CENL

Image-Language Models in the Library The Case for CLIP

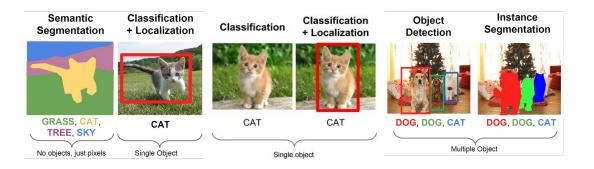




Horace Lee UNIVERSITY OF horace.lee@eng.ox.ac.uk

Computer Vision

- Ability to "see" an image and understand the content.
- Trivial for a human being, even for small children
 - A person can describe the content of a photograph that he has seen once.
 - A person can summarize a video that he has only seen once.
 - A person can recognize a face that he has only seen once before.







Images are matrices



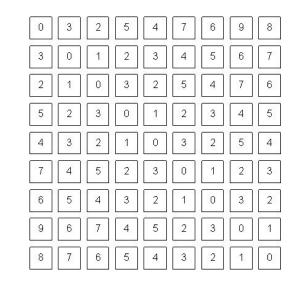


Image-Language Models in the Library: The Case for CLIP Javier de la Rosa, Jean-Philippe Moreaux, and Horace Lee

NL







Images are matrices

{BnF Bibliothèque nationale de France



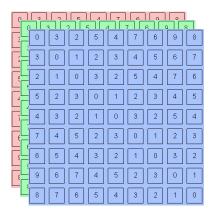
0	2	15	0	0	11	10	0	0	0	0	9	9	0	0	0
0	0	0	4	60	157	236	255	255	177	95	-61	32	0	0	29
0	10	16	119	238	255	244	245	243	250	249	255	222	103	10	0
0	14	170	255	255	244	254	255	253	245	255	249	253	251	124	1
2	98	255	228	255	251	254	211	141	116	122	215	251	238	255	49
13	217	243	255	155	33	226	52	2	0	10	13	232	255	255	36
16	229	252	254	49	12	0	0	7	7	0	70	237	252	235	62
6	141	245	255	212	25	11	9	3	0	115	236	243	255	137	0
0	87	252	250	248	215	60	0	1	121	252	255	248	144	6	0
0	13	113	255	255	245	255	182	181	248	252	242	208	36	0	19
1	0	5	117	251	255	241	255	247	255	241	162	17	0	7	0
0	0	0	4	58	251	255	246	254	253	255	120	11	0	1	0
0	0	4	97	255	255	255	248	252	255	244	255	182	10	0	4
0	22	206	252	246	251	241	100	24	113	255	245	255	194	9	0
0	111	255	242	255	158	24	0	0	6	39	255	232	230	56	0
0	218	251	250	137	7	11	0	0	0	2	62	255	250	125	3
0	173	255	255	101	9	20	0	13	3	13	182	251	245	61	0
0	107	251	241	255	230	98	55	19	118	217	248	253	255	52	4
0	18	145	250	255	247	255	255	255	249	255	240	255	129	0	5
0	0	23	113	215	255	250	248	255	255	248	248	118	14	12	0
0	0	6	1	0	52	153	233	255	252	147	37	0	0	4	1
0	0	5	5	0	0	0	0	0	14	1	0	6	6	0	0

0	2	15	0	0	11	10	0	0	0	0	9	9	0	0	0	
0	0	0	- 4	60	157	236	255	255	177	95	61	32	0	0	29	
0	10	16	119	238	255	244	245	243	250	249	255	222	103	10	0	
0	14	170	255	255	244	254	255	253	245	255	249	253	251	124	1	
2	98	255	228	255	251	254	211	141	116	122	215	251	238	255	49	
13	217	243	255	155	33	226	52	2	0	10	13	232	255	255	36	
16	229	252	254	49	12	0	0	7	7	0	70	237	252	235	62	
6	141	245	255	212	25	11	9	3	0	115	236	243	255	137	0	
0	87	252	250	248	215	60	0	1	121	252	255	248	144	6	0	
0	13	113	255	255	245	255	182	181	248	252	242	208	36	0	19	
1	0	5	117	251	255	241	255	247	255	241	162	17	0	7	0	
0	0	0	4	58	251	255	246	254	253	255	120	11	0	1	0	
0	0	4	97	255	255	255	248	252	255	244	255	182	10	0	4	
0	22	206	252	246	251	241	100	24	113	255	245	255	194	9	0	
0	111	255	242	255	158	24	0	0	6	39	255	232	230	56	0	
0	218	251	250	137	7	11	0	0	0	2	62	255	250	125	3	
0	173	255	255	101	9	20	0	13	3	13	182	251	245	61	0	
0	107	251	241	255	230	98	55	19	118	217	248	253	255	52	4	
0	18	146	250	255	247	255	255	255	249	255	240	255	129	0	5	
0	0	23	113	215	255	250	248	255	255	248	248	118	14	12	0	
0	0	6	1	0	52	153	233	255	252	147	37	0	0	4	1	
0	0	5	5	0	0	0	0	0	14	1	0	6	6	0	0	
	8 - 18 A	1	5 E	100	0.0			2.00			1.7	1.5	100	1		



Color images are tensors





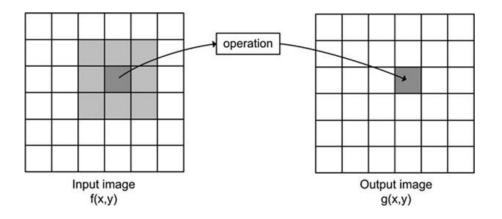
channel x height x width

Channels are usually RGB: Red, Green, and Blue Other color spaces: HSV, HSL, LUV, XYZ, Lab, CMYK, etc.

Al-lab National Library of Norway **{BnF** Bibliothèque nationale de France



We can operate on them



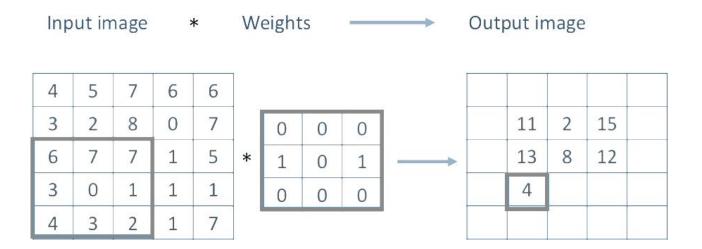


{BnF Bibliothèque nationale de France





We can operate on them



Al-lab National Library of Norway



Image-Language Models in the Library: The Case for CLIP Javier de la Rosa, Jean-Philippe Moreaux, and Horace Lee

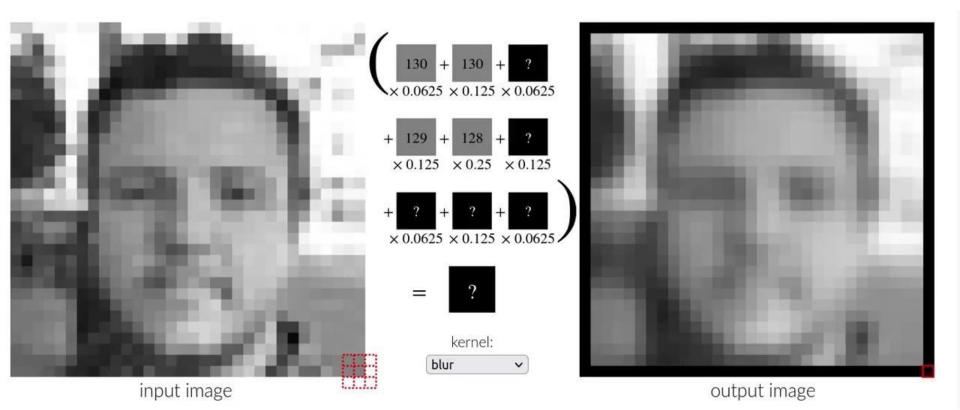
F

Convolutions

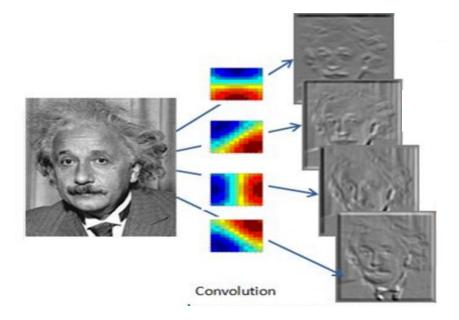
244 160 137 244 254 255 254 255 137 103 208 227 154 152 236 162 75 55 88 172 255 254 254 255 235 255 256 254 255 254 253 244 182 101 153 77 100 249 255 255 255 110 04 85 83 40 48 00 56 48 48 47 57 139 212 253 255 255 255 255 184 174 222 01 109 07 143 222 255 255 255 255 255 10 70 45 40 56 30 31 40 48 48 48 48 81 117 147 213 252 254 255 248 210 148 255 254 69 71 108 195 736 255 255 104 81 09 40 54 26 31 59 87 55 37 59 56 85 100 141 230 242 247 249 255 255 255 255 58 55 50 134 217 211 255 231 54 19 32 38 30 30 49 76 82 78 72 87 80 56 89 90 135 227 208 159 253 246 249 255 81 61 58 77 223 255 255 107 17 53 75 58 61 105 137 158 178 178 178 155 135 52 48 78 185 218 205 254 221 232 255 47 47 51 56 146 255 278 59 45 81 127 143 157 166 169 159 176 176 176 174 174 169 109 56 82 208 237 255 244 249 255 45 45 38 41 81 245 170 37 67 109 137 143 148 159 168 171 175 176 179 181 189 180 170 159 17 48 185 255 254 255 254 255 42 48 48 58 71 250 157 40 71 127 141 140 150 159 148 172 174 175 179 188 191 185 177 187 118 54 135 255 254 250 254 255 19 49 55 67 116 217 180 56 114 136 138 141 151 161 133 135 175 171 174 180 183 182 187 180 135 138 67 139 254 252 224 249 255 39 41 56 76 73 187 155 85 129 132 142 152 157 158 170 176 175 176 188 190 147 182 184 179 153 83 146 250 254 215 247 255 37 43 56 58 150 250 125 59 137 150 130 130 140 145 145 145 146 170 177 184 183 182 182 180 177 101 154 241 245 255 254 254 45 38 74 728 211 238 134 87 139 103 103 103 103 13 102 132 155 167 159 123 107 119 140 152 187 188 103 132 228 253 255 251 63 83 136 107 178 247 123 82 100 90 110 117 102 81 93 144 188 175 124 97 121 150 144 158 167 82 59 220 226 145 226 214 143 177 145 230 200 231 148 52 113 53 17 54 83 85 80 136 189 187 133 80 54 90 136 142 198 90 79 100 245 234 240 249 126 144 148 184 294 212 198 85 131 128 135 131 124 107 108 138 188 294 364 127 125 145 144 148 185 100 119 228 232 178 214 214 BB 112 101 B1 B2 B4 67 76 140 145 146 150 136 123 118 146 180 187 190 172 171 190 195 187 205 125 180 218 218 147 196 194 85 84 109 133 130 107 44 79 130 140 143 156 138 109 127 141 197 197 183 189 187 187 187 184 187 187 140 215 249 241 237 235 71 79 79 115 98 75 49 107 125 138 149 152 123 90 110 147 182 191 171 180 193 195 199 205 206 183 246 254 254 254 254 254 14 48 85 82 50 55 51 76 125 145 145 146 110 102 78 50 112 130 165 163 156 204 201 200 215 244 251 241 245 248 58 26 71 75 82 81 50 75 115 125 141 158 145 122 104 118 153 184 100 159 188 203 189 202 212 192 175 184 108 187 182 102 87 53 T0 80 51 T0 47 53 100 125 130 144 111 00 110 140 145 146 170 170 100 100 100 100 171 145 148 177 187 187 187 84 78 93 81 57 81 42 51 80 119180 134 85 78 109143 180 148 120 122 177 194 194 195 178 148 145 151 154 156 158 287 104 107 122 122 125 105 10 33 30 10 110 120 110 112 112 144 172 187 190 160 160 167 197 184 182 154 145 144 130 136 140 139 144 117 123 126 132 134 165 28 34 42 88 113 139 126 126 139 149 149 149 109 209 158 141 143 177 145 153 145 145 146 150 150 146 145 138 137 138 124 127 111 27 25 54 80 09 136 129 137 148 156 183 198 202 189 177 165 146 163 118 145 146 142 139 140 143 147 117 118 124 129 138 108 25 35 48 61 71 101 151 144 165 104 209 212 207 102 174 149 151 107 50 61 125 150 144 142 141 140 155 122 125 133 144 102 53 57 55 43 48 70 154 133 172 180 190 213 203 145 137 109 141 200 75 8 139 150 141 141 142 145 101 108 122 120 131 105 45 44 38 40 80 47 80 100 144 141 136 160 142 81 90 147 183 184 84 51 162 138 141 143 141 144 08 07 07 08 104 77 30 38 35 51 45 52 54 80 75 58 57 67 63 89 147 185 206 153 54 107 138 147 125 132 130 130 102 102 08 89 75 40 25 29 40 51 87 45 80 82 61 54 59 82 121 154 184 202 166 84 66 149 104 101 151 134 129 128



Convolutions



Convolutional Layer









AlexNet

ImageNet Classification with Deep Convolutional Neural Networks

Alex KrizhevskyIlya SutskeverUniversity of TorontoUniversity of Torontokriz@cs.utoronto.cailya@cs.utoronto.ca

Geoffrey E. Hinton University of Toronto hinton@cs.utoronto.ca

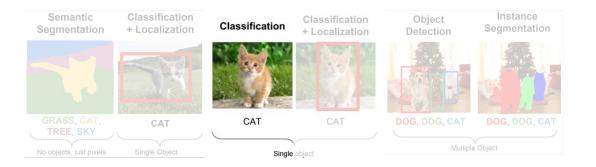
the paper that started the deep learning revolution!







Image classification





{BnF Bibliothèque nationale de France



Image classification

Classify an image into **1000** possible classes:

Abyssinian cat, Bulldog, French Terrier, Cormorant, Chickadee, red fox, banjo, barbell, hourglass, knot, maze, viaduct, etc.



Trained on the ImageNet challenge dataset with ~1.2 million images



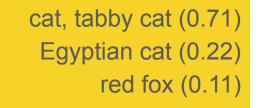


Image classification

Classify an image into **1000** possible classes:

Abyssinian cat, Bulldog, French Terrier, Cormorant, Chickadee, red fox, banjo, barbell, hourglass, knot, maze, viaduct, etc.





Trained on the ImageNet challenge dataset with ~1.2 million images





AlexNet

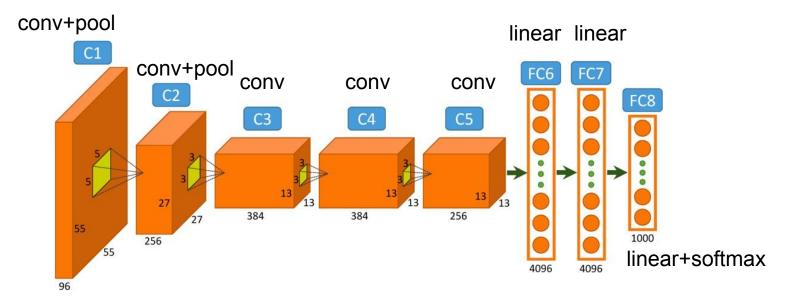


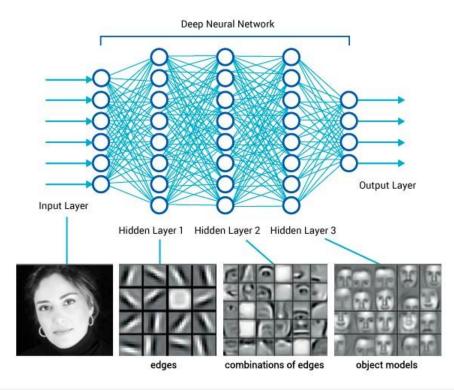




Image-Language Models in the Library: The Case for CLIP Javier de la Rosa, Jean-Philippe Moreaux, and Horace Lee

NL

But what is happening?



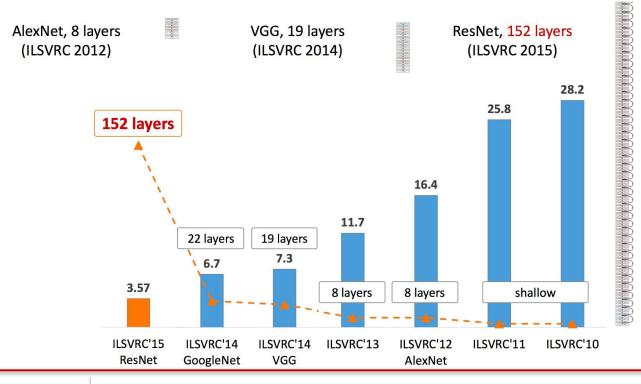


{BnF Bibliothèque nationale de France





Deeper networks \rightarrow more layers \rightarrow better performance



Al-lab National Library of Norway **{BnF** Bibliothèque nationale de France

BERT

Pre-trained language models (Devlin et al., 2019)

- Transformer-based
- Masked Language Modeling and Next Sentence Prediction
- Self-attention!



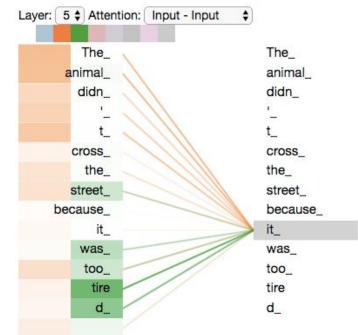






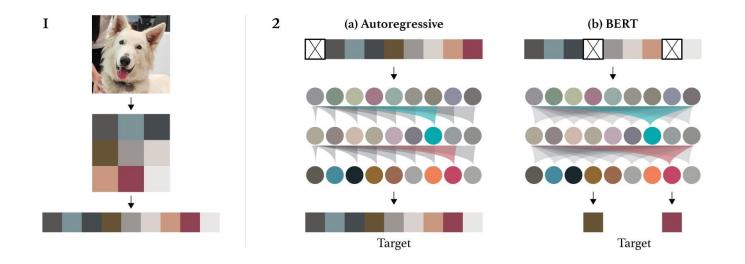
Pre-trained language models (Devlin et al., 2019)

- Transformer-based
- Masked Language Modeling and Next Sentence Prediction
- Self-attention!





Self-attention on pixels





{ BnF Bibliothèque nationale de France



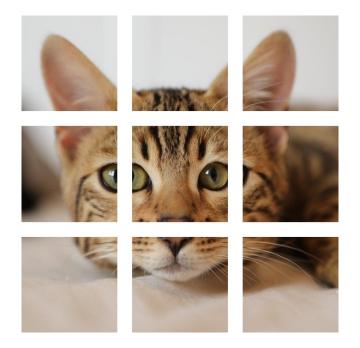


An Image is Worth 16x16 words, Dosovitskiy et al., ICLR 2021









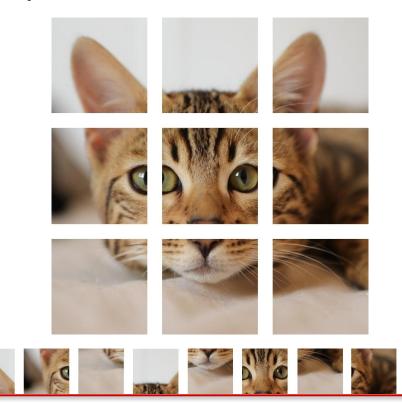
An Image is Worth 16x16 words, Dosovitskiy et al., ICLR 2021









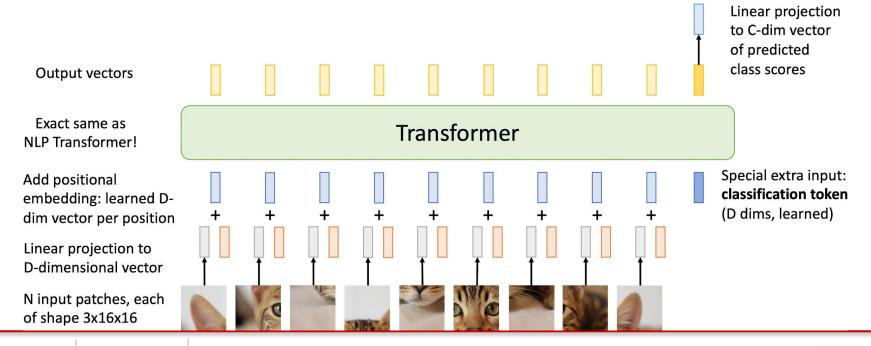


N input patches, each of shape 3x16x16





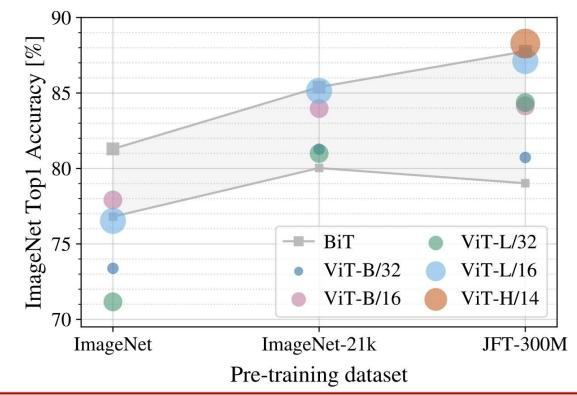




Al-lab National Library of Norway



Vision Transformers



{BnF Bibliothèque nationale de France

Text and Image Pairs



a red truck is parked on a street lined with trees



{BnF Bibliothèque nationale de France





OpenAl's CLIP: Contrastive language-image pretraining

- OpenAI collect 400 million (image, text) pairs from the web

- Image encoder + text encoder with a simple contrastive loss: given a collection of images and text, predict which (image, text) pairs actually occurred in the dataset

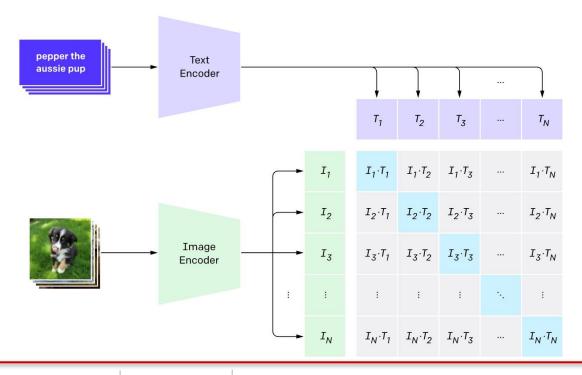






OpenAl's CLIP: Contrastive language-image pretraining

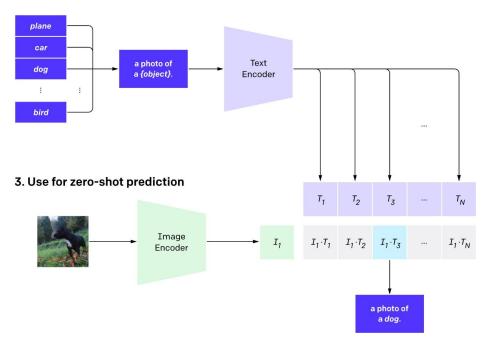
1. Contrastive pre-training







OpenAl's CLIP: Contrastive language-image pretraining



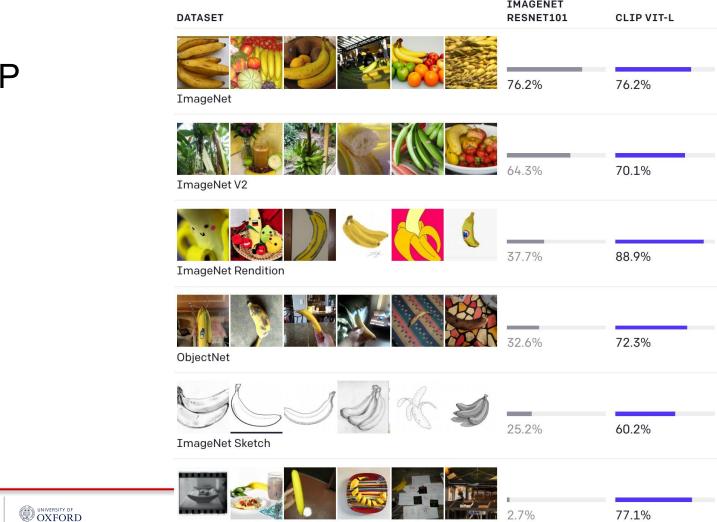
2. Create dataset classifier from label text











OpenAl's CLIP

{BnF Bibliothèque nationale de France

•••• Al-lab National Library of Norway

ImageNet Adversarial

AlexNet

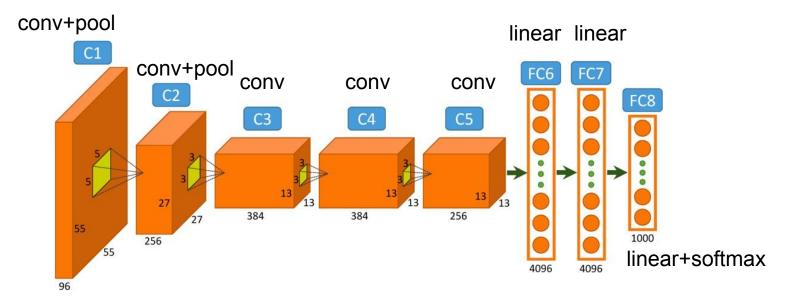
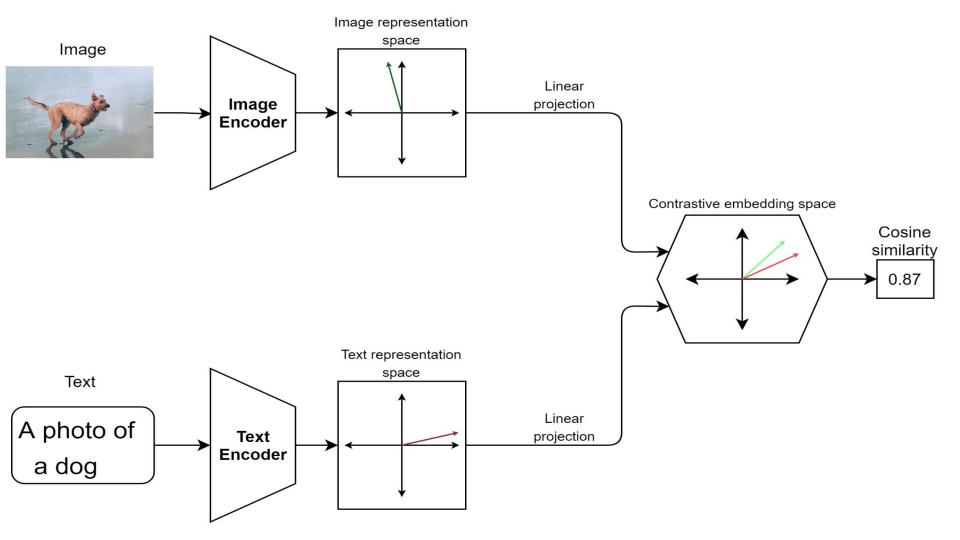




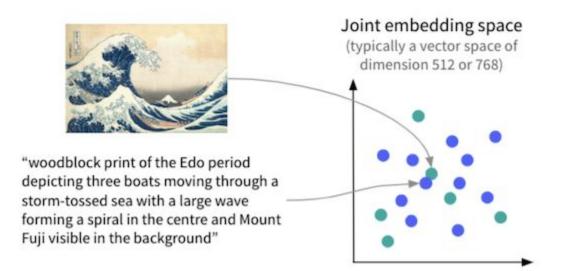


Image-Language Models in the Library: The Case for CLIP Javier de la Rosa, Jean-Philippe Moreaux, and Horace Lee

NL



Embedding Space











kids playing in the snow









kids playing in the snow









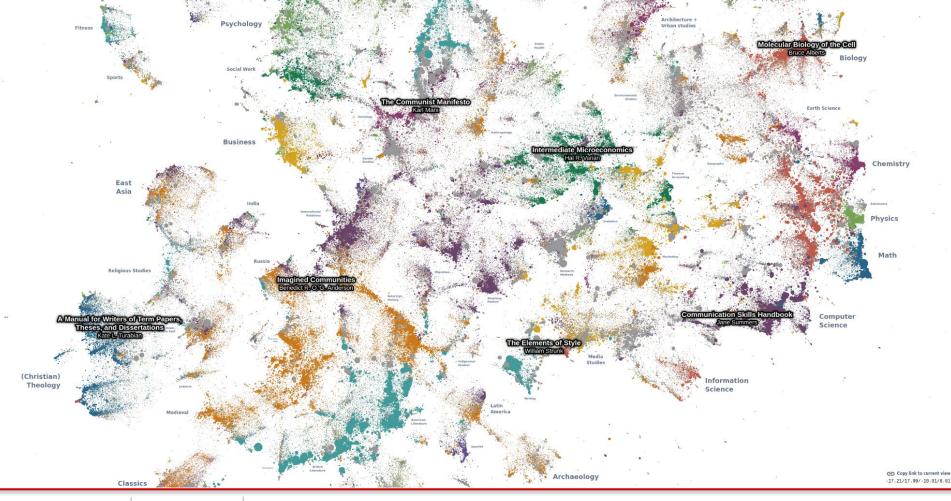


Image-Language Models in the Library: The Case for CLIP **CERL** Javier de la Rosa, Jean-Philippe Moreaux, and Horace Lee

Al-lab National Library of Norway





1. Find an image



- 1. Find an image
- 2. Get its numerical representation



{0.9, 0.12, 0.45, ...}

- 1. Find an image
- 2. Get its numerical representation
- 3. Compare to the numerical representation of all the other images

...}

- 1. Find an image
- 2. Get its numerical representation
- 3. Compare to the numerical representation of all the other images
- 4. Rank by the comparison metric



$$\{0.9, 0.12, 0.45, \ldots\} \xrightarrow{\left\{0.1, 0.23, 0.27, \ldots\right\} \rightarrow 0.1}_{\left\{0.2, 0.56, 0.87, \ldots\right\} \rightarrow 0.1}_{\left\{0.9, 0.78, 0.62, \ldots\right\} \rightarrow 0.1}_{\left\{0.9, 0.26, 0.79, \ldots\right\} \rightarrow 0.1}_{\left\{0.9, 0.26, 0.79, \ldots\right\} \rightarrow 0.1}_{\left\{0.9, 0.69, 0.16, \ldots\right\} \rightarrow 0.1}_{\left\{0.8, 0.82, 0.87, \ldots\right\} \rightarrow 0.1}_{\left\{0.5, 0.35, 0.10, \ldots\right\} \rightarrow 0.1}_{\left\{0.3, 0.72, 0.97, \ldots\right\} \rightarrow 0.1}$$

_ . . .

 $(\{0, 9, 0, 12, 0, 45, \} \rightarrow 0, 9$

- 1. Find an image
- 2. Get its numerical representation
- 3. Compare to the numerical representation of all the other images
- 4. Rank by the comparison metric



$$\{0.9, 0.12, 0.45, \ldots\} \rightarrow 0.9 \\ \{0.1, 0.23, 0.27, \ldots\} \rightarrow 0.1 \\ \{0.2, 0.56, 0.87, \ldots\} \rightarrow 0.2 \\ \{0.9, 0.78, 0.62, \ldots\} \rightarrow 0.2 \\ \{0.9, 0.29, 0.32, \ldots\} \rightarrow 0.4 \\ \{0.9, 0.26, 0.79, \ldots\} \rightarrow 0.4 \\ \{0.9, 0.69, 0.16, \ldots\} \rightarrow 0.1 \\ \{0.8, 0.82, 0.87, \ldots\} \rightarrow 0.1 \\ \{0.5, 0.35, 0.10, \ldots\} \rightarrow 0.2 \\ \{0.3, 0.72, 0.97, \ldots\} \rightarrow 0.4$$

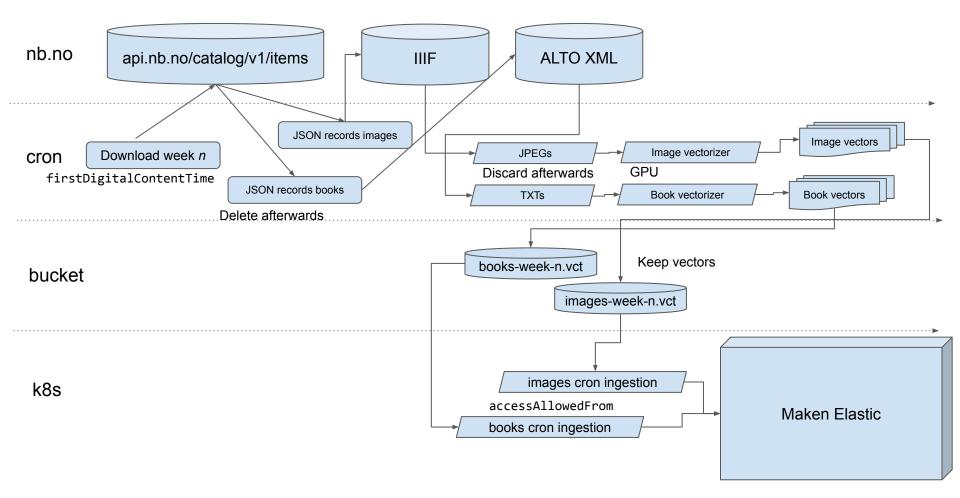


- 1. Find an image
- 2. Get its numerical representation
- 3. Compare to the numerical representation of all the other images
- 4. Rank by the comparison metric

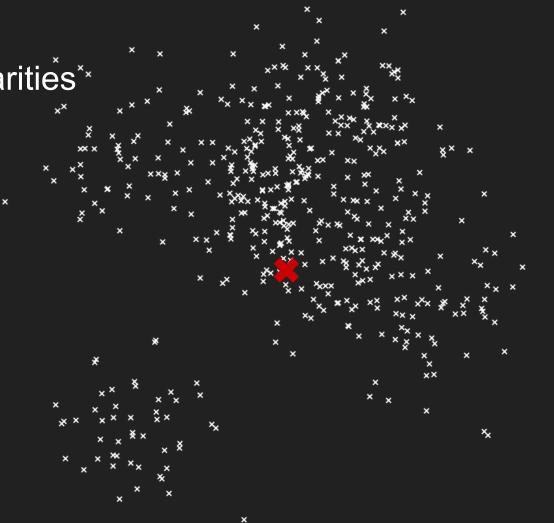
Lake Louise, Alberta, Canada Postkort fra Olden, Stryn Wilse, Anders Beer Luktvatn kommune, Sogn og Fjordane bilde Wilse, Anders Beer Mittet & Co. AS 1921 1925 001 Leden, Christian | 1909 Mer informasion Likhet: 🔵 🔵 🔵 🔘 🔿 Likhet: 💿 💿 💿 🛈 🔿 1900 Likhet: O O O O Romsdal. Romsdalsfjeldene Norge, Veblungsnes, Romsdalshorn og Vengetindene Mittet & Co. AS Mittet & Co. AS 1900 Mittet & Co. AS Likhet: O O O O Likhet: 💿 💿 💿 🛈 🔿 Likhet: O O O O Luktvatn Postkort fra Aurland kommune. Passasierskip i Eidfiorden? Sogn og Fjordane bilde 011 Mittet & Co. AS Normanns kunstforlag 1948 1950 Likhet: O O O O Likhet: 📀 🔵 🖸 🔿 Likhet: OOOOO Postkort fra Loenvatnet, Strvn Norge, Sommerkveld, Merok, Skiolden commune, Sogn og Fjordane bilde Geiranger Mittet & Co. AS 1939 015 Mittet & Co. AS Likhet: 💿 💿 💿 🛈 🔿

Lignende bilder:

https://github.com/NbAiLab/maken-es-data

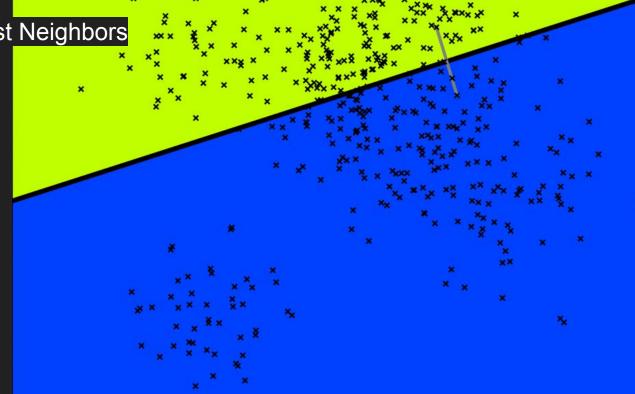


- Precise 🔽
- Slow 🗙



×

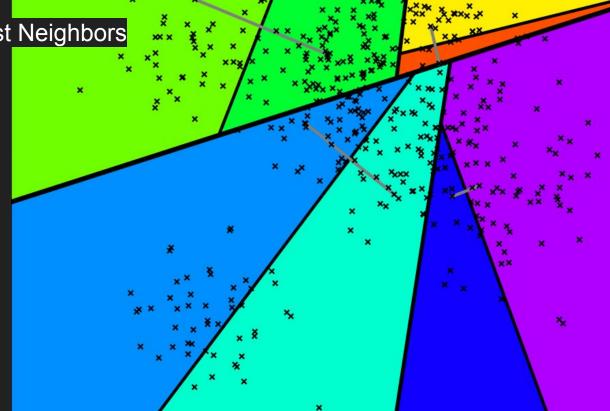
Approximate Nearest Neighbors



Approximate Nearest Neighbors

Ŷ

Approximate Nearest Neighbors

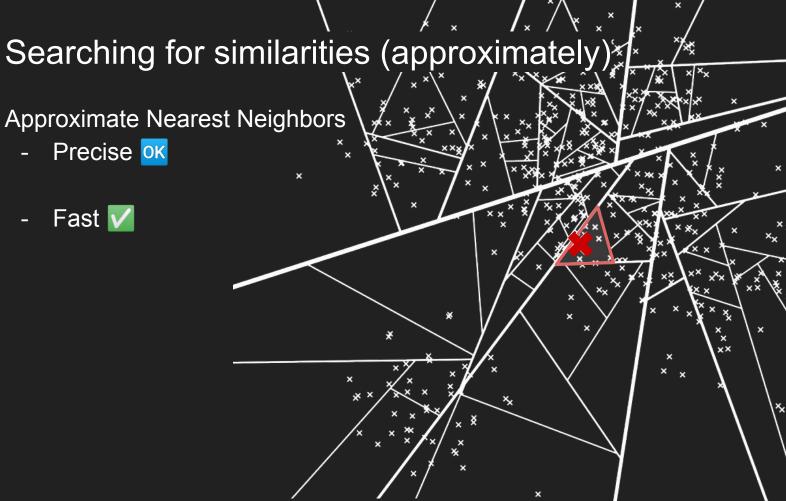


XX

××

×

Approximate Nearest Neighbors



×

Managing AI workflows for the BnF everyday life?

Commodification of AI

Target: in-house + DH (BnF Datalab users)

Why: low staffing, limited IT resources

What for: building data processing pipelines on top of AI models ("Lego way")

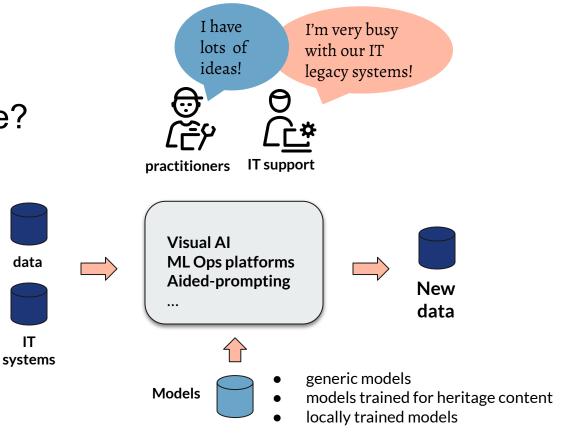


Image-Language Models in the Library: The Case for CLIP Javier de la Rosa, Jean-Philippe Moreaux, and Horace Lee





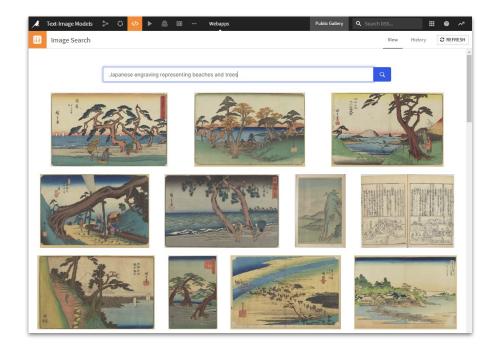
Dataiku DSS experiments

Use case #1: Gallica/classification

Semantic unsupervised image classification with CLIP model (OpenAI), using prompts:

- "japanese painting"
- "japanese ideograms"
- "book bindings"
- "blank pages"

on the Gallica Japanese engravings collection



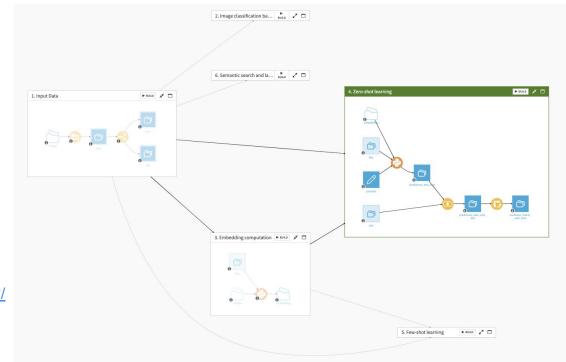
Al-lab National Library of Norway



Image-Language Models in the Library: The Case for CLIP Javier de la Rosa, Jean-Philippe Moreaux, and Horace Lee

Dataiku experiments

- Zero annotated data (we leverage CLIP "knowledge")
- Visual design of workflows (Dataiku)
- Accuracy **≃ 95%**





https://gallery.dataiku.com/projects/EX_CLIP/ (go to the </> Webapps menu)

Image-Language Models in the Library: The Case for CLIP Javier de la Rosa, Jean-Philippe Moreaux, and Horace Lee





Dataiku experiments

Use case #2: Gallica/classification

Newspaper illustrations types classification:

- GT: 9,500 tagged images from French newspapers (1910-1920)
- 4/10 classes ("prompts")
- Accuracy = 90%

Confusion matrix

GT\pred	Chart	Comic	Drawi	Engra	Games	Manus	Мар	Photo	Score
Chart	251	0	108	15	8	12	3	2	16
Comic	0	281	11	0	0	0	0	0	0
Drawi	3	23	1738	15	0	2	0	31	0
Engra	0	32	51	537	0	1	0	73	0
Games	0	0	0	41	198	15	0	20	2
Manus	0	0	4	1	0	80	0	2	1
Мар	6	0	7	0	0	0	292	2	2
Photo	11	11	73	54	1	0	3	2749	1
Score	0	0	2	0	0	0	0	0	136

Accuracy (micro average): 90.40

Accuracy (per classe): Chart / Comics / Drawing / Engraving / Games / Manuscript / Map / Photo / Score 60.48 % / 96.23 % / 95.92 % / 77.38 % / 71.74 % / 90.91 % / 94.50 % / 94.70 % / 98.55 %

- Ads, an monochrome illustrated ad printed in a heritage newspaper
- Chart, a chart or a diagram printed in a heritage newspaper
- Comics, a comics printed in a heritage newspaper
- Drawing, a monochrome drawing printed in a heritage newspaper
- Engraving, a color or grayscale old engraving printed in ...
- Games, a crossword grid or a chess game or a word game or a checkers game printed in a heritage newspaper
- Manuscript, a handwritten text
- Map, a monochrome map printed in a heritage newspaper
- Photo, a black and white picture printed in a heritage newspaper
- Score, a printed musical score printed in a heritage newspaper

$\leftrightarrow \rightarrow C$ (0) 127.0.0.1:5000	Q	Û	☆	ť	ø	æ	Ь	*		JP	÷
Use a free query OR one of these predefined classes: Drawing, Engraving, Photo, Text											
Photo											
SUBMIT											
						A Start Star	A A A A A A A A A A A A A A A A A A A				The second secon
				- ALCON		A CAR			が出る		
			North 1->	The second			E APP BA	T			1 mm

Image-Language Models in the Library: The Case for CLIP Javier de la Rosa, Jean-Philippe Moreaux, and Horace Lee





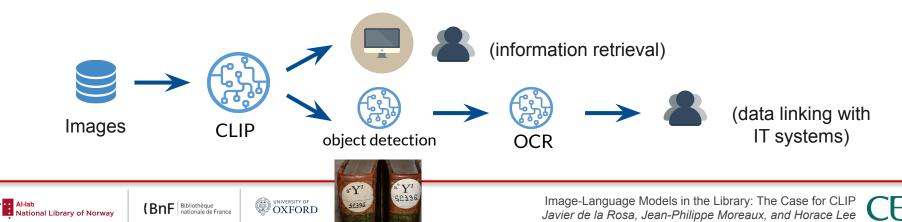
Other experiments

Use case #3: Conservation dpt

Semantic indexation of a photo bank from the BnF books restoration workshops (50k files)

Al-lab



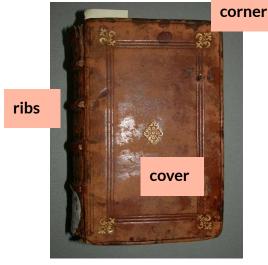


Other experiments

Use case #3: Conservation dpt

Evaluation on 500 images : **accuracy = 34%** (multi-label scenario...)

but effective for an IR task on a niche context (book restoration)



Confusi	ion mat	rix						and the second
GT\pred	Coins	Dos	Etiqu	Fleur	Illus	Nerfs	Papie	Plat
Coins	9	17	0	1	0	1	1	1
Dos	2	36	0	6	0	3	0	1
Etiqu	5	17	11	6	0	2	0	14
Fleur	2	18	0	31	0	7	6	13
Illus	9	2	1	12	19	1	3	47
Nerfs	0	46	1	0	0	26	4	1
Papie	3	4	0	0	0	2	21	4
Plat	7	30	1	5	0	11	4	11

Accuracy (micro average): 33.81

Accuracy (per classe):

Coins / Dos / Etiquette_rondage / Fleuron_decor / Illustration / Nerfs / Papier_decor / Plat 30.00 % / 75.00 % / 20.00 % / 40.26 % / 20.21 % / 33.33 % / 61.76 % / 15.94 %

(BnF Bibliothèque nationale de France OXFORD

Image-Language Models in the Library: The Case for CLIP Javier de la Rosa, Jean-Philippe Moreaux, and Horace Lee

Other experiments with CLIP

Use case #4: IR for the digital humanities

"Histoire du guartier Richelieu" project (INHA, ENC, BnF...)

- Multi-institutional data aggregation
- Visual analysis of small but heterogeneous corpora (engravings, maps, architectural drawings, ads, ...)

https://guartier-richelieu.fr/



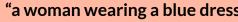


Image-Language Models in the Library: The Case for CLIP Javier de la Rosa, Jean-Philippe Moreaux, and Horace Lee





58396

WISE Image Search Engine (WISE) [10mins]

- Developed by Horace Lee, Prasanna Sridhar, Abhishek Dutta (Research Software Engineers at University of Oxford)
- An image search engine built around CLIP
 - Easy to use
 - Fast search speeds
 - Use with your own image collections
- Multimodal search (search with natural language, images, or a combination of both)

• Demo on 50+ million images from Wikimedia Commons



WISE Image Search Engine (WISE) [10mins]

- Code is available open source: <u>https://gitlab.com/vgg/wise/wise</u>
- Feel free to contact us if you would like to use WISE in your own organisation / research
 - {horacelee, prasanna, adutta} @ robots.ox.ac.uk







Thanks!

Questions?



Al-lab National Library of Norway **{BnF** Bibliothèque nationale de France

Image-Language Models in the Library: The Case for CLIP Javier de la Rosa, Jean-Philippe Moreaux, and Horace Lee