MolBerta

Models	BBBP	Tox 21	ClinTox	HIV	BACE	SIDER
GCN ^[68]	71.9 ± 0.9	70.9 ± 2.6	62.5 ± 2.8	74.0 ± 3.0	71.6 ± 2.0	53.6 ± 3.2
GIN ^[67]	65.8 ± 4.5	74.0 ± 0.8	58.0 ± 4.4	75.3 ± 1.9	70.1 ± 5.4	57.3 ± 1.6
SchNet ^[13]	84.8 ± 2.2	77.2 ± 2.3	71.5 ± 3.7	70.2 ± 3.4	76.6 ± 1.1	53.9 ± 3.7
MGCN ^[14]	85.0 ± 6.4	70.7 ± 1.6	63.4 ± 4.2	73.8 ± 1.6	73.4 ± 3.0	55.2 ± 1.8
D-MPNN ^[11]	71.2 ± 3.8	68.9 ± 1.3	90.5 ± 5.3	75.0 ± 2.1	85.3 ± 5.3	63.2 ± 2.3
Hu et al. ^[69]	70.8 ± 1.5	78.7 ± 0.4	78.9 ± 2.4	80.2 ± 0.9	85.9 ± 0.8	65.2 ± 0.9
MolCLR-GCN ^[53]	73.8 ± 0.2	74.7 ± 0.8	86.7 ± 1.0	77.8 ± 0.5	78.8 ± 0.5	66.9 ± 1.2
MolCLR-GIN ^[53]	73.6 ± 0.5	79.8 ± 0.7	93.2 ± 1.7	80.6 ± 1.1	89.0 ± 0.3	68.0 ± 1.1
MolBERT ^[33]	76.2 ± 0.0	-	-	78.3 ± 0.0	86.6 ± 0.0	-
ChemBERTa-2 ^[39]	72.8 ± 0.0	-	-	62.2 ± 0.0	79.9 ± 0.0	-
CLM ^[43]	91.5 ± 0.0	79.5 ± 0.0	-	81.3 ± 0.0	86.1 ± 0.0	61.9 ± 0.0
SELFormer ^[41]	90.2 ± 0.0	65.3 ± 0.0	-	68.1 ± 0.0	83.2 ± 0.0	74.5 ± 0.0
GPT-MolBERTa	74.1 ± 0.15	65.9 ± 0.06	49.7 ± 0.12	75.5 ± 1.29	73.4 ± 0.47	58.5 ± 0.35

GPT-MolBerta

Models	FreeSolv	ESOL	Lipophilicity
GCN ⁶⁸	2.87 ± 0.14	1.43 ± 0.05	0.85 ± 0.08
GIN ⁶⁷	2.76 ± 0.18	1.45 ± 0.02	0.85 ± 0.07
SchNet ¹³	3.22 ± 0.76	1.05 ± 0.06	0.91 ± 0.10
MGCN ¹⁴	3.35 ± 0.01	1.27 ± 0.15	1.11 ± 0.04
D-MPNN ¹¹	2.18 ± 0.91	0.98 ± 0.26	0.65 ± 0.05
Hu et al. ⁶⁹	2.83 ± 0.12	1.22 ± 0.02	0.74 ± 0.00
MolCLR-GCN ⁵³	2.39 ± 0.14	1.16 ± 0.00	0.78 ± 0.01
MolCLR-GIN ⁵³	2.20 ± 0.20	1.11 ± 0.01	0.65 ± 0.08
MolBERT ³³	0.948 ± 0.33	0.531 ± 0.04	0.561 ± 0.03
ChemBERTa-2 ³⁹	-	-	0.798 ± 0.00
ChemFormer ³⁵	1.23 ± 0.00	0.633 ± 0.00	0.598 ± 0.00
SELFormer ⁴¹	2.797 ± 0.00	0.682 ± 0.00	0.735 ± 0.00
GPT-MolBERTa	0.896 ± 0.02	0.477 ± 0.01	0.758 ± 0.01

GPT-MoLBerta Interpretability

The given SMILES molecule, NC12C3CC1(C3)OC2=N, represents a complex organic compound. Here are some important features about this molecule:
\n\n1 . Molecular Formula: The molecular formula of this compound is C₁₀H₁₂N₂O
\n\n2 . Aromatic Ring: The molecule contains a fused aromatic ring system, represented by the numbers 1, 2, and 3 in the SMILES string. This ring system consists of two benzene rings fused together.
\n\n3 . Nitrogen Atom: The presence of the nitrogen atom (N) in the SMILES string indicates the substitution of one of the carbon atoms in the aromatic ring with a nitrogen atom.
\n\n4 . Oxygen Atom: The molecule also contains an oxygen atom (O) attached to one of the carbon atoms in the aromatic ring. This oxygen atom is connected to the carbon atom represented by the number 2 in the SMILES string.
\n\n5 . Heterocyclic Structure: The combination of the nitrogen atom and the oxygen atom in the molecule creates a heterocyclic structure. Heterocyclic compounds are organic compounds that contain at least one atom other than carbon in the ring.
\n\n6 . Stereochemistry: The SMILES string does not provide information about the stereochemistry of the molecule. Therefore, the exact spatial arrangement of atoms and groups in the compound cannot be determined solely from the given SMILES representation.
\n\nIt is important to note that while the SMILES string provides a concise representation of the molecule's structure, it does not provide information about the specific functional groups or the overall chemical properties of the compound.

Challenges in distinguishing subtle differences in adsorption configurations

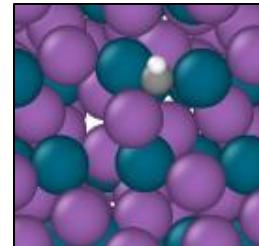
Adsorption energy:

$$\Delta E_i = E_{\text{sys},i} - E_{\text{slab}} - E_{\text{gas}},$$

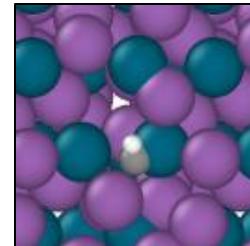
$$\Delta E_{\text{ads}} = \min_i(\Delta E_i).$$

Adsorption configurations:

- Adsorption site
- Adsorbate orientation
- Relaxed states of surface
-

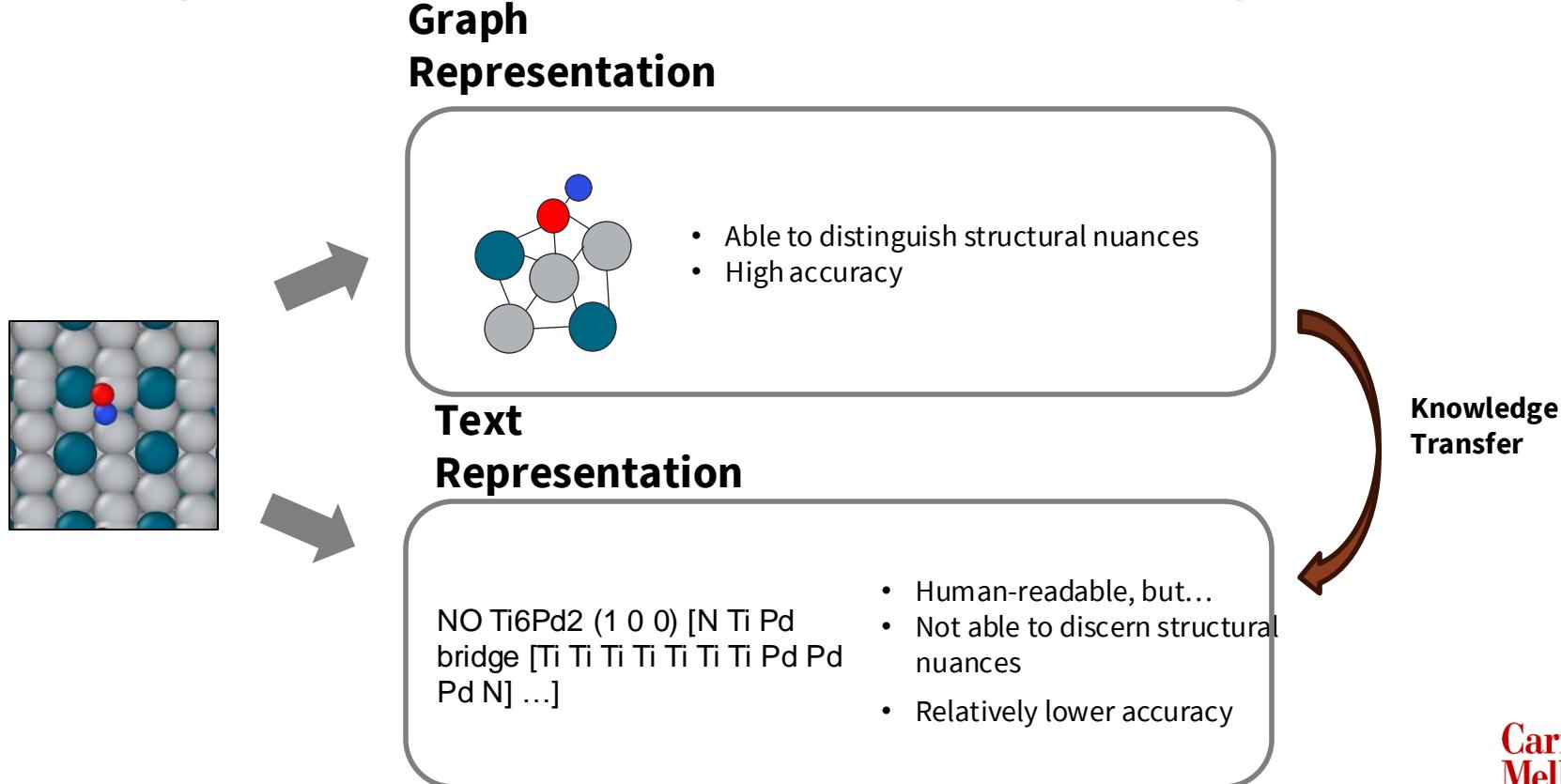


$$\Delta E = 0.78 \text{ eV}$$

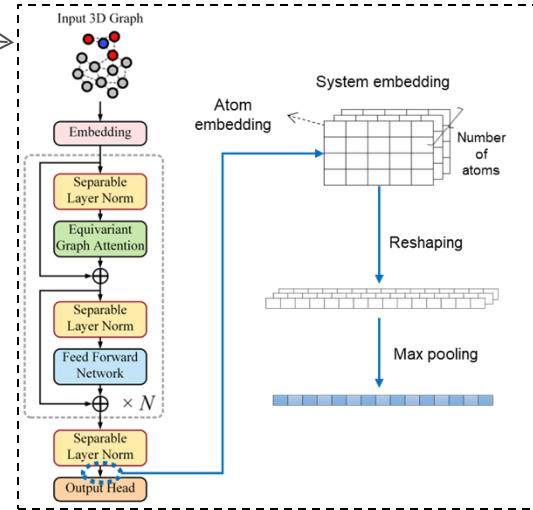
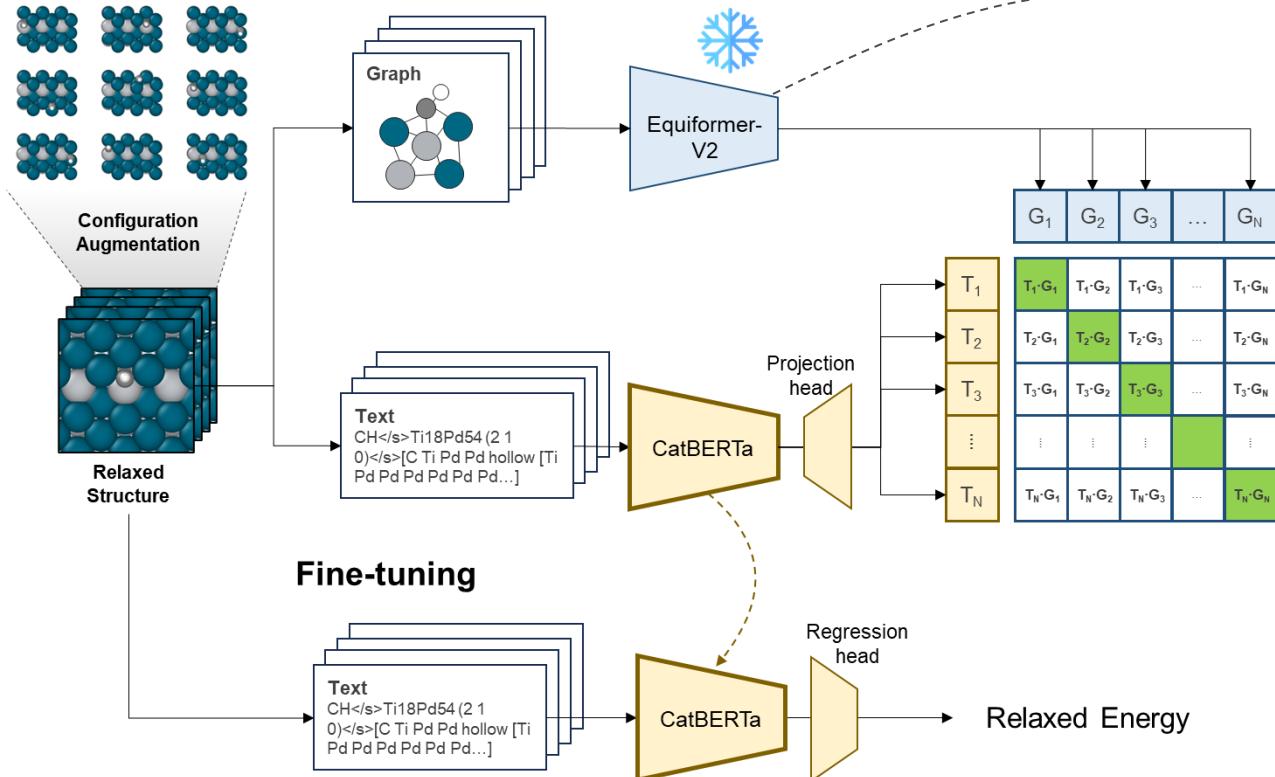


$$\Delta E = 0.63 \text{ eV}$$

Text representation has limitations in accurate prediction



Graph-Assisted Pretraining



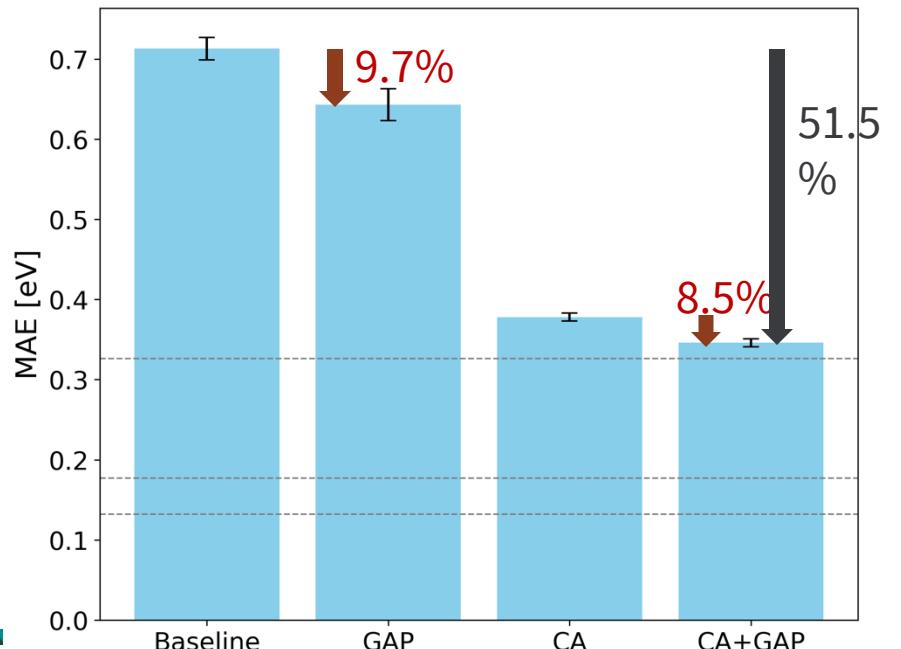
- Trained on DFT-relaxed structures
- Tested on ML-relaxed structures
- Prediction only relying on texts

Improved prediction accuracy

Baseline: Fine-tuned with energy prediction without any enhancement

GAP: Graph-assisted pretraining

CA: Configuration augmentation



DimeNet++

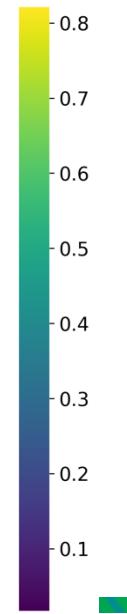
GemNet-OC
EquiformerV2



Graph-assisted pretraining redirects attention toward the configuration section

Baseline

	<S>	Ads.	Cat.	Conf.
Head 1	0.025	0.821	0.038	0.115
Head 2	0.020	0.520	0.055	0.406
Head 3	0.078	0.484	0.123	0.315
Head 4	0.082	0.306	0.202	0.411
Head 5	0.036	0.737	0.066	0.161
Head 6	0.028	0.209	0.203	0.561
Head 7	0.059	0.678	0.071	0.192
Head 8	0.083	0.482	0.117	0.318
Head 9	0.056	0.278	0.288	0.378
Head 10	0.091	0.518	0.107	0.284
Head 11	0.017	0.036	0.357	0.590
Head 12	0.037	0.339	0.144	0.480
AVG	0.051	0.451	0.148	0.351

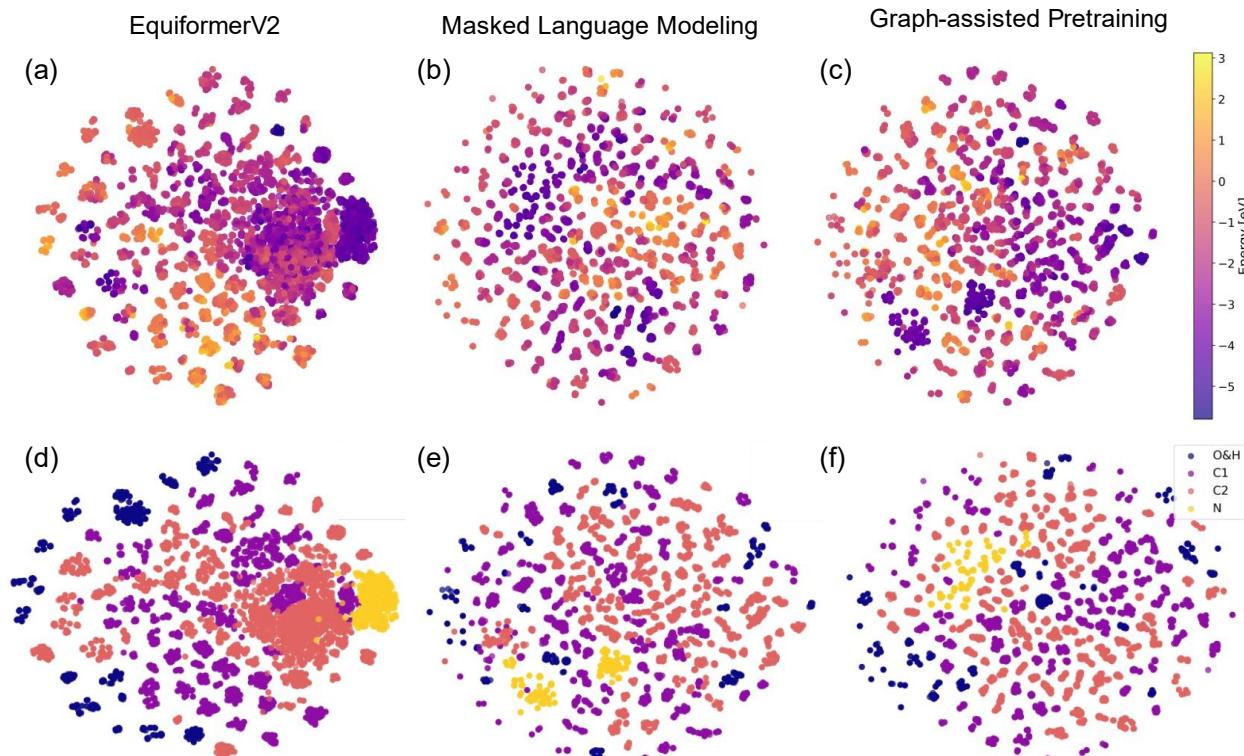


Graph-assisted pretraining

	<S>	Ads.	Cat.	Conf.
Head 1	0.015	0.133	0.239	0.613
Head 2	0.017	0.249	0.273	0.460
Head 3	0.007	0.208	0.358	0.427
Head 4	0.015	0.225	0.281	0.480
Head 5	0.003	0.152	0.366	0.480
Head 6	0.023	0.111	0.241	0.625
Head 7	0.004	0.124	0.367	0.505
Head 8	0.003	0.219	0.331	0.447
Head 9	0.003	0.037	0.322	0.638
Head 10	0.005	0.135	0.349	0.511
Head 11	0.016	0.121	0.206	0.657
Head 12	0.014	0.083	0.211	0.692
AVG	0.011	0.150	0.295	0.544

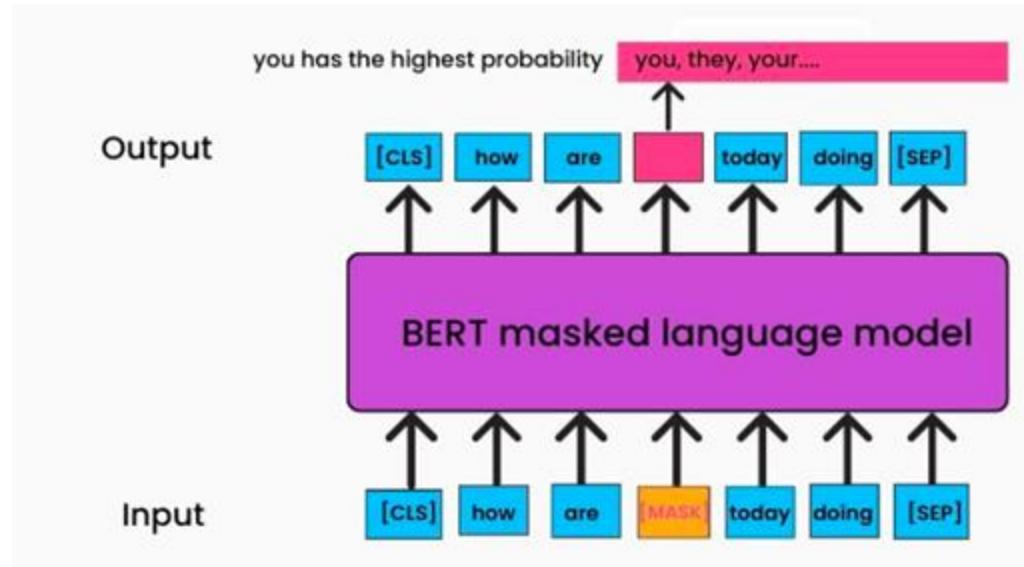


Latent space is restructured to provide a more effective starting point for fine-tuning



BERT

Encoder-only architecture. 340 million parameters.



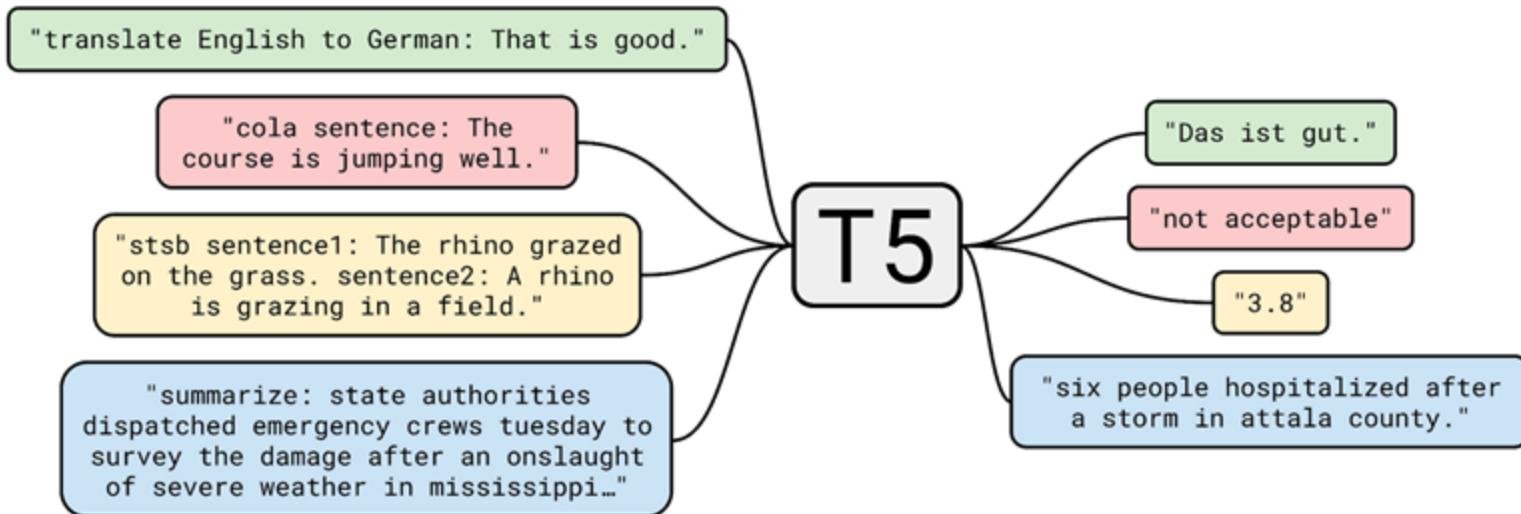
Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018)

Taken from <https://www.turing.com/kb/how-bert-nlp-optimization-model-works>

T5

ENCODER-DECODER MODEL

SHUFFLING, MASKED LANGUAGE MODELING. 3B PARAMETERS



Raffel, Colin, et al. "Exploring the limits of transfer learning with a unified text-to-text transformer." *The Journal of Machine Learning Research* 21.1 (2020): 5485-5551.

GPT

DECODER ONLY ARCHITECTURE

TRAINED FOR THE NEXT WORD PREDICTION (TURNS OUT, THIS SCALES WELL WITH MORE DATA AND MODEL SIZE, THUS WE HAVE THE MODERN GPTs NOW)

GPT2 - 1.75B PARAMETERS

...

GPT4 - 275B PARAMETERS (ESTIMATION)

SSL for Vision

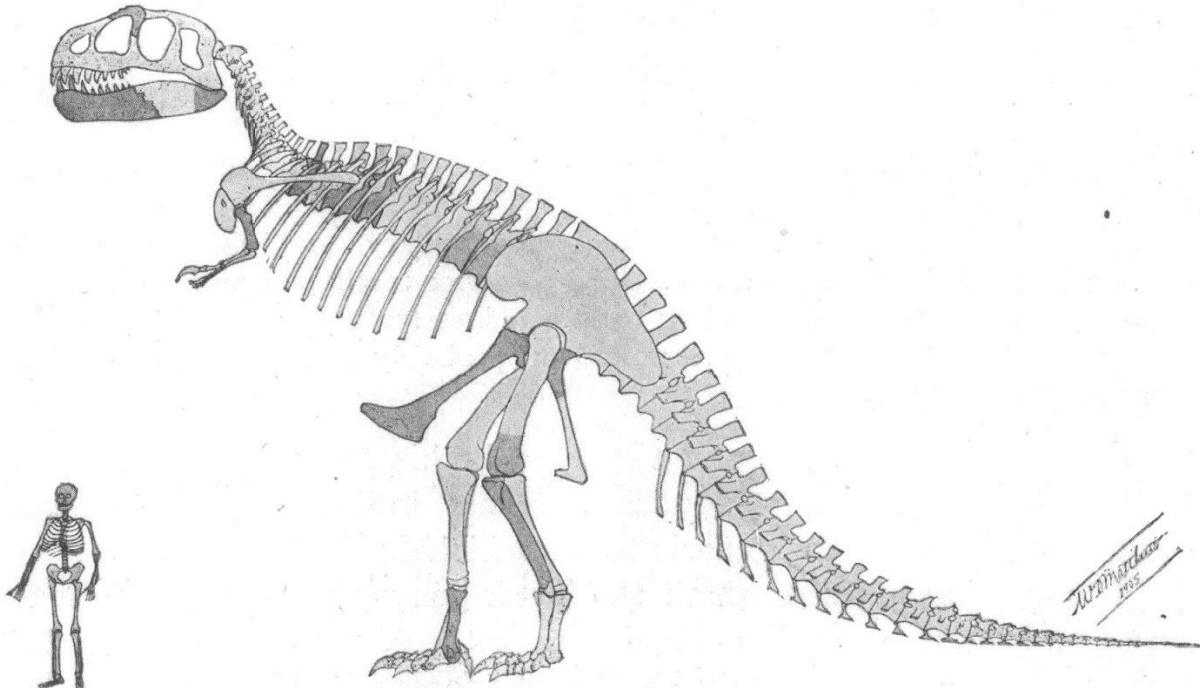
HOW CAN WE IMPLEMENT SSL TECHNIQUES IN VISION?



Foundational Models



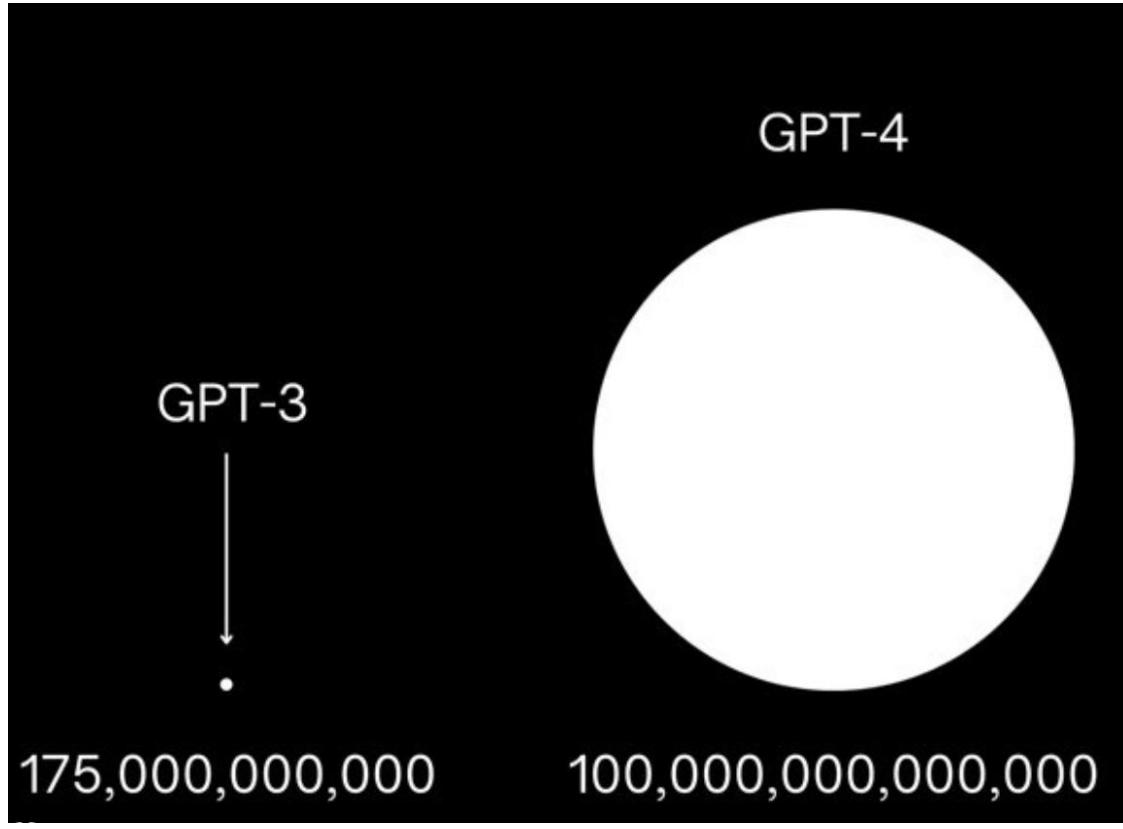
Foundational Models



GPT-2
1.5B Parameters

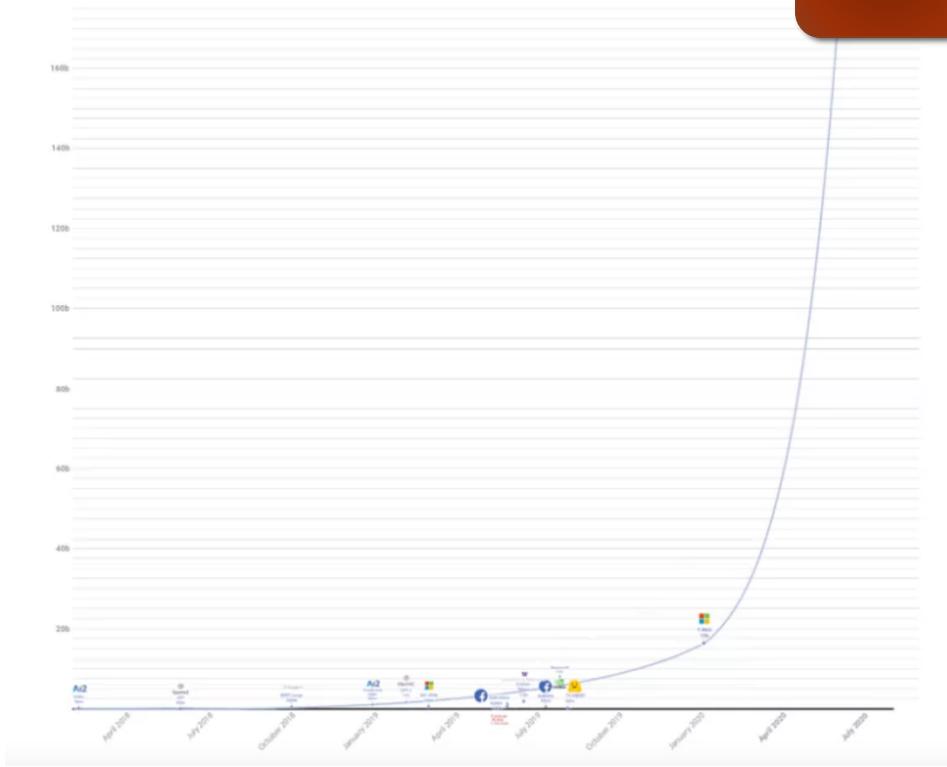
GPT-3
175B Parameters

Foundational Models

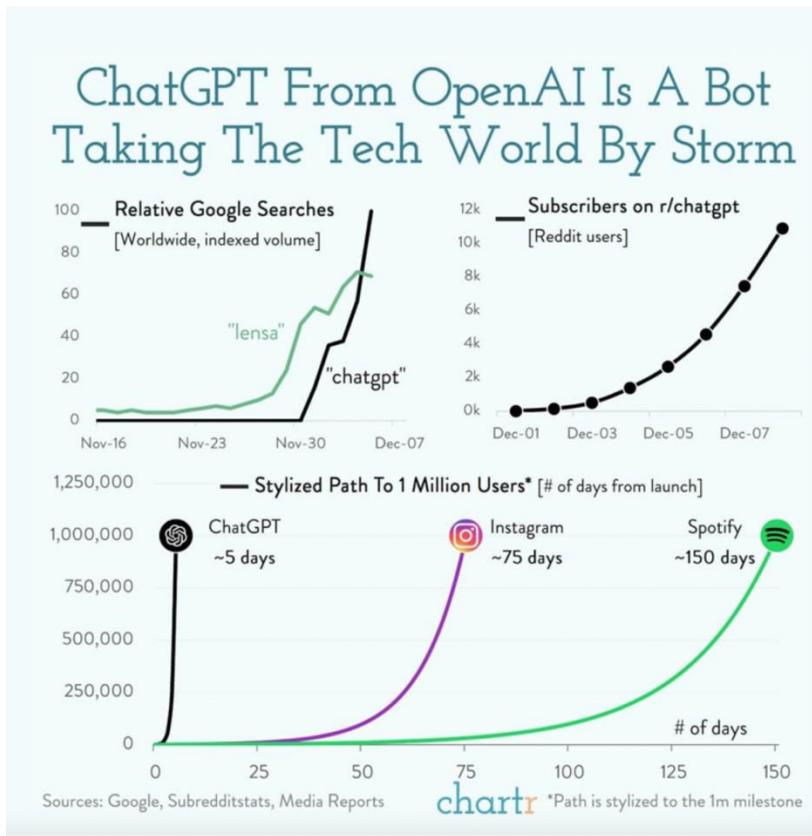


Foundational Models

GPT-4

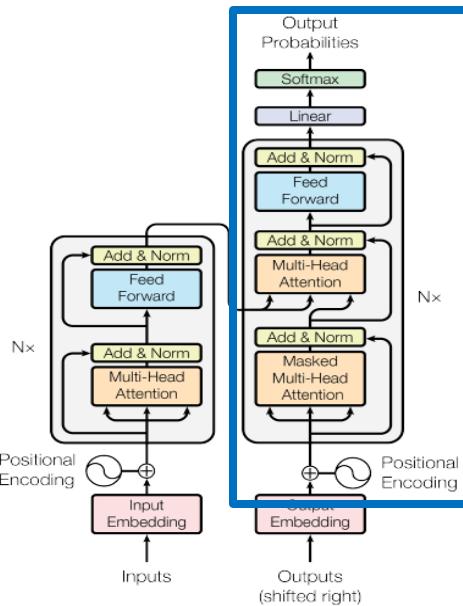


Foundational Models



Generative Pre-Training (GPT)

https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

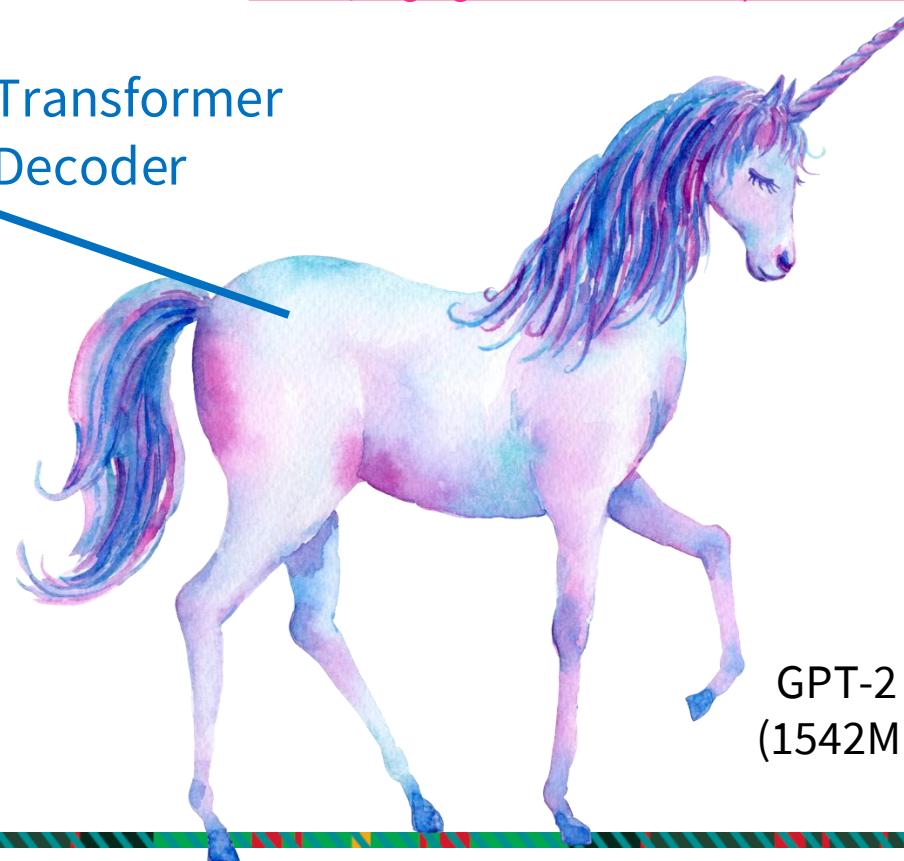


BERT
(340M)

ELMO
(94M)



Transformer
Decoder



GPT-2
(1542M)

Source of image: <https://huaban.com/pins/1714071707/>

Carnegie
Mellon
University

Extensions

Autoregressive

- GPT (Generative Pre-training)
- CTRL (Conditional Transformer LM for Controllable Generation)
- Reformer
- XLNet

GPT-1

- **Model:** Transformer Decoders we just described
- **Objective:** next word prediction (cross-entropy loss)
- **Data:** BooksCorpus (7k books from a variety of genres, such as Adventure, Fantasy, and Romance)

GPT-1

Authors were primarily focused on demonstrating that you could fine-tune this LM on supervised tasks and get SOTA results

Improving Language Understanding by Generative Pre-Training

Alec Radford
OpenAI
alec@openai.com

Karthik Narasimhan
OpenAI
karthikn@openai.com

Tim Salimans
OpenAI
tim@openai.com

Ilya Sutskever
OpenAI
ilyasu@openai.com



Carnegie
Mellon
University

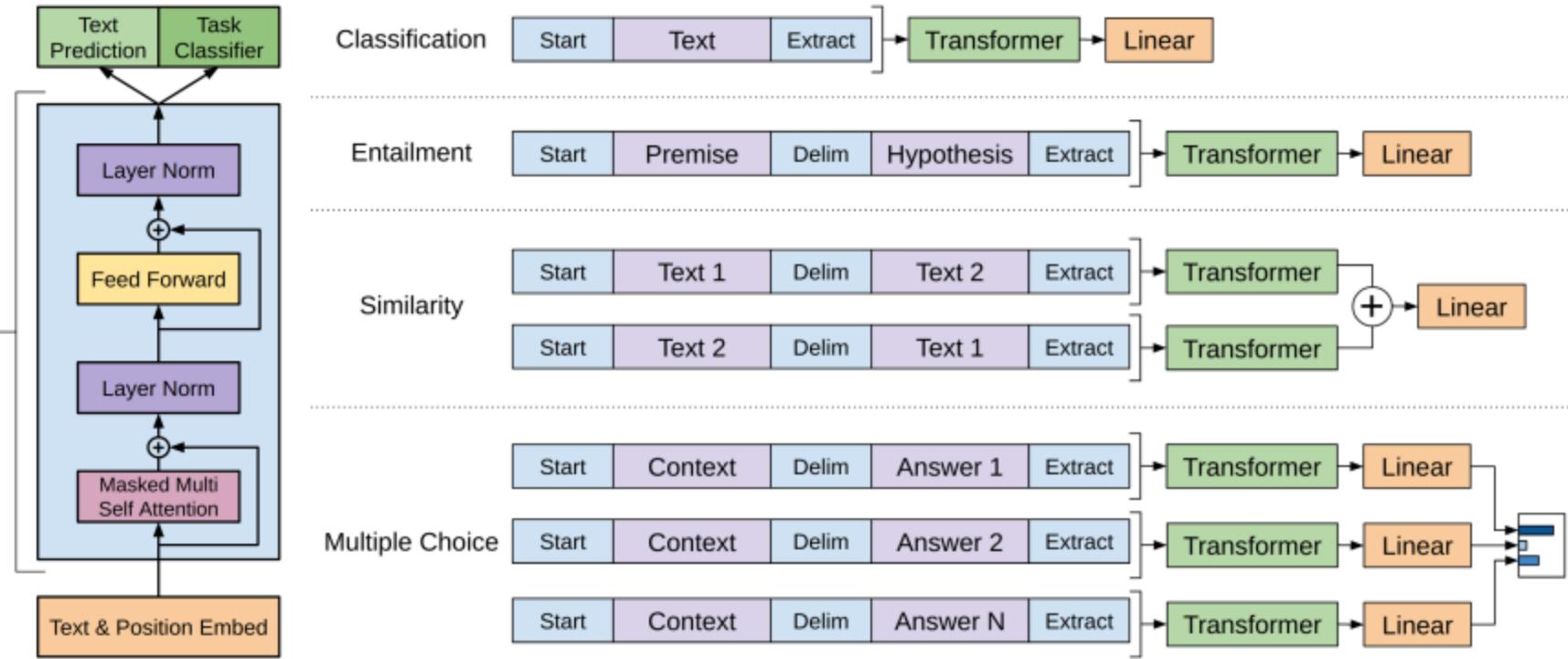


Figure 1: **(left)** Transformer architecture and training objectives used in this work. **(right)** Input transformations for fine-tuning on different tasks. We convert all structured inputs into token sequences to be processed by our pre-trained model, followed by a linear+softmax layer.

Given an unsupervised corpus of tokens $\mathcal{U} = \{u_1, \dots, u_n\}$, we use a standard language modeling objective to maximize the following likelihood:

$$L_1(\mathcal{U}) = \sum_i \log P(u_i | u_{i-k}, \dots, u_{i-1}; \Theta) \quad (1)$$

After training the model with the objective in Eq. 1, we adapt the parameters to the supervised target task. We assume a labeled dataset \mathcal{C} , where each instance consists of a sequence of input tokens, x^1, \dots, x^m , along with a label y . The inputs are passed through our pre-trained model to obtain the final transformer block's activation h_l^m , which is then fed into an added linear output layer with parameters W_y to predict y :

$$P(y|x^1, \dots, x^m) = \text{softmax}(h_l^m W_y). \quad (3)$$

This gives us the following objective to maximize:

$$L_2(\mathcal{C}) = \sum_{(x,y)} \log P(y|x^1, \dots, x^m). \quad (4)$$

We additionally found that including language modeling as an auxiliary objective to the fine-tuning helped learning by (a) improving generalization of the supervised model, and (b) accelerating convergence. This is in line with prior work [50, 43], who also observed improved performance with such an auxiliary objective. Specifically, we optimize the following objective (with weight λ):

$$L_3(\mathcal{C}) = L_2(\mathcal{C}) + \lambda * L_1(\mathcal{C}) \quad (5)$$

GPT-1

Table 2: Experimental results on natural language inference tasks, comparing our model with current state-of-the-art methods. 5x indicates an ensemble of 5 models. All datasets use accuracy as the evaluation metric.

Method	MNLI-m	MNLI-mm	SNLI	SciTail	QNLI	RTE
ESIM + ELMo [44] (5x)	-	-	<u>89.3</u>	-	-	-
CAFE [58] (5x)	80.2	79.0	<u>89.3</u>	-	-	-
Stochastic Answer Network [35] (3x)	<u>80.6</u>	<u>80.1</u>	-	-	-	-
CAFE [58]	78.7	77.9	88.5	<u>83.3</u>		
GenSen [64]	71.4	71.3	-	-	<u>82.3</u>	59.2
Multi-task BiLSTM + Attn [64]	72.2	72.1	-	-	82.1	61.7
Finetuned Transformer LM (ours)	82.1	81.4	89.9	88.3	88.1	56.0

NLI is when you predict if the hypothesis phrase is entailed, neutral, or contradicts the preceding premise phrase.

GPT-1

Table 3: Results on question answering and commonsense reasoning, comparing our model with current state-of-the-art methods.. 9x means an ensemble of 9 models.

Method	Story Cloze	RACE-m	RACE-h	RACE
val-LS-skip [55]	76.5	-	-	-
Hidden Coherence Model [7]	<u>77.6</u>	-	-	-
Dynamic Fusion Net [67] (9x)	-	55.6	49.4	51.2
BiAttention MRU [59] (9x)	-	<u>60.2</u>	<u>50.3</u>	<u>53.3</u>
Finetuned Transformer LM (ours)	86.5	62.9	57.4	59.0

Story Cloze is like MLM, by predicting the blank

GPT-1

Method	Classification		Semantic Similarity		GLUE	
	CoLA (mc)	SST2 (acc)	MRPC (F1)	STS-B (pc)	QQP (F1)	
Sparse byte mLSTM [16]	-	93.2	-	-	-	-
TF-KLD [23]	-	-	86.0	-	-	-
ECNU (mixed ensemble) [60]	-	-	-	<u>81.0</u>	-	-
Single-task BiLSTM + ELMo + Attn [64]	<u>35.0</u>	90.2	80.2	55.5	<u>66.1</u>	64.8
Multi-task BiLSTM + ELMo + Attn [64]	18.9	91.6	83.5	72.8	63.3	<u>68.9</u>
Finetuned Transformer LM (ours)	45.4	91.3	82.3	82.0	70.3	72.8

Overall, our approach achieves new state-of-the-art results in 9 out of the 12 datasets we evaluate on, outperforming ensembles in many cases. Our results also indicate that our approach works well across datasets of different sizes, from smaller datasets such as STS-B ($\approx 5.7k$ training examples) – to the largest one – SNLI ($\approx 550k$ training examples).

Transformer

What if we want to generate a new output sequence?

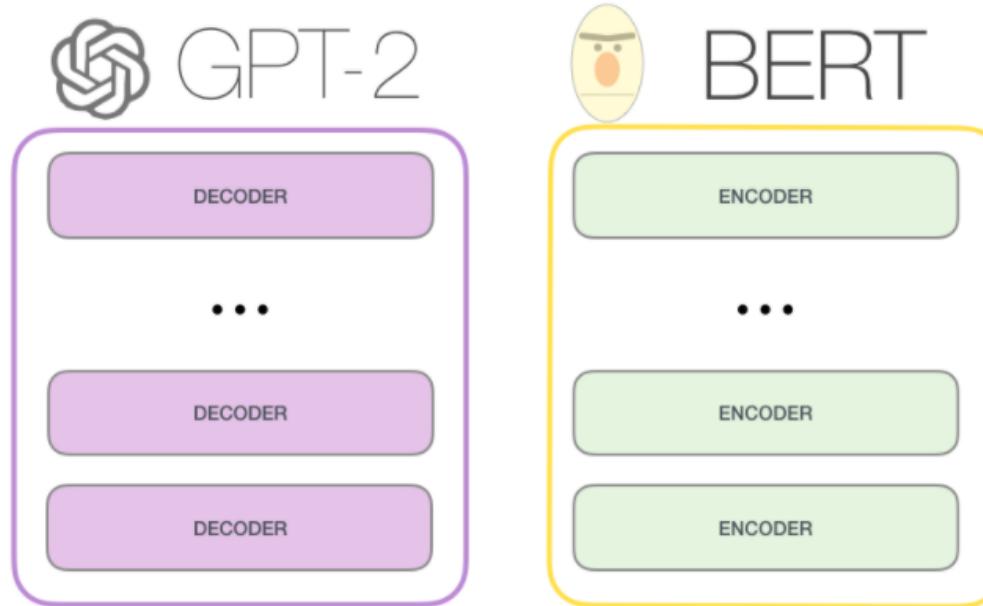
GPT-2 model to the rescue!

Generative Pre-trained Transformer 2



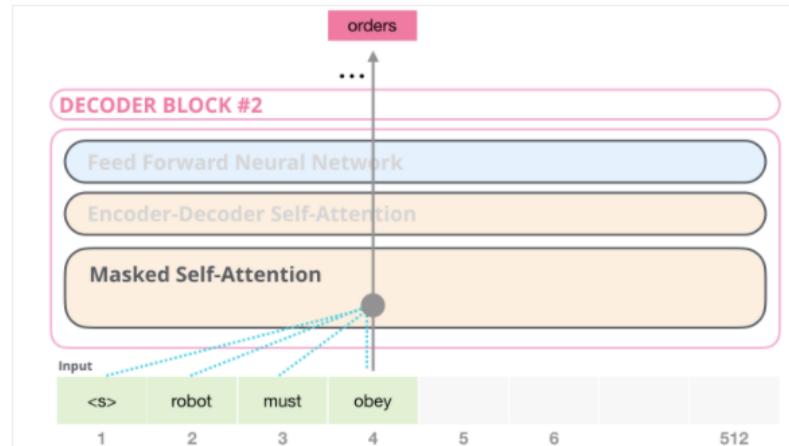
GPT-2 (a Transformer)

GPT-2 uses only **Transformer Decoders** (no Encoders) to generate new sequences from scratch or from a starting sequence



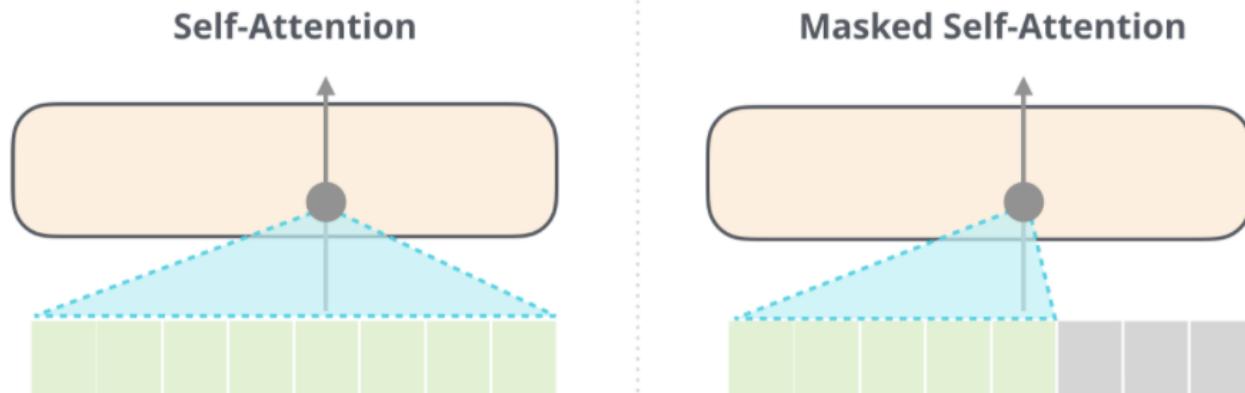
GPT-2 (a Transformer)

- There is **no Attention** (since there is no **Transformer Encoder** to attend to). So, there is only **Self-Attention**.
- As it processes each word/token, it **masks** the “future” words and conditions on and attends to the previous words

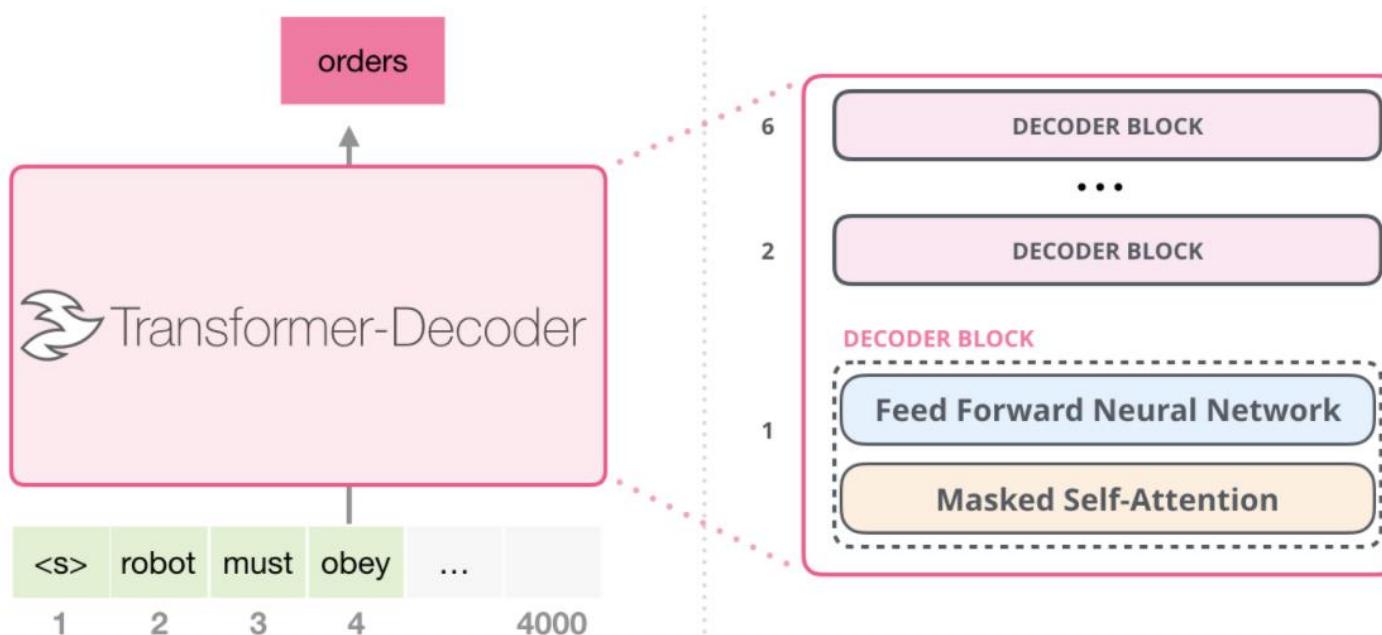


GPT-2 (a Transformer)

As it processes each word/token, it **masks** the “future” words and conditions on and attends to the previous words



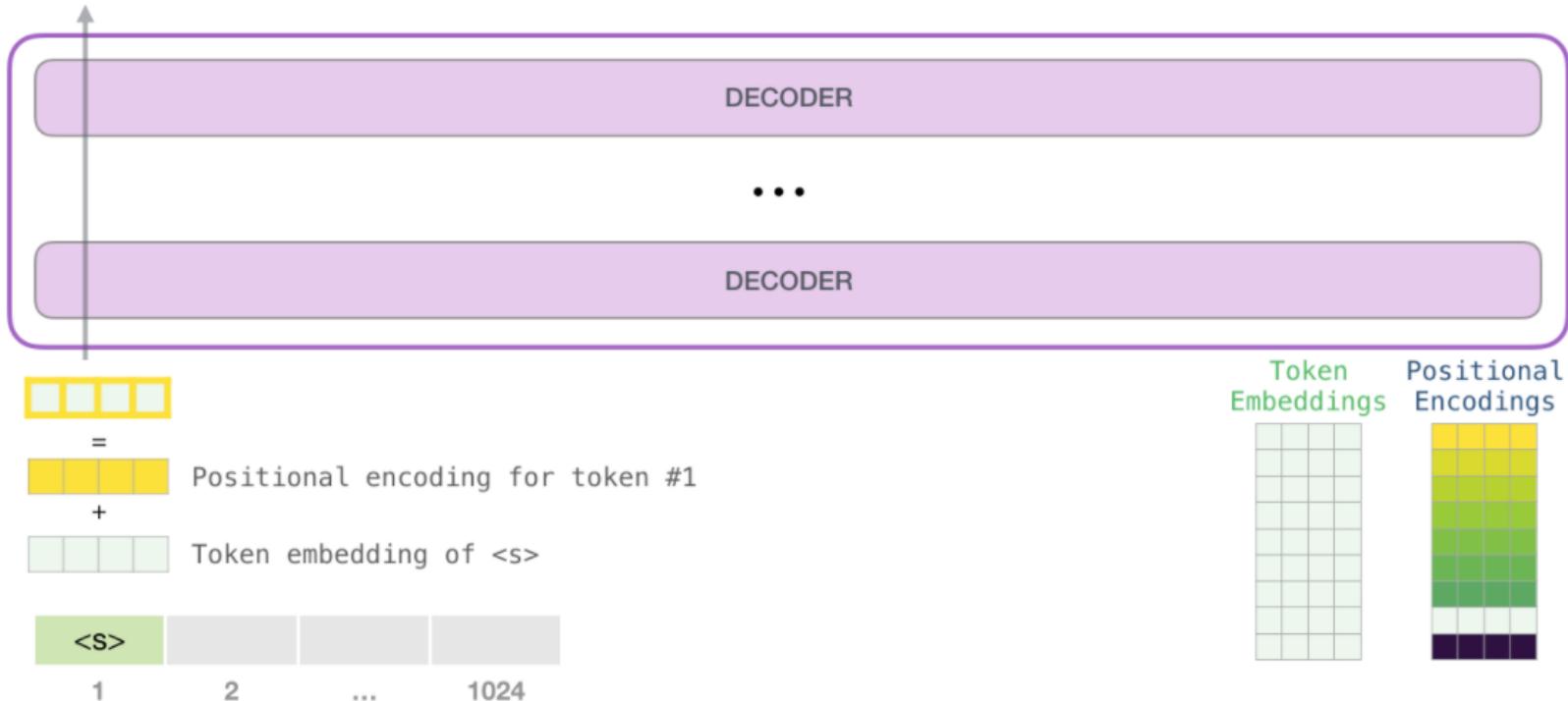
GPT-2 (a Transformer)



GPT-2 (a Transformer)

- Technically, it doesn't use words as input but **Byte Pair Encodings** (subwords), similar to BERT's WordPieces.
- Includes **positional embeddings** as part of the input, too.
- Easy to fine-tune on your own dataset (language)

GPT-2 (a Transformer)



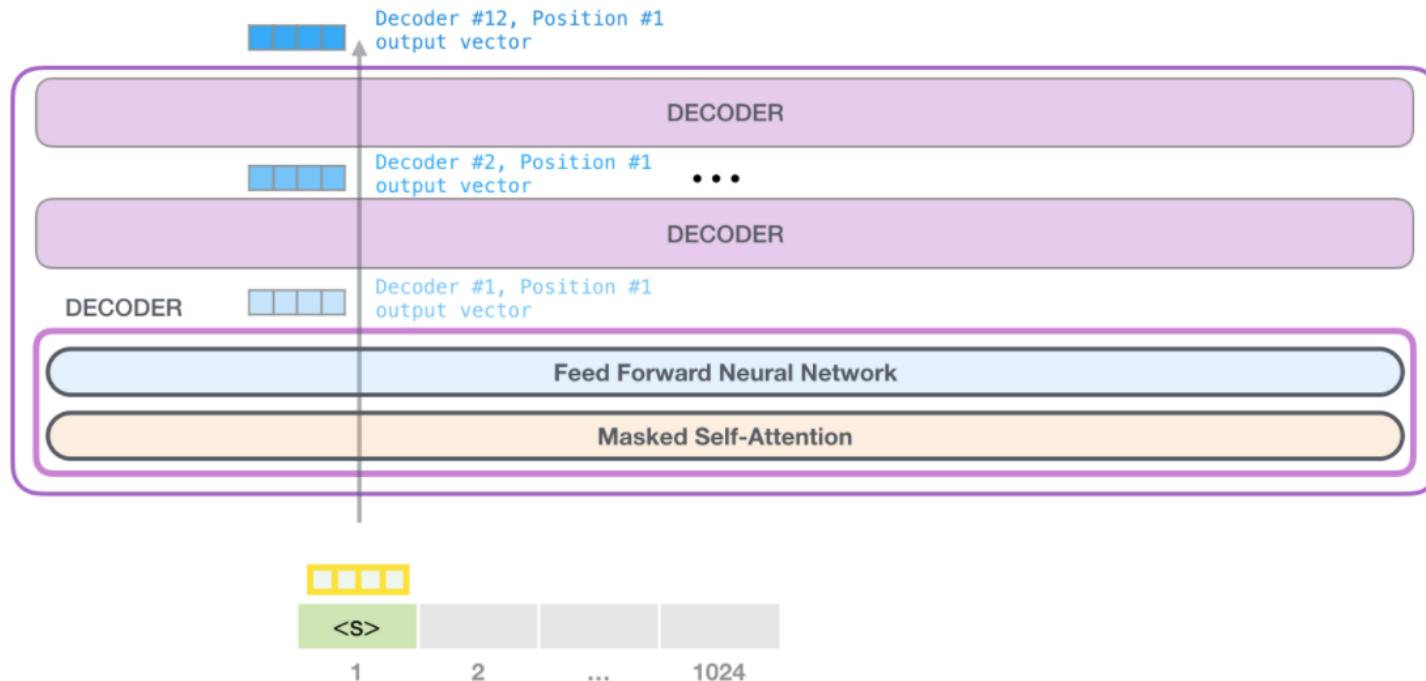
Byte Pair Encodings (BPE)

- Invented in 1994 ([Gage](#)) and updated in 2015 ([Sennrich et al.](#))
- Looks at the individual symbols (e.g., characters) and repeated merges the most frequent pairs (a la agglomerative clustering)
- Stop after **N** merges (you specify **N**). GPT uses **N =40k**

Philip Gage. 1994. A New Algorithm for Data Compression. *C Users J.*, 12(2):23–38, February

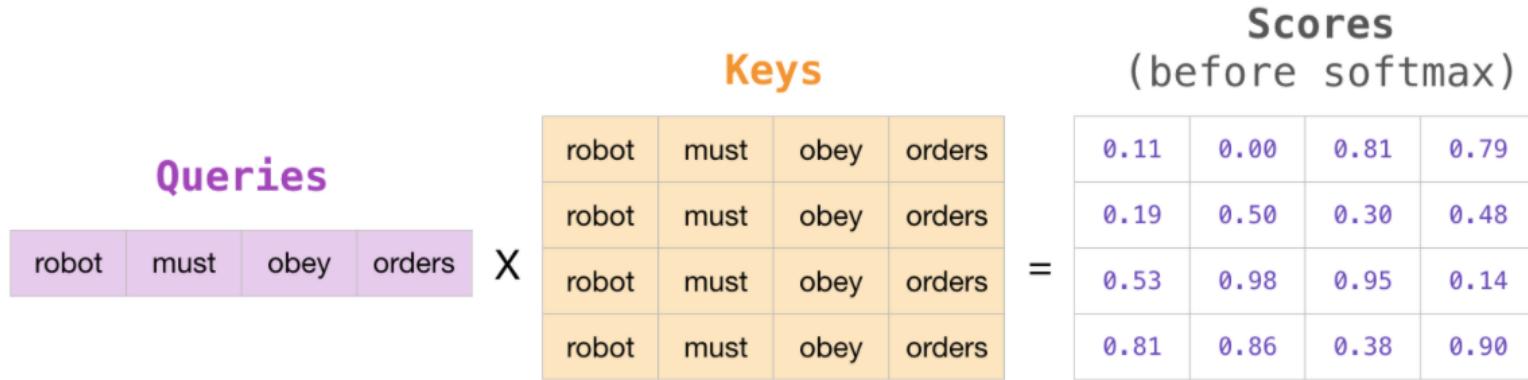
R. Sennrich, B. Haddow, and A. Birch. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*, 2015

GPT-2 (a Transformer)



GPT-2's Masked Attention

For efficiency, we can still calculate all query-key calculations with matrix multiplications, then mask before softmax'ing.



GPT-2's Masked Attention

For efficiency, we can still calculate all query-key calculations with matrix multiplications, then mask before softmax'ing.



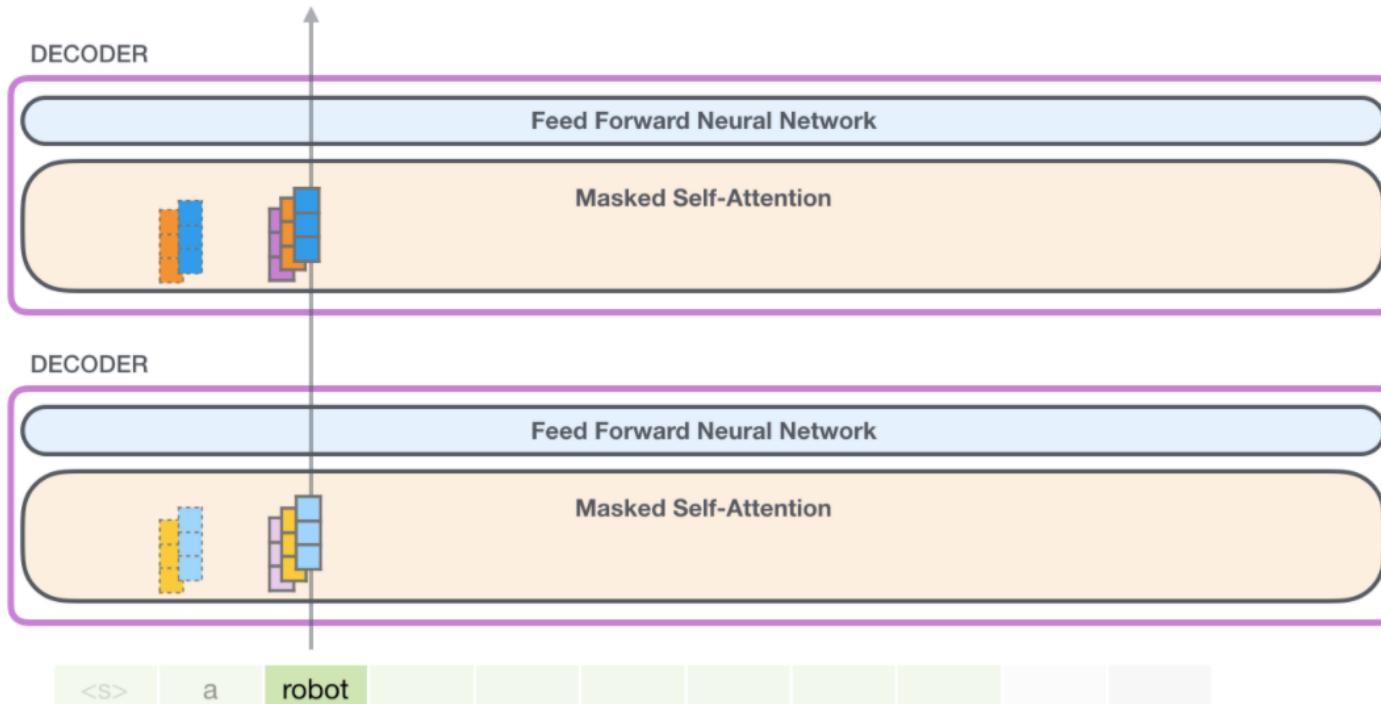
GPT-2's Masked Attention

For efficiency, we can still calculate all query-key calculations with matrix multiplications, then mask before softmax'ing.



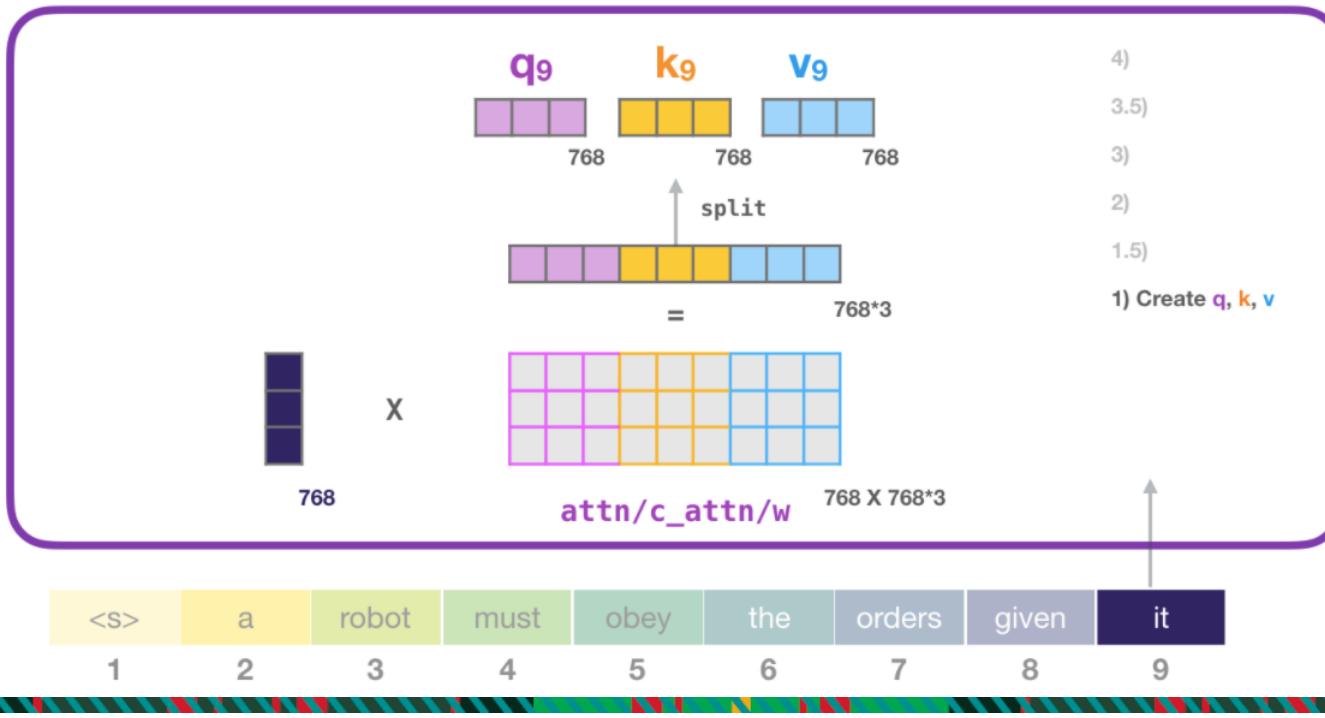
GPT-2's

Representations are propagated upwards through the network



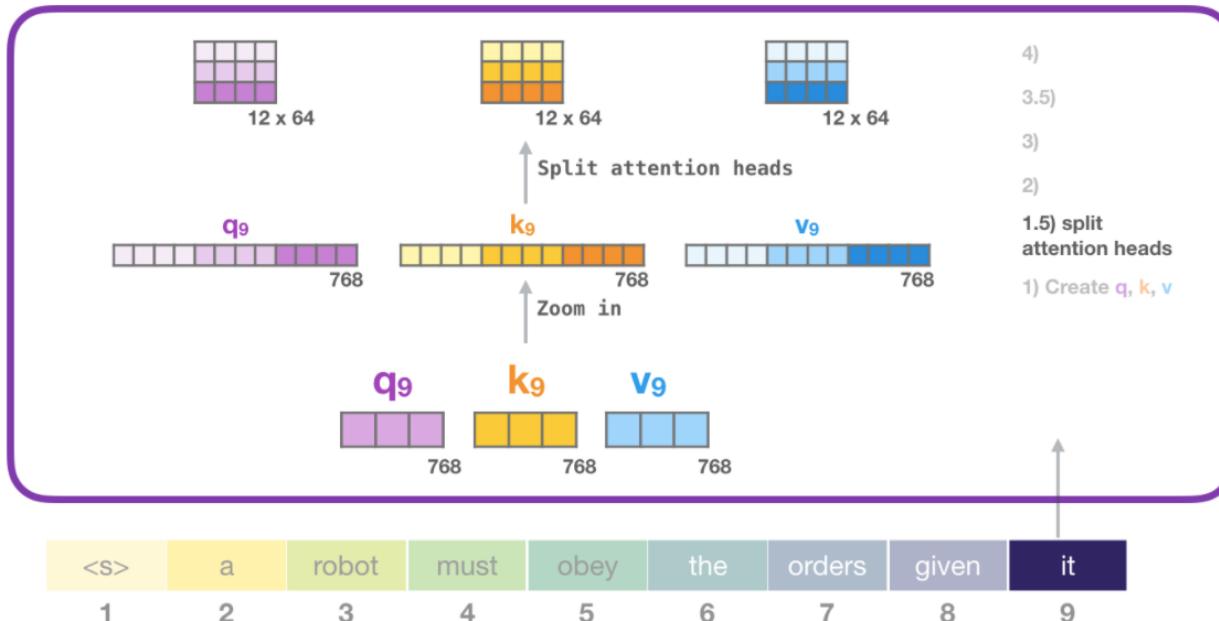
GPT-2's

Self-attention is otherwise identical to what we saw in BERT



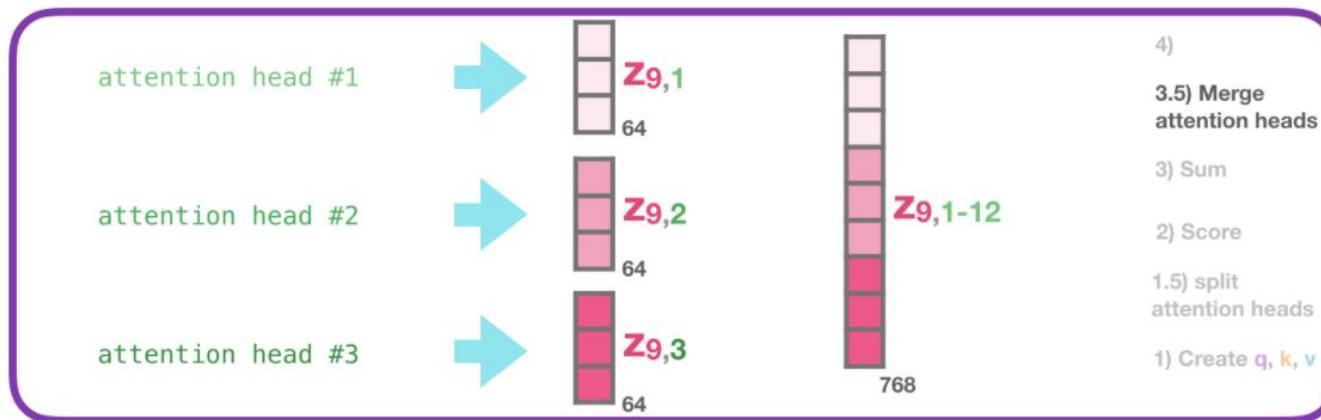
GPT-2's

Can have Multiple Self-Attention heads



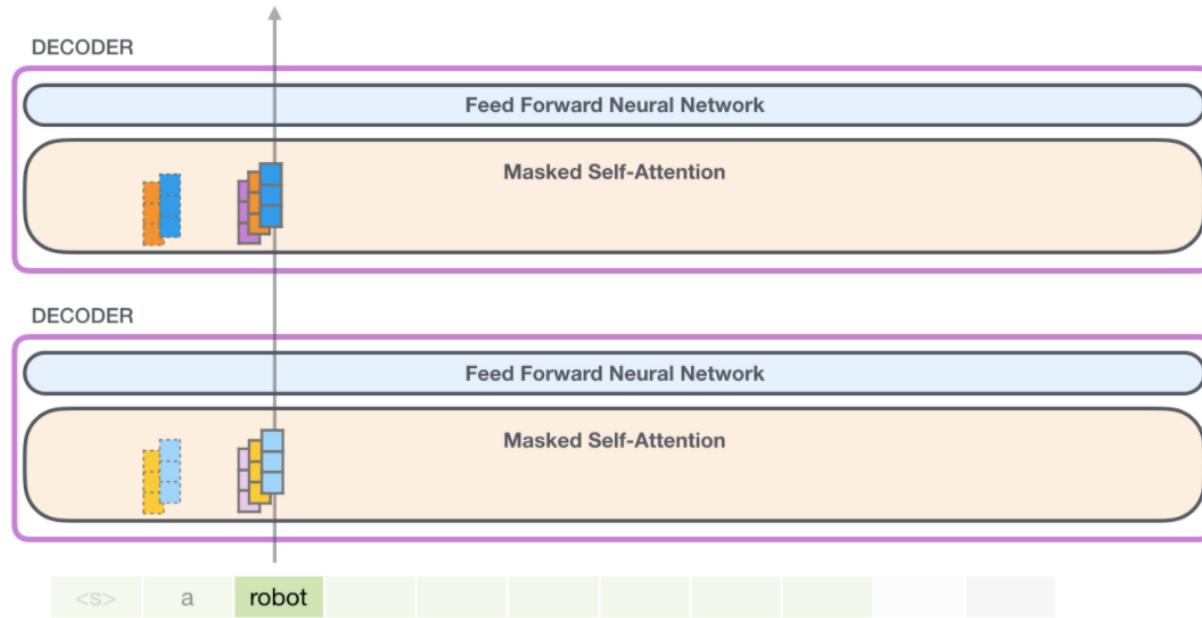
GPT-2's

Each Self-Attention head is responsible for exactly 1 resulting, output embedding



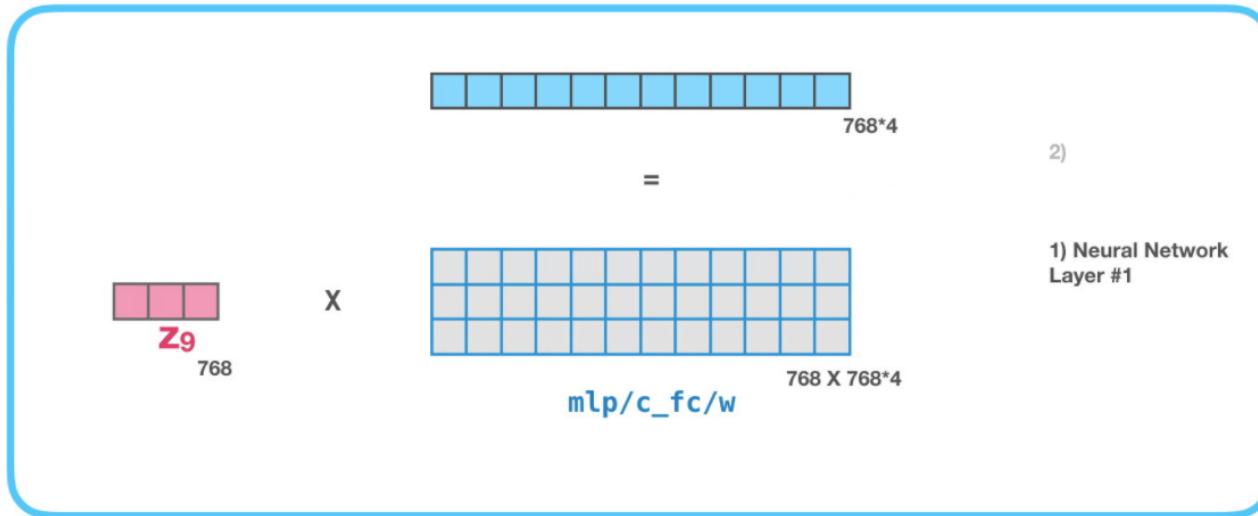
GPT-2's

Remember, these Masked Self-Attention layers are fed into a FFNN



GPT-2's

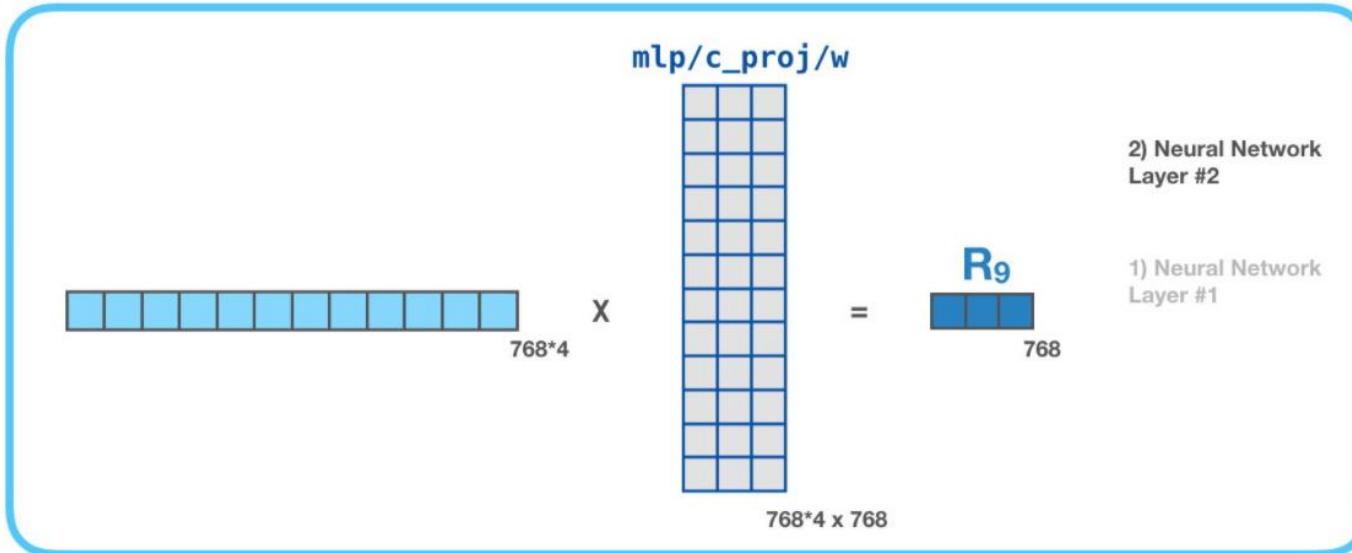
Remember, these Masked Self-Attention layers are fed into a FFNN



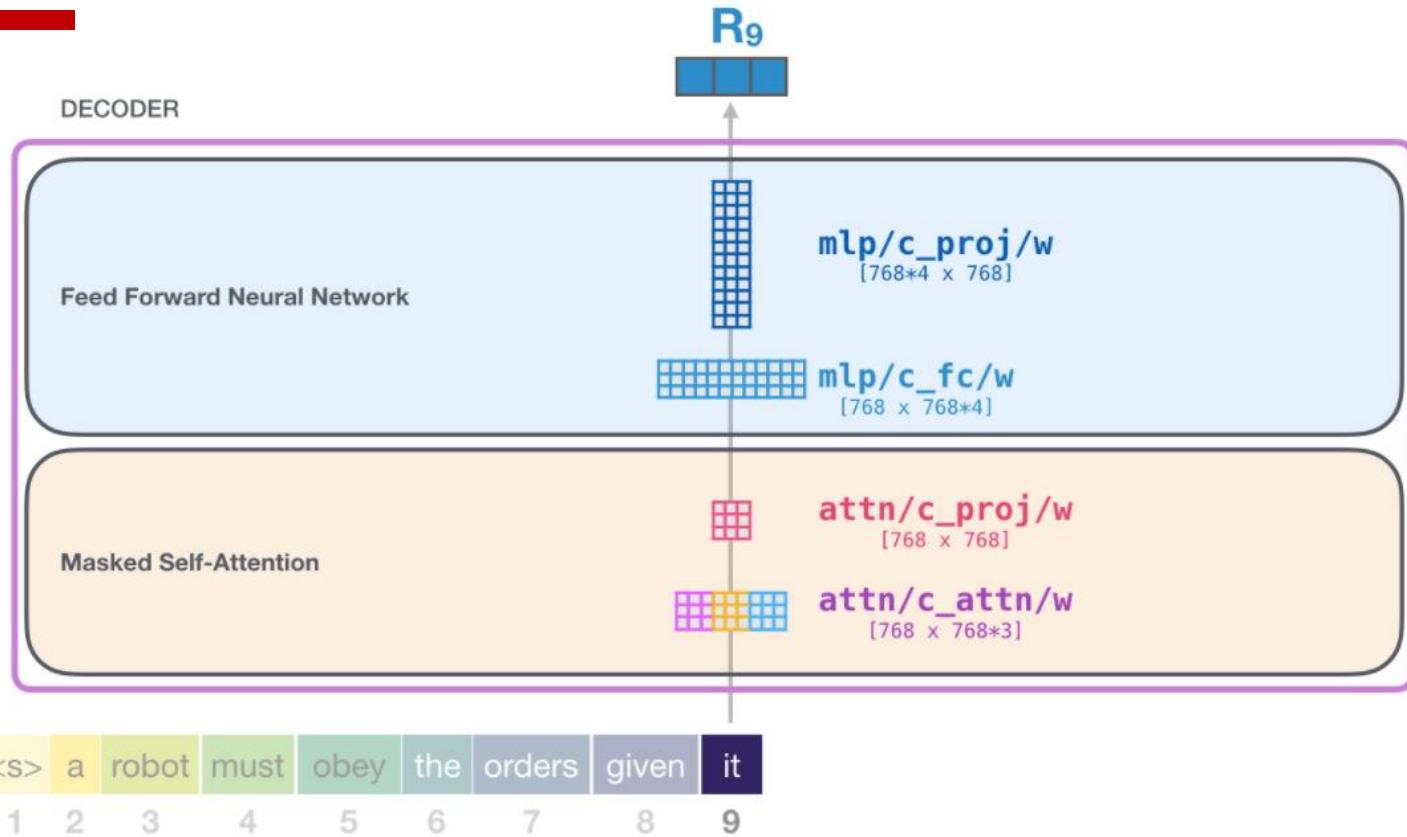
First hidden layer expands to 4x in size of the input

GPT-2's

2nd (final) layer of the FFNN projects it back to the original size



GPT-2's



Each Decoder block has its own weights (e.g., W_k , W_q , W_v)

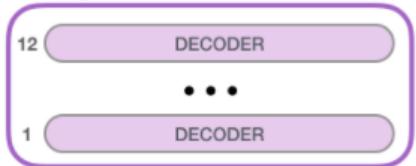
But the entire model only has 1 token-embedding weight matrix and positional encoding weight matrix. This helps all the blocks to work together and supplement their captured aspects



The authors of GPT-2 created 4 different version (sizes) of the model



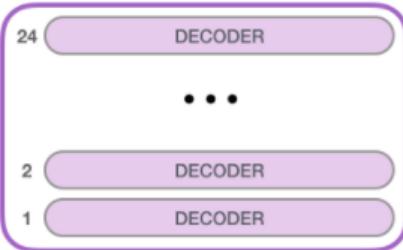
GPT-2
SMALL



Model Dimensionality: 768



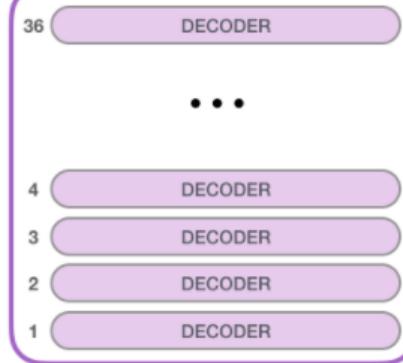
GPT-2
MEDIUM



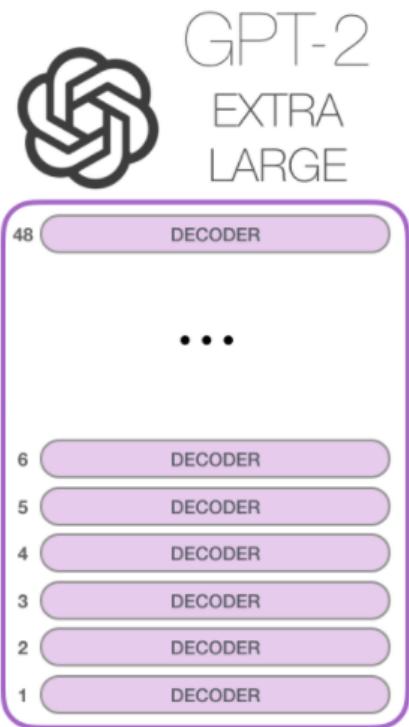
Model Dimensionality: 1024



GPT-2
LARGE



Model Dimensionality: 1280



Model Dimensionality: 1600

GPT-2 is identical to **GPT-1**, but:

- has Layer normalization in between each sub-block (as we've already seen)
- Vocab extended to 50,257 tokens and context size increased from 512 to 1024
- **Data:** 8 million docs from the web (Common Crawl), minus Wikipedia

Language Models are Unsupervised Multitask Learners

Alec Radford *¹ Jeffrey Wu *¹ Rewon Child¹ David Luan¹ Dario Amodei **¹ Ilya Sutskever **¹

You can finagle the system to yield synthetic predictions.

Children's Book Test (CBT) is a classification task. Fill-in-the-blank, and you predict which of the 10 possible choices is correct.

You can compute the probability of each choice + its ending.

You can finagle the system to yield synthetic predictions.

LAMBADA dataset tests model's ability to understand long-range dependencies.

Task: predict the final word of sentences which humans need 50+ tokens of context in order to accurately predict.

GPT-2 Results

Language Models are Unsupervised Multitask Learners

	LAMBADA (PPL)	LAMBADA (ACC)	CBT-CN (ACC)	CBT-NE (ACC)	WikiText2 (PPL)	PTB (PPL)	enwik8 (BPB)	text8 (BPC)	WikiText103 (PPL)
SOTA	99.8	59.23	85.7	82.3	39.14	46.54	0.99	1.08	18.3
117M	35.13	45.99	87.65	83.4	29.41	65.85	1.16	1.17	37.50
345M	15.60	55.48	92.35	87.1	22.76	47.33	1.01	1.06	26.37
762M	10.87	60.12	93.45	88.0	19.93	40.31	0.97	1.02	22.05
1542M	8.63	63.24	93.30	89.05	18.34	35.76	0.93	0.98	17.48

Table 3. Zero-shot results on many datasets. No training or fine-tuning was performed for any of these results. PTB and results are from (Gong et al., 2018). CBT results are from (Bajgar et al., 2016). LAMBADA accuracy result is from (Hoang and LAMBADA perplexity result is from (Grave et al., 2016). Other results are from (Dai et al., 2019).

You can finagle the system to yield synthetic predictions.

Summarization. The add the text “TL;DR:” after an article, then generate 100 tokens with top-2 random sampling, then extract the first 3 sentences.

GPT-2 Results

	R-1	R-2	R-L	R-AVG
Bottom-Up Sum	41.22	18.68	38.34	32.75
Lede-3	40.38	17.66	36.62	31.55
Seq2Seq + Attn	31.33	11.81	28.83	23.99
GPT-2 TL; DR:	29.34	8.27	26.58	21.40
Random-3	28.78	8.63	25.52	20.98
GPT-2 no hint	21.58	4.03	19.47	15.03

Table 4. Summarization performance as measured by ROUGE F1 metrics on the CNN and Daily Mail dataset. Bottom-Up Sum is the SOTA model from (Gehrmann et al., 2018)

GPT-2 Results

Question	Generated Answer	Correct	Probability
Who wrote the book the origin of species?	Charles Darwin	✓	83.4%
Who is the founder of the ubuntu project?	Mark Shuttleworth	✓	82.0%
Who is the quarterback for the green bay packers?	Aaron Rodgers	✓	81.1%
Panda is a national animal of which country?	China	✓	76.8%
Who came up with the theory of relativity?	Albert Einstein	✓	76.4%
When was the first star wars film released?	1977	✓	71.4%
What is the most common blood type in sweden?	A	✗	70.6%
Who is regarded as the founder of psychoanalysis?	Sigmund Freud	✓	69.3%
Who took the first steps on the moon in 1969?	Neil Armstrong	✓	66.8%
Who is the largest supermarket chain in the uk?	Tesco	✓	65.3%
What is the meaning of shalom in english?	peace	✓	64.0%
Who was the author of the art of war?	Sun Tzu	✓	59.6%
Largest state in the us by land mass?	California	✗	59.2%
Green algae is an example of which type of reproduction?	parthenogenesis	✗	56.5%
Vikram samvat calender is official in which country?	India	✓	55.6%
Who is mostly responsible for writing the declaration of independence?	Thomas Jefferson	✓	53.3%

GPT-2 Results

Easy to fine-tune on your own dataset (language)

SYSTEM PROMPT (HUMAN-WRITTEN)

A train carriage containing controlled nuclear materials was stolen in Cincinnati today. Its whereabouts are unknown.

MODEL COMPLETION (MACHINE-WRITTEN, FIRST TRY)

The incident occurred on the downtown train line, which runs from Covington and Ashland stations.

In an email to Ohio news outlets, the U.S. Department of Energy said it is working with the Federal Railroad Administration to find the thief.

“The theft of this nuclear material will have significant negative consequences on public and environmental health, our workforce and the economy of our nation,” said Tom Hicks, the U.S. Energy Secretary, in a statement. “Our top priority is to secure the theft and ensure it doesn’t happen again.” The stolen material was taken from the University of Cincinnati’s Research Triangle Park nuclear research site, according to a news release from Department officials.

The Nuclear Regulatory Commission did not immediately release any information.

Context (WebText test)

Pinky

The pink ghost's AI is designed to "feel" opposite of the red ghost's behavior. Pinky actually attempts to get out in front of Pac-Man. This is accomplished by setting the target 4 tiles ahead of Pac-Man's current location in the direction that Pac-Man is travelling. One exception to this is when Pac-Man is traveling up. Due to an overflow bug in the code, the calculation includes a left offset equal to the expected up offset.

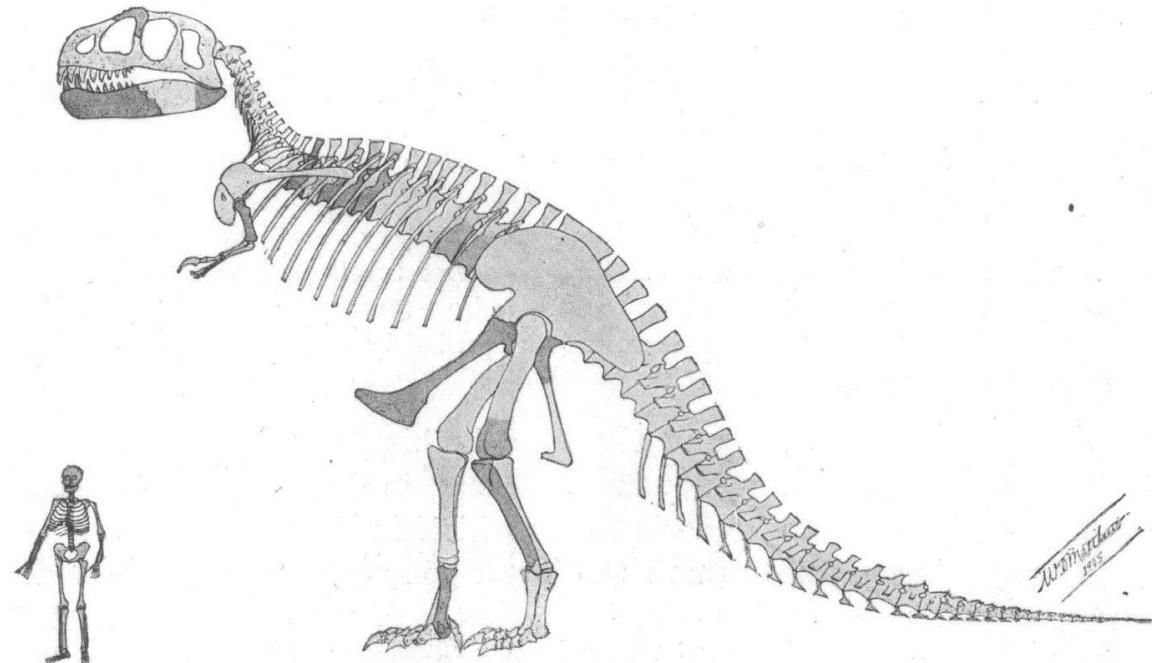
The use of Pac-Man's expected position allows the player to exploit the navigation code for Pinky. This was most often seen in situations where Pac-Man and Pinky are in close proximity to each other. For example, if they were headed toward one another, once they are less than 4 tiles apart, the target tile is now behind Pinky. If he encounters an intersection while in this configuration, he will attempt to double-back (by turning) in order to reach the target tile.

Player's would often perform what was referred to as a "head fake" by changing directions rapidly just as Pinky was approaching an intersection. By displacing the target tile just as Pinky

Pinky

was about to cross the intersection, Pac-Man would be able to dodge the ghosts' projectiles and return to the safe location of the safe house. Buster The yellow ghost's AI is specifically designed to try to avoid Pac-Man from leaving the safe house. The player must press the SPACE bar to reveal the safe house's doors, then go into them to find the yellow ghost hiding in the next room. The player must also note that the ghost will not attempt to exit the safe house if it sees Pac-Man already there. In fact, the ghost will not attempt an exit in the first place unless they're already in the safe house. These two behaviors are based on the fact that the player does not always attempt to reach the safe house before Pac-Man (or Blinky). The player must also beware of this glitch if a ghost (or another player) is already in the safe house and attempting to get out as soon as Pac-Man or Blinky is gone.

Comparison between GPT2 and GPT3



GPT-2
1.5B Parameters

GPT-3
175B Parameters

GPT-3 (Brown et al. 2020)

Language Models are Few-Shot Learners

Tom B. Brown*

Benjamin Mann*

Nick Ryder*

Melanie Subbiah*

Jared Kaplan[†]

Prafulla Dhariwal

Arvind Neelakantan

Pranav Shyam

Girish Sastry

Amanda Askell

Sandhini Agarwal

Ariel Herbert-Voss

Gretchen Krueger

Tom Henighan

Rewon Child

Aditya Ramesh

Daniel M. Ziegler

Jeffrey Wu

Clemens Winter

Christopher Hesse

Mark Chen

Eric Sigler

Mateusz Litwin

Scott Gray

Benjamin Chess

Jack Clark

Christopher Berner

Sam McCandlish

Alec Radford

Ilya Sutskever

Dario Amodei

OpenAI



GPT-3 various architectures

Model Name	n_{params}	n_{layers}	d_{model}	n_{heads}	d_{head}	Batch Size	Learning Rate
GPT-3 Small	125M	12	768	12	64	0.5M	6.0×10^{-4}
GPT-3 Medium	350M	24	1024	16	64	0.5M	3.0×10^{-4}
GPT-3 Large	760M	24	1536	16	96	0.5M	2.5×10^{-4}
GPT-3 XL	1.3B	24	2048	24	128	1M	2.0×10^{-4}
GPT-3 2.7B	2.7B	32	2560	32	80	1M	1.6×10^{-4}
GPT-3 6.7B	6.7B	32	4096	32	128	2M	1.2×10^{-4}
GPT-3 13B	13.0B	40	5140	40	128	2M	1.0×10^{-4}
GPT-3 175B or “GPT-3”	175.0B	96	12288	96	128	3.2M	0.6×10^{-4}

Table 2.1: Sizes, architectures, and learning hyper-parameters (batch size in tokens and learning rate) of the models which we trained. All models were trained for a total of 300 billion tokens.

Training compute

List price of compute to train
GPT-3 175B: ~\$4.5M

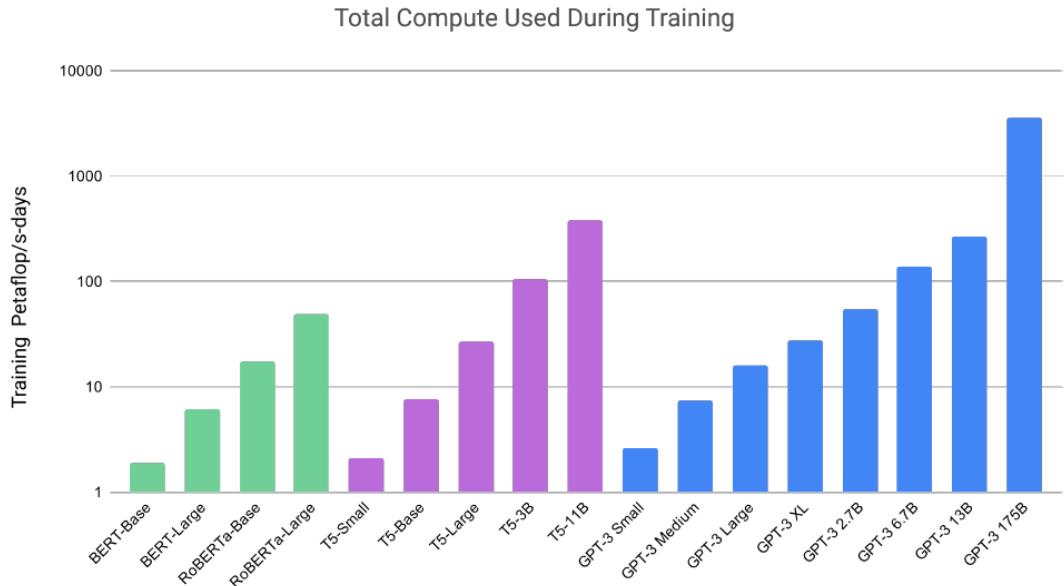
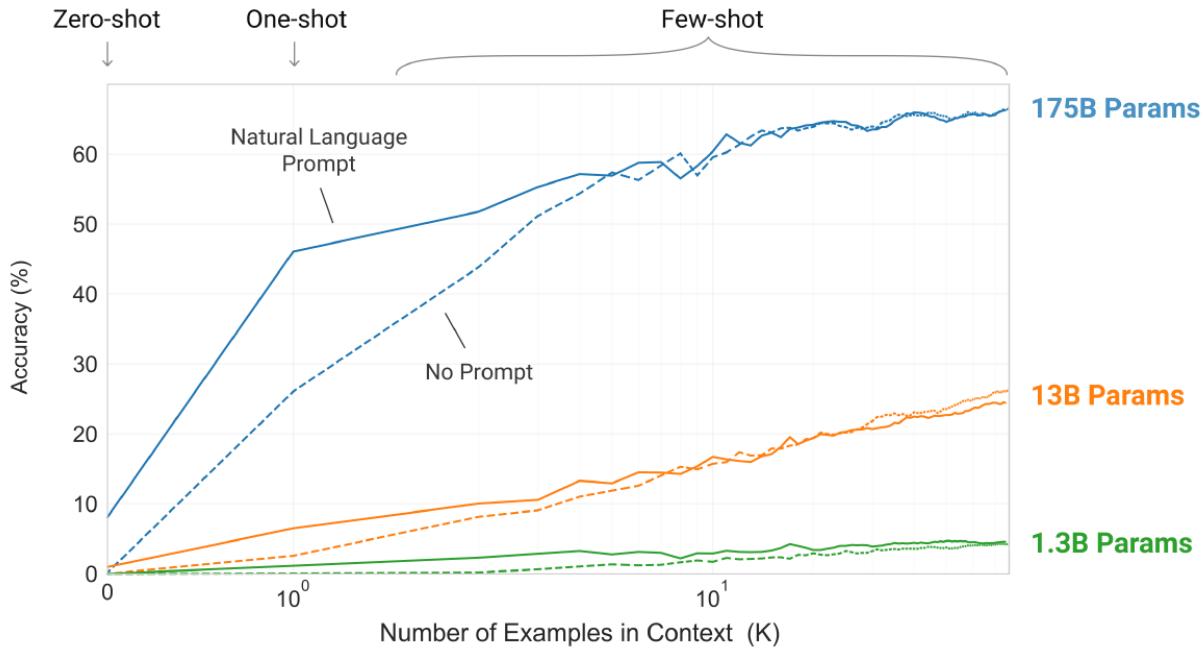


Figure 2.2: Total compute used during training. Based on the analysis in Scaling Laws For Neural Language Models [KMH⁺20] we train much larger models on many fewer tokens than is typical. As a consequence, although GPT-3 3B is almost 10x larger than RoBERTa-Large (355M params), both models took roughly 50 petaflop/s-days of compute during pre-training. Methodology for these calculations can be found in Appendix D.

GPT3 performance



To Sum Up

1. GPT 3 WITH 175 B PARAMETERS IS CAPABLE OF LEARNING CONTEXT
2. GPT3 IS USED IN CHATGPT WITH POWERFUL CAPABILITY OF PERFORMING DIFFERENT TASKS