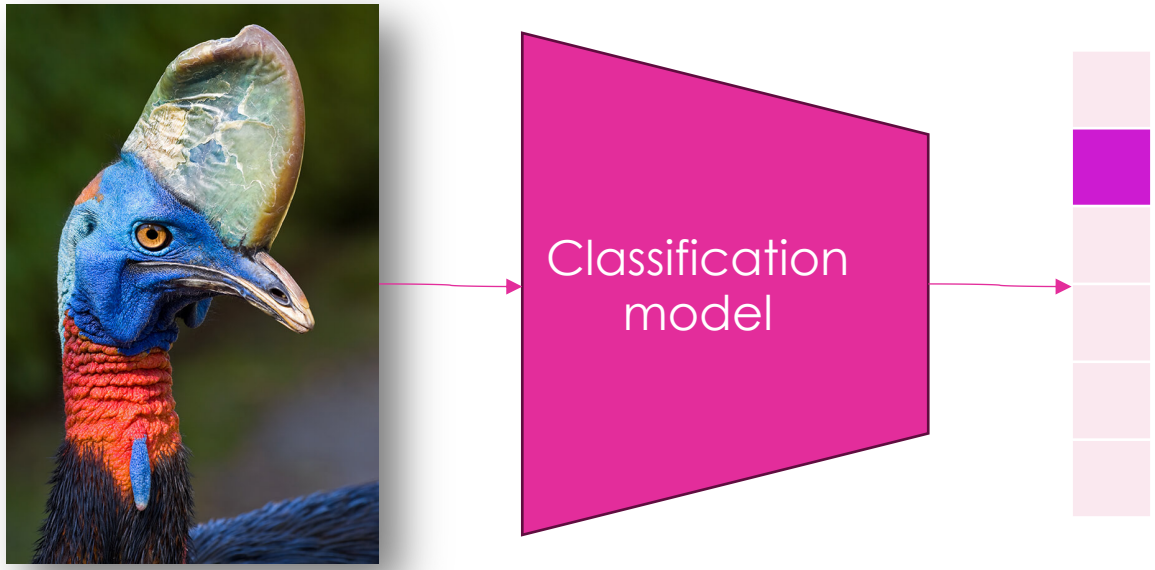
Democratizing Language and Vision Technology

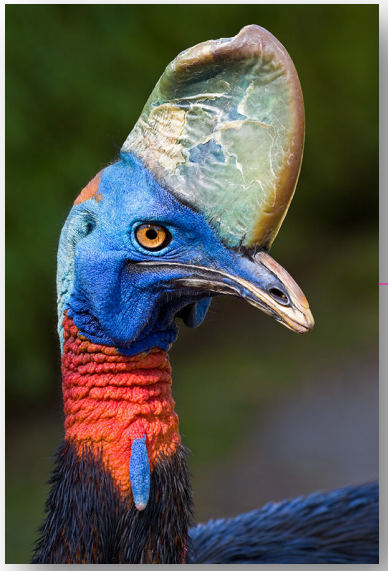
PAOLO ROTA

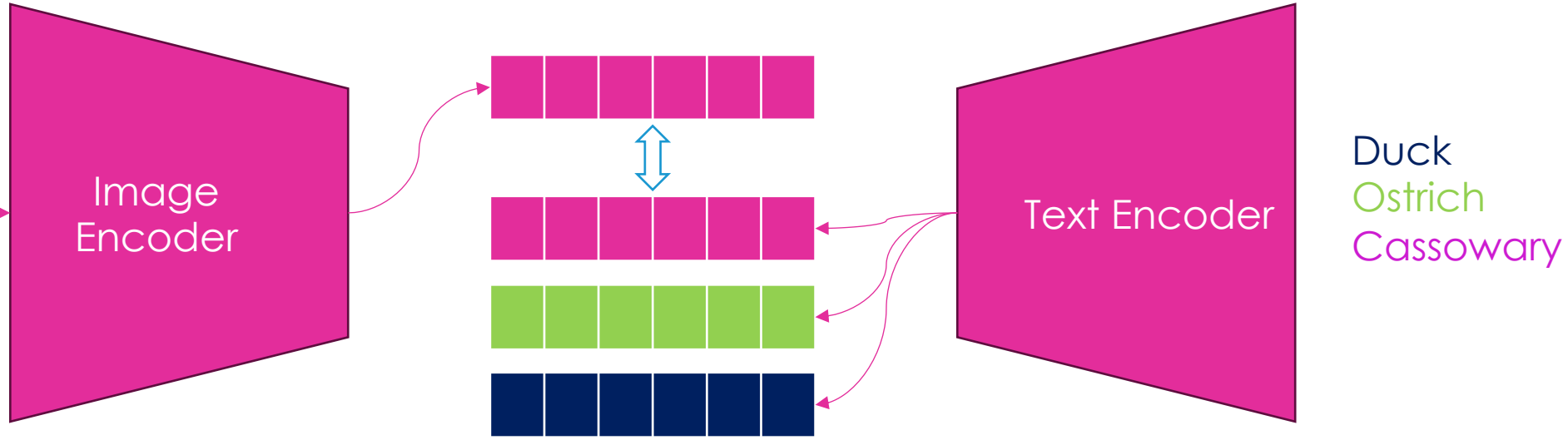
Traditional Image Classification



- We need to know all classes at Training time

Zero-shot image classification

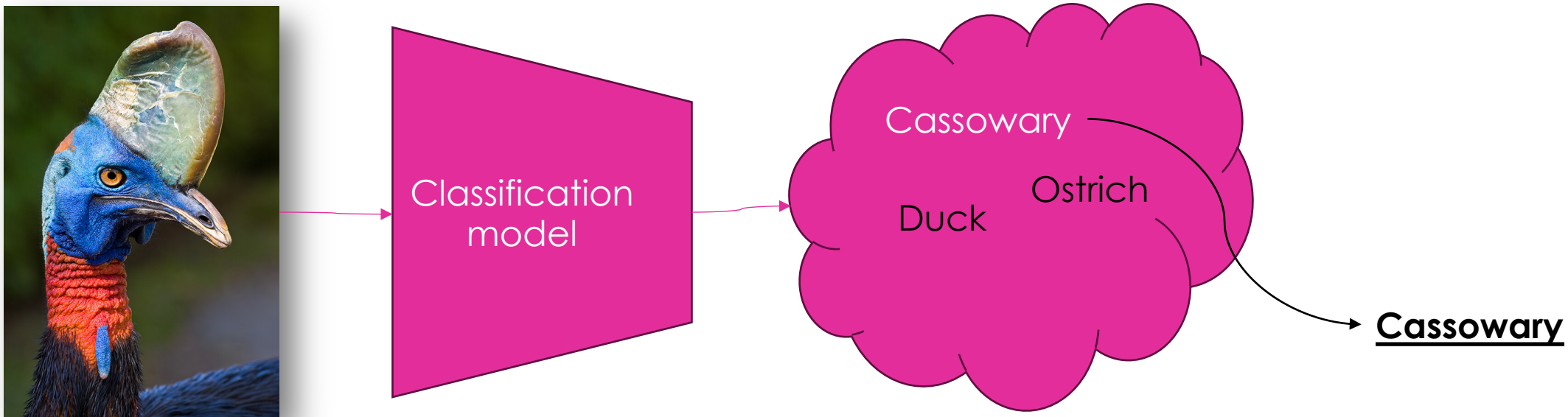




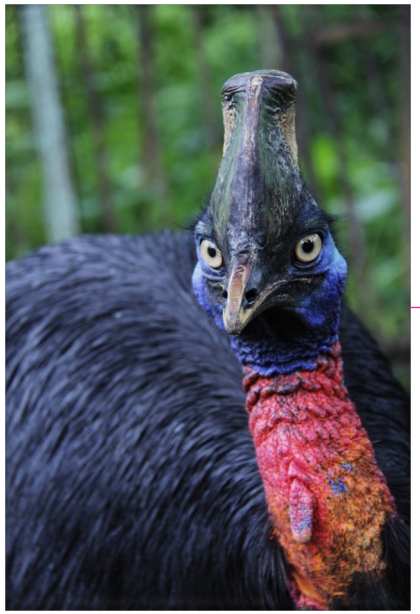
- We need to know all classes at <u>test time</u>

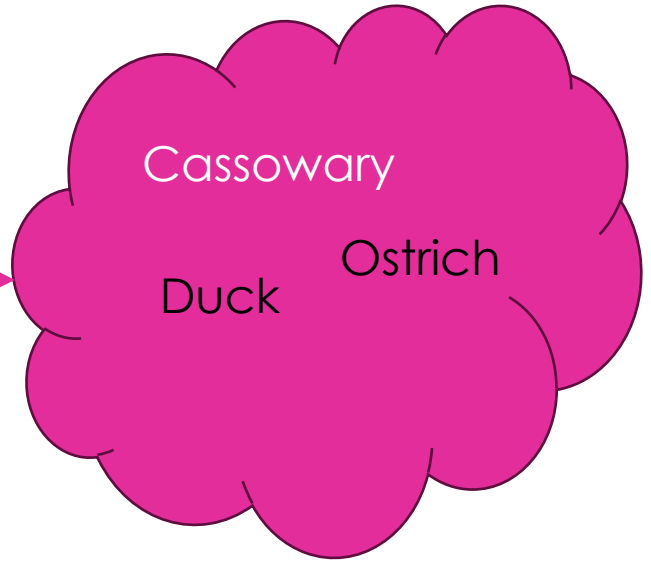
Vocabulary-free Image Classification (VIC)

Aims to assign to an input image a class that resides in an unconstrained language-induced semantic space **without** the prerequisite of a known vocabulary



Vocabulary-free Image Classification (VIC)





- Huge search space.
- Fine-grained concepts.
- Categories at test time are not defined beforehand.

So much power comes with a cost…

So much power comes with a cost…

Florence

Microsoft

match the dimensions of image and language features. Our *Florence* pretrained model has in total $893M$ parameters, including the language transformer with $256M$ parameters and the *CoSwin*-H transformer with $637M$ parameters. The model takes 10 days to train on 512 NVIDIA-A100 GPUs with 40GB memory per GPU.

So much power comes with a cost…

Florence Microsoft

match the dimensions of image and language features. Our *Florence* pretrained model has in total $893M$ parameters, including the language transformer with $256M$ parameters and the *CoSwin*-H transformer with $637M$ parameters. The model takes 10 days to train on 512 NVIDIA-A100 GPUs with 40GB memory per GPU.

DALL-E OpenAI

term is a joint distribution over the 32×32 positions in the spatial grid output by the encoder. We divide the overall loss by $256 \times 256 \times 3$, so that the weight of the KL term becomes $\beta/192$, where β is the KL weight. The model is trained in mixed-precision using standard (i.e., global) loss scaling on 64 16 GB NVIDIA V100 GPUs, with a per-GPU batch size of 8, resulting in a total batch size of 512. It is trained for a total of 3,000,000 updates.

So much power comes with a cost…

CLIP OpenAI

train a ViT-B/32, a ViT-B/16, and a ViT-L/14. The largest ResNet model, RN50x64, took 18 days to train on 592 V100 GPUs while the largest Vision Transformer took 12 days on 256 V100 GPUs. For the ViT-L/14 we also pre-train at a higher 336 pixel resolution for one additional epoch to boost

Florence Microsoft

match the dimensions of image and language features. Our *Florence* pretrained model has in total $893M$ parameters, including the language transformer with $256M$ parameters and the *CoSwin*-H transformer with $637M$ parameters. The model takes 10 days to train on 512 NVIDIA-A100 GPUs with 40GB memory per GPU.

DALL-E OpenAI

term is a joint distribution over the 32×32 positions in the spatial grid output by the encoder. We divide the overall loss by $256 \times 256 \times 3$, so that the weight of the KL term becomes $\beta/192$, where β is the KL weight. The model is trained in mixed-precision using standard (i.e., global) loss scaling on 64 16 GB NVIDIA V100 GPUs, with a per-GPU batch size of 8, resulting in a total batch size of 512. It is trained for a total of 3,000,000 updates.

So much power comes with a cost…

CLIP OpenAI

train a ViT-B/32, a ViT-B/16, and a ViT-L/14. The largest ResNet model, RN50x64, took 18 days to train on 592 V100 GPUs while the largest Vision Transformer took 12 days on 256 V100 GPUs. For the ViT-L/14 we also pre-train at a higher 336 pixel resolution for one additional epoch to boost

Florence Microsoft

match the dimensions of image and language features. Our *Florence* pretrained model has in total $893M$ parameters, including the language transformer with $256M$ parameters and the *CoSwin*-H transformer with $637M$ parameters. The model takes 10 days to train on 512 NVIDIA-A100 GPUs with 40GB memory per GPU.

CoCa Google Research

warm up the learning rate for the first 2% of training steps to a peak value of 8×10^{-4}, and linearly decay it afterwards. Pretraining CoCa takes about 5 days on 2,048 CloudTPUv4 chips. Following [12, 13, 14], we continue pretraining for one epoch on a higher resolution of 576×576. For finetuning evaluation, we mainly follow simple protocols and directly train CoCa on downstream tasks without further metric-specific tuning like CIDEr scores (details in Appendix A and B).

DALL-E OpenAI

term is a joint distribution over the 32×32 positions in the spatial grid output by the encoder. We divide the overall loss by $256 \times 256 \times 3$, so that the weight of the KL term becomes $\beta/192$, where β is the KL weight. The model is trained in mixed-precision using standard (i.e., global) loss scaling on 64 16 GB NVIDIA V100 GPUs, with a per-GPU batch size of 8, resulting in a total batch size of 512. It is trained for a total of 3,000,000 updates.

So much power comes with a cost…

CLIP OpenAI

train a ViT-B/32, a ViT-B/16, and a ViT-L/14. The largest ResNet model, RN50x64, took 18 days to train on 592 V100 GPUs while the largest Vision Transformer took 12 days on 256 V100 GPUs. For the ViT-L/14 we also pre-train at a higher 336 pixel resolution for one additional epoch to boost

Florence Microsoft

match the dimensions of image and language features. Our *Florence* pretrained model has in total $893M$ parameters, including the language transformer with $256M$ parameters and the *CoSwin*-H transformer with $637M$ parameters. The model takes 10 days to train on 512 NVIDIA-A100 GPUs with 40GB memory per GPU.

Flamingo Google DeepMind

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] Details can be found in Appendix B.1.2. In short, our largest run was trained on 1536 TPU chips for 15 days.

CoCa Google Research

warm up the learning rate for the first 2% of training steps to a peak value of 8×10^{-4}, and linearly decay it afterwards. Pretraining CoCa takes about 5 days on 2,048 CloudTPUv4 chips. Following [12, 13, 14], we continue pretraining for one epoch on a higher resolution of 576×576. For finetuning evaluation, we mainly follow simple protocols and directly train CoCa on downstream tasks without further metric-specific tuning like CIDEr scores (details in Appendix A and B).

DALL-E OpenAI

term is a joint distribution over the 32×32 positions in the spatial grid output by the encoder. We divide the overall loss by $256 \times 256 \times 3$, so that the weight of the KL term becomes $\beta/192$, where β is the KL weight. The model is trained in mixed-precision using standard (i.e., global) loss scaling on 64 16 GB NVIDIA V100 GPUs, with a per-GPU batch size of 8, resulting in a total batch size of 512. It is trained for a total of 3,000,000 updates.

So much power comes with a cost…

CLIP OpenAI

train a ViT-B/32, a ViT-B/16, and a ViT-L/14. The largest ResNet model, RN50x64, took 18 days to train on 592 V100 GPUs while the largest Vision Transformer took 12 days on 256 V100 GPUs. For the ViT-L/14 we also pre-train at a higher 336 pixel resolution for one additional epoch to boost

Florence Microsoft

match the dimensions of image and language features. Our *Florence* pretrained model has in total $893M$ parameters, including the language transformer with $256M$ parameters and the *CoSwin*-H transformer with $637M$ parameters. The model takes 10 days to train on 512 NVIDIA-A100 GPUs with 40GB memory per GPU.

SAM Meta

initialize SAM from an MAE [47] pre-trained ViT-H. We distribute training across 256 GPUs, due to the large image encoder and 1024×1024 input size. To limit GPU mem-

Flamingo Google DeepMind

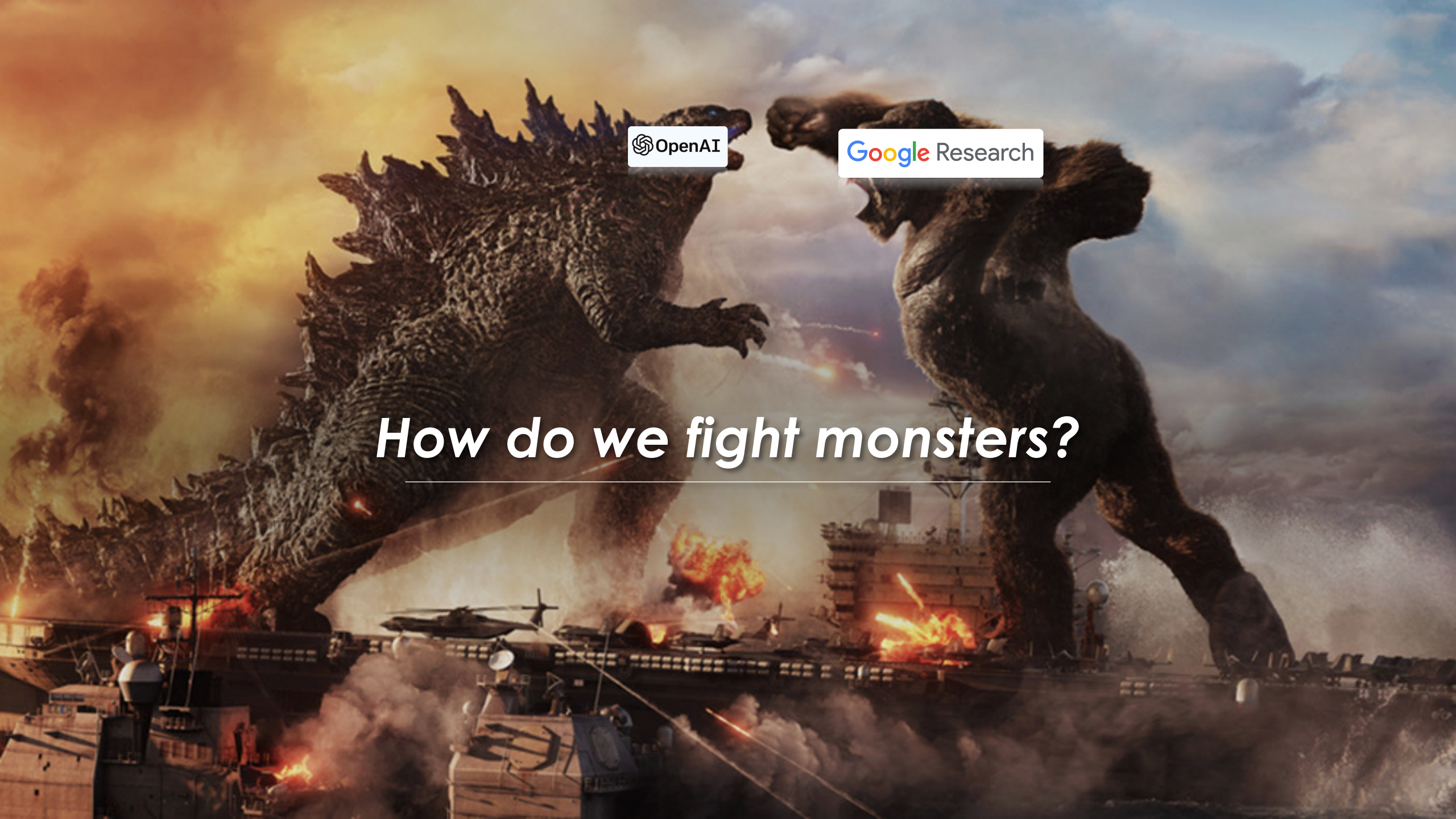
(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] Details can be found in Appendix B.1.2. In short, our largest run was trained on 1536 TPU chips for 15 days.

CoCa Google Research

warm up the learning rate for the first 2% of training steps to a peak value of 8×10^{-4}, and linearly decay it afterwards. Pretraining CoCa takes about 5 days on 2,048 CloudTPUv4 chips. Following [12, 13, 14], we continue pretraining for one epoch on a higher resolution of 576×576. For finetuning evaluation, we mainly follow simple protocols and directly train CoCa on downstream tasks without further metric-specific tuning like CIDEr scores (details in Appendix A and B).

DALL-E OpenAI

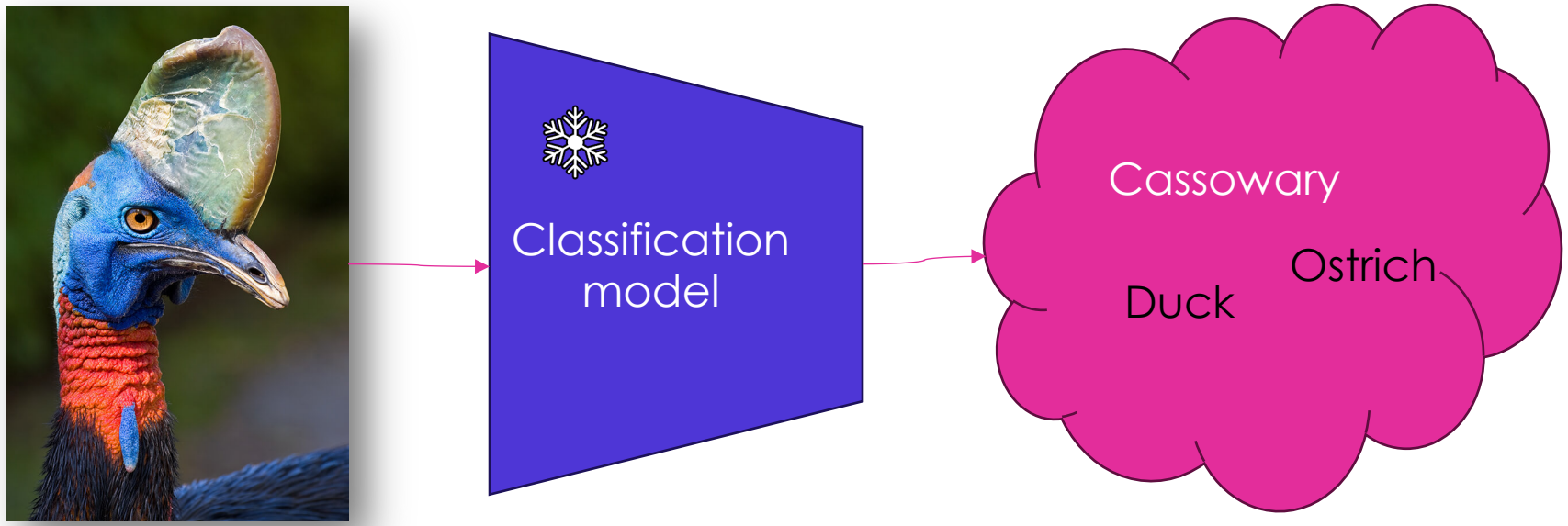
term is a joint distribution over the 32 × 32 positions in the spatial grid output by the encoder. We divide the overall loss by 256 × 256 × 3, so that the weight of the KL term becomes $\beta/192$, where β is the KL weight. The model is trained in mixed-precision using standard (i.e., global) loss scaling on 64 16 GB NVIDIA V100 GPUs, with a per-GPU batch size of 8, resulting in a total batch size of 512. It is trained for a total of 3,000,000 updates.


OpenAI
Google Research
How do we fight monsters?


OpenAI
Google Research
How do we fight monsters?
You are here
UNIVERSITAS ATHESINA STUDIORUM

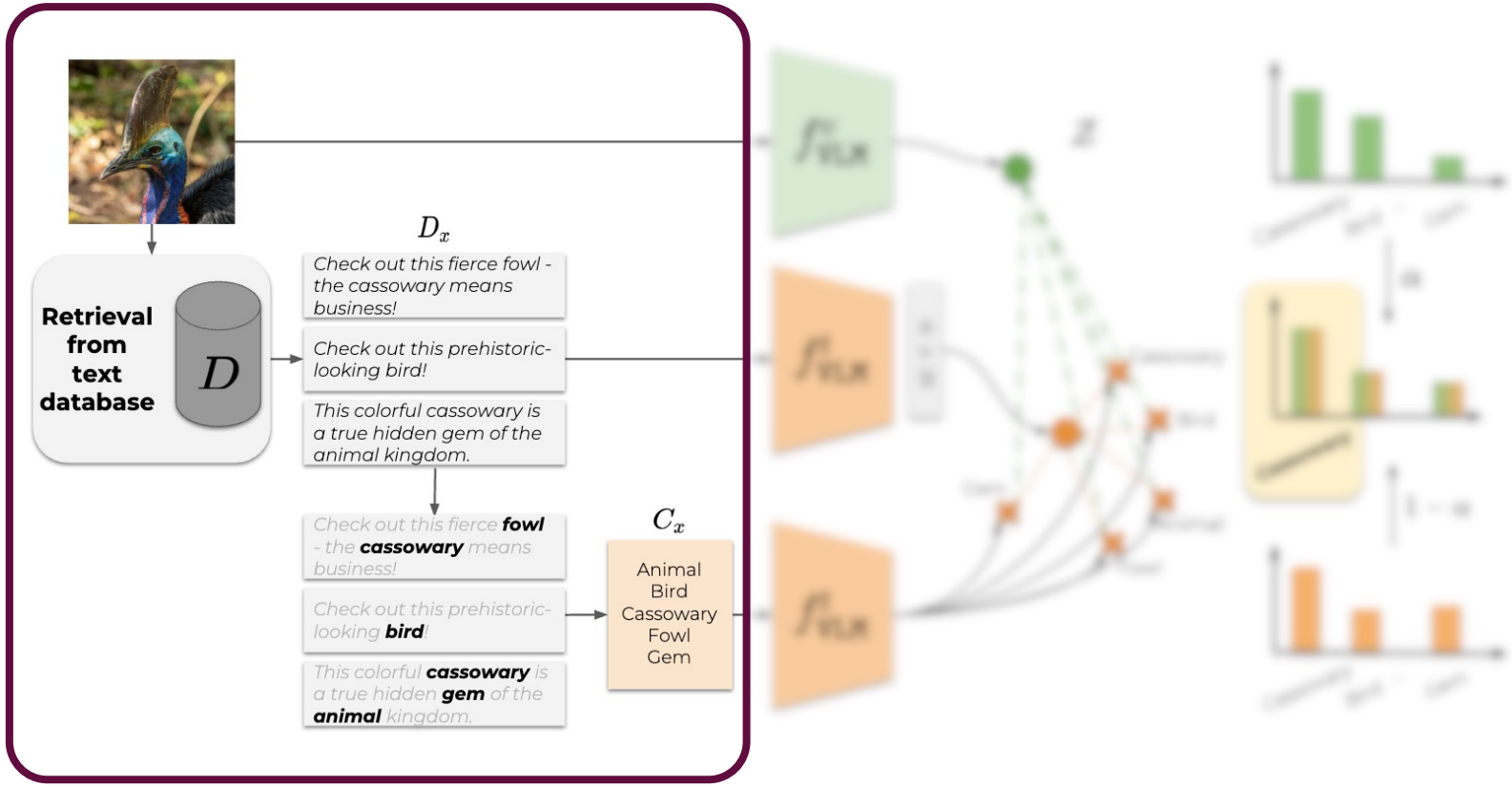

...but we had an Idea!!

Why don't we train it at all?



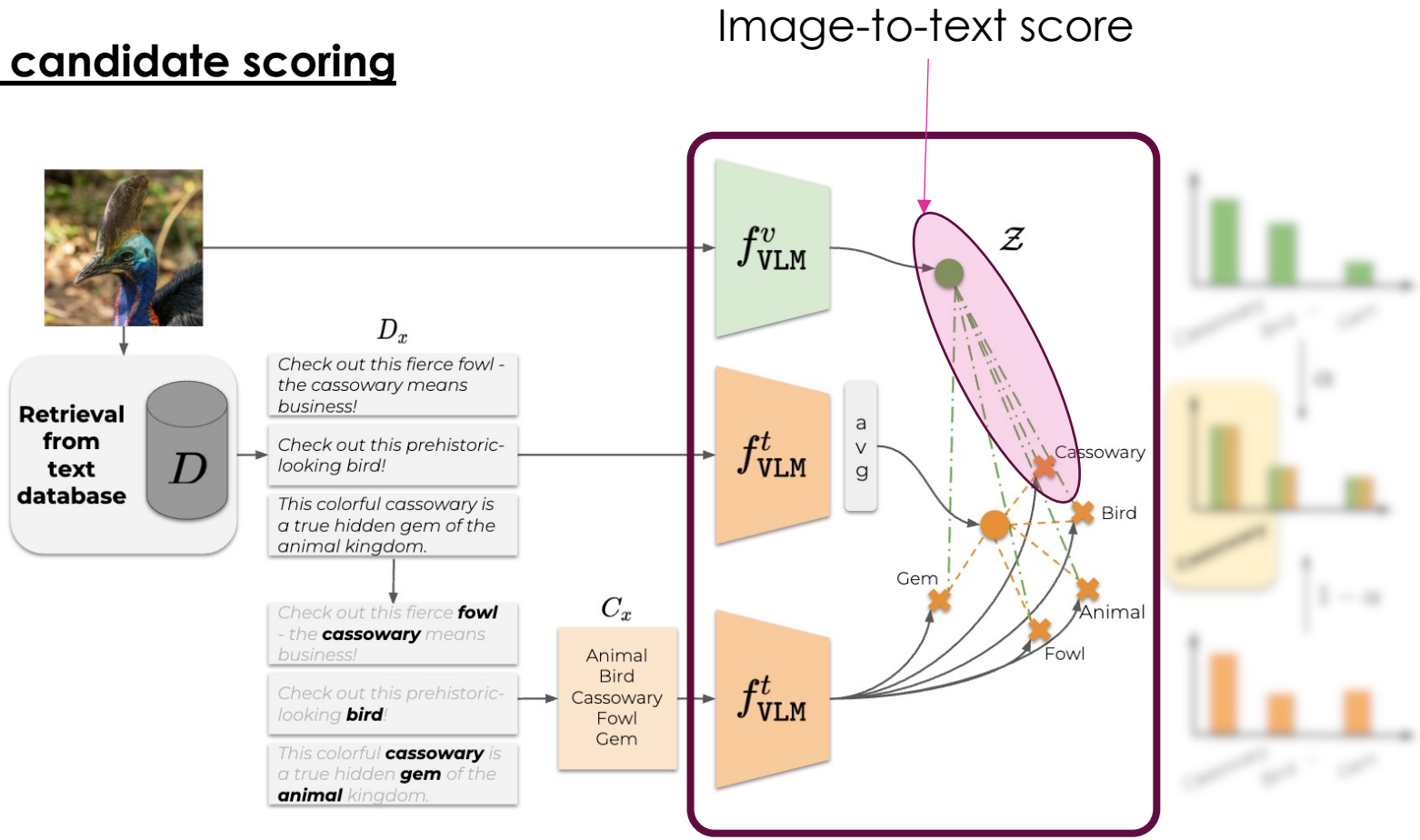
Category Search from External Databases (CaSED)

Generating candidate categories



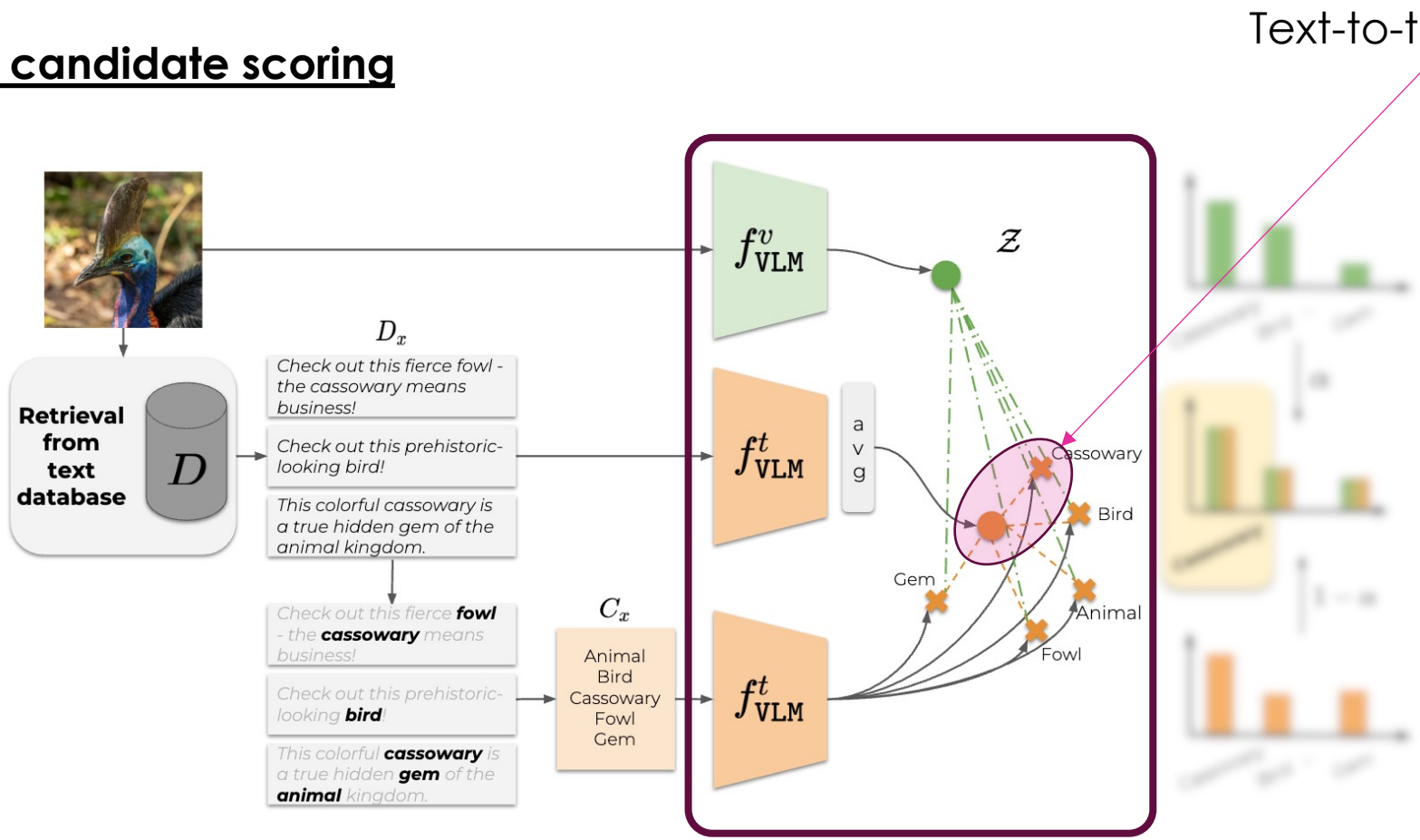
Category Search from External Databases (CaSED)

Multimodal candidate scoring



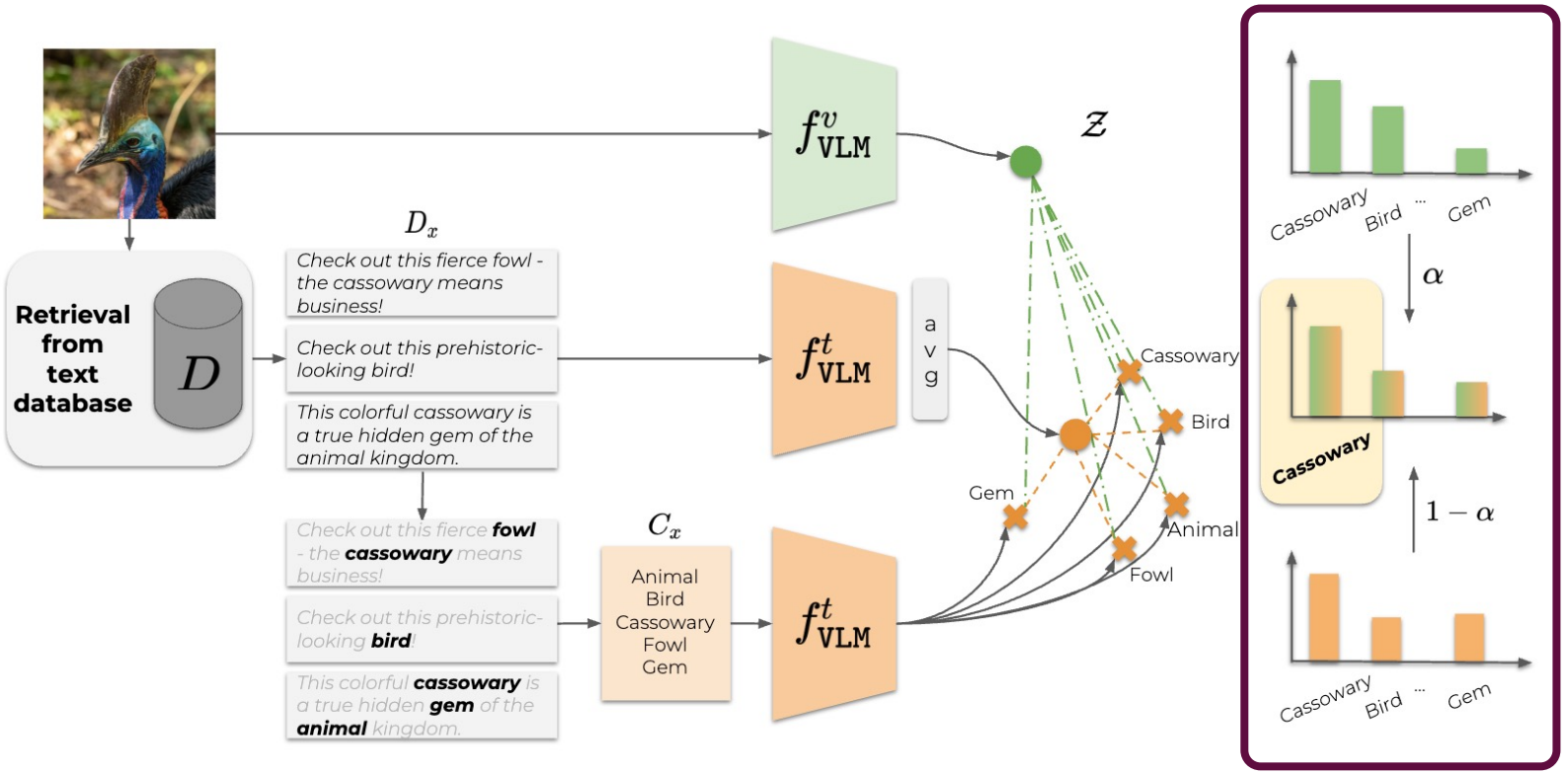
Category Search from External Databases (CaSED)

Multimodal candidate scoring



Category Search from External Databases (CaSED)

Final predicted candidate



Results

Method		Cluster Accuracy (%) ↑										
		C101	DTD	ESAT	Airc.	Flwr	Food	Pets	SUN	Cars	UCF	**Avg.**
CLIP	WordNet	34.0	20.1	16.7	16.7	58.3	40.9	52.0	29.4	18.6	39.5	32.6
	English Words	29.1	19.6	22.1	15.9	64.0	30.9	44.4	24.2	19.3	34.5	30.4
Caption	Closest Caption	12.8	8.9	16.7	13.3	28.5	13.1	15.0	8.6	20.0	17.8	15.5
	BLIP-2 (ViT-L)	26.5	11.7	23.3	5.4	23.6	12.4	11.6	19.5	14.8	25.7	17.4
	BLIP-2 (ViT-g)	37.4	13.0	**25.2**	10.0	29.5	19.9	15.5	21.5	27.9	32.7	23.3
VQA	BLIP-2 (ViT-L)	60.4	20.4	21.4	8.1	36.7	21.3	14.0	32.6	28.8	44.3	28.8
	BLIP-2 (ViT-g)	**62.2**	23.8	22.0	15.9	57.8	33.4	23.4	36.4	**57.2**	**55.4**	38.7
CaSED		51.5	**29.1**	23.8	**22.8**	**68.7**	**58.8**	**60.4**	**37.4**	31.3	47.7	**43.1**
CLIP upper bound		87.6	52.9	47.4	31.8	78.0	89.9	88.0	65.3	76.5	72.5	69.0

Conclusion

In Summary:

- Vocabulary-free image classification
- Training-free approach
- Open-source and Open-data approach.

Limitations:

- Classes present in the database (scarcity and bias)
- No track of the output history
- Slightly different labels (e.g., cassowary, Casuarius)
- Granularity problem

Thanks

Alessandro Conti

Enrico Fini

Massimiliano Mancini

Paolo Rota

Yiming Wang

Elisa Ricci

and…

Scan Me!

https://alessandroconti.me/papers/2306.00917

Acknowledgements

This work was supported by the MUR PNRR project **FAIR** - Future AI Research (PE00000013) funded by the **NextGenerationEU**. E.R. is partially supported by the **PRECRISIS**, funded by the EU Internal Security Fund (ISFP-2022-TFI-AG-PROTECT-02-101100539), the EU project **SPRING** (No. 871245), and the by the PRIN project **LEGO-AI** (Prot. 2020TA3K9N). The work was carried out in the Vision and Learning joint laboratory of FBK and UNITN. We also thank the Deep Learning Lab of the **ProM Facility** for the GPU time.

Scan Me!

https://alessandroconti.me/papers/2306.00917