

Automated Image and Video Analysis: A practical introduction

Felicia Loecherbach

Automated Image and Video Data Analysis with Python
OPINION Training School Elbasan
May 29, 2025

Giving credit

- ▶ A lot (if not most) of the material I will be using for this workshop has been developed originally by Andreu Casas and Nora Webb Williams who came up with this course a few years ago
- ▶ Especially for the earlier parts (except multimodal modelling) they have a great book explaining things more in detail (**Images as Data for Social Science Research**) as well as a special issue "Images as Data" in Computational Communication Research (2022) with many interesting examples of how these methods can be used in the social sciences
- ▶ I will occasionally show results and insights into research projects that are not published yet, please treat them confidentially

Why do Images Matter?

People are more likely to pay attention to visuals

IMMIGRATION

How America Got to 'Zero Tolerance' on Immigration

Battles have raged within the White House over family separations, ICE raids and President Trump's obsession with a wall.

Together, they have remade homeland security.

15m ago 393 comments

Mr. Trump's approach follows a model from Europe and Australia, our Interpreter columnists write.

3h ago



Kirsten Luce

Dahmen (2012) "*Photographic Framing in the Stem Cell Debate*"

Why do Images Matter?

People are more likely to recall information learned through visuals



apple



banana



cherry



mango



orange



pear



pineapple



tangerine



watermelon



strawberry

Paivio et al. (1968) "Why are pictures easier to recall than words?"

Why do Images Matter?

Visuals evoke stronger emotional reactions



Grabe Bucy (2009) “*Images Bite Politics*”

Why do Images Matter?

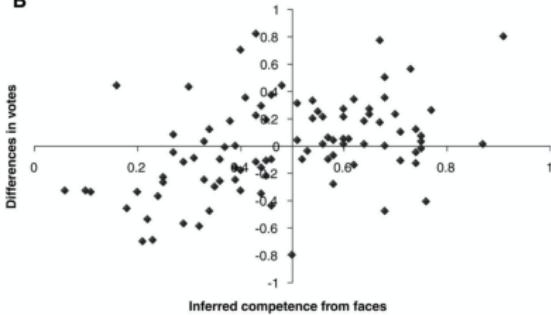
Image effects in **politics**: images → inference of competence → voting

A



Which person is the more competent?

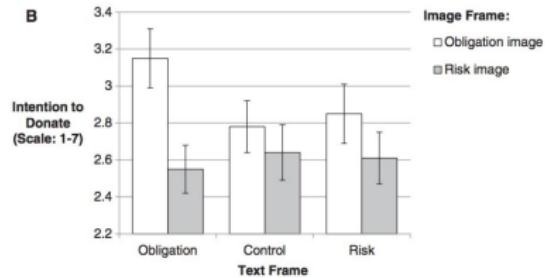
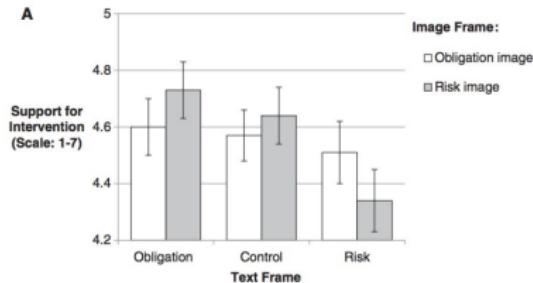
B



Todorov et al. (2009) "Inferences of Competence from Faces Predict Election Outcomes"

Why do Images Matter?

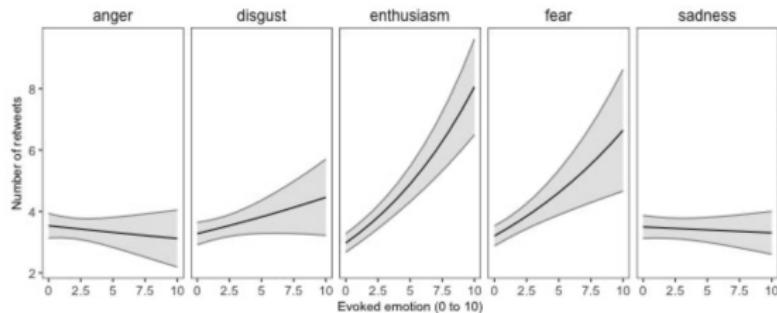
Image effects in **politics**: images → framing → attitudes



Powell et al. (2015) "A Clearer Picture"

Why do Images Matter?

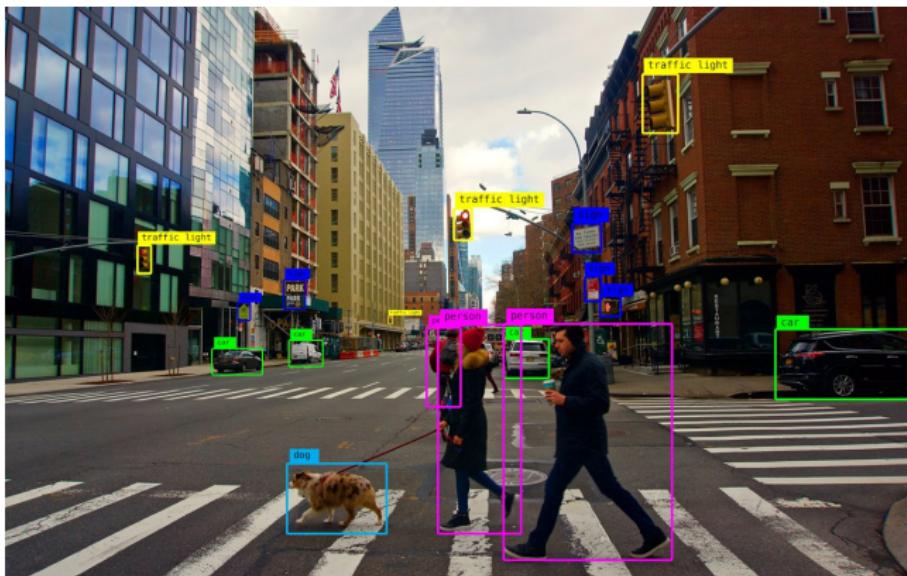
Image effects in **politics**: images → emotions → **mobilization**



Casas & Webb Williams (201) "Images That Matter"

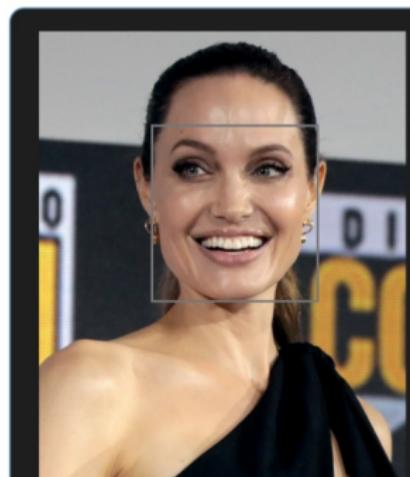
Available Automated Image Analysis Methods

Object detection & recognition



Available Automated Image Analysis Methods

Face detection & recognition



target: img1.jpg

found

- #1
id: img4.jpg
distance: 0.205
- #2
id: img2.jpg
distance: 0.234
- #3
id: img6.jpg
distance: 0.254

Available Automated Image Analysis Methods

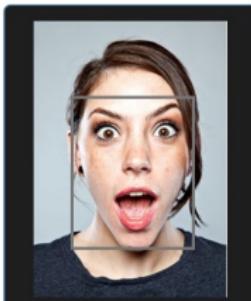
Face analysis



```
{  
  "age": 28.66,  
  "emotion": "neutral",  
  "gender": "Woman",  
  "race": "latino hispanic"  
}
```



```
{  
  "age": 29.27,  
  "emotion": "happy",  
  "gender": "Woman",  
  "race": "white"  
}
```



```
{  
  "age": 29.27,  
  "emotion": "surprise",  
  "gender": "Woman",  
  "race": "white"  
}
```



```
{  
  "age": 29.74,  
  "emotion": "neutral",  
  "gender": "Woman",  
  "race": "white"  
}
```

Available Automated Image Analysis Methods

Image Similarity



Available Automated Image Analysis Methods

Unsupervised Clustering



Available Automated Image Analysis Methods

And many others...

- ▶ Text extraction (OCR)
- ▶ Caption generation
- ▶ Sentiment analysis (evoked emotions)
- ▶ Visual aesthetics analysis
- ▶ etc...

Available Automated Image Analysis Methods

In this workshop we'll focus on...

- ▶ Supervised image classification
- ▶ Unsupervised image classification
- ▶ Multimodal classification (text + image)

Outline

- 1 What are supervised models useful for?
- 2 The basics of supervised classification
- 3 Pre-trained models and zero-shot classification
- 4 Fine-tuning a pre-trained model for a new supervised task

What are supervised models useful for?

When you know what you're looking for in an image

- ▶ Binary: is this a picture of a street protest?
- ▶ Categorical: is this picture a meme, cartoon, or other?
- ▶ Continuous: how much violence is there in the image?

What are supervised models useful for?

Parallelism to text analysis

- ▶ Binary: is this a political message/document?
- ▶ Categorical: topic: Environment, healthcare, etc.
- ▶ Continuous: how negative is the message?

What are supervised models useful for?

Some examples: Binary

Cantú (2019, *APSR*)

Does the vote tally have an alteration?

A

VOTACION RECIBIDA EN LA URNA (con numero)	VOTOS ENCONTRADOS EN OTRAS URNAS (con numero)
191	131
07	7
108	138
00	
128	138

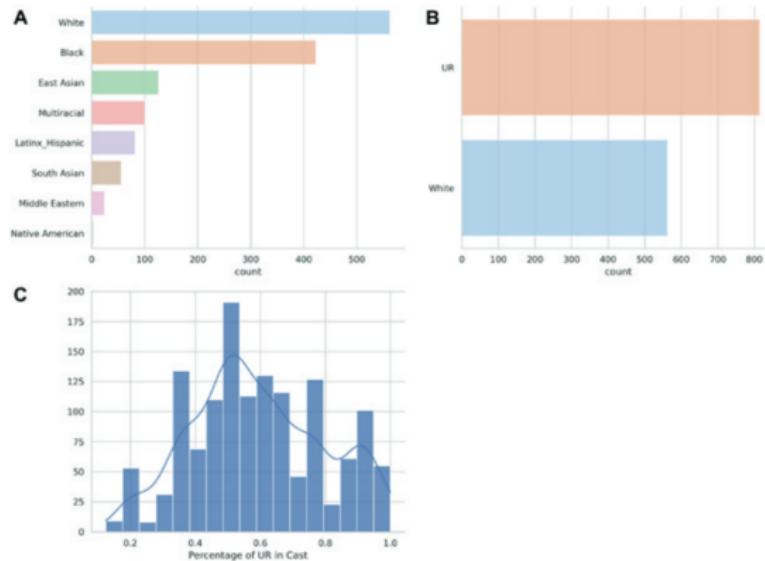
B

VOTACION RECIBIDA EN LA URNA (con numero)	VOTOS ENCONTRADOS EN OTRAS URNAS (con numero)
19	
120	
121	
1	
10	
37	
1	
22	
2	
273	
14	
287	

What are supervised models useful for?

Some examples: Categorical

Malik et al. (2022, *CCR*)
Ethnicity of characters in movies



What are supervised models useful for?

Some examples: Continuous

Steinert-Threlkeld et al. (2022, *JOP*)

How much police/protester violence is there in the image?

(b) State Violence



Seoul .031



Hong Kong .145



Barcelona .654



Caracas .849

(c) Protester Violence



Seoul .021



Barcelona .255



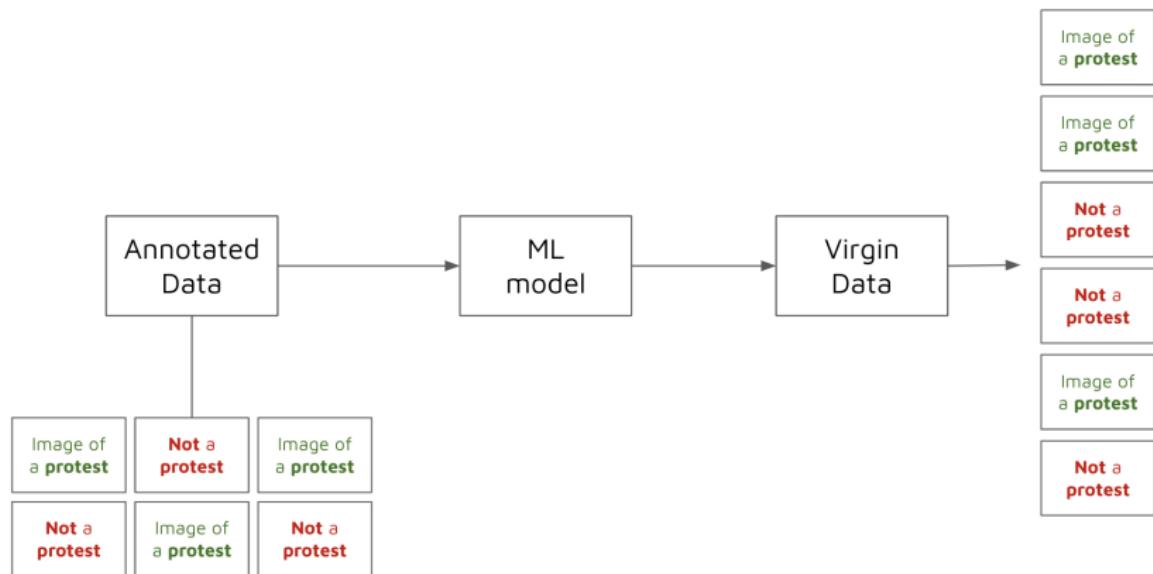
Hong Kong .478



Caracas .998

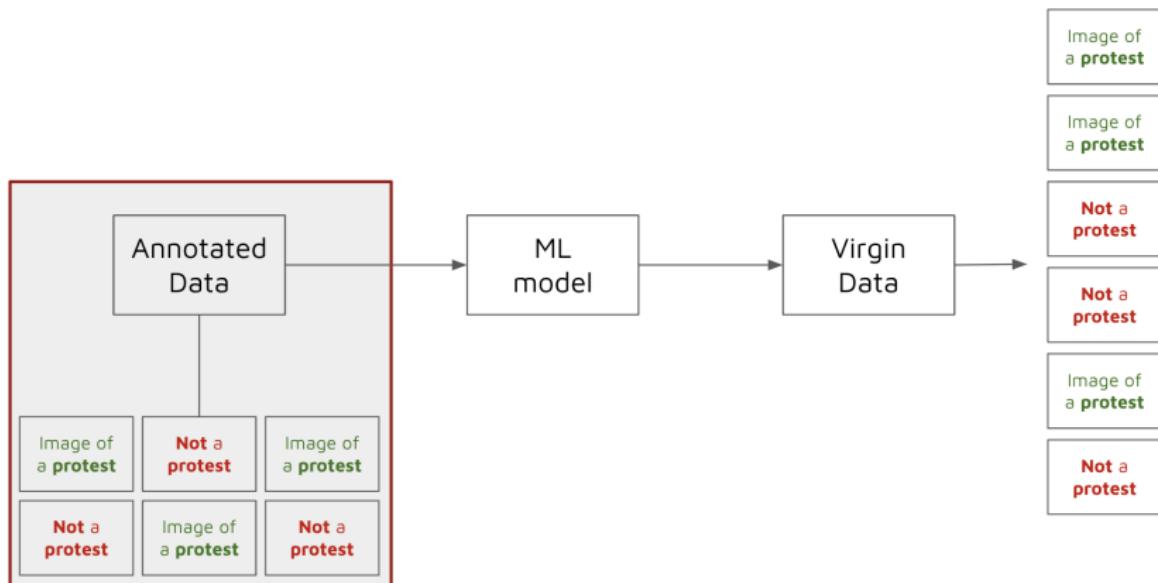
The Basics of Supervised Classification

Basic logic



The Basics of Supervised Classification

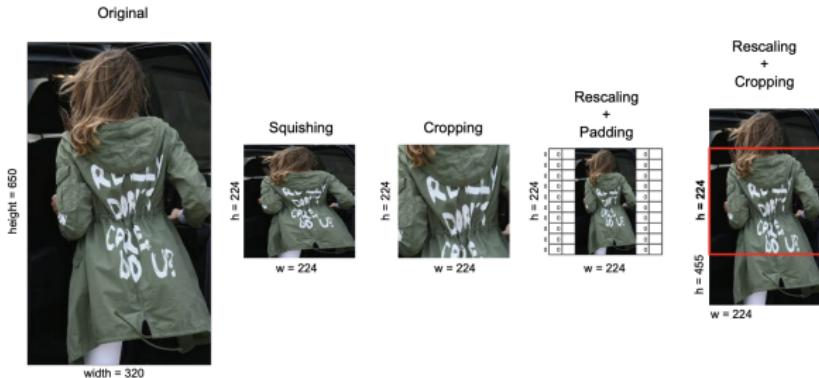
Practical considerations: Pre-processing



The Basics of Supervised Classification

Practical considerations: Pre-processing

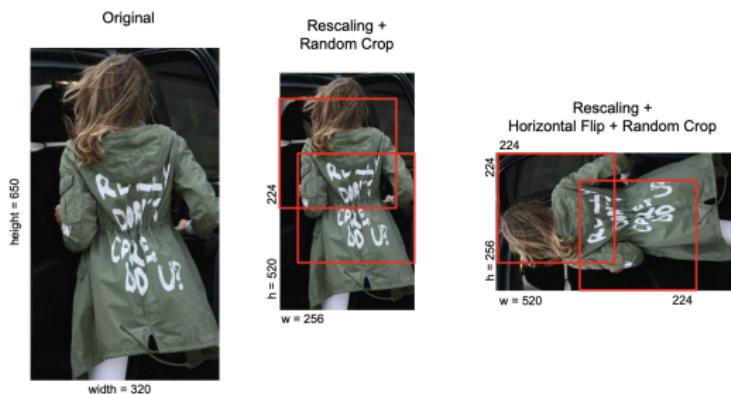
- ▶ Input images need to be of the same particular size (224x224 pixels)



The Basics of Supervised Classification

Practical considerations: Pre-processing

- ▶ Input images need to be of the same particular size (224x224 pixels)
- ▶ Data augmentation



The Basics of Supervised Classification

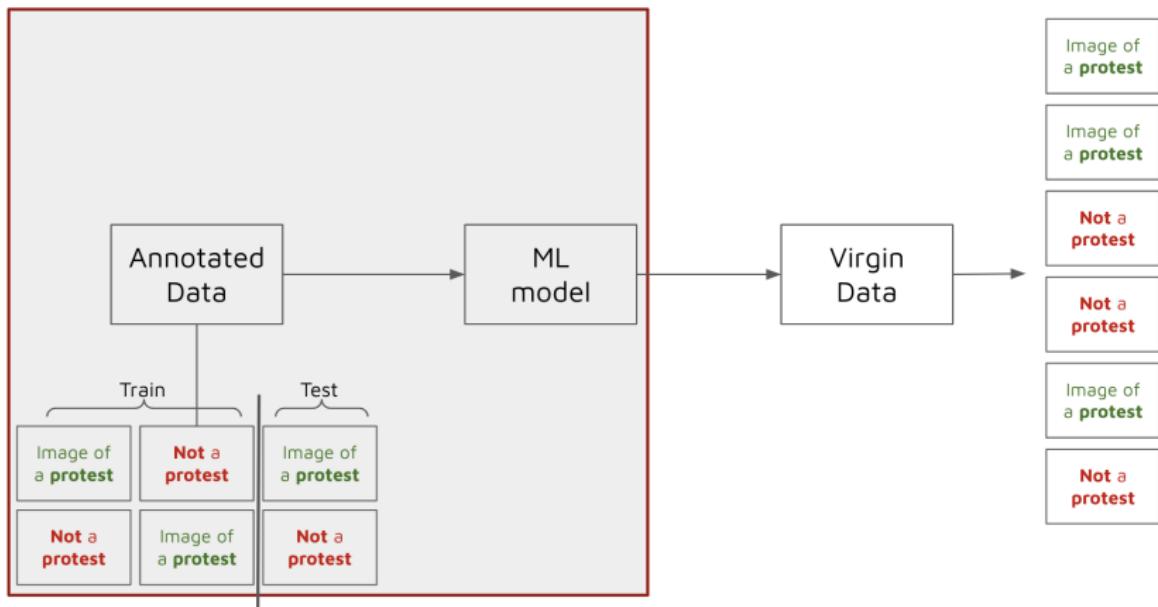
Practical considerations: Pre-processing

- ▶ Input images need to be of the same particular size (224x224 pixels)
- ▶ Data augmentation
- ▶ Data organization

- train
 - class01
 - class02
 - ...
 - classN
- test
 - class01
 - class02
 - ...
 - classN

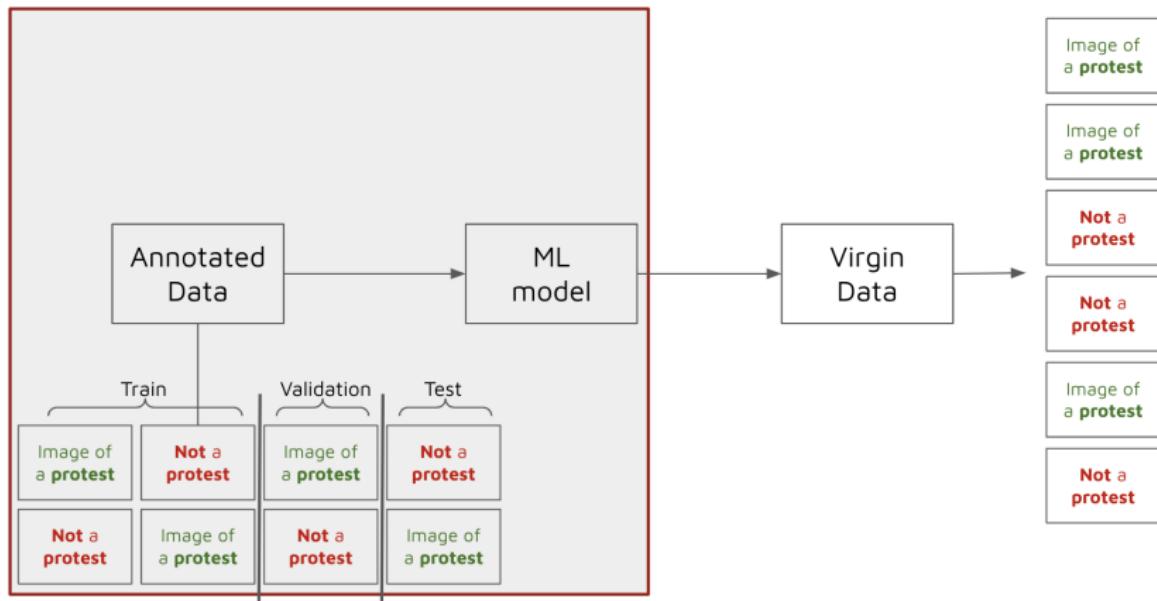
The Basics of Supervised Classification

Training the model



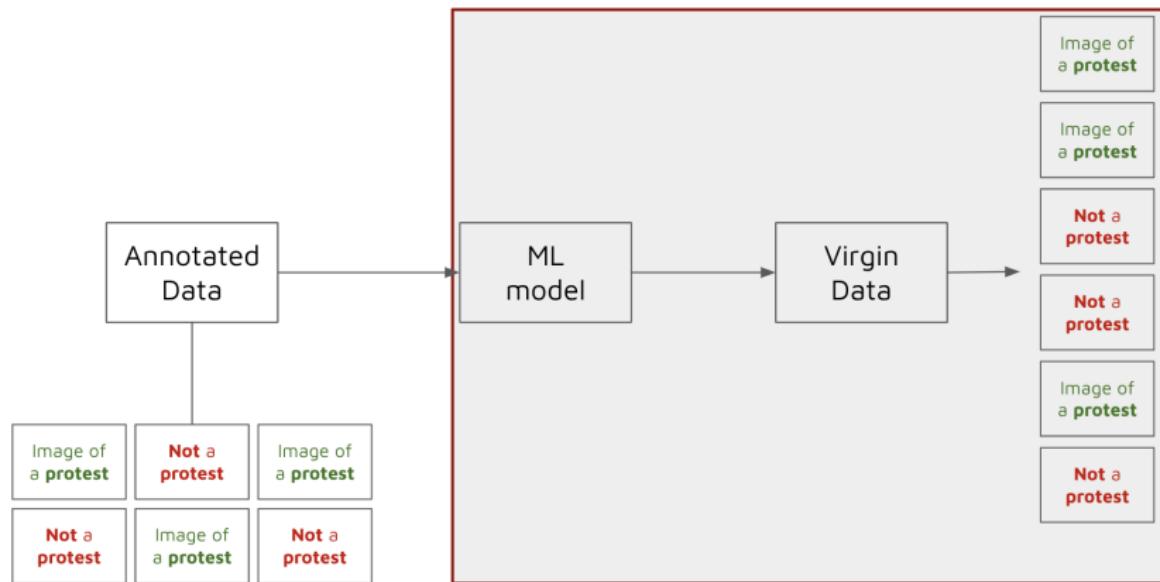
The Basics of Supervised Classification

Training the model



Pre-trained models and zero-shot classification

Applying an existing model to your data



Pre-trained models and zero-shot classification

Common benchmark datasets in computer vision

- ▶ Cifar-10 & Cifar-100 (cat, dog, horse, ...)
- ▶ MNIST (handwritten digits: 1, 2, 3, 4, ...)
- ▶ MS COCO: many categories and subcategories
- ▶ ImageNet: 14 million examples. 1,000 objects.
- ▶ ... many many others!

Pre-trained models and zero-shot classification

Many are available through deep learning libraries in python

Here a link to all the ones available through pytorch

Pre-trained models and zero-shot classification

The shortcomings and the dangers

- ▶ No model trained to predict our quantity of interest
- ▶ Even if there is, it may not generalize well to our own data

Pre-trained models and zero-shot classification

Most common zero-shot classification

- ▶ Ethnicity
- ▶ Gender
- ▶ Facial expressions/emotions

Entering 2021: Vision Transformers (ViT)

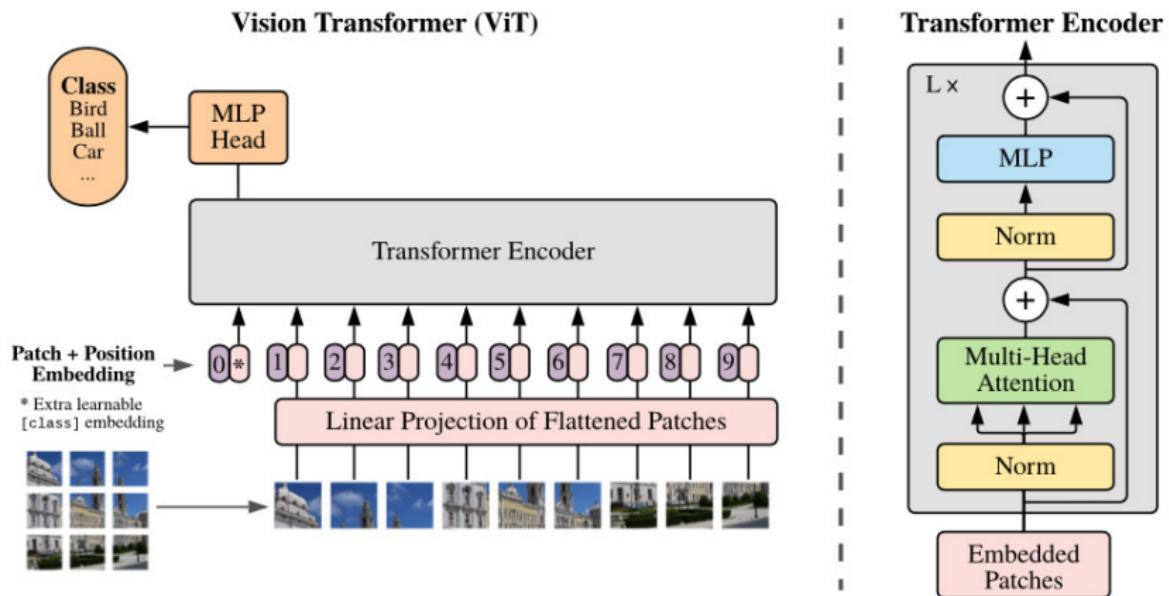
Originally introduced by Google Brain

Simple summary: The model breaks the image into small pieces, processes the relationships between them, and makes a prediction about what the image represents.

- ▶ Very close to using transformer models in NLP
- ▶ Splitting an image into a grid of sub-image patches
- ▶ Embed each patch with a linear projection (= make embedding)
- ▶ Each embedded patch becomes a token, and the resulting sequence of embedded patches is the sequence you pass to the model

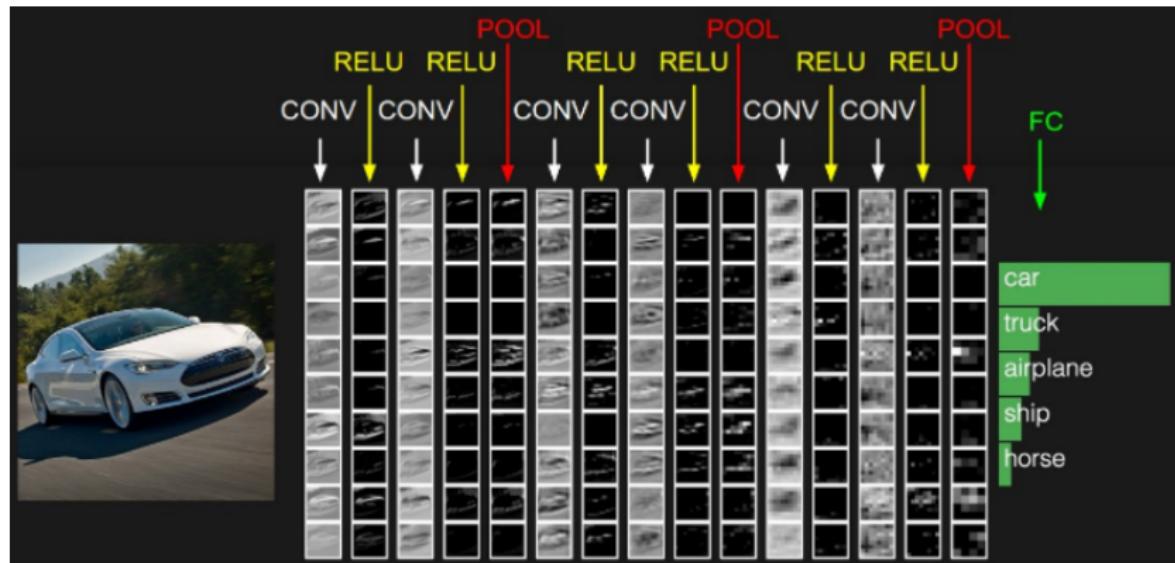
Entering 2021: Vision Transformers (ViT)

Originally introduced by Google Brain



Fine-tuning a pre-trained model

Leveraging an existing model for our own task



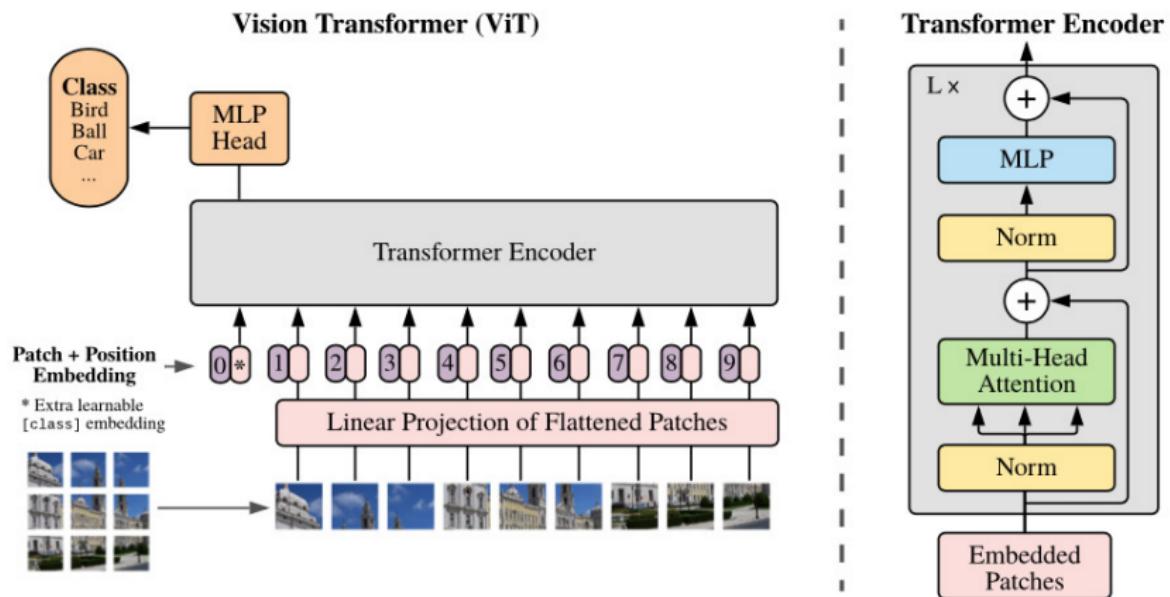
Fine-tuning a pre-trained model

Leveraging an existing model for our own task

layer name	output size	18-layer	34-layer	50-layer	101-layer	152-layer
conv1	112×112			7×7, 64, stride 2		
				3×3 max pool, stride 2		
conv2_x	56×56	$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix} \times 2$	$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix} \times 3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix} \times 3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix} \times 3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix} \times 3$
conv3_x	28×28	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix} \times 2$	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix} \times 4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix} \times 4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix} \times 4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix} \times 8$
conv4_x	14×14	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix} \times 2$	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix} \times 6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix} \times 6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix} \times 23$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix} \times 36$
conv5_x	7×7	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix} \times 2$	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix} \times 3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix} \times 3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix} \times 3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix} \times 3$
	1×1	average pool, 1000-d fc, softmax				
FLOPs		1.8×10^9	3.6×10^9	3.8×10^9	7.6×10^9	11.3×10^9

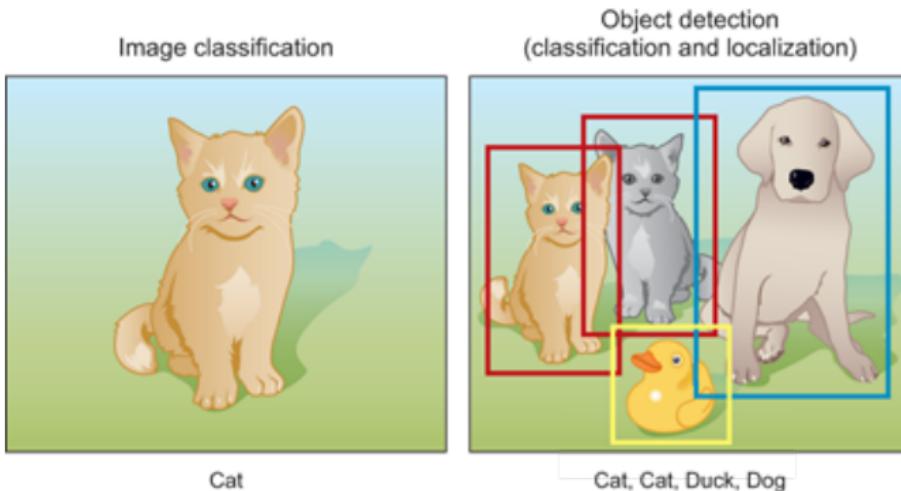
Fine-tuning a pre-trained model

Leveraging an existing model for our own task



Fine-tuning a pre-trained model

Additional consideration: detection v. recognition



*We will not cover object detection in the tutorial (apart from when working with faces). See this [link](#) for sample pytorch code on how to fine-tune a widely-used object detection + recognition model (RCNN).

Outline

- 1 What are unsupervised models useful for?
- 2 The logic of unsupervised classification
- 3 Some illustrative results

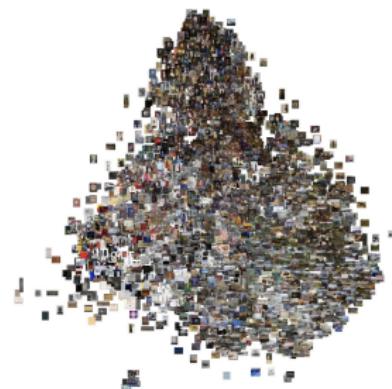
What are unsupervised models useful for?

- ▶ Data exploration: learn more about your data



What are unsupervised models useful for?

- ▶ Data exploration: learn more about your data
- ▶ “Deduplication”



What are unsupervised models useful for?

- ▶ Data exploration: learn more about your data
- ▶ “Deduplication”
- ▶ Stratified sampling



What are unsupervised models useful for?

- ▶ Data exploration: learn more about your data
- ▶ “Deduplication”
- ▶ Stratified sampling
- ▶ Descriptive analysis



What are unsupervised models useful for?

An example: diffusion of #FamiliesBelongTogether on Twitter



What are unsupervised models useful for?

An example: diffusion of #FamiliesBelongTogether on Twitter

$$f(\text{online_engagement}) = \text{image} \times \text{viewer}$$

Image features:

- ▶ social identity
- ▶ expectation of success
- ▶ **evoked emotions**

View features:

- ▶ gender
- ▶ race/ethnicity
- ▶ ideology/**partisanship**

The Logic of Unsupervised Classification

An example: diffusion of #FamiliesBelongTogether on Twitter

Goal: to cluster images with similar content, to reduce and balance data-annotation

A



B



C



D

To: Representative Giffords

Trump is going off the rails! He must stop obstructing justice! Please do your job as my representative in Washington and check his unbalance! America needs to live up to its values and reunite immigrant children with their families!

Scott in Billings, MT

resistbot

TEXT RESIST TO 80409

E



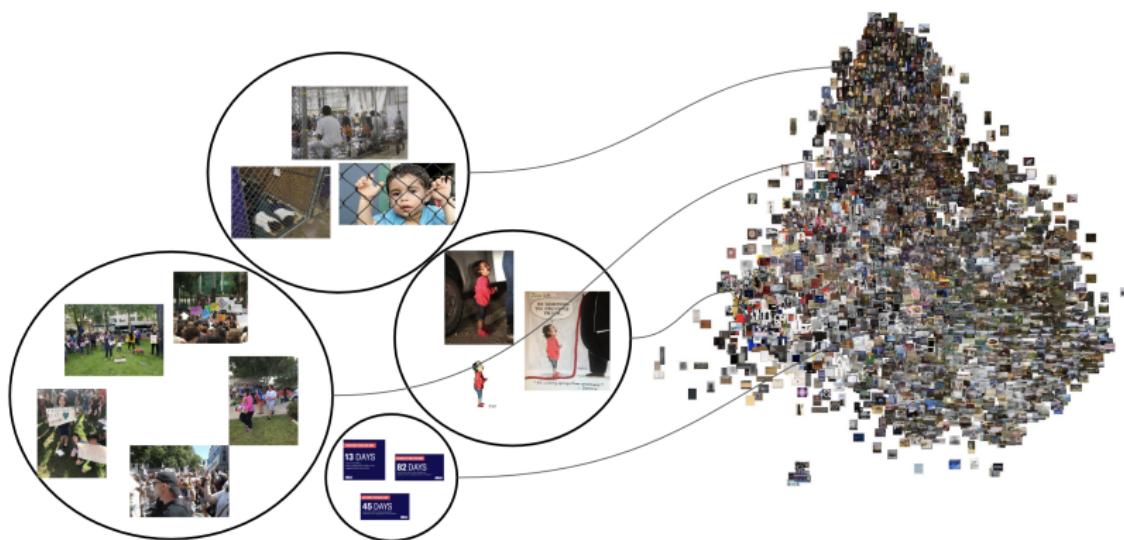
F



The Logic of Unsupervised Classification

An example: diffusion of #FamiliesBelongTogether on Twitter

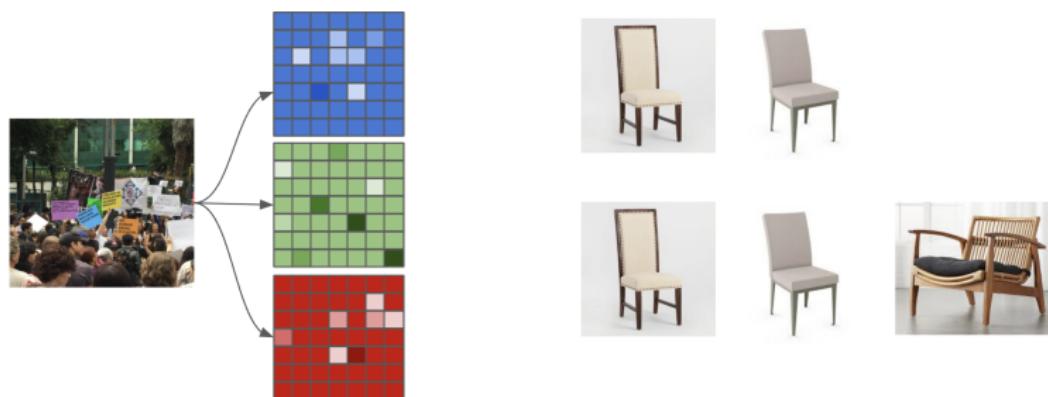
Goal: to cluster images with similar content, to reduce and balance data-annotation



The Logic of Unsupervised Classification

Three main challenges

1. How to numerically represent image content in order to cluster based on content similarity



The Logic of Unsupervised Classification

Three main challenges

1. How to numerically represent image content in order to cluster based on content similarity
2. How to select the best fitting number of clusters
 - ▶ Datasets are likely to be highly unbalanced, complicating the cluster-discovering process
 - ▶ “True” number of clusters is unknown
 - ▶ Want to make sure truly cohesive clusters are found

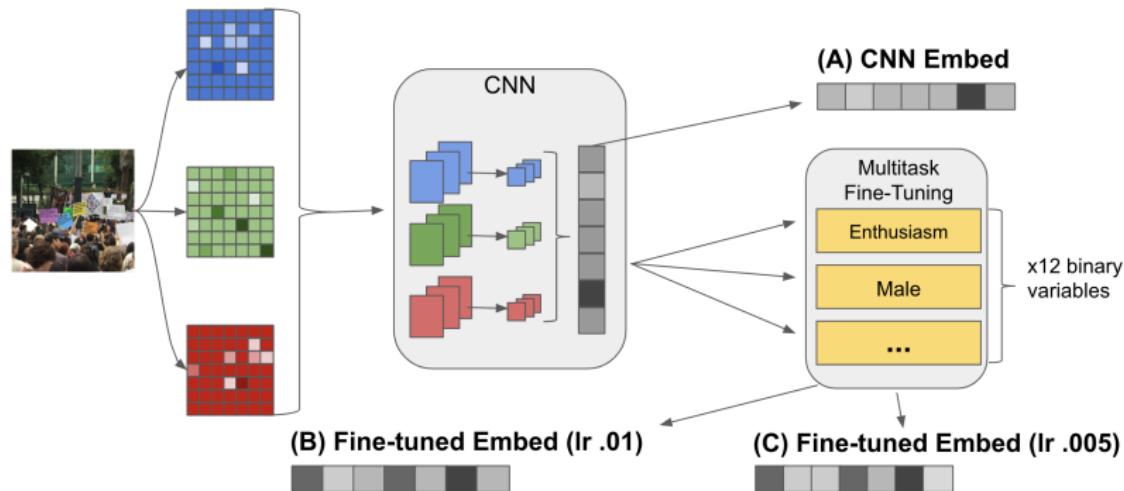
The Logic of Unsupervised Classification

Three main challenges

1. How to numerically represent image content in order to cluster based on content similarity
2. How to select the best fitting number of clusters
 - ▶ Datasets are likely to be highly unbalanced, complicating the cluster-discovering process
 - ▶ “True” number of clusters is unknown
 - ▶ Want to make sure truly cohesive clusters are found
3. How do we evaluate...
 - ▶ Are we using the “best” numeric image representation possible? (1)
 - ▶ The performance of different clustering configurations? (2)

The Logic of Unsupervised Classification

1. Three approaches to reducing image dimensionality



The Logic of Unsupervised Classification

2. Clustering process

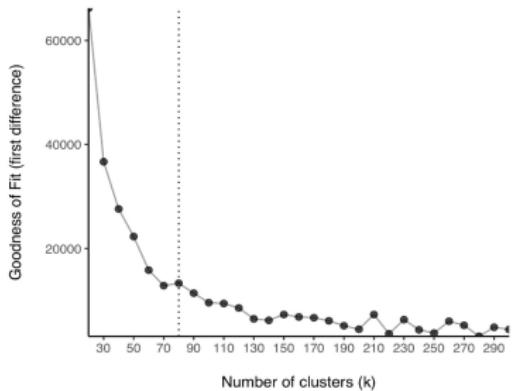
- ▶ **Iterative process** find big cohesive clusters first, and then see if there are smaller ones
- ▶ Intra-cluster similarity (silhouette scores) to judge cluster cohesiveness

Hyperparameters	Algorithm
→ Step Size	X = input image matrix (20,000 images x 512 embed size)
→ Similarity Threshold	1. Choose number of K clusters to predict
→ Sample Size	a. Randomly sample <u>1,000</u> images from X
→ Stop Size	b. Fit several k-means algorithms (increase K by <u>5</u> each time)...
→ Converge Window	... until goodness of fit doesn't improve (avg. across <u>3</u> iterations) ... STOP if $K >$ images in X
	2. Fit k-means algorithm predicting K image clusters
	3. Evaluate intra-cluster similarity (silhouette score)
	4. Find cohesive clusters (silhouette score > <u>0.04</u>)
	5. Remove cohesive images from X, and keep classifying the rest ... STOP if only <u>20</u> images left

The Logic of Unsupervised Classification

3. Building a validation set

- ▶ Number of clusters is unknown but clusters are likely to be unbalanced
- ▶ Fit a first clustering algorithm (e.g. k-means)



The Logic of Unsupervised Classification

3. Building a validation set

- ▶ Number of clusters is unknown but clusters are likely to be unbalanced
- ▶ Fit a first clustering algorithm (e.g. k-means)
- ▶ Select random pairs from each cluster (e.g. n=554) and label them for cohesiveness

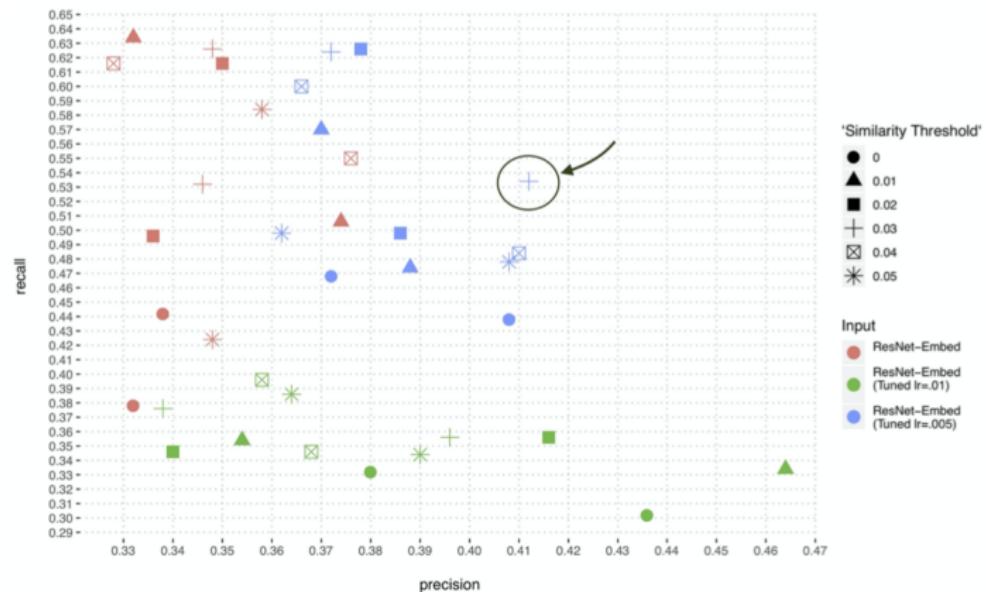


	coder1	coder2	Agreement 87%
	1	1	
	1	1	
	0	0	
	0	1	
	1	1	

Kappa
0.64

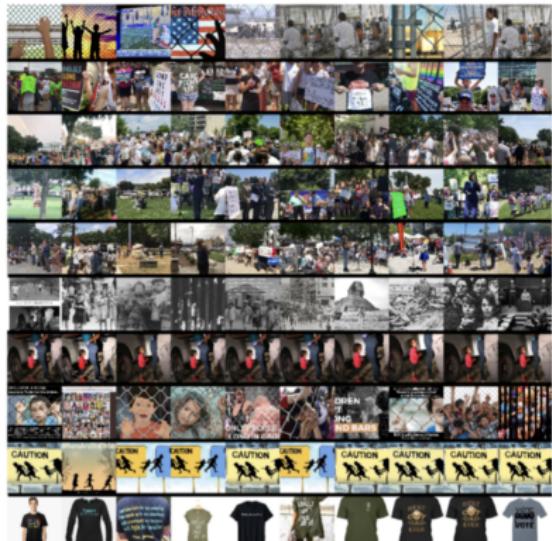
The Logic of Unsupervised Classification

3. Comparing performance of clustering hyperparameters



Some illustrative results

1. Image exploration



(1)

(2)

(3)

(4)

(5)

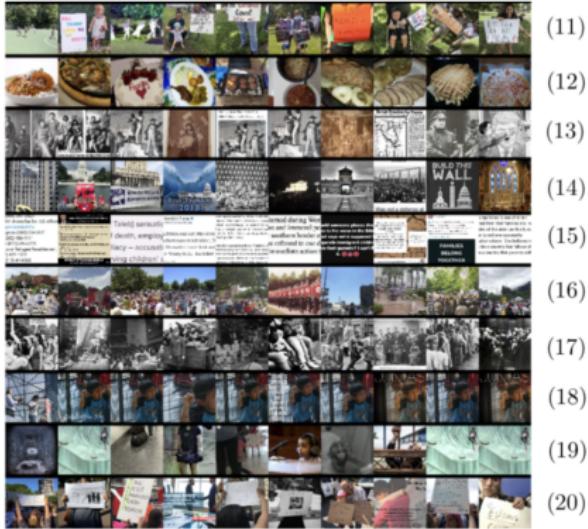
(6)

(7)

(8)

(9)

(10)



(11)

(12)

(13)

(14)

(15)

(16)

(17)

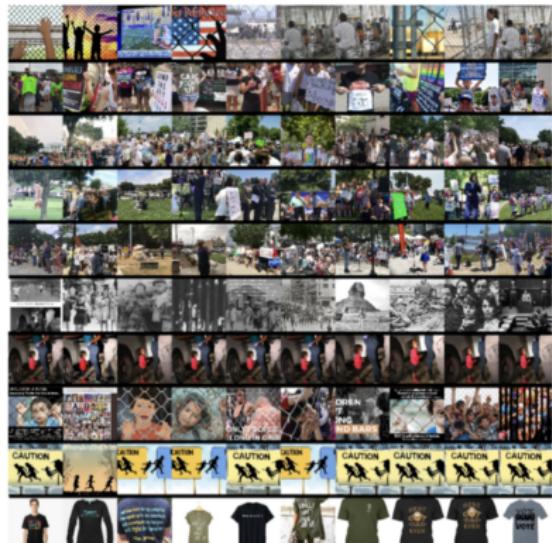
(18)

(19)

(20)

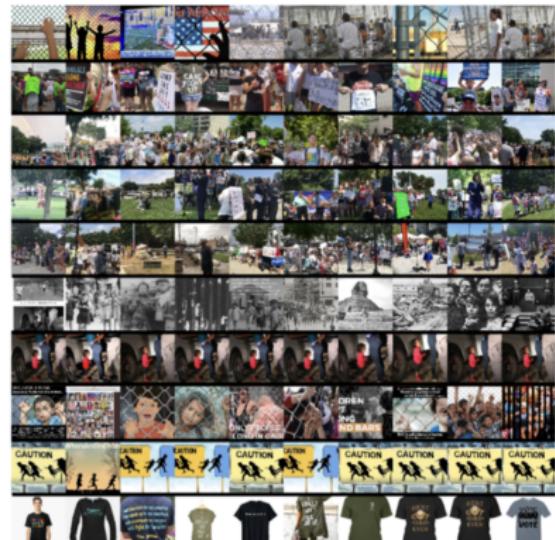
Some illustrative results

1. Image exploration

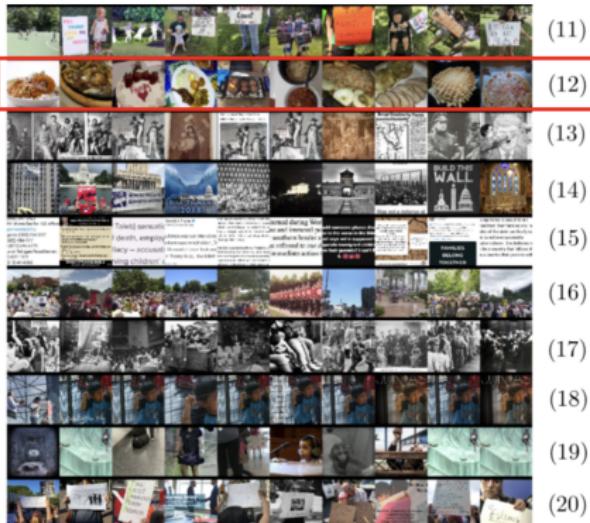


Some illustrative results

1. Image exploration



(1) (2) (3) (4) (5) (6) (7) (8) (9) (10)



(11) (12) (13) (14) (15) (16) (17) (18) (19) (20)

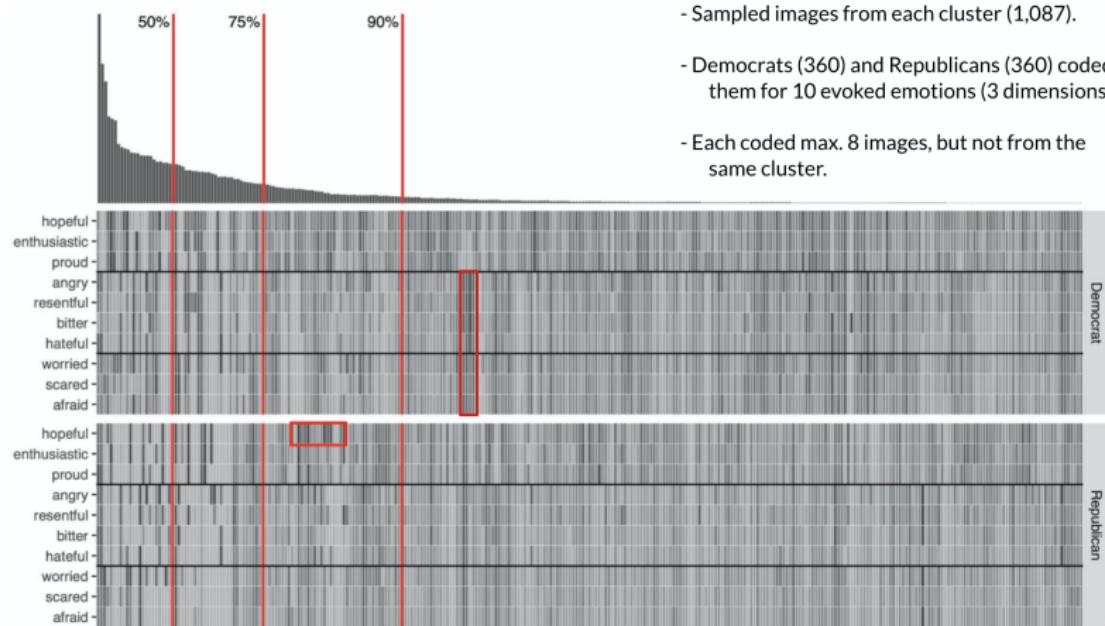
Some illustrative results

1. Image exploration



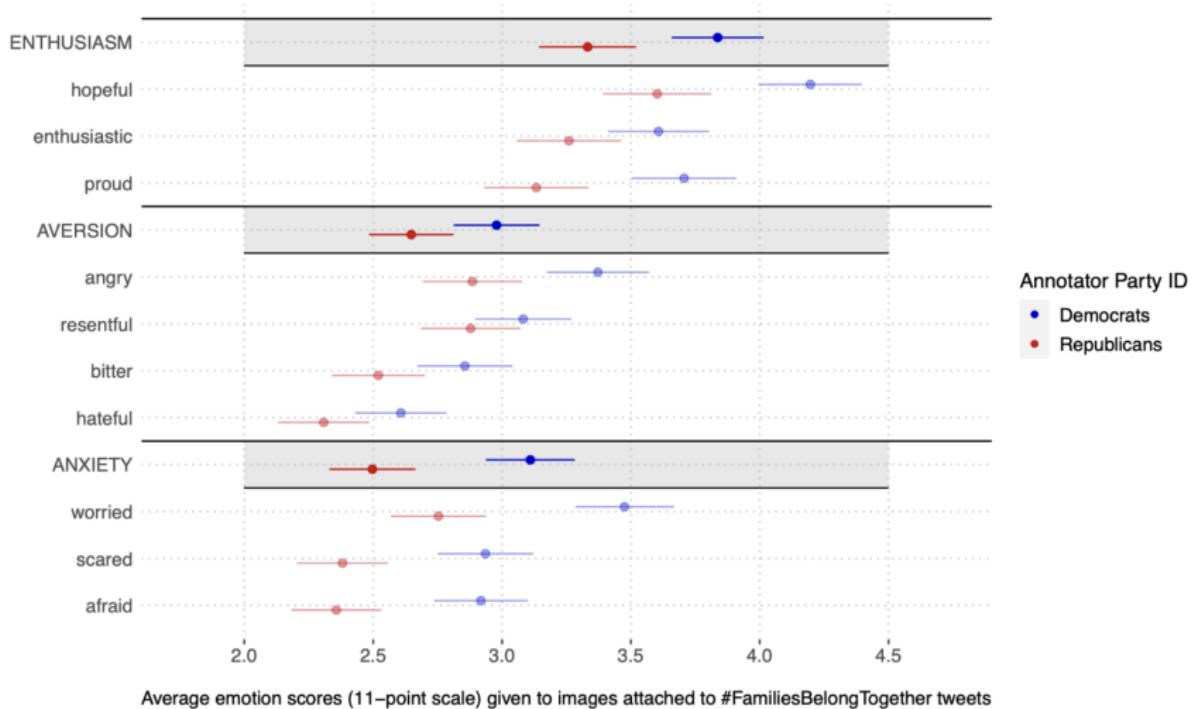
Some illustrative results

2. Purposive sampling



Some illustrative results

3. Differences in annotation

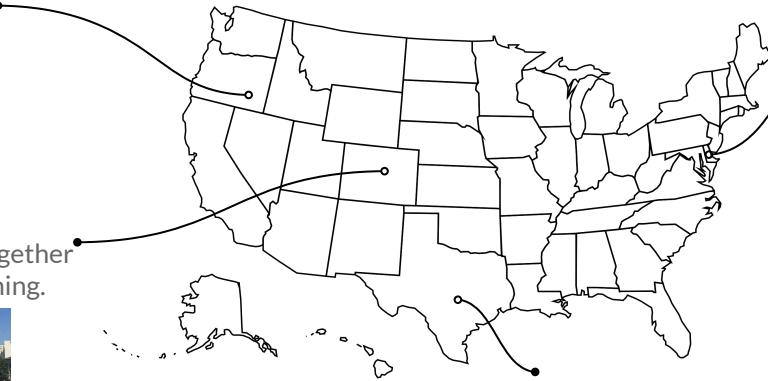


Multimodal Modeling for Political Science Research



@xxxx: F*** Trump. #FamiliesBelongTogether

@xxxx:
#FamiliesBelongTogether
in Denver this morning.



@ResistSnow: #FreeTheChildren
#FamiliesBelongTogether
Our children long for realistic maps of
the future that they can be...

@fams2gether: Nobody,
no matter who you are,
where you are from, you
should spend an hour like
this!



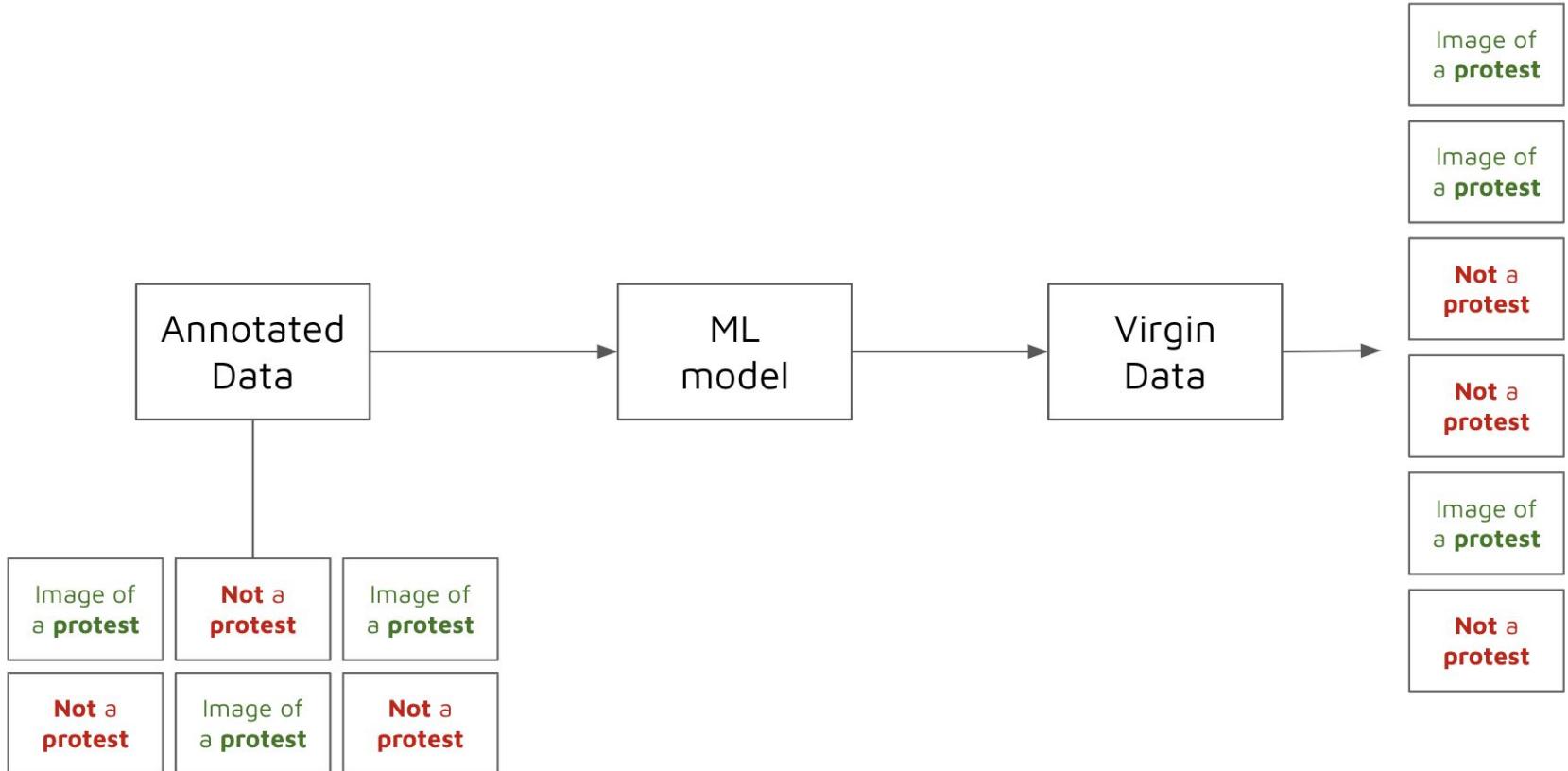
Andreu Casas

Royal Holloway Univ. of London

Freek Cool

Vrije Universiteit Amsterdam

Supervised machine learning is often used to identify a defined concept in large amounts of data



Often this data is **multimodal**: with text, visuals, audio

 **Donald J. Trump**  @realDonaldTrump · Sep 21

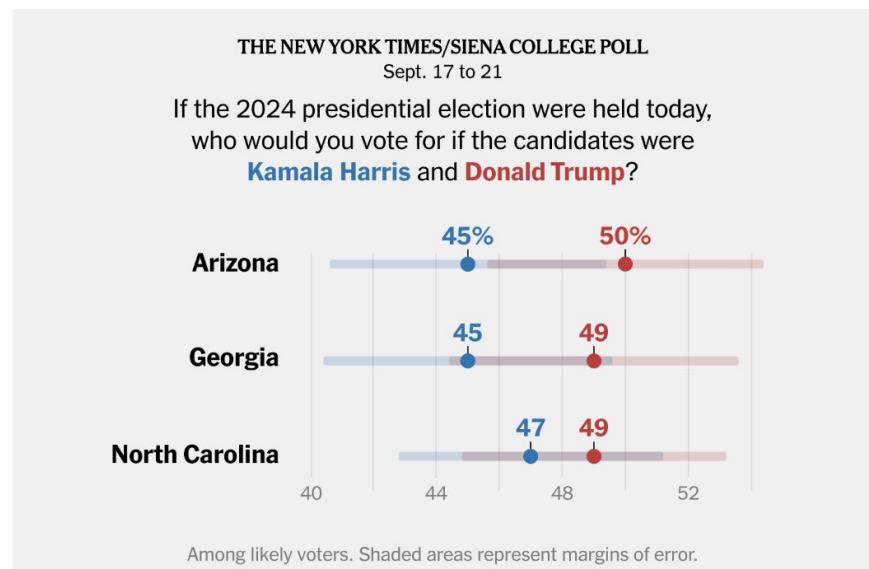
Hello everyone! I have something incredible to share today, as we are introducing the launch of our Official Trump Coins! The ONLY OFFICIAL coin designed by me—and proudly minted here in the U.S.A. The President Donald J. Trump First Edition Silver Medallion will be available

Show more



0:00 / 1:25

24K 27K 144K 26M



Yet we mostly only use one data modality to train supervised ML (e.g. text)

Donald J. Trump  @realDonaldTrump · Sep 21

Hello everyone! I have something incredible to share today, as we are introducing the launch of our Official Trump Coins! The ONLY OFFICIAL coin designed by me—and proudly minted here in the U.S.A. The President Donald J. Trump First Edition Silver Medallion will be available

Show more



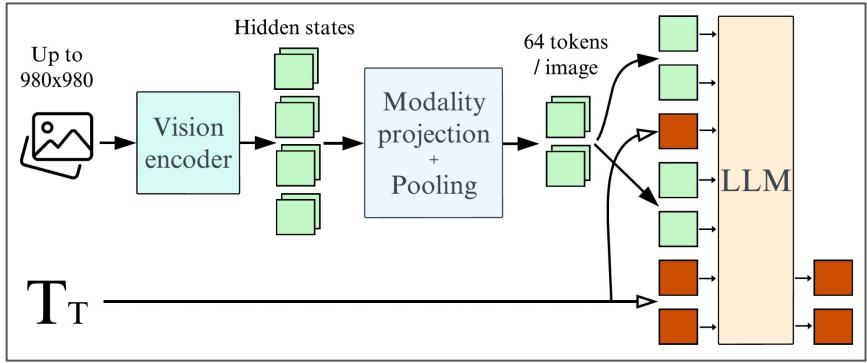
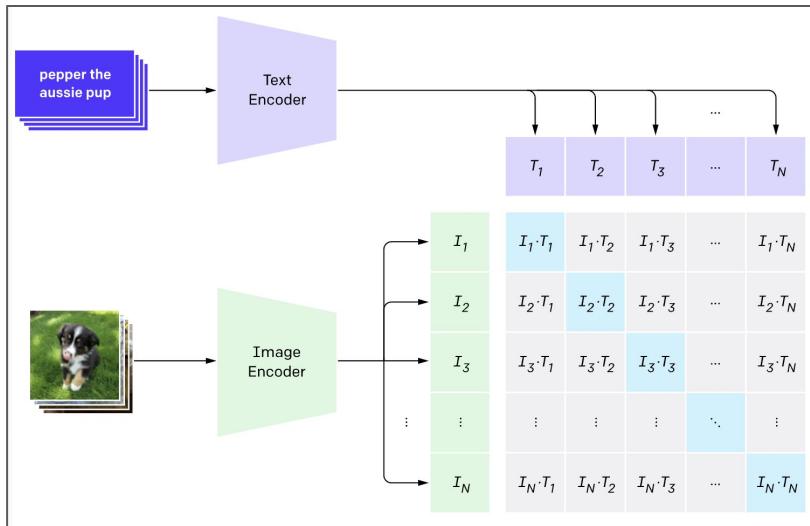
0:00 / 1:25

24K 27K 144K 26M



Recent computational advances make **multimodal modeling** easier

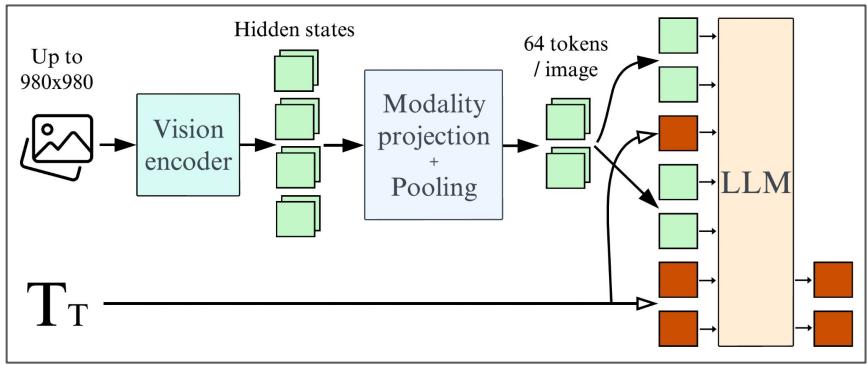
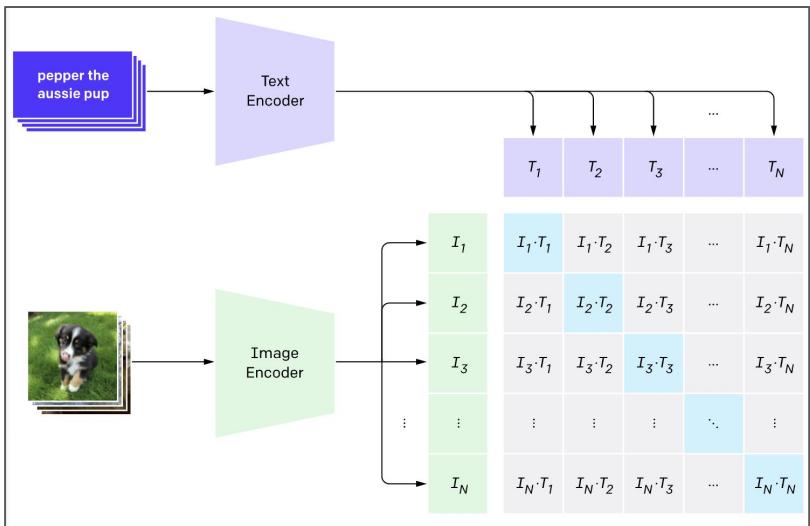
CLIP



Idefics2

But we don't know much about **whether nor the conditions under which multimodality can help improve the performance of supervised ML**

CLIP



Idefics2

Two **original annotated dataset** (10 tasks)

- (1) **YouTube Videos** from channels posting on US politics (N = ~4,000) → 6 tasks
- (2) **Twitter Posts** from interest groups from US, ES, DK, GE (N = ~4,000) → 4 tasks

Seven **models**:

- (1) **Text only**: SVM, BERT, Llama2, and Llama3
- (2) **Image only**: CNN
- (3) **Text + Image**: CLIP, Idefics2

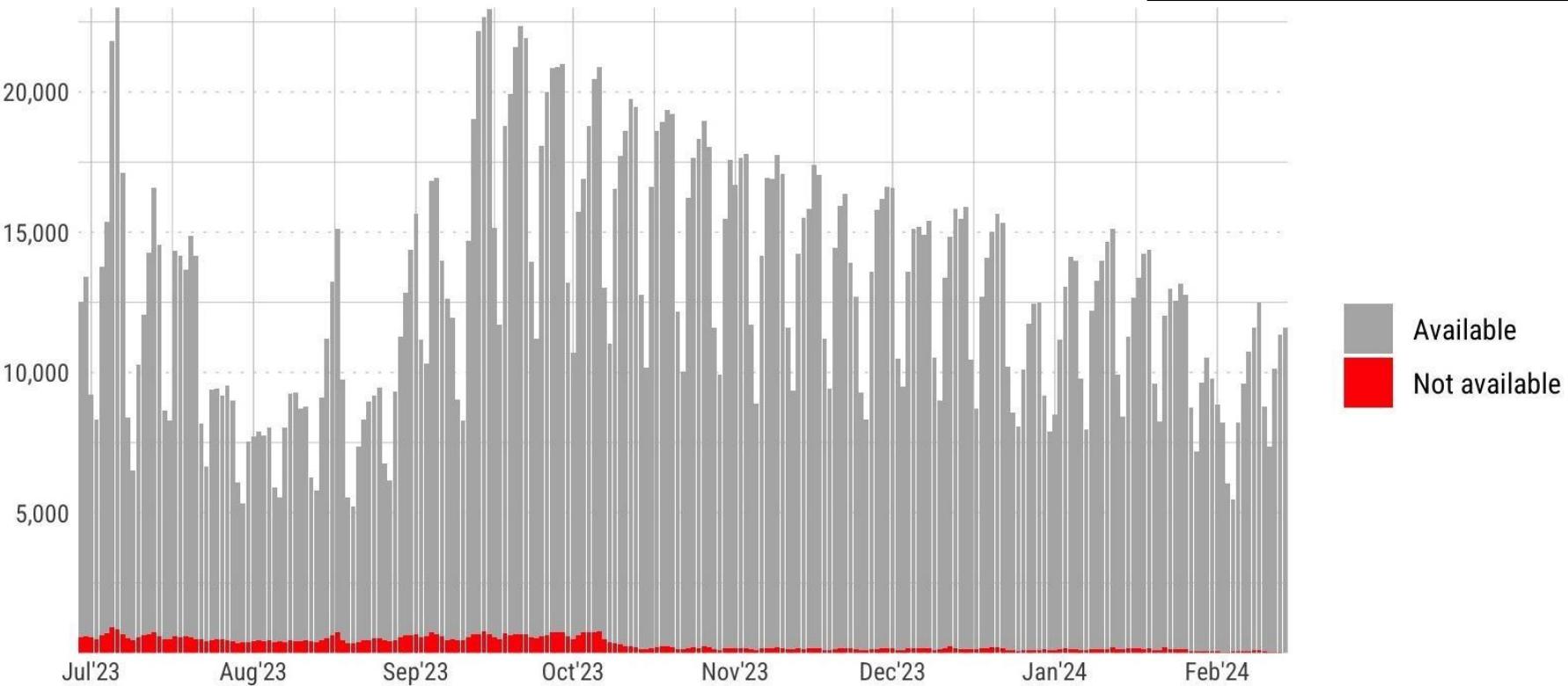
- (1) **YouTube Videos** from channels that post about US politics (N = ~4,000)
 - (a) *US politics*: is the video about US politics?
 - (b) *Hateful*: does the video have hateful content?
 - (c) *Typology*: what type of video? (e.g. opinion, high-quality news, etc.)
 - (d) *Ideology*: ideology of the video (liberal, moderate, conservative, neutral)

Seven **models**:

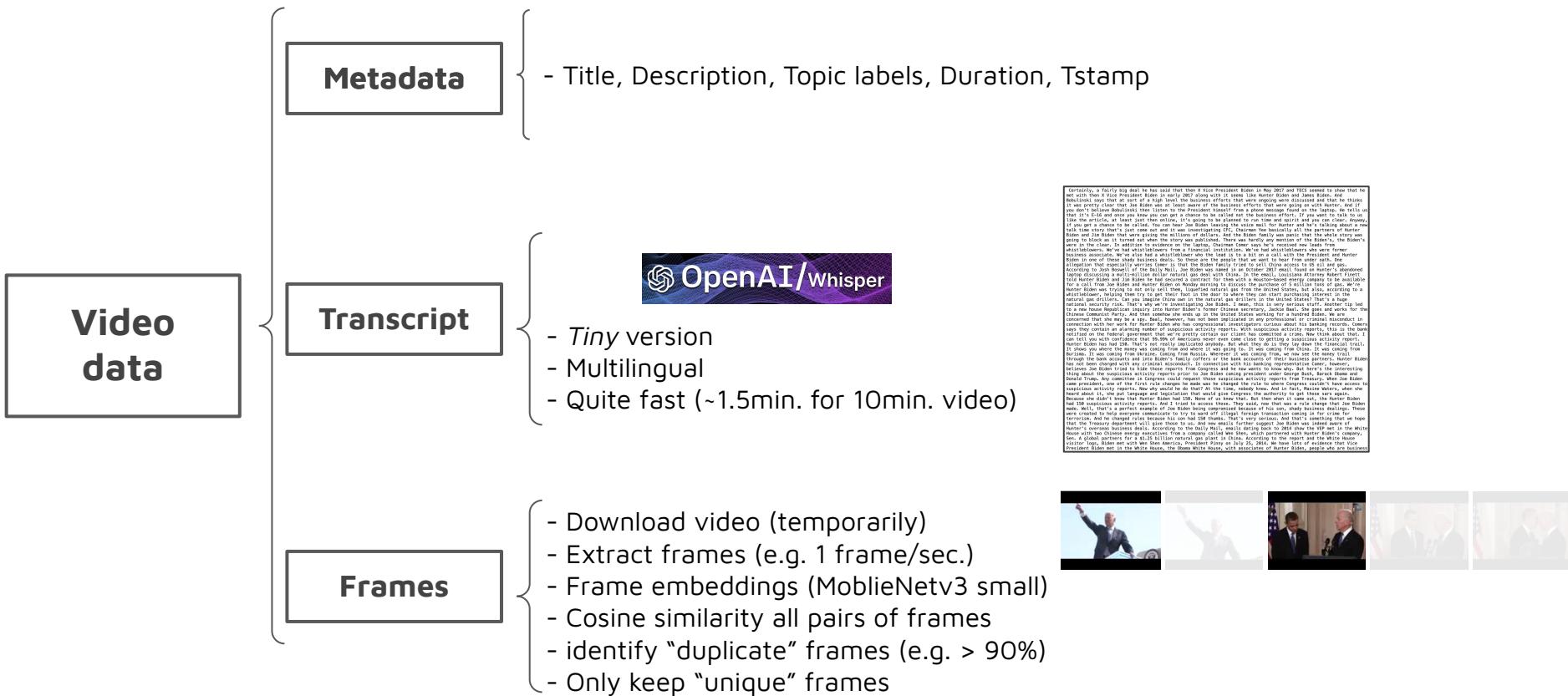
- (1) **Text only**: SVM, BERT, Llama2, and Llama3
- (2) **Image only**: CNN
- (3) **Text + Image**: CLIP, Idefics2

Data

- about **6 million videos**
- from about **12k channels**
- **3% channels** not available
- **2% videos** not available



Data



Data

Task	Description	Values	N	%
US Politics	Whether the video is about, or relevant to, US politics	0	1,935	49.8%
		1	1,945	50.2%
		<i>N</i>	3,880	100.0%
Hateful	Whether the video contains hateful language/behavior	0	3,431	88.4%
		1	449	11.6%
		<i>N</i>	3,880	100%
Idiology	The ideological leaning of the video	Neutral	238	21.7%
		Conservative	476	41.9%
		Moderate	176	15.5%
		Liberal	247	20.9%
		<i>N</i>	1,137	100%
		Campaign	16	1.4%
Typology	Type of video	Educational	61	5.2%
		Satire	73	6.2%
		Low-Qual News	108	9.2%
		High-Qual News	332	28.4%
		Opinion	581	49.6%
		<i>N</i>	1,171	100%

Models

Text only:

- (1) SVM
- (2) BERT
- (3) Llama2
- (4) Llama3

Image only

- (5) CNN

Text + Image

- (6) CLIP
- (7) ldefics2

Models

Text only:

- Ngram-based, **no prior language knowledge**, fully task dependent.

(1) SVM

(2) BERT

(3) Llama2

(4) Llama3

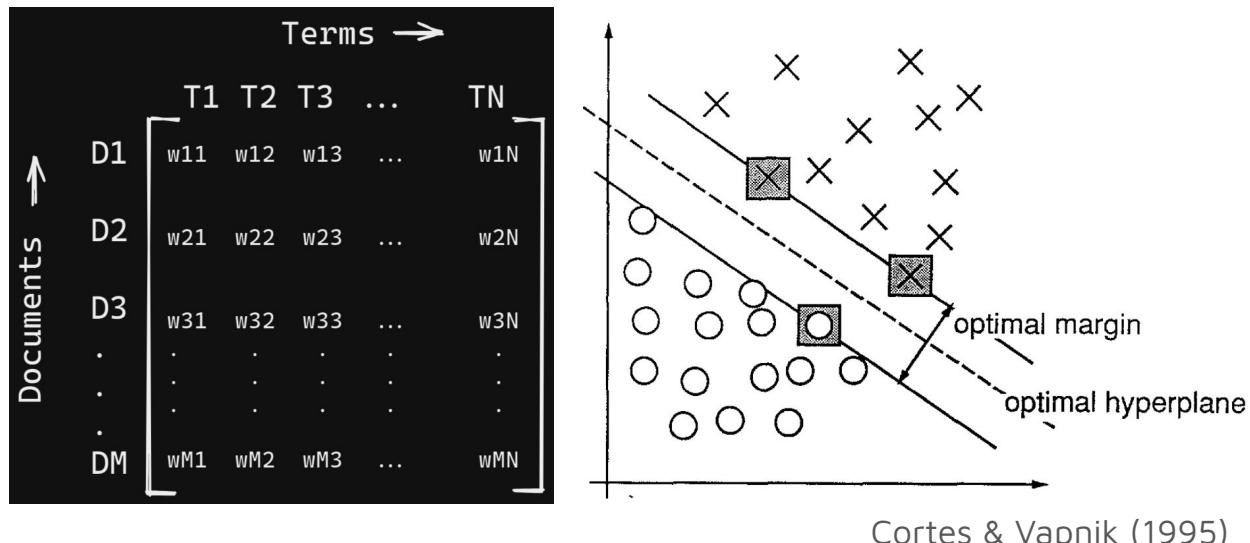
Image only

(5) CNN

Text + Image

(6) CLIP

(7) Idefics2



Cortes & Vapnik (1995)

Models

Text only:

(1) SVM

(2) BERT

(3) Llama2

(4) Llama3

Image only

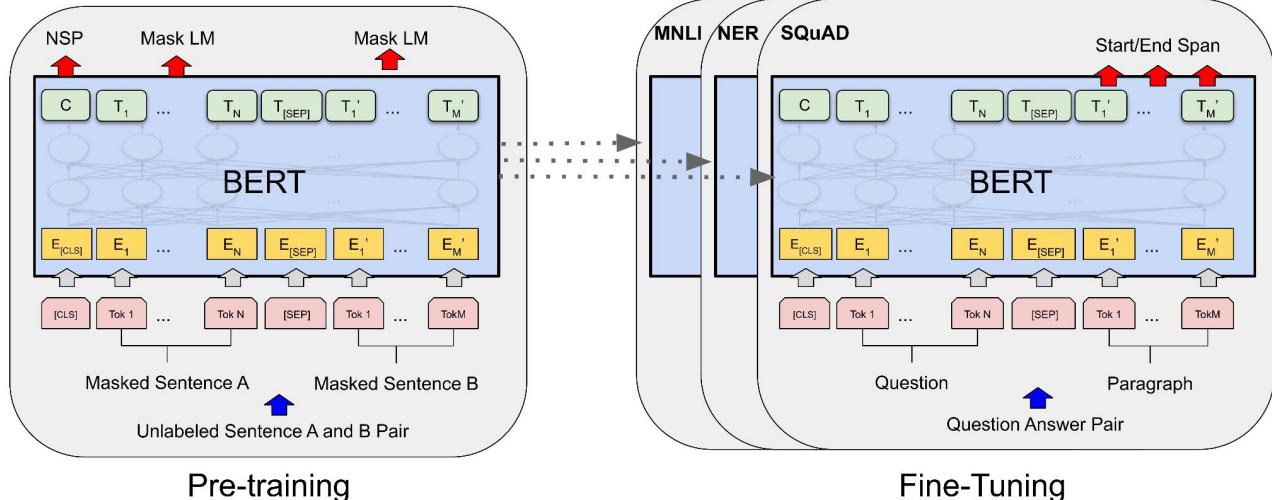
(5) CNN

Text + Image

(6) CLIP

(7) Idefics2

- Transformed-based, self-supervised, language model (prior knowledge), **fine-tuned on next sentence**, can be fine-tuned to do new tasks.
- Trained on: 11k books (800mil tokens) + English wikipedia (**2.5 bil** tokens)
- **bert-base-uncased: 110 mil** parameters
- Context length: **512** tokens



Models

Text only:

(1) SVM

(2) BERT

(3) Llama2

(4) Llama3

- Transformed-based, self-supervised, language model (prior knowledge), **fine-tuned on instruction task**, can be fine-tuned on new instructions
- Trained on: **2 tril** tokens, publicly available sources (although *unknown*)
- **7Bil/13bil/70bil** parameters → **Llama-2-7b-chat**
- Context length: **4,096 tokens**

Image only

(5) CNN

Text + Image

(6) CLIP

(7) Idefics2

Models

Text only:

- (1) SVM
- (2) BERT
- (3) Llama2
- (4) Llama3**

- Transformed-based, self-supervised, language model (prior knowledge), **fine-tuned on instruction task**, can be fine-tuned on new instructions
- Trained on: **15 tril** tokens, publicly available sources (although *unknown*)
- **8Bil/70/405 bil** parameters → **Llama-3-8b-instruct**
- Context length: **8,000**

Image only

- (5) CNN

Text + Image

- (6) CLIP
- (7) Idefics2

Models

Text only:

- (1) SVM
- (2) BERT
- (3) Llama2
- (4) Llama3

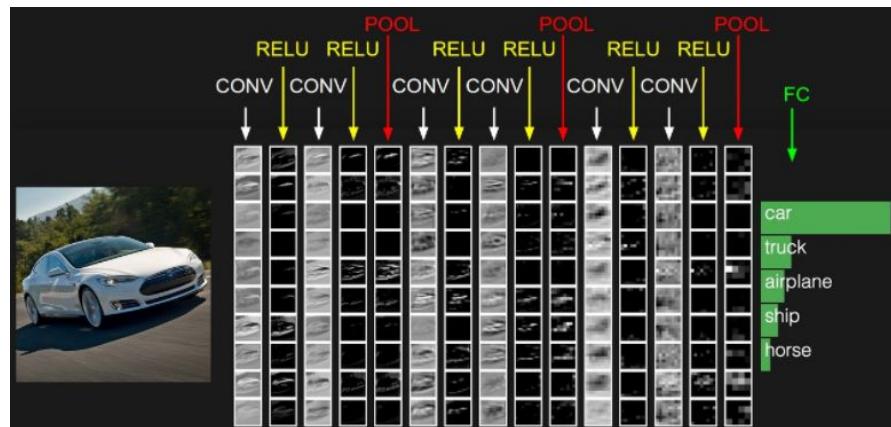
- Pre-trained for object recognition: 1,000 ImageNet object classes
- Trained on: **1.28 mil** images
- **25.6 mil** parameters → **ResNet50**
- Input size: 224 x 224 x 3

Image only

- (5) CNN**

Text + Image

- (6) CLIP
- (7) Idefics2



Models

Text only:

- (1) SVM
- (2) BERT
- (3) Llama2
- (4) Llama3

Image only

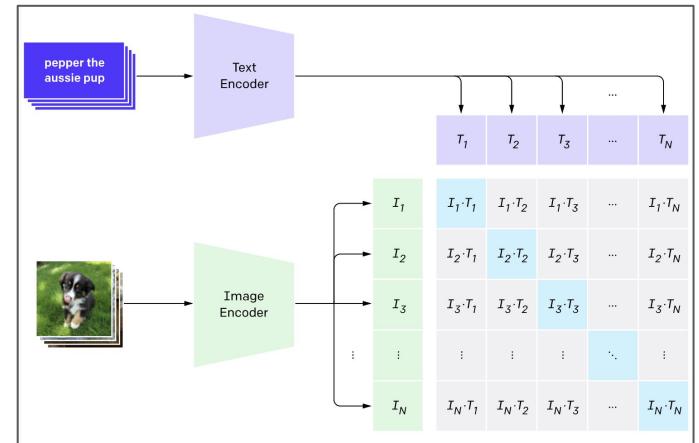
- (5) CNN

Text + Image

- (6) CLIP**

- (7) Idefics2

- Text and image encoder
- Trained on **400 mil** text-image pairs: e.g. image and its caption
- **Self-trained**: predicting correct text-image pair
- Image input size: **224 x 224 x 3**
- **150/400 mil** parameters → **ViT-B/32**
- Context length: **77 tokens**



Models

- Text and image **transformer** encoder
- Trained on:

Text only:

- (1) SVM
 - (2) BERT
 - (3) Llama2
 - (4) Llama3
- **interleaved image-text document: 350 mil images and 115 bil text tokens.**
 - **Image-text pairs**
 - **PDF OCR extraction: 40 mil**
 - **instruction/chat: 50 open-source datasets**
 - Total: **1.5 bil** images and **225 bil** text tokens

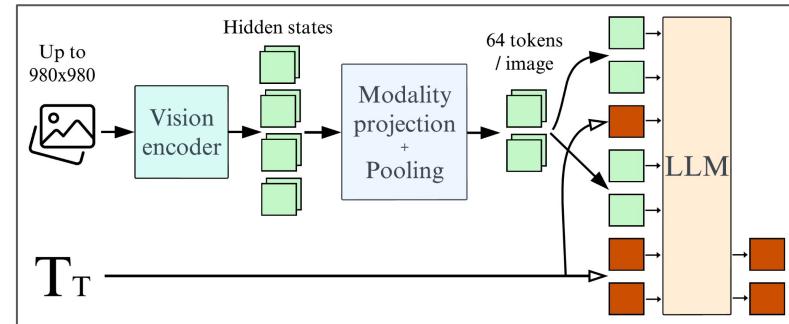
Image only

- (5) CNN
- **8 bil** parameters → **idefics2-8b**
 - Image input size: **native resolution (up to 980 x 980)**

Text + Image

- (6) CLIP
- (7) Idefics2**

- Context length: ?



Set Up

Text only:

(1) SVM

(2) BERT

(3) Llama2

(4) Llama3

- Same **train** (80%) - **test** (20%) sets across models
- Further split train set 80/20 into train/**validation** per **fold**
- **3 folds** (today results only for 1 fold) and **10 epochs/fold**
- Image processing/input:
 - Resize + center_crop all images: 224 x 224 x3
- Text processing/input:
 - 20,000 token vocabulary; no stopwords; linear kernel
 - 512 tokens/video
 - 800 tokens/video
 - 77 tokens/video

Image only

(5) CNN

- Fine-tuning:
 - Transcript_v + Frame_{v,f} → Label_v
 - BERT, CNN: new prediction head for each of our tasks
 - Llama2, Llama3, Idefics: same prompts across models
- Evaluation of image and text + image models:
 - Binary: Pred = 1 if sum(Frame_{v,f}) > threshold (=0 otherwise)
 - Multiclass: Pred = mode(Frame_{v,f})

Text + Image

(6) CLIP

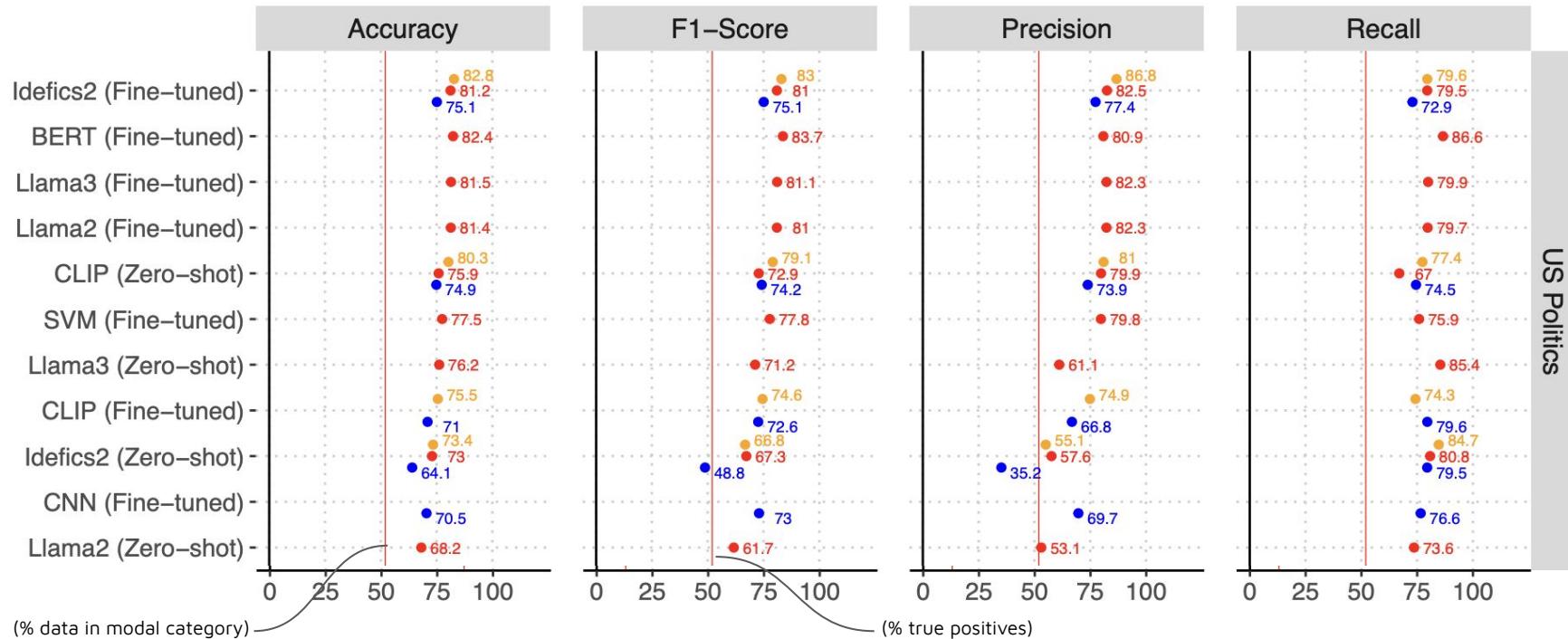
(7) Idefics2

A quick look at how to fine tune a VLM (ldefics2)

```
messages = [
    {
        "role": "user",
        "content": [
            {"type": "image"}, # frame_v1_f1
            {"type": "image"}, # frame_v1_f2
            {"type": "image"}, # frame_v1_f3
            {"type": "text", "text": transcript_v1,
             {"type": "text", "text": "Is the previous text and images about or relevant to US politics? Answer YES or NO."}
        ]
    },
    {
        "role": "assistant",
        "content": [
            {"type": "text", "text": "YES"}
        ]
    }
]
```

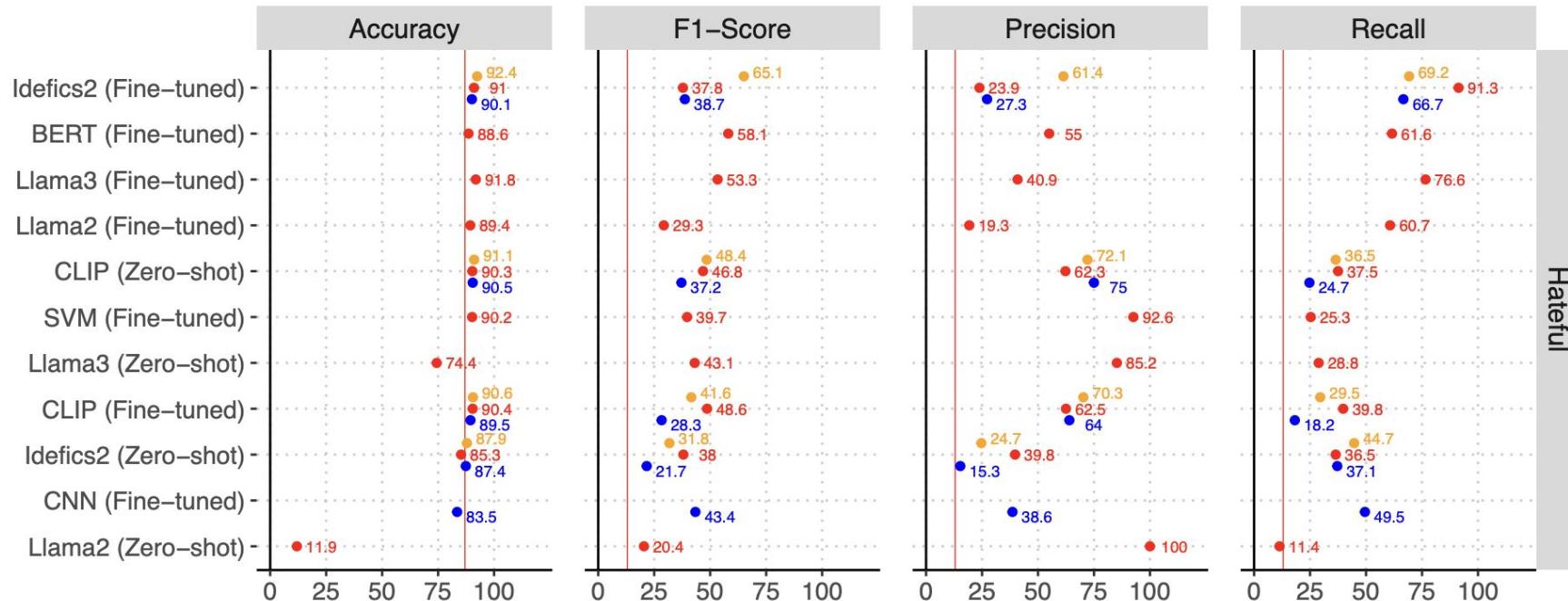
Results: overview

• Image • Multimodal • Text



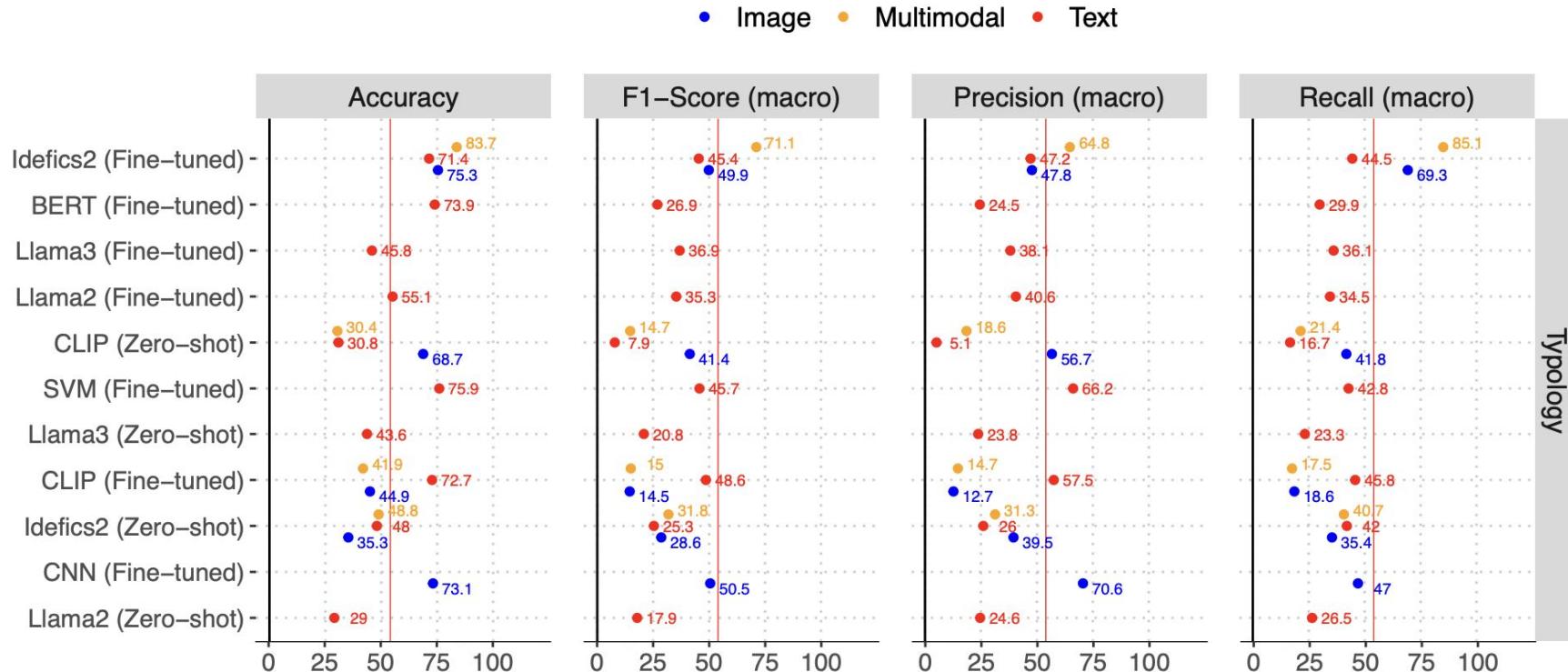
Results: overview

• Image • Multimodal • Text



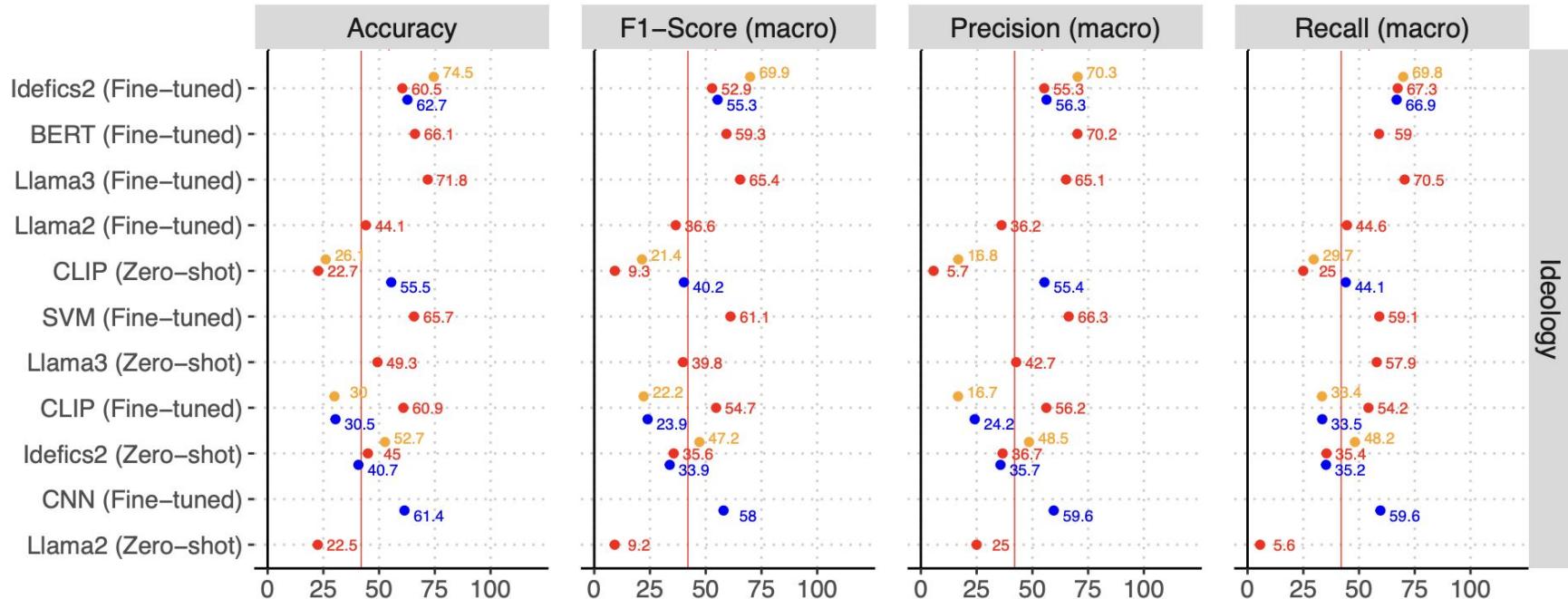
Hateful

Results: overview

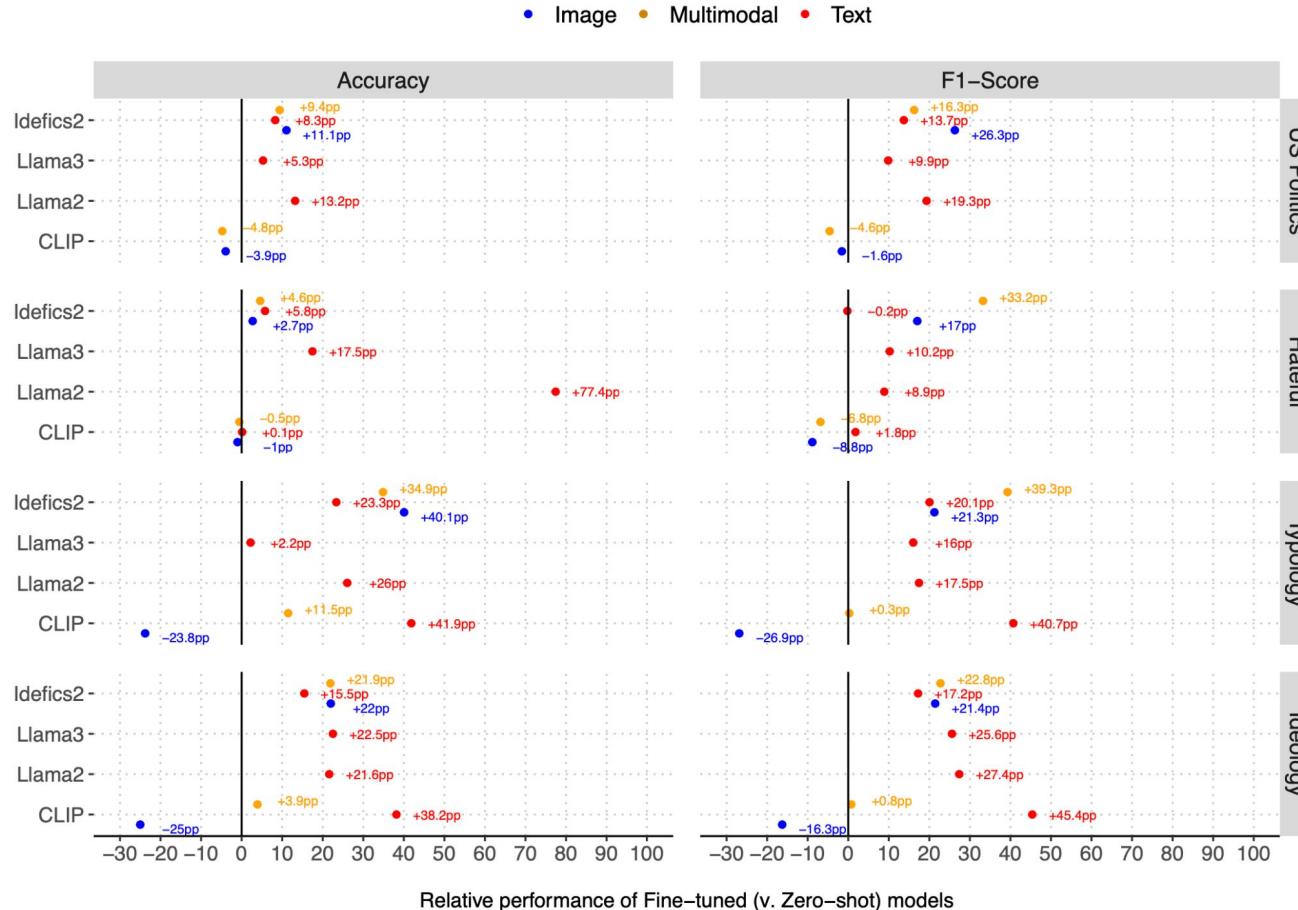


Results: overview

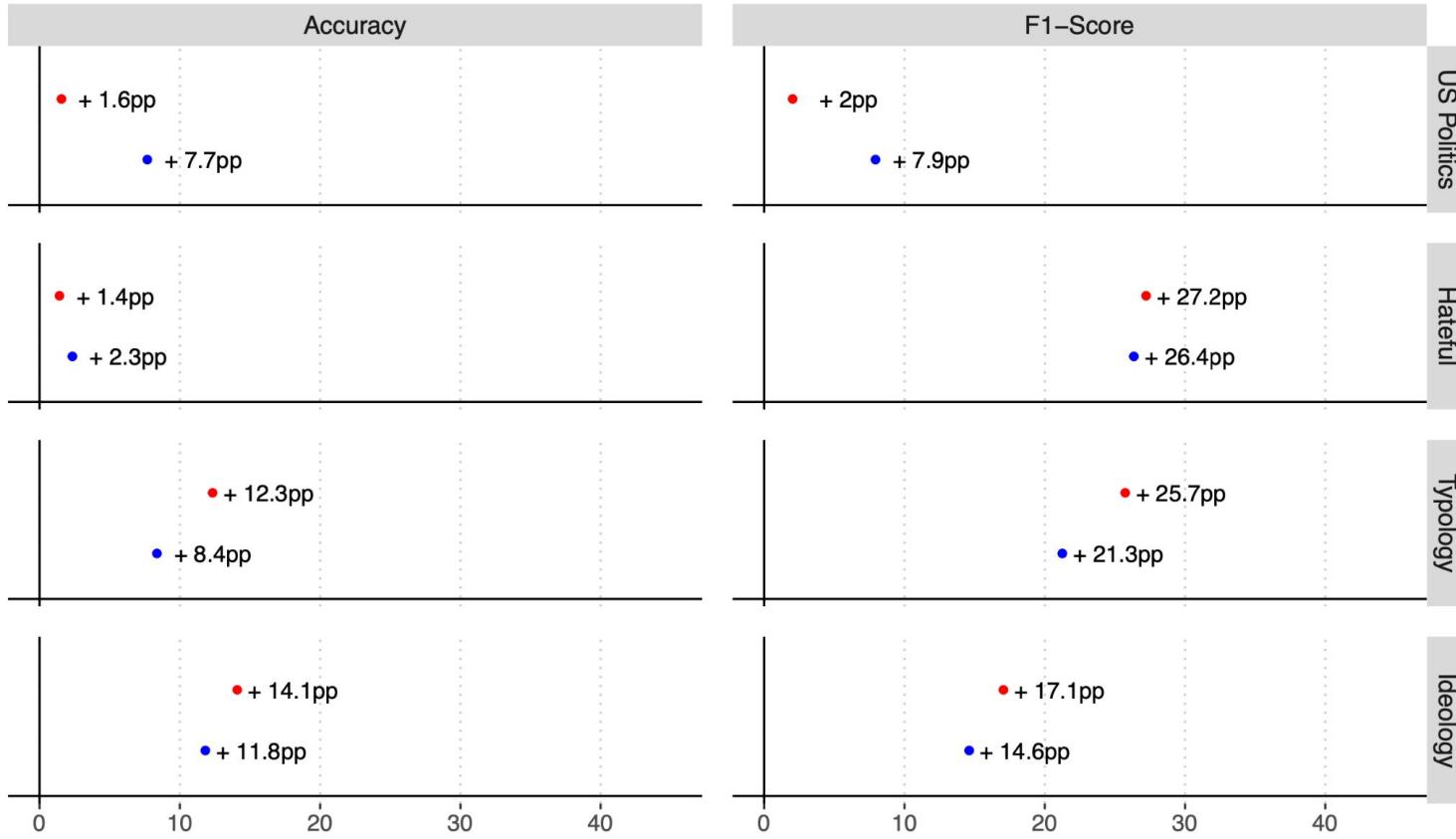
• Image ● Multimodal ● Text



Results: fine-tuning makes a big difference

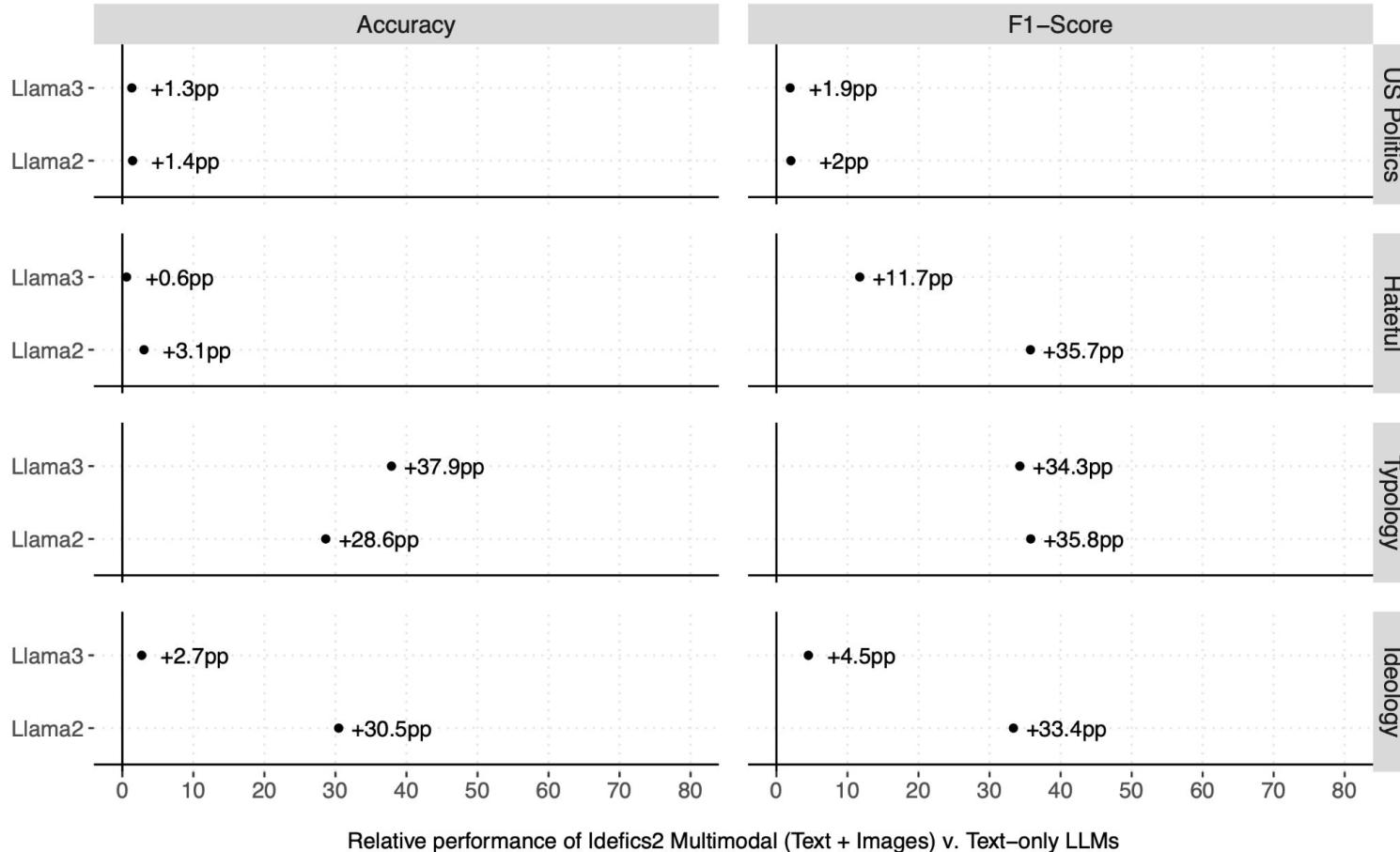


Results: VLM performs better in multimodal settings

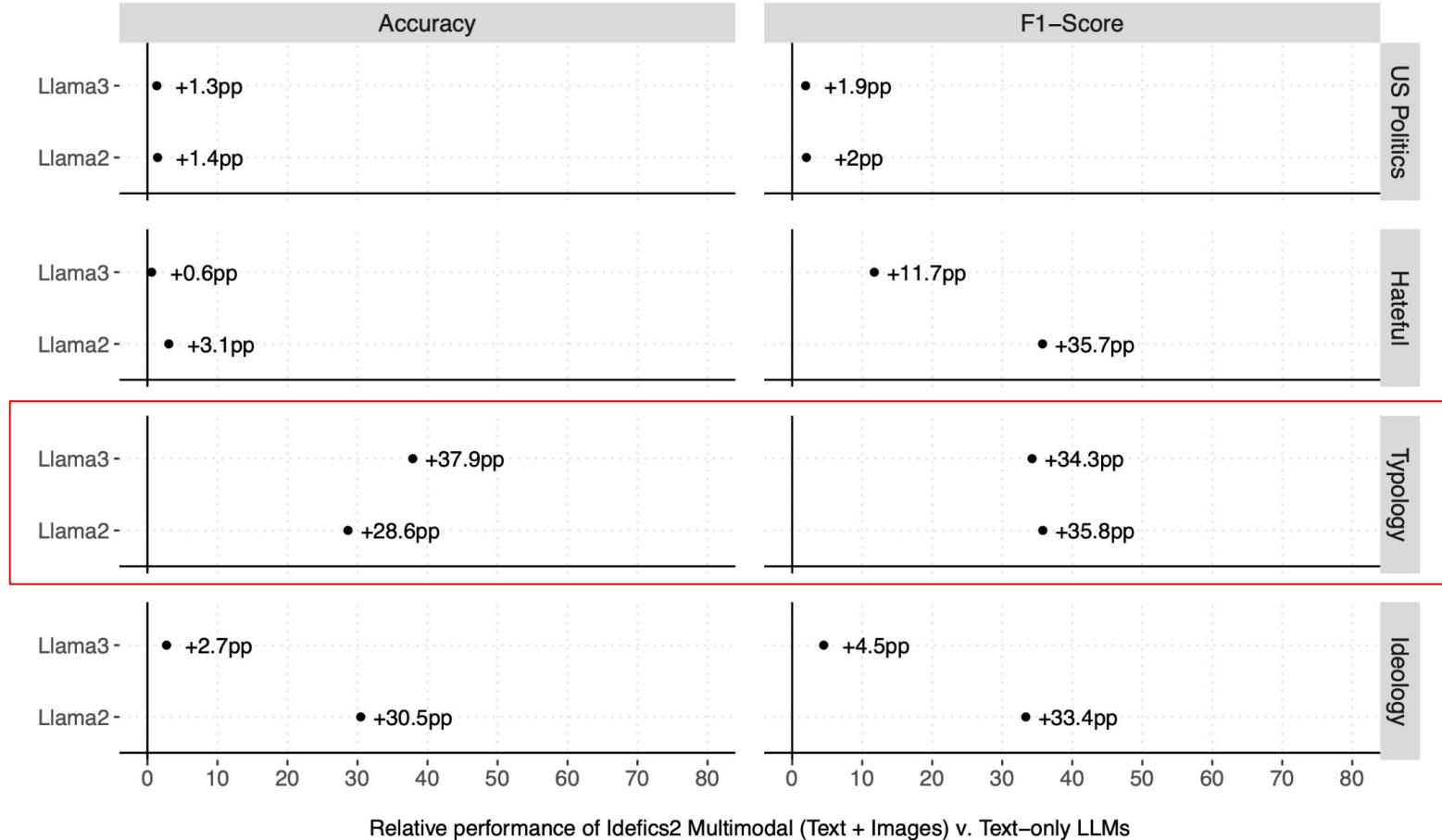


Idefics2: relative performance of multimodal v. text- and image-only

Results: VLM outperforms SOTA open-source LLMs



Results: ... particularly on more visual-dependent task



Limitations

- Current results based on only 1 fold
- Smaller Llama2 and Llama3 models (7B v. 40-400B parameters)
- Only max. of 800 text tokens per video
- Probably some more prompt engineering is needed
- NEW promising open-source VLM: Idefics3-8B-llama3

Conclusions and next steps

- Multimodality helps to improve performance (v. text-only LLMs)
- Particularly on more visual-dependent tasks
- Next steps:
 - address limitations discussed in previous slide (folds, prompts, add larger/new models)
 - run the same computational experiments on the missing target variables for the YouTube dataset; and also on the Interest Group Twitter data
 - potentially adding one more dataset (TikTok) with higher visual dependence