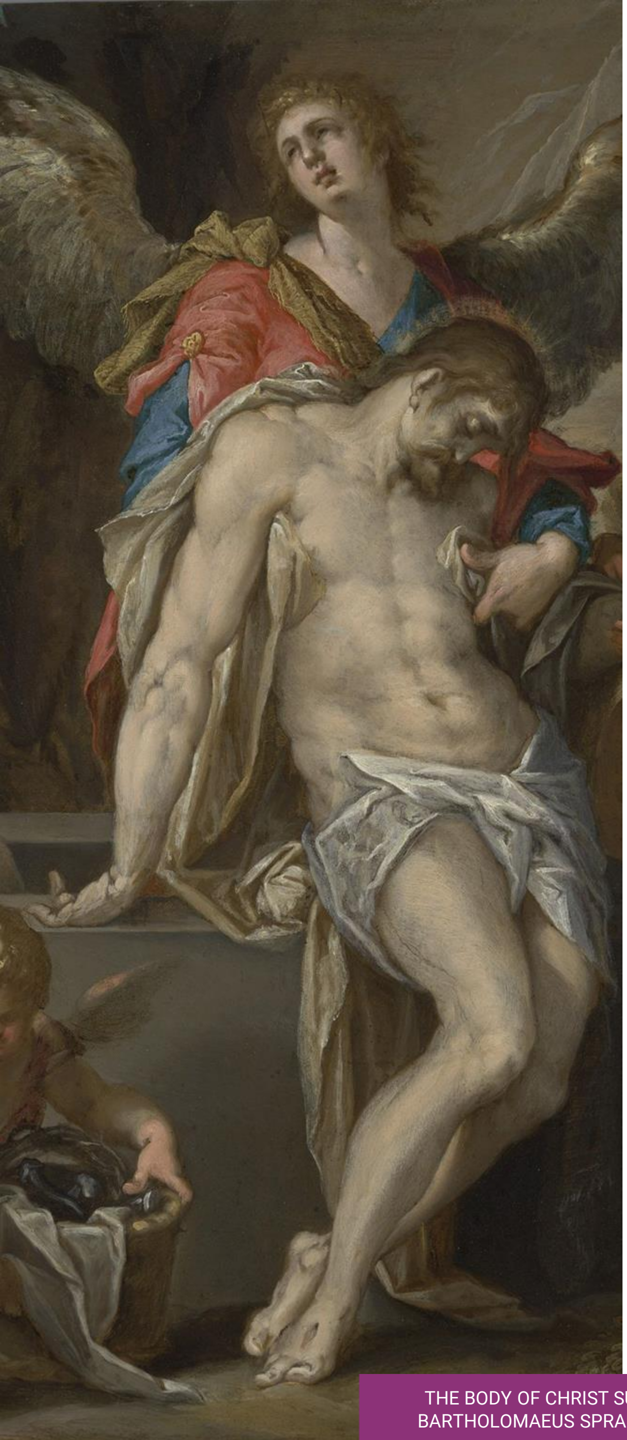
Saint George on a Bike: issues in small data and how to mix bottom-up and top-down approaches to compensate for the lack of big data.

BSC team members: Artem Reshetnikov, Joaquim Moré, Cedric Bhihe, Sergio Mendoza,
Maria-Cristina Marinescu



THE BODY OF CHRIST SUPPORTED BY ANGELS
BARTHOLOMAEUS SPRANEERS. PUBLIC DOMAIN

Motivation

Focus on cultural heritage as a way to understand our past, approach the future, find inspiration, innovate.
An area with a lot of metadata issues!
Good labels and descriptions enable research, education / cultural / social projects, and can improve web accessibility for the blind.

Goal: Contextualize the objects and image composition to ultimately endow AI with culture, symbols and tradition insight (and generate rich metadata).
Focus on (figurative) paintings of XII-XVIII centuries (especially iconography). Europe!

Automatic metadata annotation

New forms of interaction with users through web-pages and apps

Interaction with minorities such as visually impaired citizens

Improve search and browse

Why not use current tools?

The main challenge:
Current approaches are very successful for everyday images, but fail for cultural heritage. They work well for recent pictures, give that they were trained on very large datasets with these characteristics.

Main challenge

E.g.

- Old objects not in use anymore, e.g. inkwell
- Objects with different shapes in the past, e.g. plow
- New objects, different but with similar shape as old ones, e.g. cell phone/book
- Unusual actions for everyday life, e.g. man killing a horse

Use jointly techniques from different (AI) fields to apply them to images or (image, text) pairs: deep learning, natural language-based models, [semantic metadata extraction and reasoning]

Data input & output

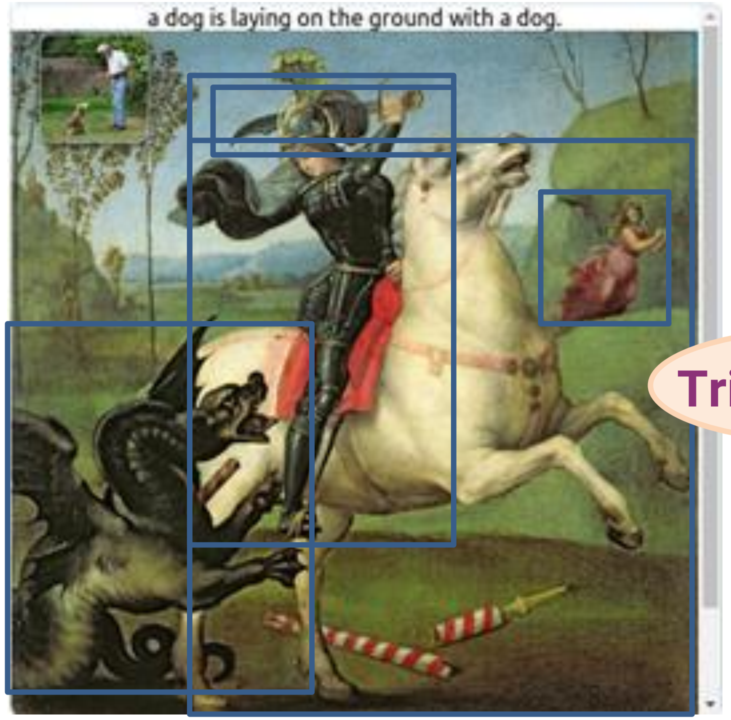
Input: image and possibly metadata

Output: different semantic levels

Semantic Level	*Examples*
Semantic resources (tags) *From vocabularies, preferably with linked data URIs.* **Class annotation**	Adoration of the Magi: • Jesus Christ, Virgin Mary, Wise Man (as subjects coming from a vocabulary). • http://iconclass.org/rkd/73B57/: "Adoration of the kings: the Wise Men present their gifts to the Christ-child (gold, frankincense and myrrh)."
Textual captions **Description generation**	"Man reading a book in a dark room." "Woman plays a guitar outdoors during sunny weather."
Semantic/knowledge graph *Graphs with relationships between semantic resources, where the link can also have a URI.*	(St. George, kill, dragon) (Woman, sits) Adoration of the Magi: (Wise Man, adore, Jesus Christ), (Virgin Mary, hold, Jesus Christ)

Triples (s,p,o)

Our two approaches to generating rich annotations



Saint George riding a horse kills the dragon. The princess runs in the background.

Triples (s,p,o)

Class annotation

Simple triple-like caption seeds:
objects: knight, sword, horse, dragon, woman
caption seeds: (knight kill dragon), (knight ride horse), (woman run)
Why this approach: no visual descriptions of image content is available
What this alternative implies: 1. obtaining object annotations to train for detecting objects 2. generating likely relationships between objects

Natural language visual descriptions
What this alternative implies: obtaining description annotations for images

Description generation

Pipeline

Time context

Object detection

Mask R-CNN based on a dataset (prepared for cultural heritage domain) and transfer learning

Class annotation

Pose classification

candidate object labels for each bounding box

Visual Relationships

Inferring visual relationships between detected objects in images with complex scenes using bounding box analysis

NLP for verb generation and refinement of classes

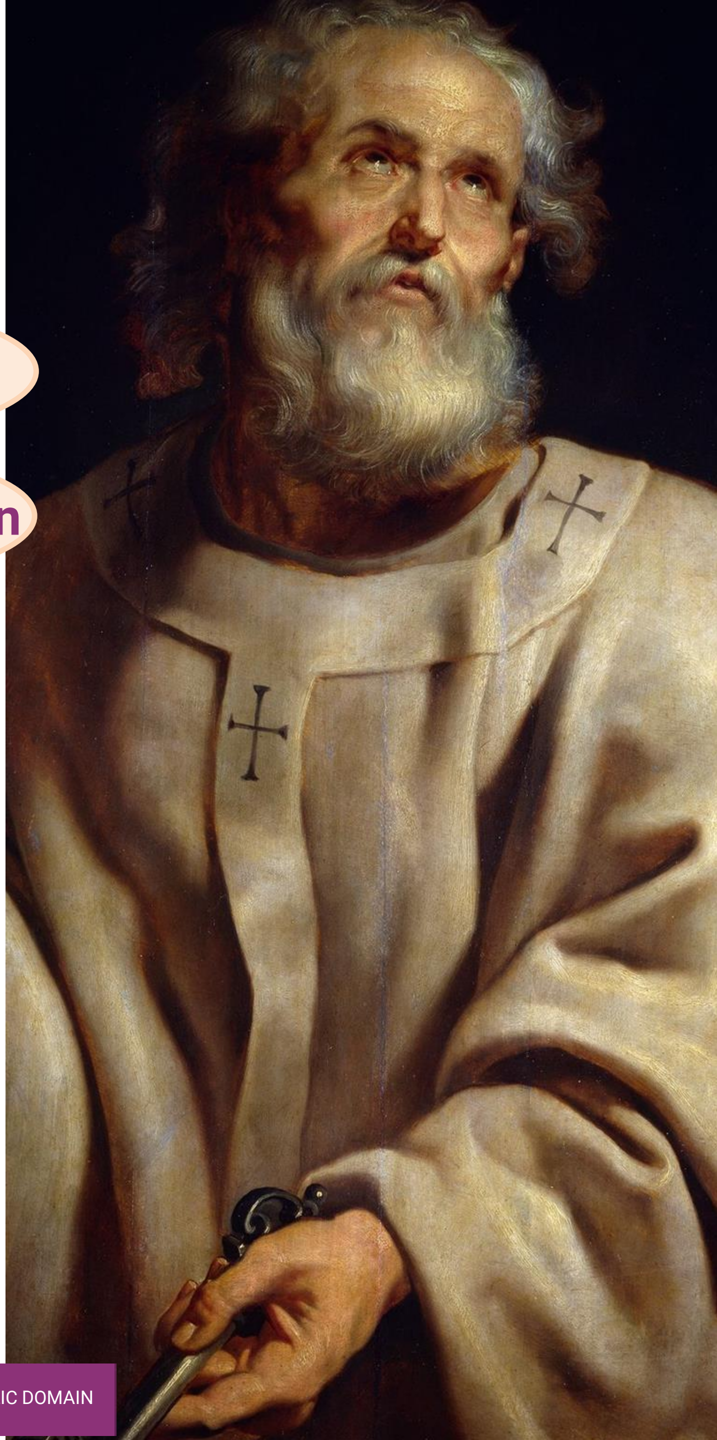
The model attempts to predict the original value of a masked (relationship, object) word. Prediction is based on the semantic context. Semantic context is provided by the other, non-masked, words in the sequence.

Triples (s,p,o)

Description classifier

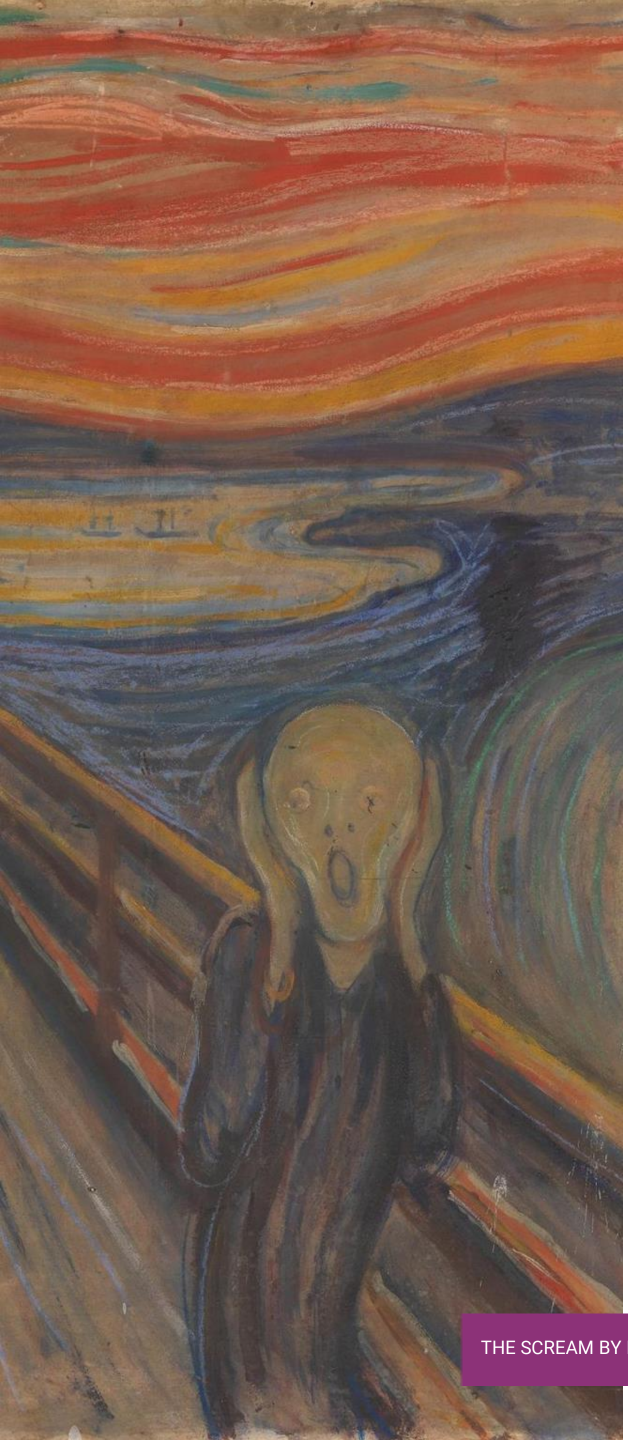
Crowdsourced descriptions

Description generation



SAINT PETER BY RUBENS. PUBLIC DOMAIN

Object detection + Pose classification

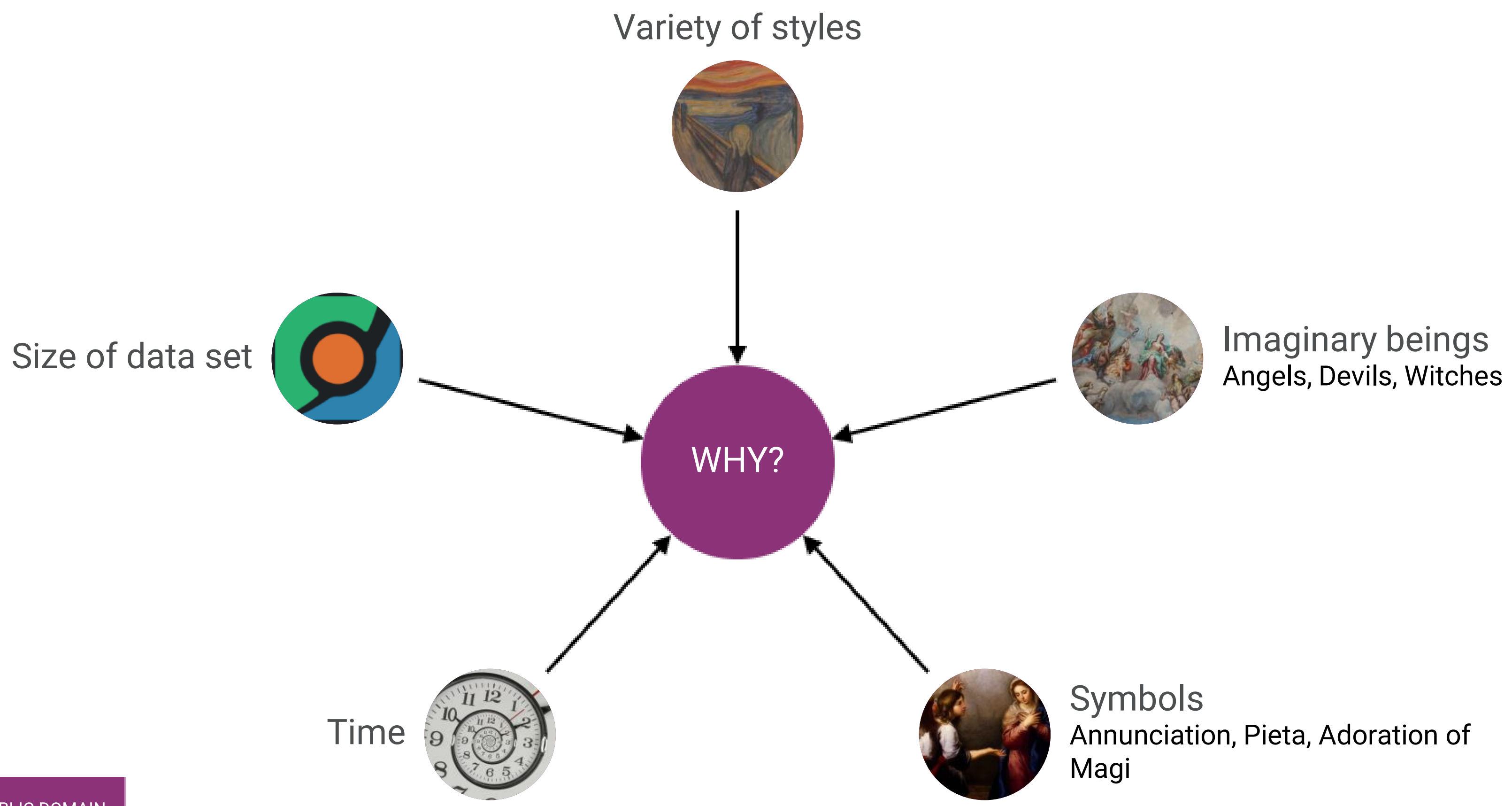


THE SCREAM BY EDVARD MUNCH. PUBLIC DOMAIN

Identifying the problem

Current approaches are very successful for everyday images, but fail for cultural heritage. They work well for recent pictures, given that the were trained on very large datasets with these characteristics.

And cultural Heritage?



DEArt Dataset

Images from…

Europeana Collection

WIKIART

British Museum

MS COCO

Pharos

Getty Museum

IconClass AI Testset

Museum d'Orsay

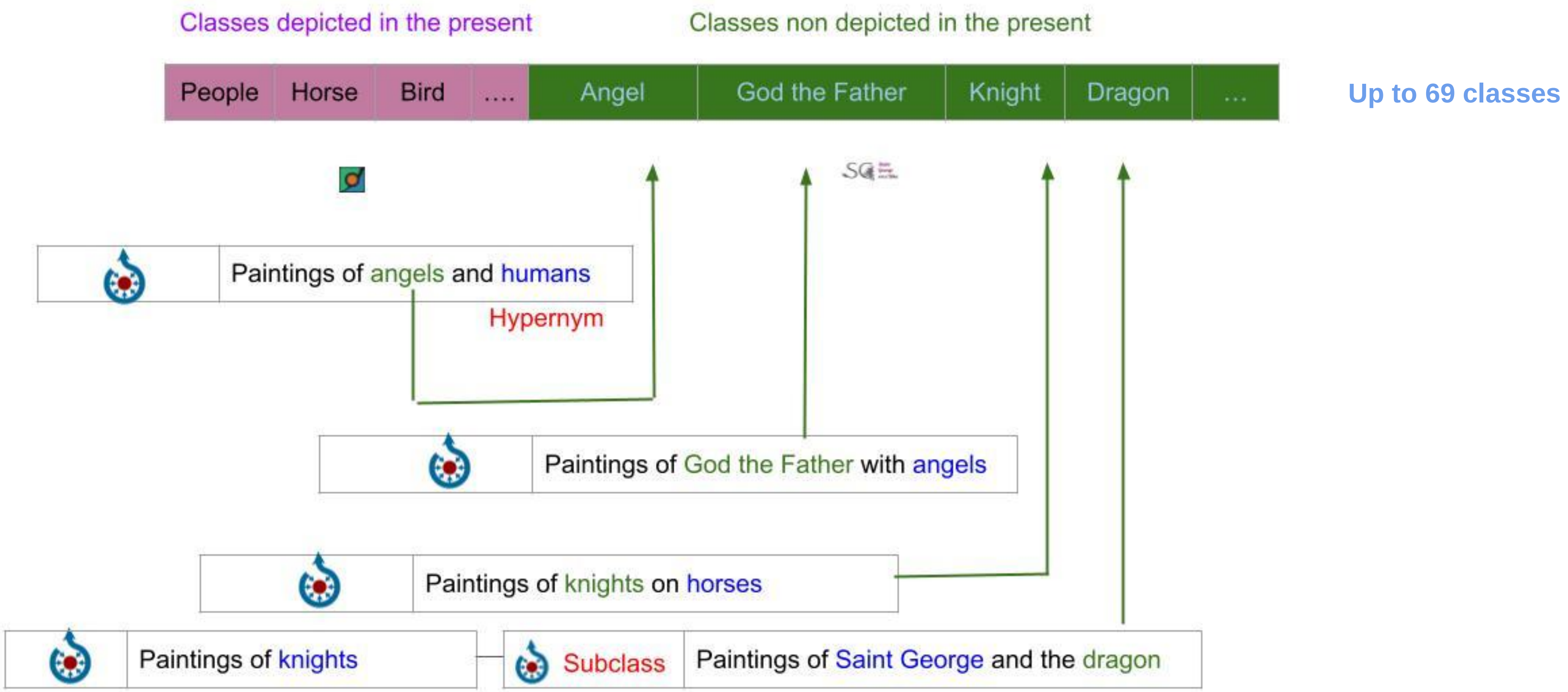
Web Gallery of Art

Wikimedia Commons, WikiData, Wikipedia

Prado Museum

Rijksmuseum

Selection of classes to be detected



THE ANNUNCIATION BY MURILLO. PUBLIC DOMAIN

DEArt

69 classes

15k images (publishing)
21 (training)

13 poses

30+k captions

Bounding boxes with class and pose labels (for human-like objects)

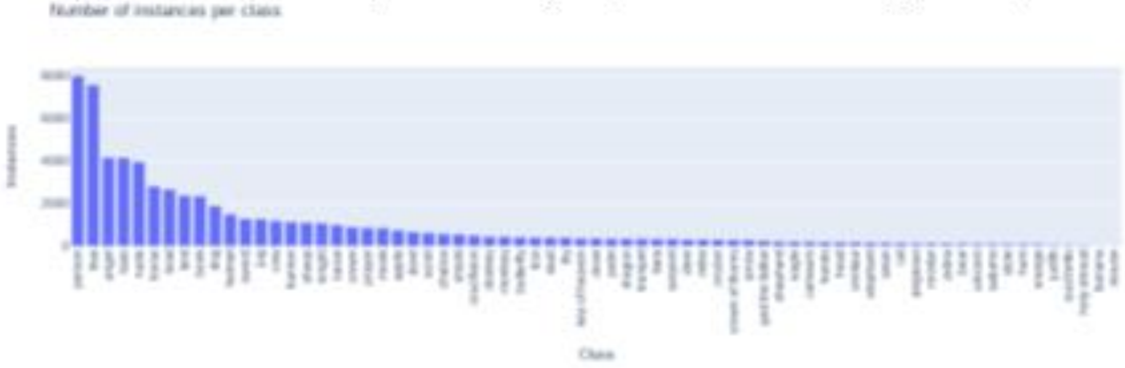
Annotation process:

- Follows PASCAL Visual Object Classes (VOC) Challenge: consistency (guidelines), accuracy (manual check), exhaustiveness (manual check)
- 10K images manually; 5K images with semi-supervised approach in 3 batches followed by retraining (over 70% of dataset).
- Double-check annotation quality every 2K images: random check of 100 images for each of top 10 classes

Object detection via deep learning

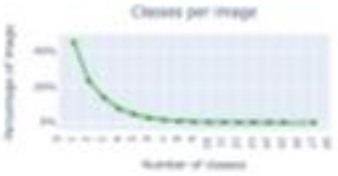
Some statistics

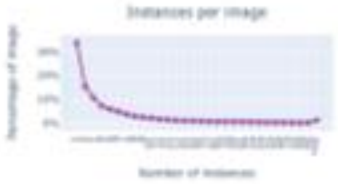
Approximately 38K `person` objects, 105799 bounding boxes, 56230 human-like creatures



Instances and poses per class

Contextual information (average number of object classes and instances per image)





Object detection via deep learning

a. Transfer learning using Resnet-152 V1 object detection model, pre-trained on MS COCO 2017
b. Faster R-CNN architecture for training: 70% training, 15% validation, and 15% test sets.
c. Choice of images is random within each class; we use annotated-images2 (Python library) to select images such that these percentages are as closely as possibly met for each of the 69 classes.
d. We place detected objects in temporal context to choose most probable class, e.g. horse vs motorcycle, book vs cell phone

AP@0.5 per class
mAP@0.5 = 31.2

Object detection via deep learning

Results: Testing existing models over DEArt

	apple	banana	bear	bird	boat	book	cat	cow	dog
MS COCO	0.04	0.008	0.03	0.12	0.24	0.05	0.04	0.23	0.12
Open Images	0.008	0.005	0.12	0.01	0.07	0.00007	0.04	-	0.08
Pascal VOC	-	-	-	0.02	0.05	-	0.09	0.13	0.02
DEArt	0.13	0.12	0.39	0.15	0.42	0.34	0.21	0.33	0.45

0.44 in the COCO dataset

0.36 in the COCO dataset

0.28 in the COCO dataset

- Results for the 16 classes that are included both in MS COCO and our dataset.
- 53 cultural heritage-specific classes not covered.

elephant	horse	mouse	orange	person	sheep	zebra
0.21	0.2	0	0.4	0.25	0.15	0.89
0.38	0.09	0	0	0.07	0.04	0.5
-	0.03	-	-	0.05	0.004	-
0.56	0.34	0.09	0.09	0.68	0.2	0.91

0.33 in the Pascal dataset

0.22 in the Pascal dataset

0.17 in the Pascal dataset

Pose classification of human-like objects

Results: Pose classification

Xception network, trained from scratch: 70% training,15% validation, and 15% test sets.

KerasTunner for hyper-parameter tuning.

F1 score due to high imbalance between pose labels: F1=0.471, weighted F1=0.89.

Class	bend	fall	kneel	lie down	partial	pray	push / pull	ride	sit / eat	squat	stand (up)	unrecognized	walk
F1	0.33	0.08	0.32	0.8	0.9	0.33	0.09	0.1	0.83	0.12	0.84	0.89	0.5

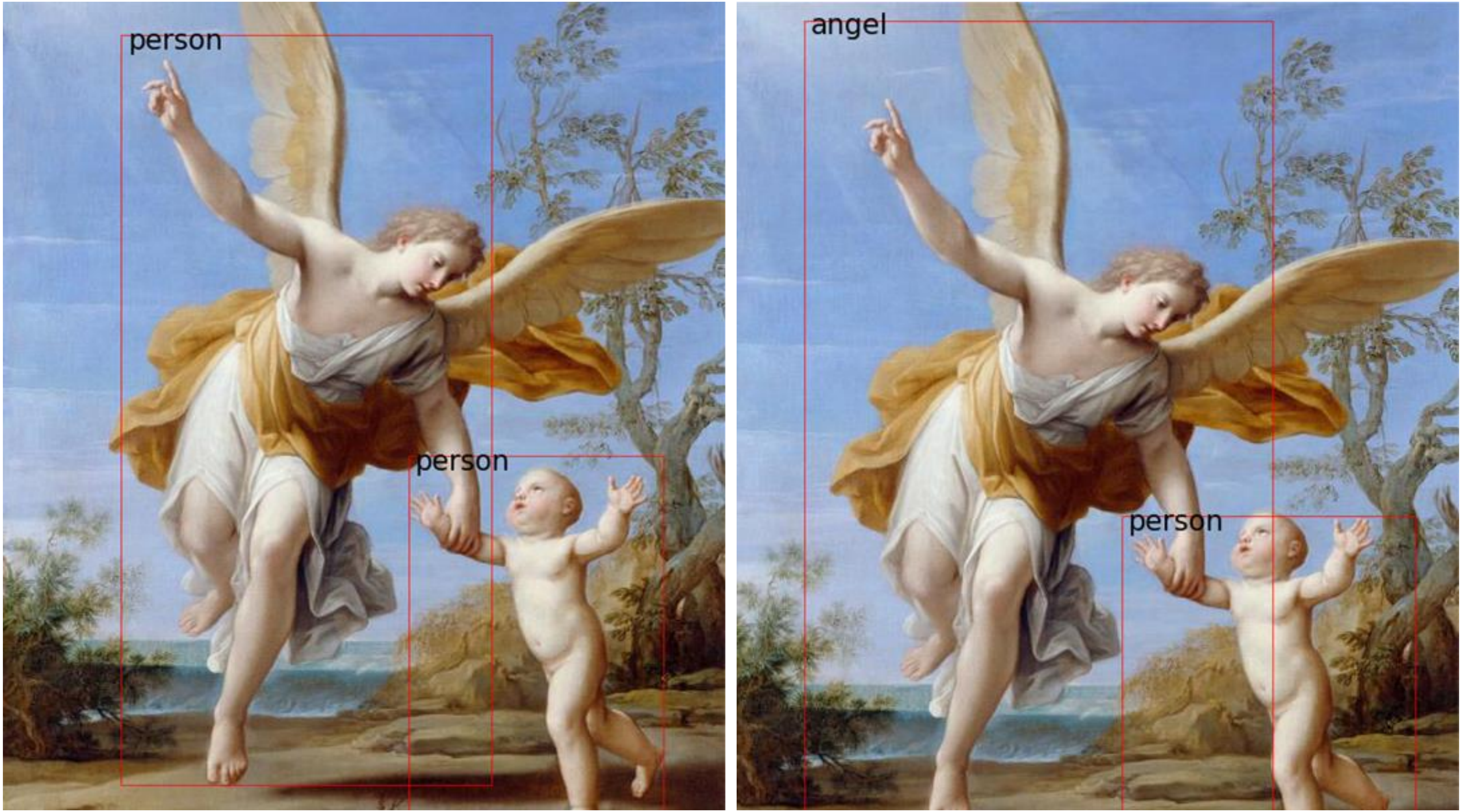
Minority classes

Current **list** of **classes**

crucifixion, angel, person, crown of thorns, horse, dragon, bird, dog, boat, cat, book, sheep, shepherd, elephant, zebra, crown, tiara, camauro, zucchetto, mitre, saturno, skull, orange, apple, banana, nude, monk, lance, key of heaven, banner, chalice, palm, sword, rooster, knight, scroll, lily, horn, prayer, tree, arrow, crozier, deer, devil, dove, eagle, hands, head, lion, serpent, stole, trumpet, judith, halo, helmet, shield, jug, holy shroud, god the father, swan, butterfly, bear, centaur, pegasus, donkey, mouse, monkey, cow, unicorn

Current **list** of **poses**

bend, fall, kneel, lie down, partial, pray, push/pull, ride, sit/eat, squats, stand, walk/move/run


person
person
angel
person

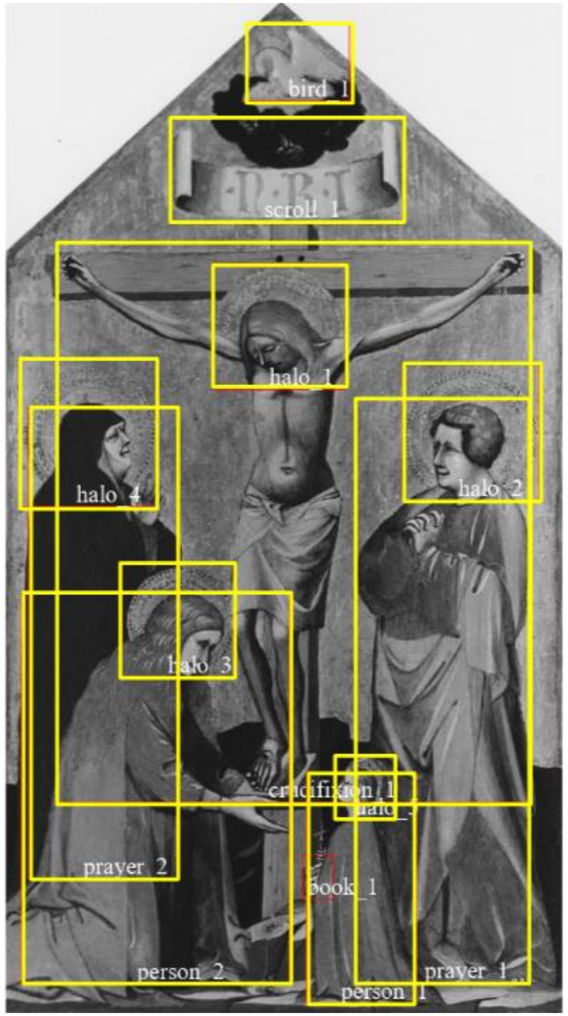
Object detection via deep learning

chalice
jug
jug
jug
person
zucchetto
book
boat
banner
boat
banner
boat
banner
mitre
person
crucifixion
crown of thorns

Saint George on a Bike

Triple generation: Bounding box analysis (VizRel)

Analysis methodology



1 Parametrize images

Calculation of parameters such as object label, label identifier, unique label, bbx center point, object location, relative surface area, orientation and form factor, etc.

2 Pick criteria

Choose relevant parameters based on main topic candidate, hypernym, symbolic content).

3 Elaborate rules

Inference of visual relationships between co-occurring objects is rule-based (i.e. heuristic). It allows for the elucidation of relative positions of pairs of detected objects, detection of bbx overlaps, general ordering of objects in the composition.

4 Propose visual relationships

Final output

Rule-based visual relationships

Detected objects:

person_1, crown_1, person_2, halo_1

Reference objects:

('person_1', (449.5, 700.0), 'cc', 20.61)

Visual relationships:

person_1+crown_1:

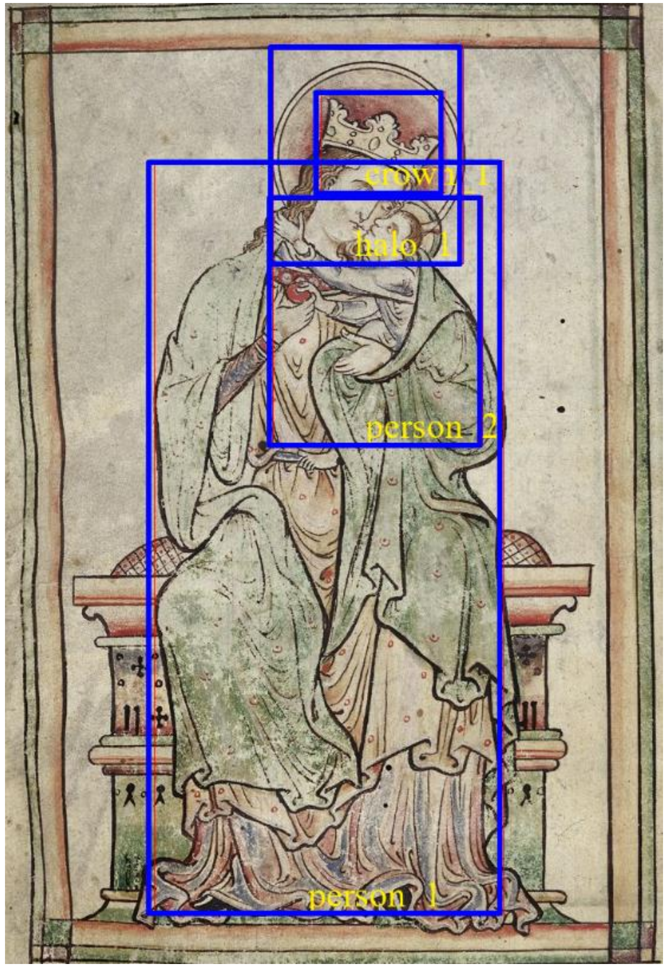
('person_1 stands', 'person_1 wears crown_1', 'person_1 coiffed_with crown_1) is

king/queen/saint_mary

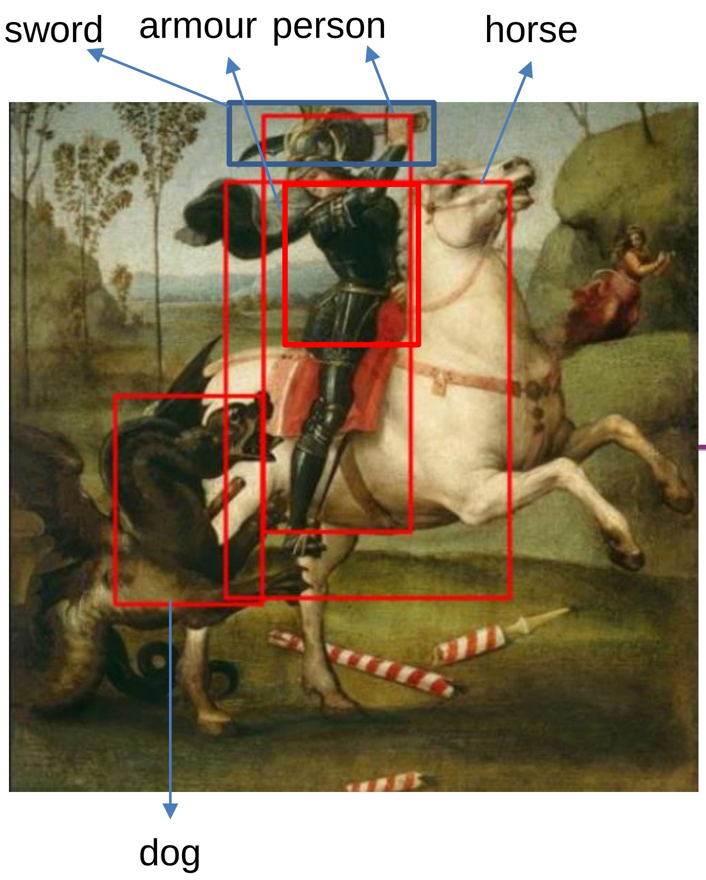
person_1+person_2:

('person_1 stands', 'person_2 is (child)/(infant)/(dwarf))',

person_1 holds person_2



Refining classes and generating/refining relationships via a language model



Dog

Knight ____ dog → ????

Class refinement

Knight ← Person with Armour is a …

Knight __on__ horse → Knight rides horse

Relationship refinement

Horse

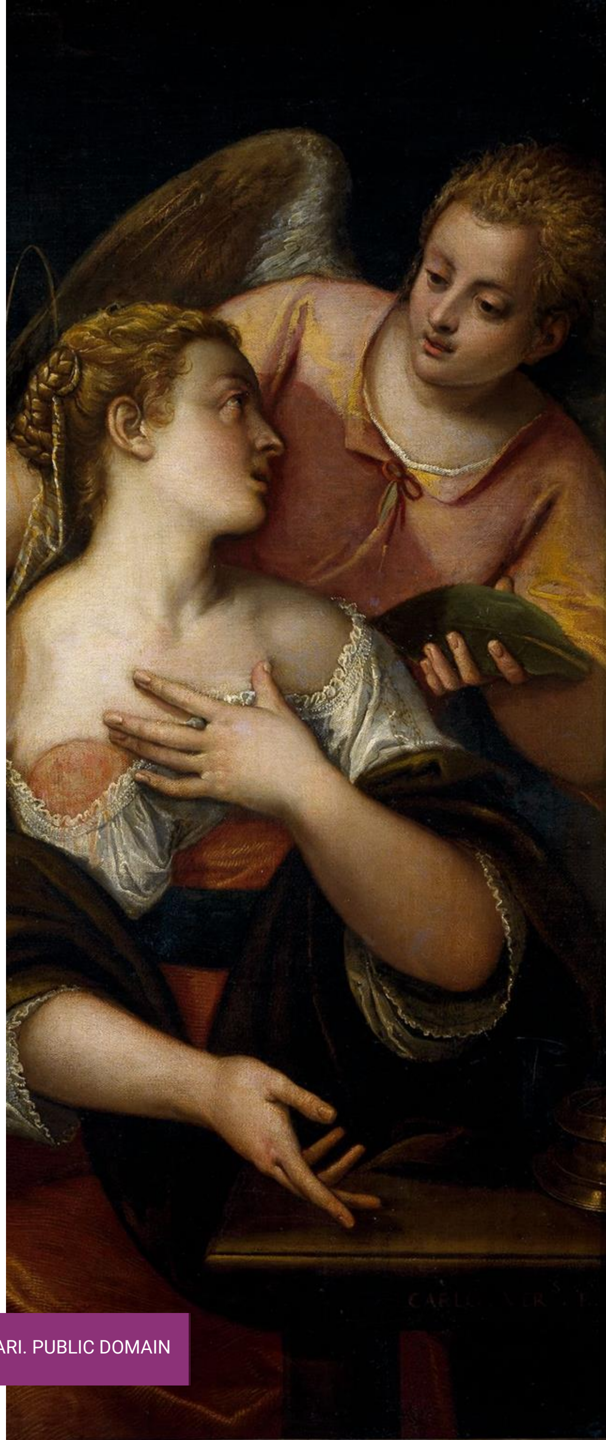
Knight ____ sword → Knight holds a sword

Relationship generation

Sword

Sword ____ dog → ????

Dog



SAINT AGATHA PAINTING BY CARLO CALIARI. PUBLIC DOMAIN

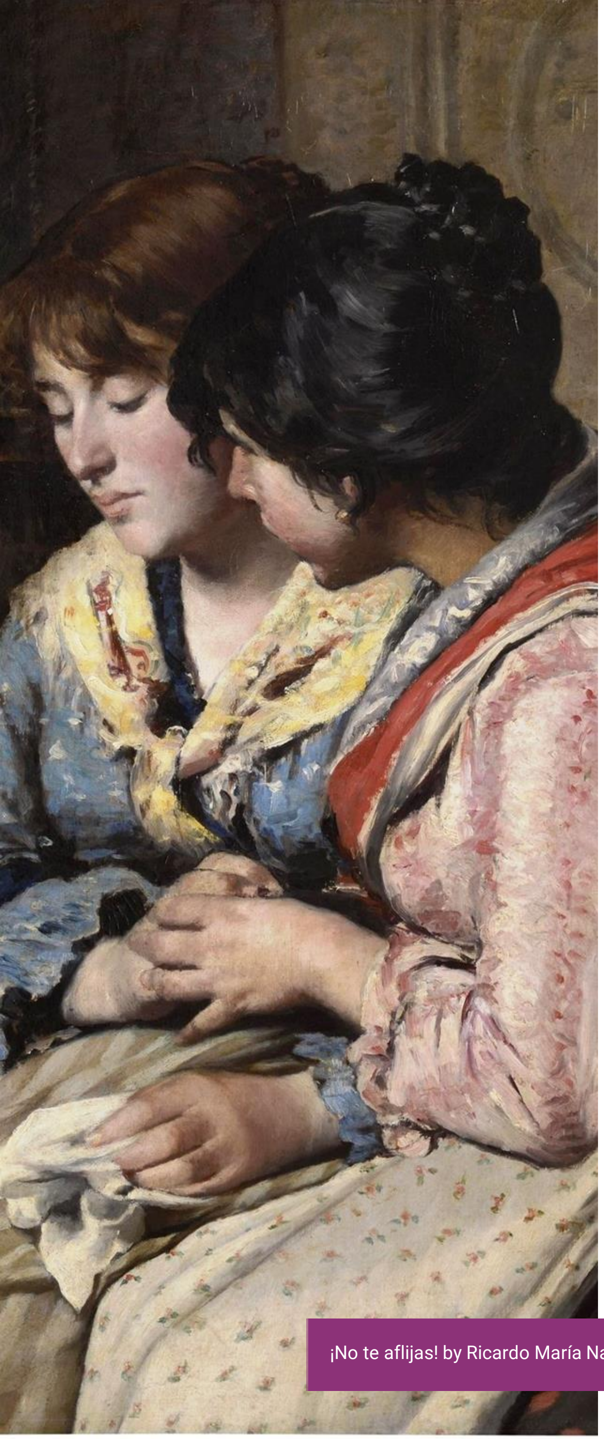

ÍÑIGO LÓPEZ DE MENDOZA BY FRANCISCO DÍAZ CARREÑO. PUBLIC DOMAIN

- Based on CLOZE test
- Transformer-based language model
- Model attempts to predict the original value of a masked word
- Prediction is based on the semantic context
- Semantic context is provided by the other, non-masked, words in the sequence

Caption generation

Dataset for visual description generation

- With previous approach we can generate sets of triples (object, relationship, object) or actions such as standing, eating, etc.
- To generate full descriptions in natural language, we need a sizeable dataset of aligned paintings / descriptions
- Use deep learning



¡No te aflijas! by Ricardo María Navarrete Fos. Public Domain

Use Zooniverse crowdsourcing platform

7543 images annotated with 4-5 descriptions.

Our goal is annotation of all 15K images with 5 annotations per image

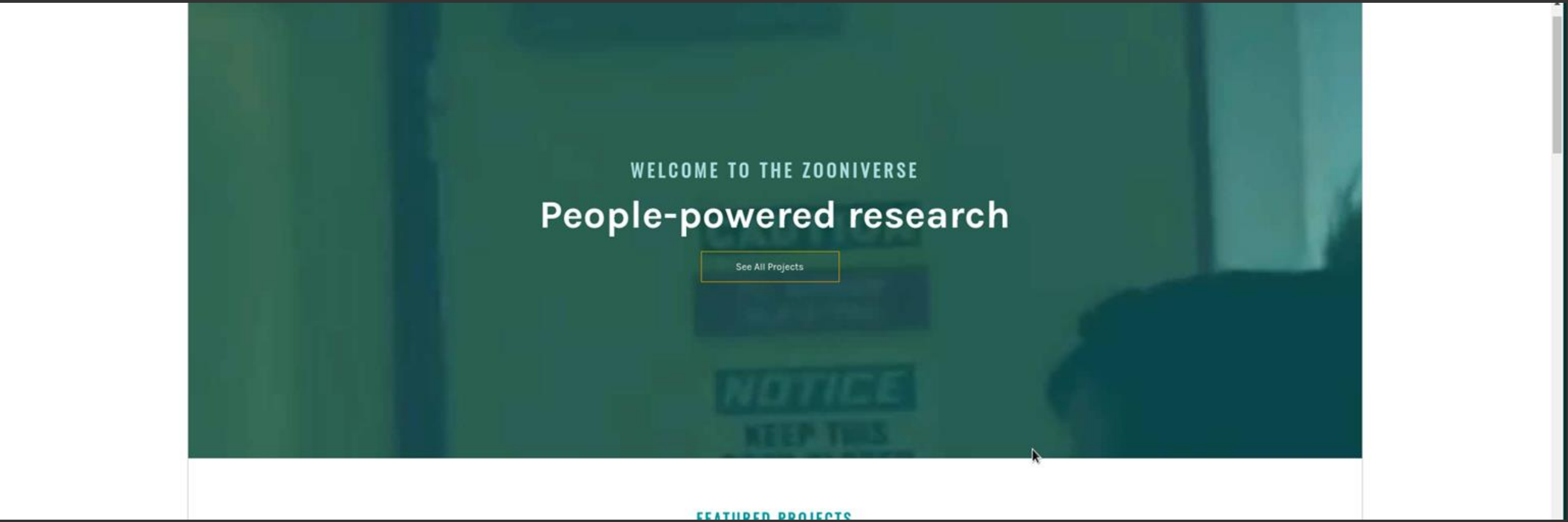
1,859 volunteers

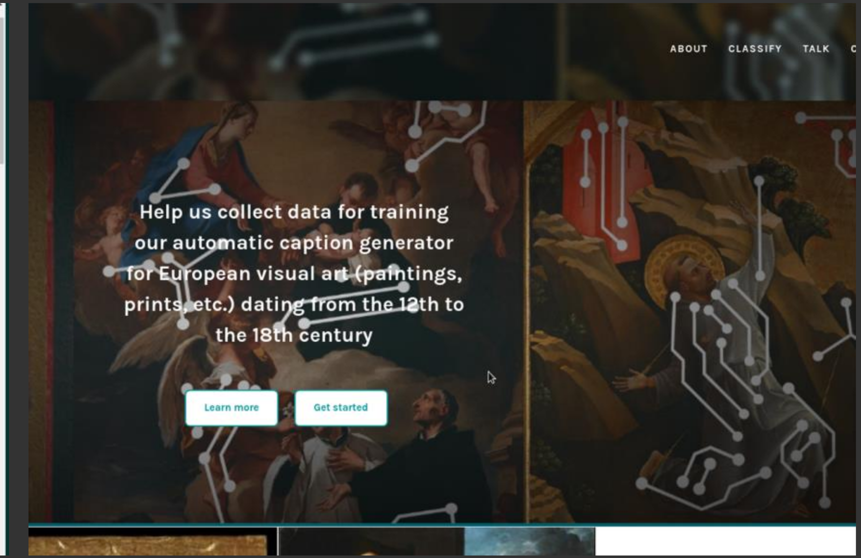
154 discussion threads

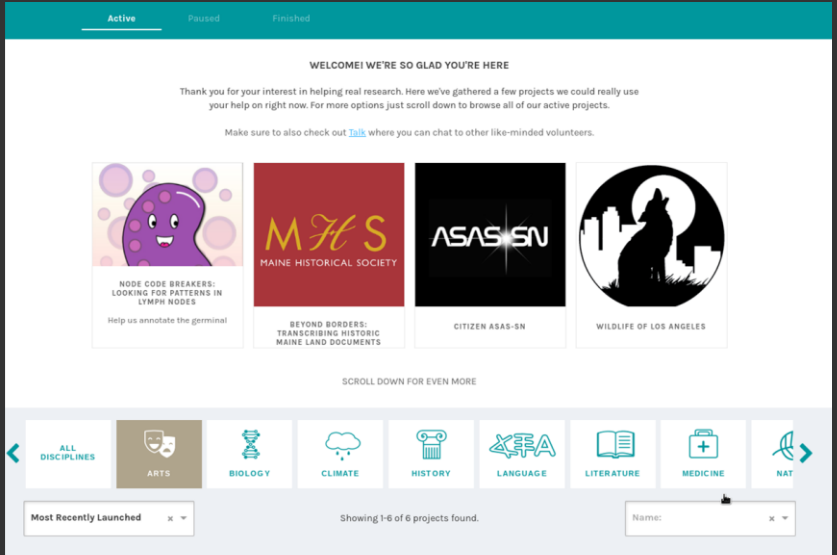
362 comments

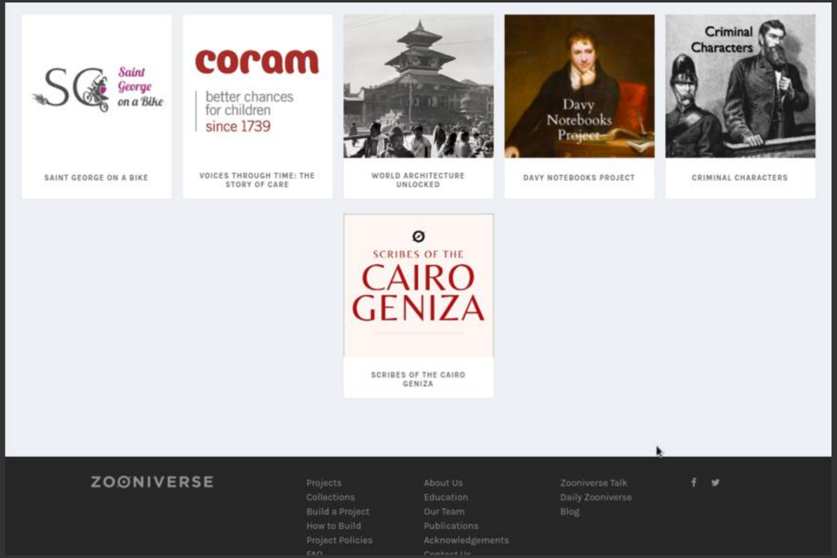
17 media and web mentions

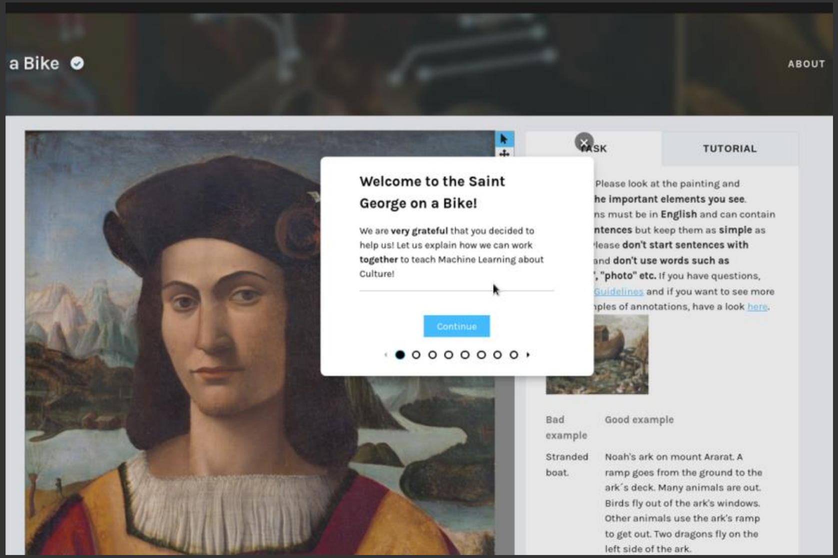
Developed and implemented a set of guidelines











https://www.zooniverse.org/

https://www.zooniverse.org/projects/artem-dot-reshetnikov/saint-george-on-a-bike/

Generation of more complex (natural language) descriptions

Training using attention mechanism

- Own trained model for encoder: detecting features specific to iconography (e.g. angel, monk, sword, Christ) was a key factor necessary for a good decoder performance.
- Decoder: Recurrent Neural Network (RNN) with attention mechanism (GRU or LSTM). This approach is efficient only if the encoder can correctly detect features that enable labeling objects with names that can help the decoder make the correlation between specific areas of the image and description words.

halflength elderly man in fur coat and jacket looks at the viewer

Currently we generate good captions for not very complex paintings (portraits, biblical scenes with few details, iconic paintings).

Intention to follow up on the crowdsourcing campaign.

mother mary sits with the baby jesus on her lap jesus holds fruit in his hand

mother mary sits with the baby jesus on her lap jesus holds fruit in his hand

halflength elderly man in fur coat and jacket looks at the viewer

Evaluation of enrichments resulting from object detection

Evaluation results

evaluated images	generated enrichments	correct and precise	merely acceptable	relevant
ca.700	ca.2100	78%	5%	70%

The **recall** measured was between 58-77%.

Limitations: many classes that are relevant (e.g. Jesus Christ, Virgin Mary) were excluded from the final list of target classes and may be be detected only in other enrichment steps (caption generation).

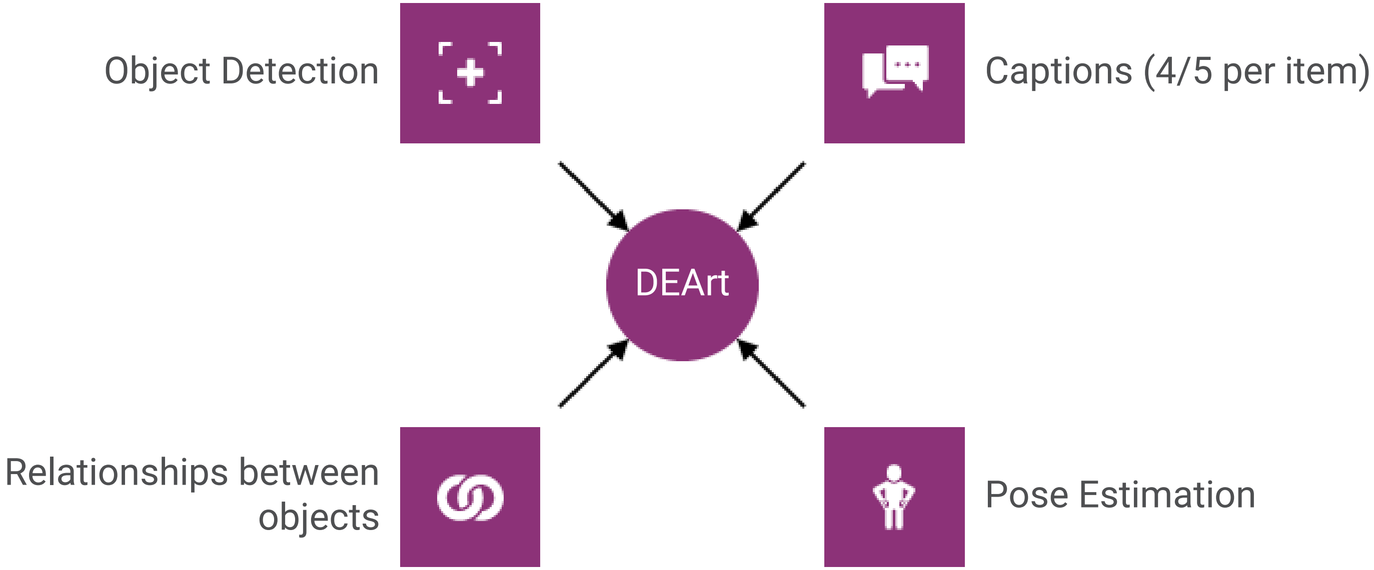
Evaluation of description generation

Compare automatically generated descriptions with human references

- n-gram based methods (BLEU): metrics strongly dependent on exact matching return weak results
  - Underlines the issue of the added difficulty of artworks compared to photographs: artwork objects and actions may be seen in different perspectives

- Semantic similarity score: scores the similarity between an automatic description (candidate) and the description from a set of human references whose semantic content is the closest. The score is computed with a transformer language model- returns better results than n-gram (approx 0.3)

DEArt (Dataset of European Art)



SAINT GEORGE AND THE DRAGON BY RAPHAEL. PUBLIC DOMAIN

Challenges we faced

- Data collection
- Poor metadata
- Evaluation method

Data collection issues

E.g. (for images):

- Some classes are represented only in a few images
- Style, medium, color may differ significantly between artists
- Not so many paintings anyway and can't produce them when needed

Approach to solve them:

- Small dataset by data mining standards requires complementary techniques, particularly to detect unusual / imaginary / symbolic objects
- Data augmentation was not very successful

Poor metadata (text data) issues

E.g. (for text):

- Not many descriptions of images (nor exhaustive object annotations)
- Image descriptions contain context and form information, much less content - assumption that one *sees* what is in the image
- No formal knowledge of which are “visual” relationships - e.g. an ontology

Approach to solve them:

- Caption classification - issue of what is NOT a visual description
- Approximation of visual relationships from COCO and IconClass
- Crowdsourcing

Evaluation method issues

Evaluation of automatically generated metadata vs. human references

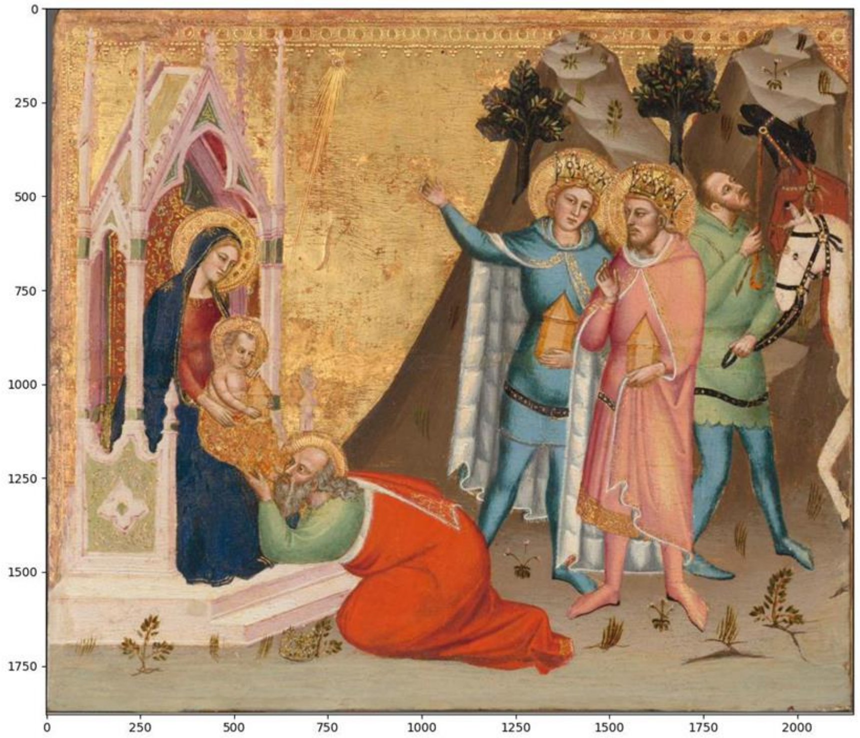
- Automatic evaluation with existing scoring methods is problematic for captions, especially given the diversity of cultural heritage descriptions (e.g. different symbolic levels, named entities, levels of detail in the description by annotators with different knowledge of iconography, art history, etc)
- Quantifying enrichments quality and usefulness to the user
- The question becomes: is pure deep learning (bottom up) enough to generate descriptive texts of paintings?

Can a description generation model ALONE work well for CH?

Good caption

woman in white dress holds baby

But here… halucination!



jesus christ on the cross soldiers are on the cross soldiers are on the cross

Adoration of the shephards, not crucifixion.
Jesus Christ on the cross. Soldiers are on the cross.

Very likely need a mix of bottom-up (deep learning) and top-down approaches to correctly model CH knowledge! E.g. Caption seeds, knowledge graphs + inference, NLP

Next steps

- Further iteration(s) of the description generator based on Zooniverse campaign
- Automated application of NL model for refinement of objects or relationships
- Inference over triple sets / Knowledge graph creation
- Improve the accuracy of the caption classifier
- Test an approach that generates “new artworks” from textual descriptions to increase the dataset size, especially for minority classes

Conclusions

Concrete outcomes:

- DEArt has 15K+ images, 80% non-canonical, annotated with all BBx instances of 69 classes (53 specific to cultural heritage).
- We gathered a dataset of visual descriptions for 7500 images, fully annotated with 4-5 descriptions.
- We achieved good accuracy of the object detection model. The description generation model not very accurate yet, following evaluation.

At a more abstract level, the project uncovered some *new challenges in CH* and generated *new research questions*.

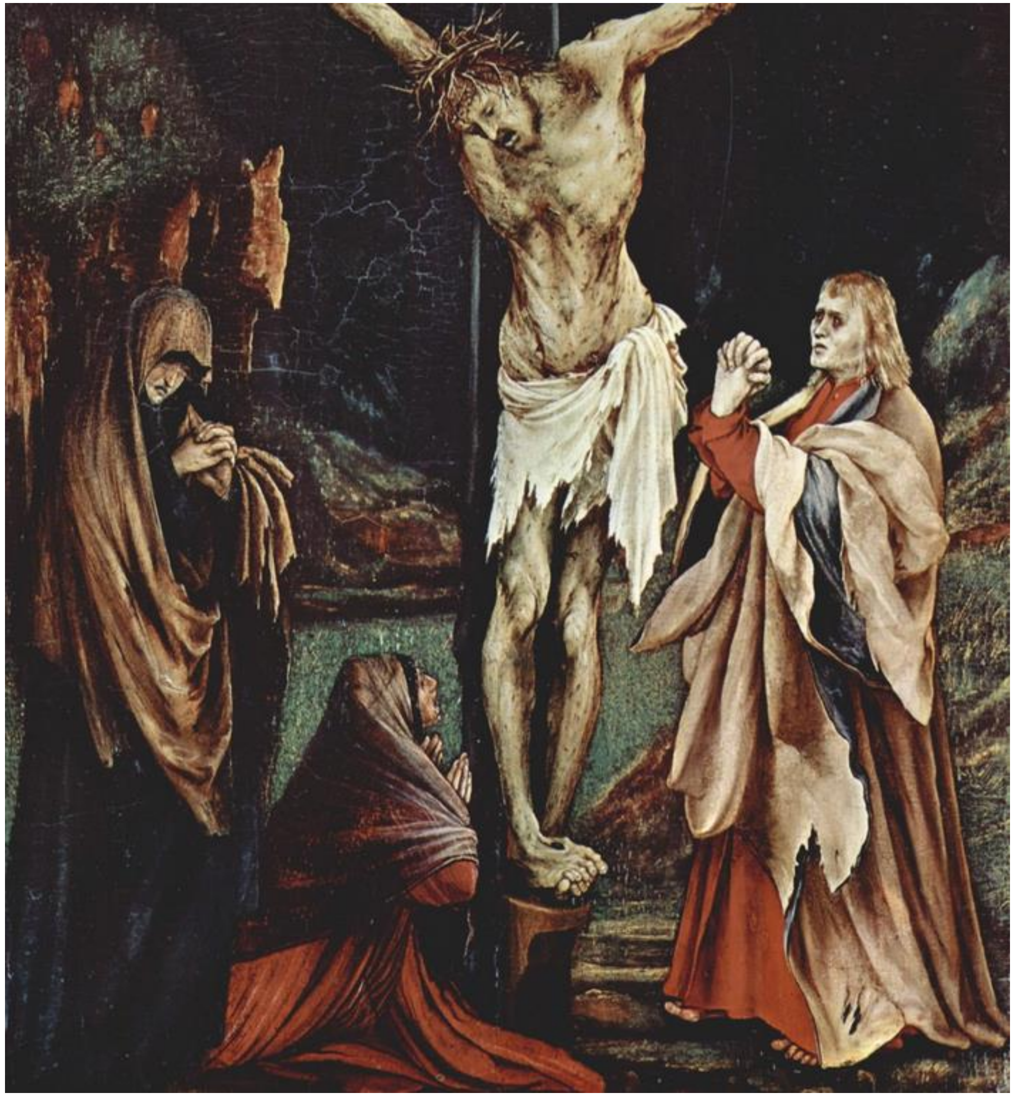
- E.g. Can a description generation model work well for CH, given the size of the datasets and the levels of manual descriptions? Complementary top down knowledge may prevent hallucinations in images and texts and spark the idea for a future project: hallucination prevention?
- E.g. Are triple-like descriptions redundant if we have descriptions?
- E.g. Art institutions assume that the descriptions are for people who ‘see’ the paintings. But, what about visually impaired people, or machines? SGoaB helps to increase the inclusion of citizenry in cultural heritage.

What about the recent large language models (LLM)?

Could we just generate image descriptions simply by asking an LLM to describe it?

Descriptions of typical/unsurprising scenes are very good now (as opposed to 1.5 / 1 year ago), but we still found hallucinations:

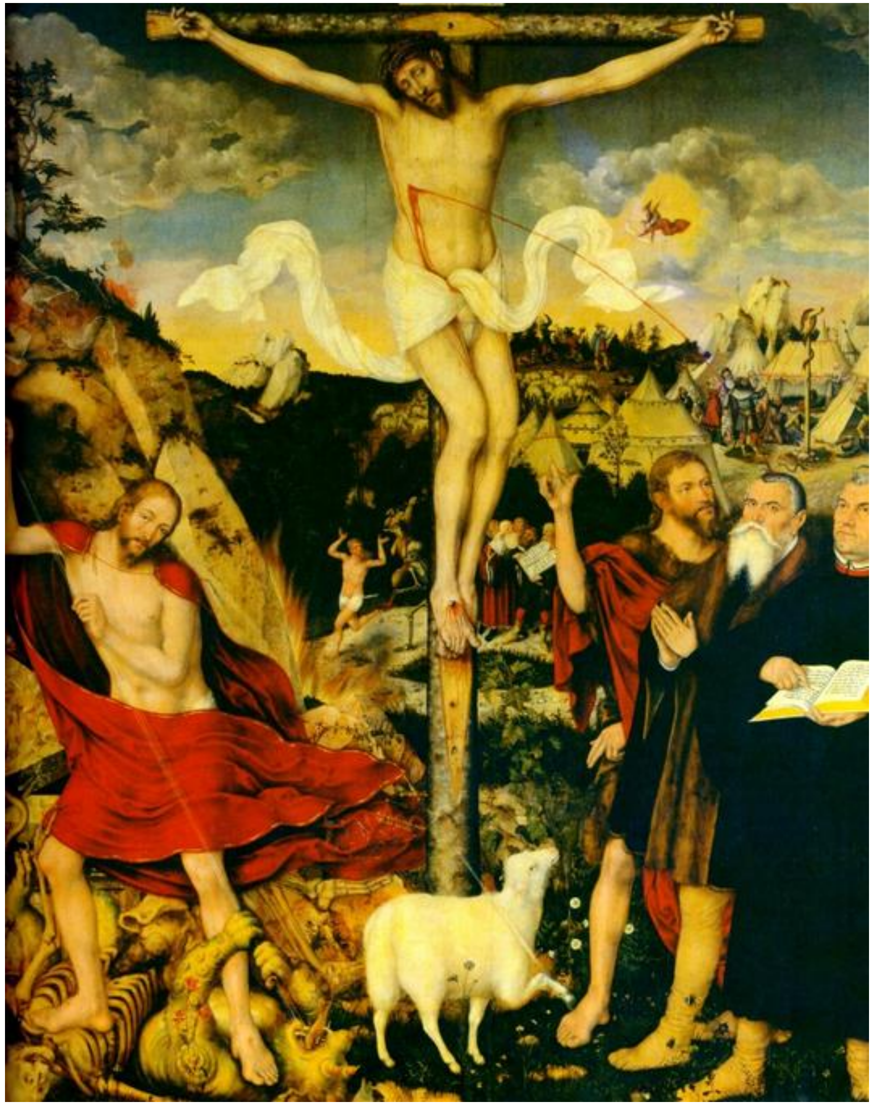
What about the recent large language models (LLM)?



"… The figures are shown with halos, indicating their sacred status."

Description of inexistent objects, hypothetically because they are usually present in the theme it recognizes.

What about the recent large language models (LLM)?



“To the left, a figure is kneeling with one breast exposed, which is Saint John the Evangelist, often shown in a youthful and compassionate manner. To the right stand two figures…”

Does not describe objects that ARE there, hypothetically because they are usually absent in the theme it recognizes.

What about the recent large language models (LLM)?

Could we just generate image descriptions simply by asking an LLM to describe it?

Descriptions of typical/unsurprising scenes are very good now (as opposed to 1.5 / 1 year ago), but we still found hallucinations.

Would passing the LLM a prompt based on our objects / triples improve the description? ... currently under investigation.

Example: FrAI Angelico (Quim Moré)

Proof of concept with paitings from El Prado

- Object detector
- Position-based object relationships (Visrel)
- (subject, predicate, object) tuples extracted from **Iconclass** annotations and painting descriptions from El Prado collections
- Labels from the XMLs of the metadata associated with the Prado images

Reconocimiento de entidades específicas

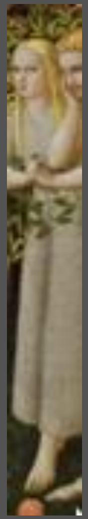
Dado un objeto etiquetado como una entidad general (e.g: persona), reetiquetarlo como una entidad más específica

Reconocimiento de entidades específicas

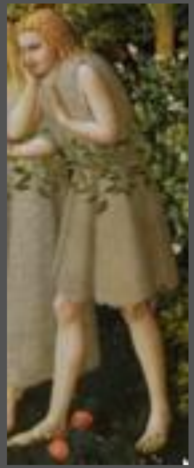
Detección de entidades generales

persona_1

persona_2



persona_3



Reconocimiento de entidades específicas

Detección de identificadores de atributo

halo_1

halo_2

Reconocimiento de entidades específicas

Detección de identificadores de composición y grupos de composición

Reconocimiento de entidades específicas

Para cada entidad general:
a) comprobar si una entidad general mantiene una relación con un identificador de atributo

Reconocimiento de entidades específicas

a.1 Tomar las tuplas que tienen
pred = relación de atributo
dobj = identificador de atributo
Tuplas separadas por temas

```
Attribute indentifiers for: ('person_1', 'is_with')
halo_1
```

	subj	pred	dobj	pobj	topic (with Iconclass code)
279	Christ	is_with	halo		Christ(11D)

	subj	pred	dobj	pobj	topic (with Iconclass code)
818	the_virgin_Mary	is_with	halo		The_Virgin_Mary(11F)

Reconocimiento de entidades específicas

a.2 Entrenar al predictor de entidades con las tuplas obtenidas en cada tema
a.3 Predecir la entidad que ocupa la posición de sujeto en el tema *t*

< ?, is_with,halo, Christ(11D)>

< ?, is_with,halo, The_Virgin_Mary(11F)>

```
ENTITIES PREDICTED FOR person_1

Christ IN TOPIC Christ(11D) WITH PROBABILITY 1.0
the_virgin_Mary IN TOPIC The_Virgin_Mary(11F) WITH PROBABILITY 1.0
```

The_Virgin_Mary

Christ

Descubrimiento de entidades no reconocidas

El reconocedor de objetos no siempre reconoce todos los objetos relevantes en una pintura.

Sin embargo, gracias a los temas relacionados con los objetos detectados, sí que podemos preguntar al usuario si puede ver objetos relacionados con este tema y, si es así, incorporarlos en la lista de objetos representados

Descubrimiento de entidades no reconocidas

Temas reconocidos

After_the_Fall (71A5)

Angel

Adam_and_Eve

The_Annunciation (73A5)

Angel

The_Virgin_Mary

Descubrimiento de entidades no reconocidas

Encontrar en las tuplas de entrenamiento una relación entre dos objetos identificados con un objeto no identificado (en un tema reconocido)

	subj	pred	dobj	pobj	topic (with Iconclass code)
49	angel	chase	Adam and Eve	with_sword	After_the_Fall(71A5)

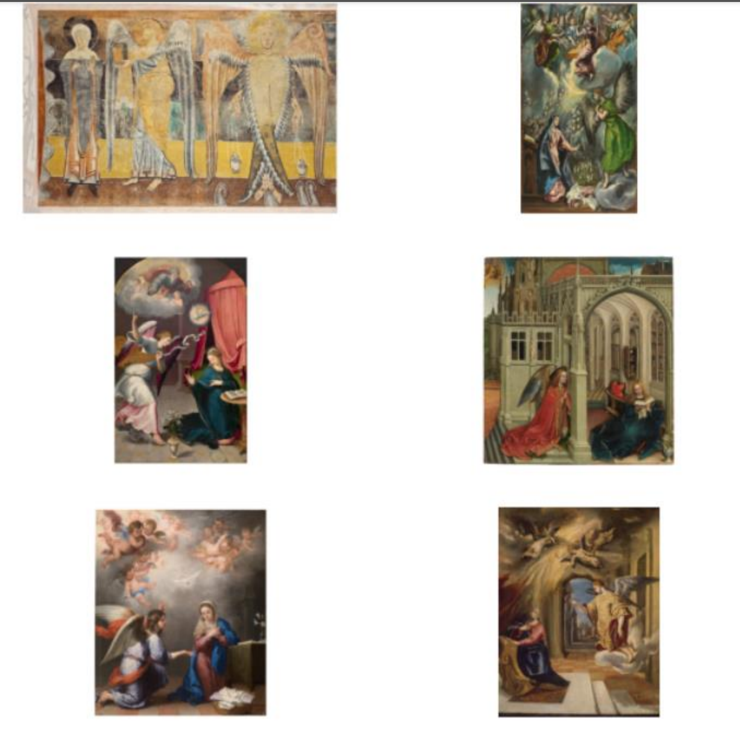
Mira la pintura detenidamente y señala las entidades que puedes ver

- [ ] Espada

Descubrimiento de entidades no reconocidas

Importar los tags de las pinturas de la colección del Prado etiquetados con un tema reconocido. Los tags están recogidos en los archivos .rdf de cada pintura

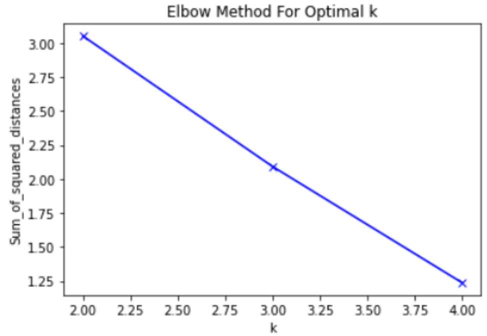
	File	Tags
0	P000827.rdf	La_Anunciación Paloma_blanca_simbólica Ángel V...
1	P003888.rdf	Paloma_blanca_simbólica Zarzamora Ángel Virgen...
2	P002828.rdf	La_Anunciación Paloma_blanca_simbólica Azucena...
3	P000015-001.rdf	La_Anunciación Golondrina Paloma_blanca_simból...
4	P001915.rdf	Libro La_Anunciación Azucena Virgen_María San_...
5	P000970.rdf	La_Anunciación Paloma_blanca_simbólica Azucena...
6	P007282.rdf	La_Anunciación Ángel_San_Gabriel



Descubrimiento de entidades no reconocidas

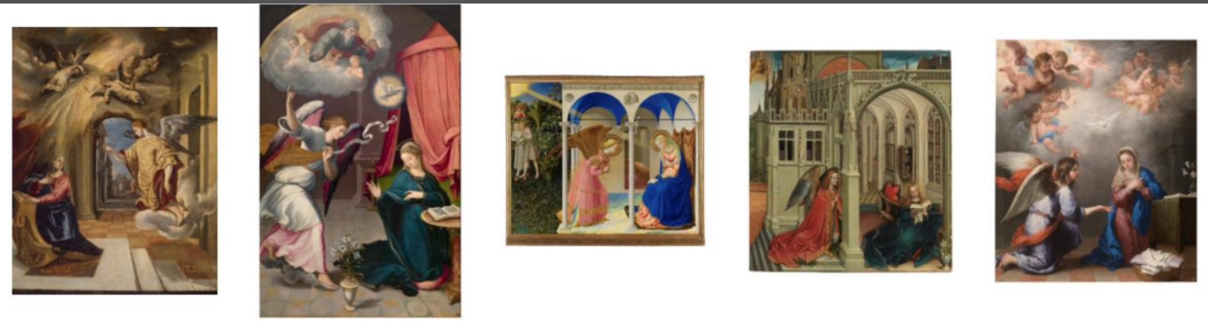
Las pinturas se vectorizan según sus tags. Luego se agrupan en clusters. Las pinturas con tags más representativos de un cluster se toman como referencias a entidades que se pueden encontrar en la pintura de un tema.

Cálculo del número de clusters óptimo




```
Put the number of clusters 3
[0 1 0 0 0 0 2]
['P000827.rdf', 'P003888.rdf', 'P002828.rdf', 'P000015-001.rdf', 'P001915.rdf', 'P000970.rdf', 'P007282.rdf']
0 is the largest cluster
```

Pinturas parecidas según sus tags en el cluster 0

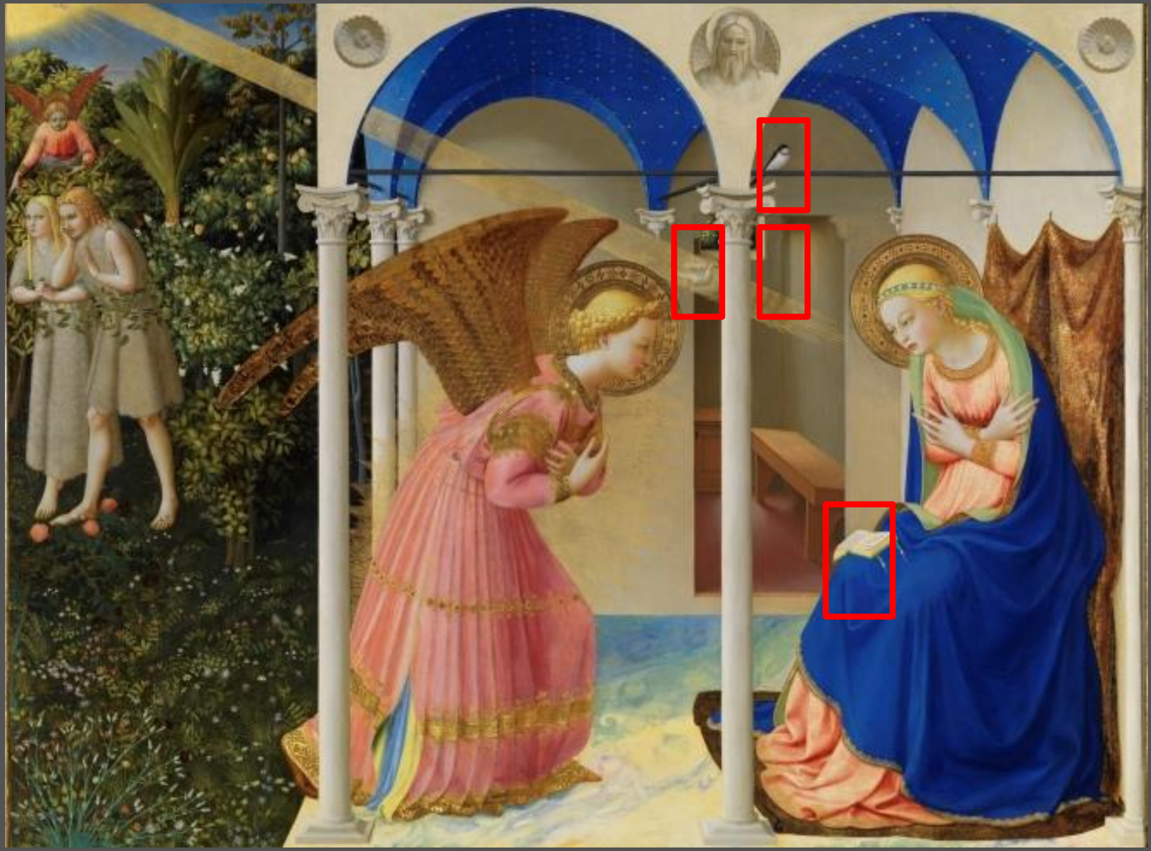


Tags más representativos del cluster 0

	Tag	Score
0	virgen_maría	0.305184
1	la_anunciación	0.305184
2	san_gabriel	0.302439
3	paloma_blanca_simbólica	0.282619
4	azucena	0.237041
5	libro	0.237041
6	ángel	0.179572
7	florero	0.178763
8	golondrina	0.151512
9	vidriera_artística	0.118728
10	cesto_de_costura	0.096211

Descubrimiento de entidades no reconocidas

Se pide al visitante de la web que mire detalladamente el cuadro y marque las entidades que ve y que no han sido identificadas



Look at the painting more closely and tick the entities you can see

- [x] paloma_blanca_simbólica
- [ ] azucena
- [x] libro
- [x] golondrina
- [x] vidriera_artística

Questions?