Compatibility analysis of fine-tuned models for fine-grained text-based zero-shot image retrieval

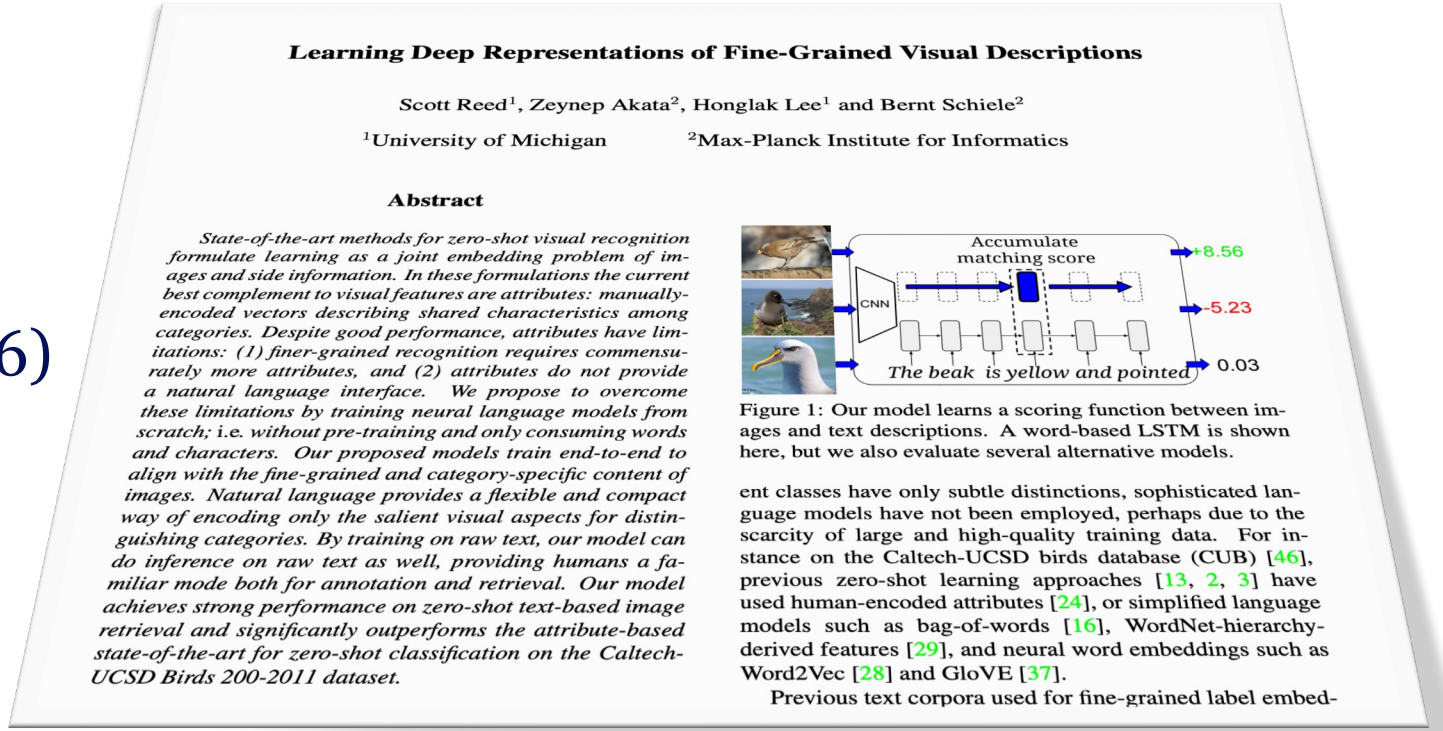
András Schmelczer | MIR - 2022

Universiteit Leiden
The Netherlands

Motivation

- Text-based image retrieval is useful
- If it's generalisable and highly domain-specific, even more so

- Excellent work by Reed et al. (2016)
- Let's revisit it using current SOTA
- Combine & fine-tune networks



Learning Deep Representations of Fine-Grained Visual Descriptions

Scott Reed[1], Zeynep Akata[2], Honglak Lee[1] and Bernt Schiele[2]

[1]University of Michigan [2]Max-Planck Institute for Informatics

Abstract

State-of-the-art methods for zero-shot visual recognition formulate learning as a joint embedding problem of images and side information. In these formulations the current best complement to visual features are attributes: manually-encoded vectors describing shared characteristics among categories. Despite good performance, attributes have limitations: (1) finer-grained recognition requires commensurately more attributes, and (2) attributes do not provide a natural language interface. We propose to overcome these limitations by training neural language models from scratch; i.e. without pre-training and only consuming words and characters. Our proposed models train end-to-end to align with the fine-grained and category-specific content of images. Natural language provides a flexible and compact way of encoding only the salient visual aspects for distinguishing categories. By training on raw text, our model can do inference on raw text as well, providing humans a familiar mode both for annotation and retrieval. Our model achieves strong performance on zero-shot text-based image retrieval and significantly outperforms the attribute-based state-of-the-art for zero-shot classification on the Caltech-UCSD Birds 200-2011 dataset.

Figure 1: Our model learns a scoring function between images and text descriptions. A word-based LSTM is shown here, but we also evaluate several alternative models.

ent classes have only subtle distinctions, sophisticated language models have not been employed, perhaps due to the scarcity of large and high-quality training data. For instance on the Caltech-UCSD birds database (CUB) [46], previous zero-shot learning approaches [13, 2, 3] have used human-encoded attributes [24], or simplified language models such as bag-of-words [16], WordNet-hierarchy-derived features [29], and neural word embeddings such as Word2Vec [28] and GloVE [37].

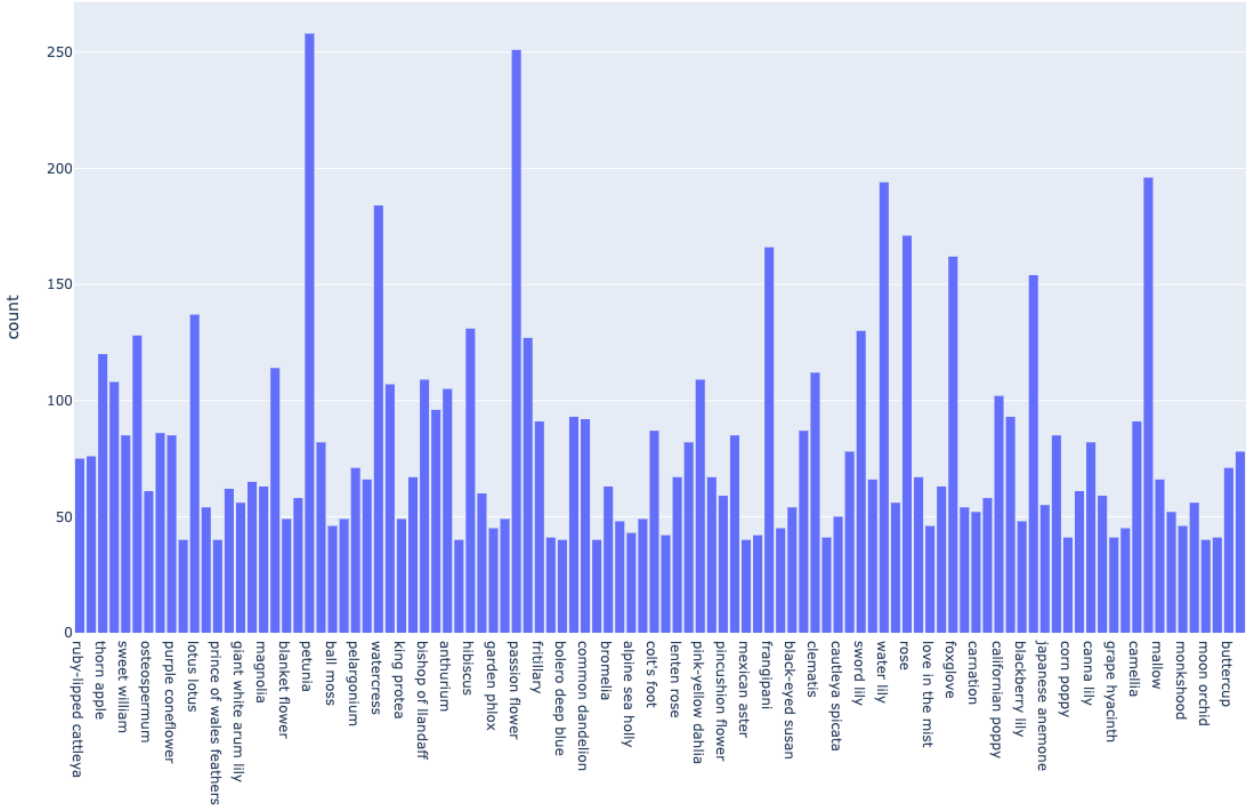
Previous text corpora used for fine-grained label embed-

Datasets

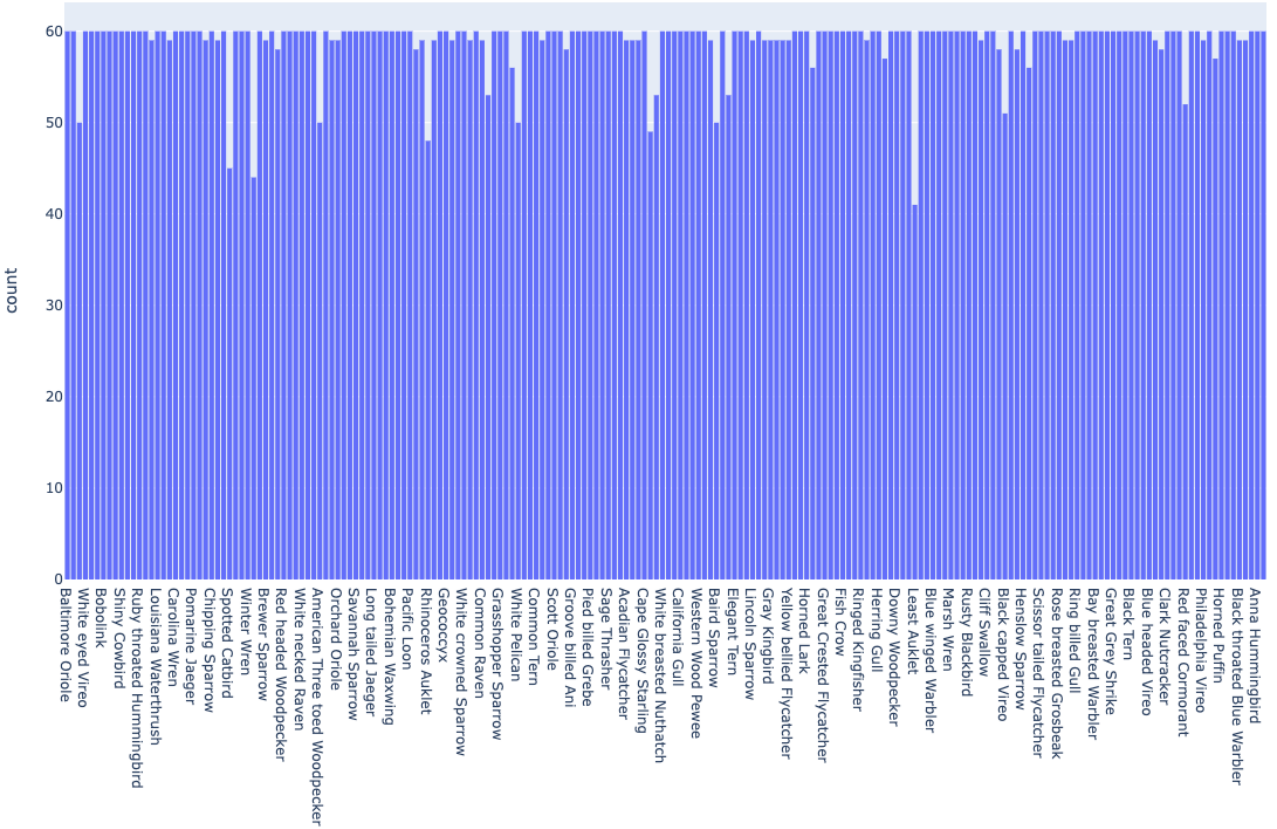
- Caltech-UCSD Birds (+ 10 textual descriptions per image)
  200 categories, 11,788 images
- Oxford-102 Flowers (+ 10 textual descriptions per image)
  102 categories, 8,189 images
- ~200k (image, caption, species)
- Very fine-grained

Datasets – class distribution

Flowers

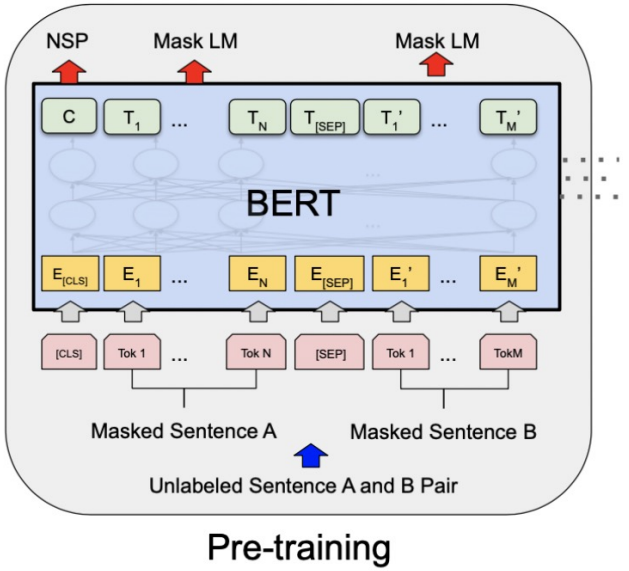


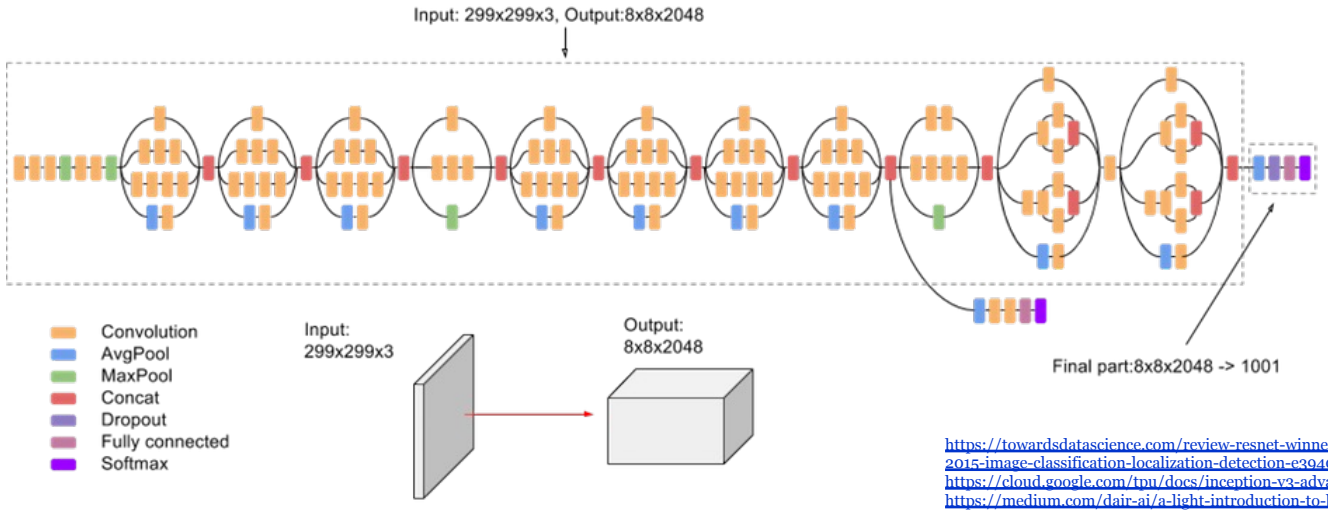
Birds

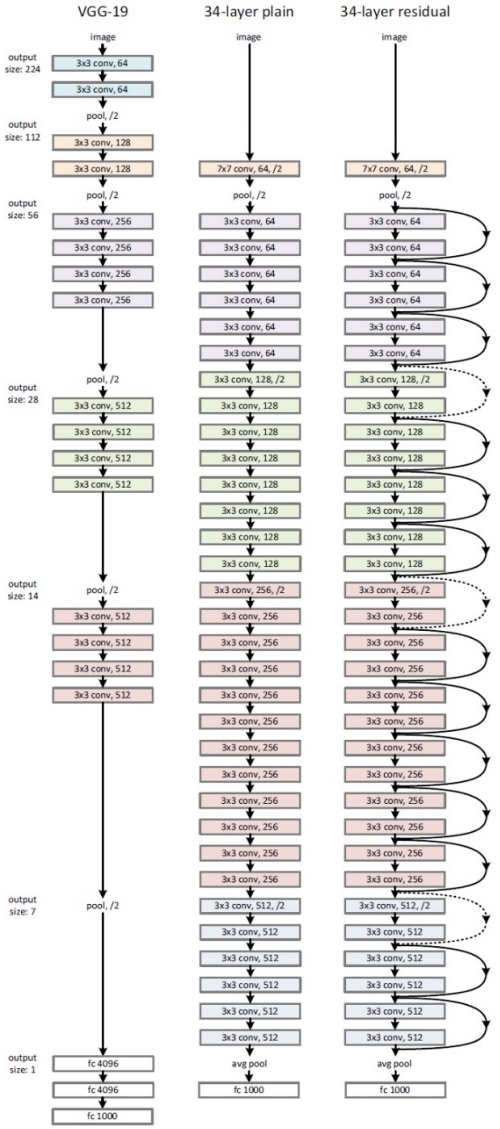


Background

- Text encoders:
  - DistilBERT
  - TF-IDF
- Image classifiers:
  - VGG
  - ResNet
  - Inception V3

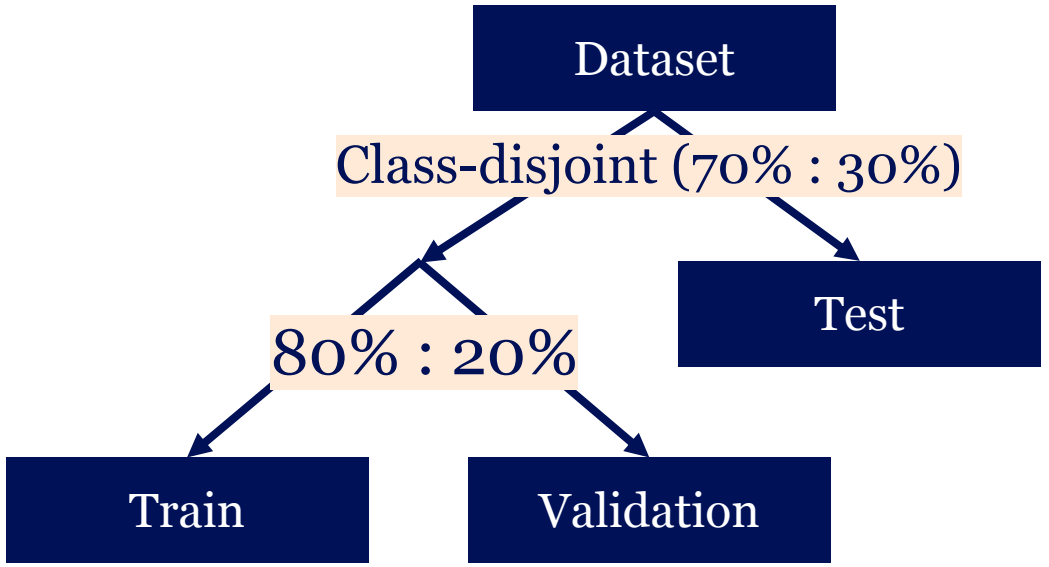






https://towardsdatascience.com/review-resnet-winner-of-ilsvrc-2015-image-classification-localization-detection-e39402bfa5d8
https://cloud.google.com/tpu/docs/inception-v3-advanced
https://medium.com/dair-ai/a-light-introduction-to-bert-2da54f96b68c

Methods



PyTorch

scikit learn

Hugging Face

GENSIM
topic modelling for humans

Methods – steps

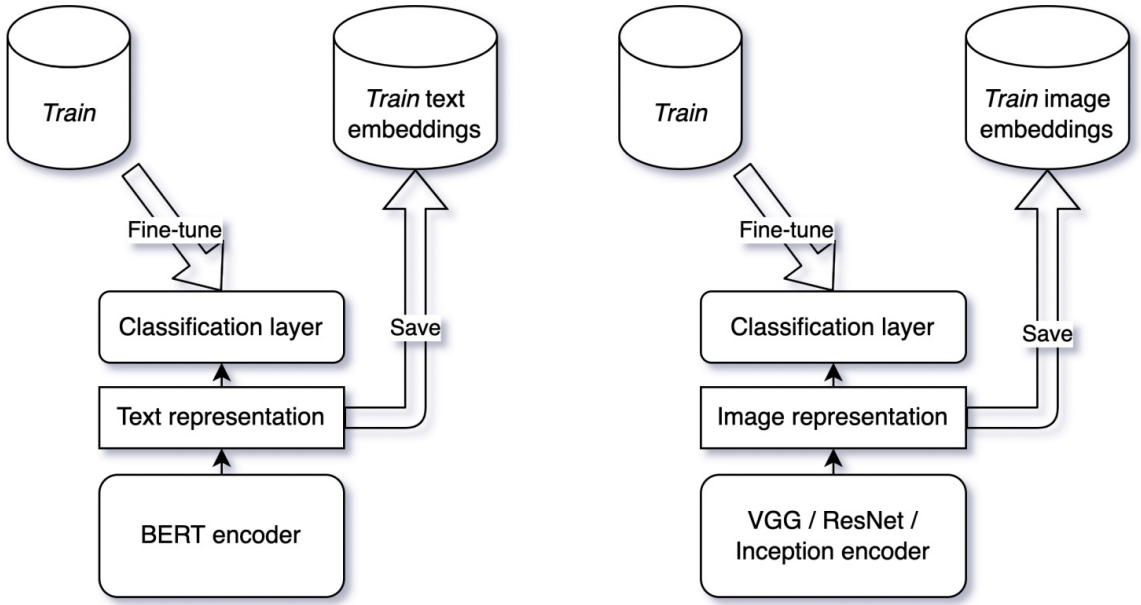


Fig. 3. The first training step is to separately fine-tune the networks via text-classification and image-classification. *Train* denotes a class-disjoint split of either *birds* or *flowers*.

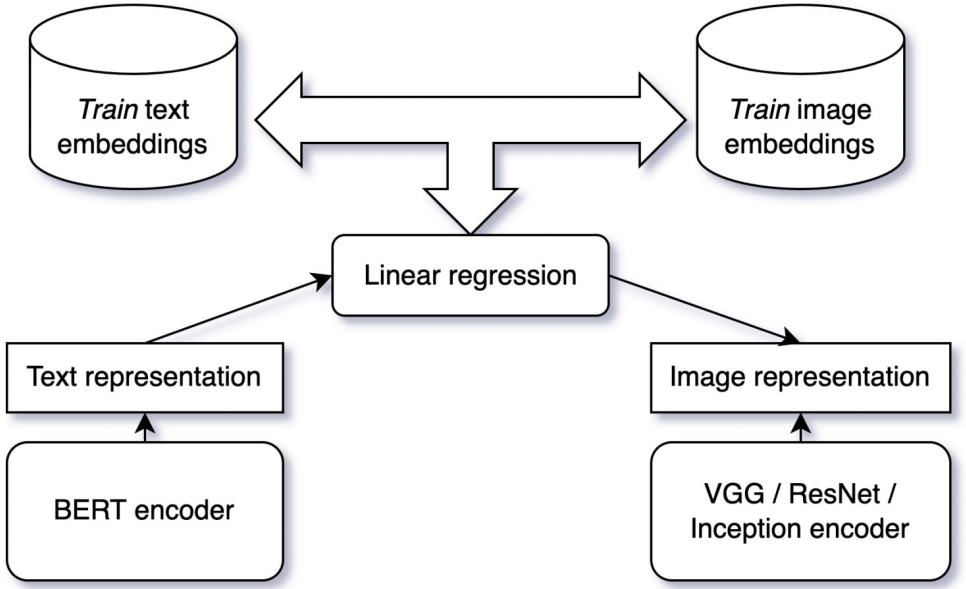


Fig. 4. For the second-stage of training, the last, classification layers are removed from both models. The exposed, raw text embeddings are mapped to the image-embeddings using a linear projection learned from the actual embeddings of the *train* split.

Methods – steps

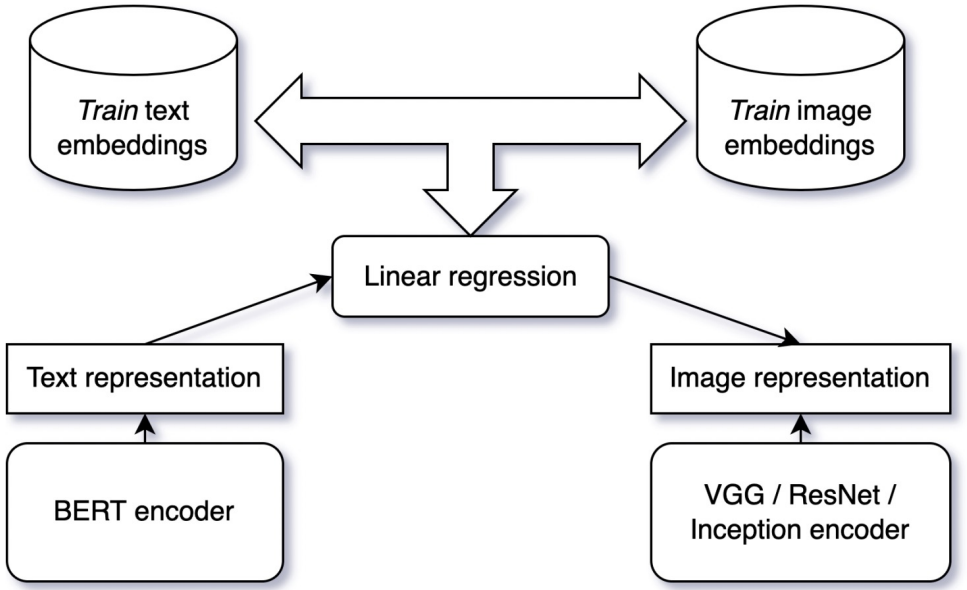


Fig. 4. For the second-stage of training, the last, classification layers are removed from both models. The exposed, raw text embeddings are mapped to the image-embeddings using a linear projection learned from the actual embeddings of the *train* split.

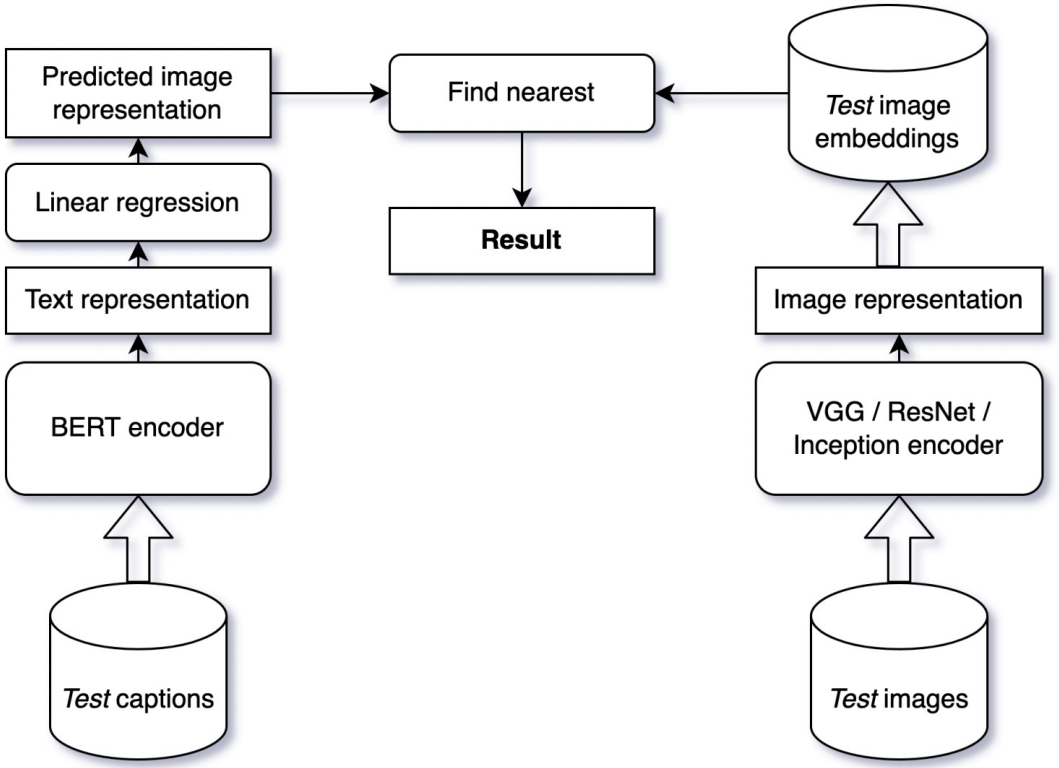


Fig. 5. First, each image from the *test* split is encoded using the fine-tuned and subsequently beheaded image-classifier. For each caption, an image embedding is predicted using the fine-tuned text encoder and a linear projection. Finally, the images with the most similar embedding to the predicted image embedding are returned.

Results – fine-tuning

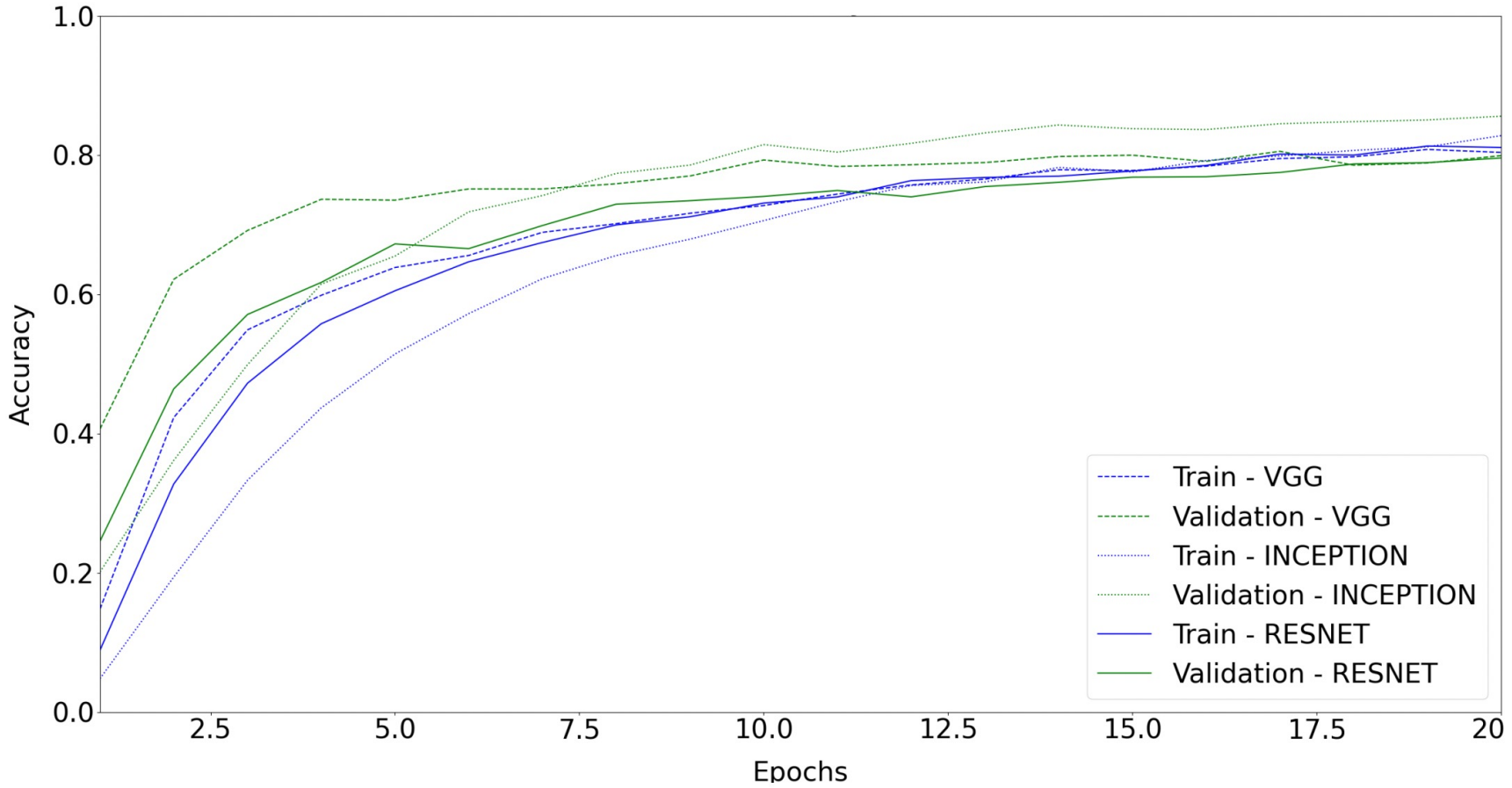


Fig. 6. The accuracy metrics over time of the image classifiers on the *birds* train and validation datasets.

Results – fine-tuning

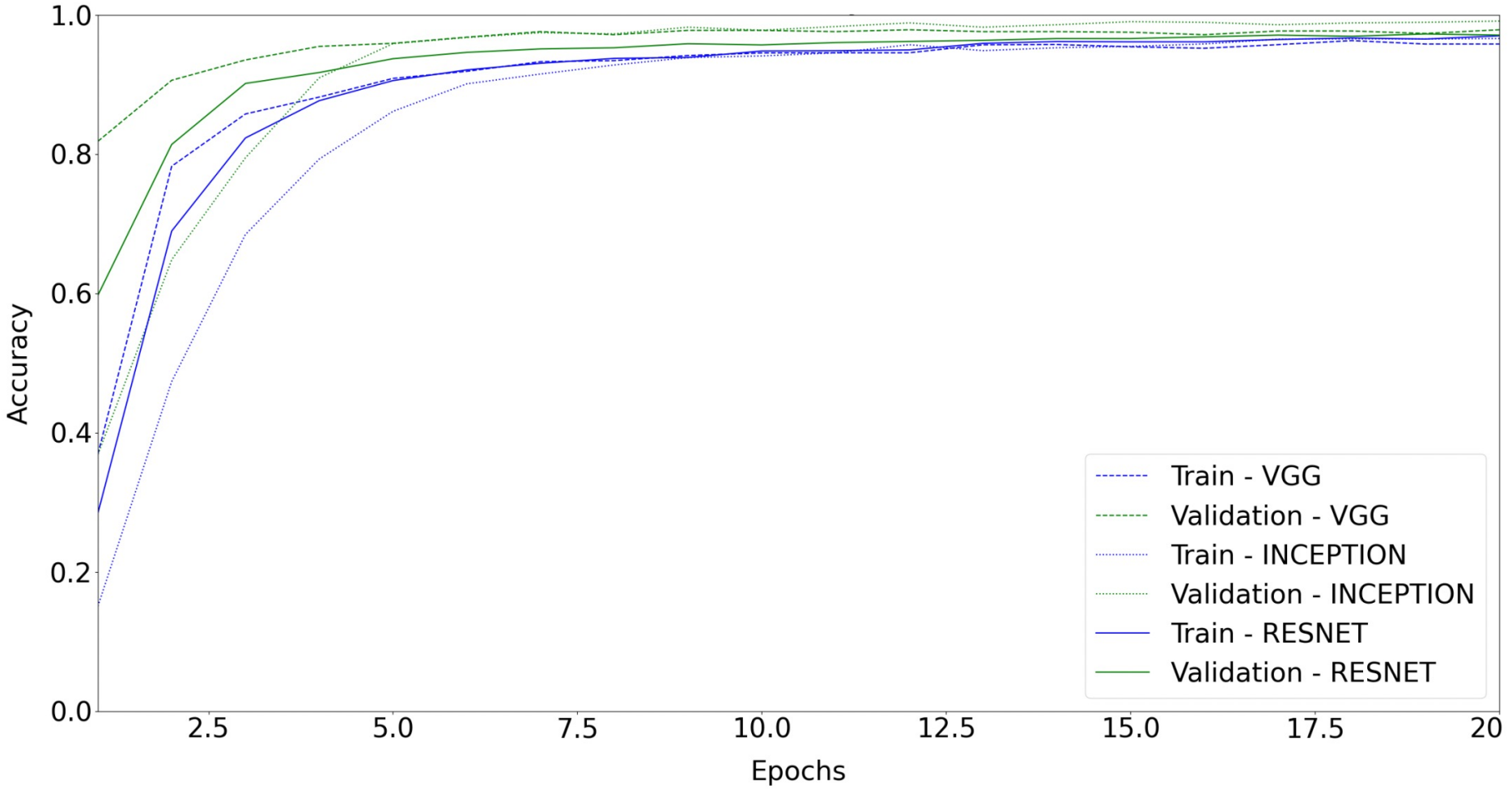


Fig. 7. The accuracy metrics over time of the image classifiers on the *flowers* train and validation datasets.

Results tables

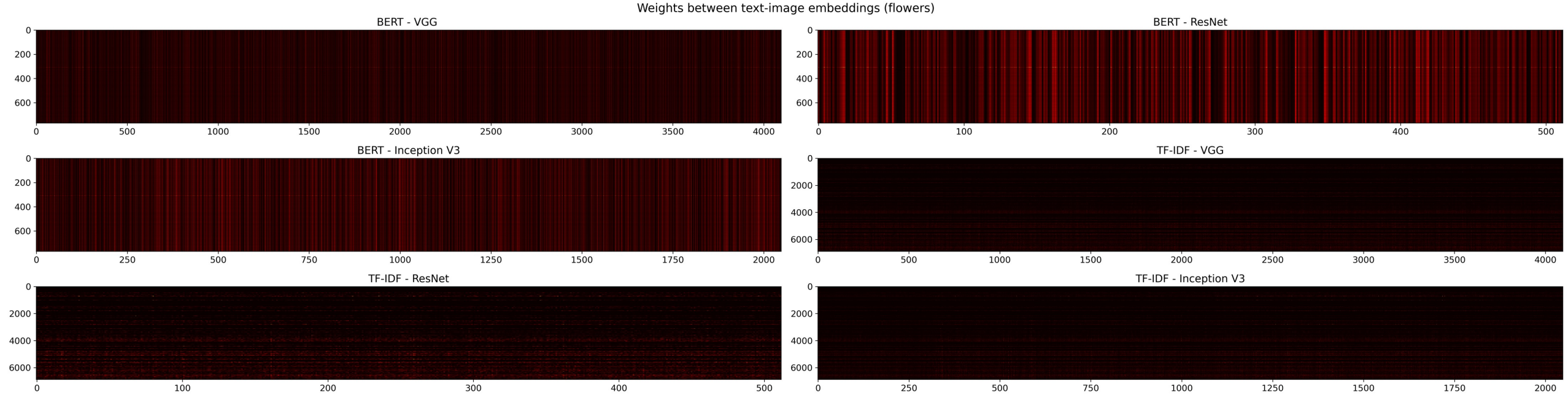
Table 1. Comparison of the 6 presented models on the task of zero-shot text-based image retrieval using the Caltech-UCSD *birds* 200 [10] dataset.

	Top-1 Accuracy	Precision	Recall	F1-score	Top-5 Accuracy
DistilBERT - VGG	0.19	0.25	0.19	0.19	0.43
DistilBERT - ResNet	0.11	0.17	0.11	0.10	0.32
DistilBERT - Inception v3	0.12	0.33	0.11	0.11	0.29
TF-IDF - VGG	0.11	0.28	0.11	0.11	0.33
TF-IDF - ResNet	0.10	0.33	0.10	0.09	0.24
TF-IDF - Inception v3	0.10	0.38	0.10	0.09	0.21

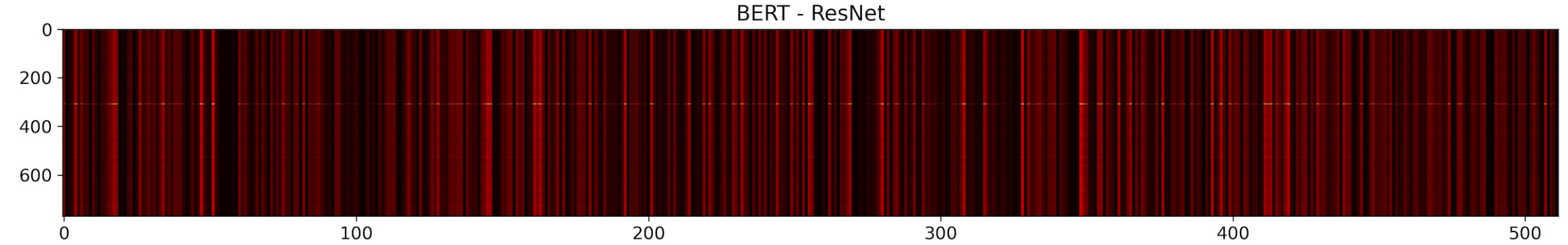
Table 2. Comparison of the 6 presented models on the task of zero-shot text-based image retrieval using the Oxford-102 *flowers* [11] dataset.

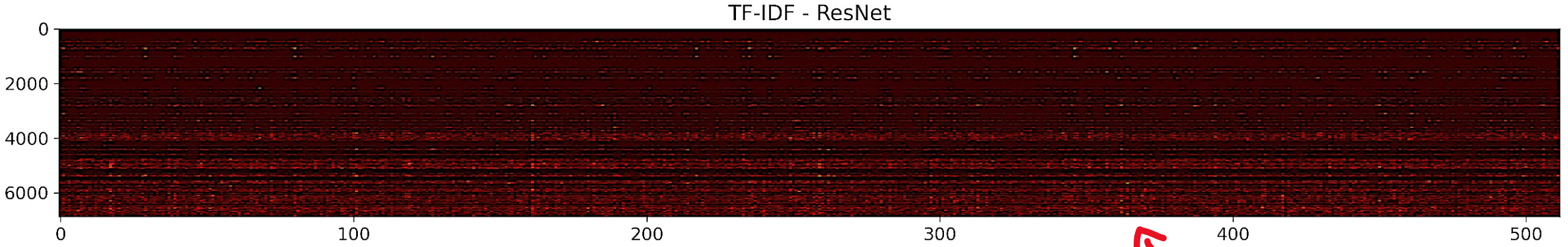
	Top-1 Accuracy	Precision	Recall	F1-score	Top-5 Accuracy
DistilBERT - VGG	0.37	0.30	0.27	0.23	0.57
DistilBERT - ResNet	0.36	0.31	0.26	0.24	0.53
DistilBERT - Inception v3	0.40	0.29	0.29	0.27	0.58
TF-IDF - VGG	0.20	0.28	0.16	0.13	0.46
TF-IDF - ResNet	0.18	0.28	0.13	0.12	0.33
TF-IDF - Inception v3	0.24	0.24	0.17	0.15	0.43

Results – weights



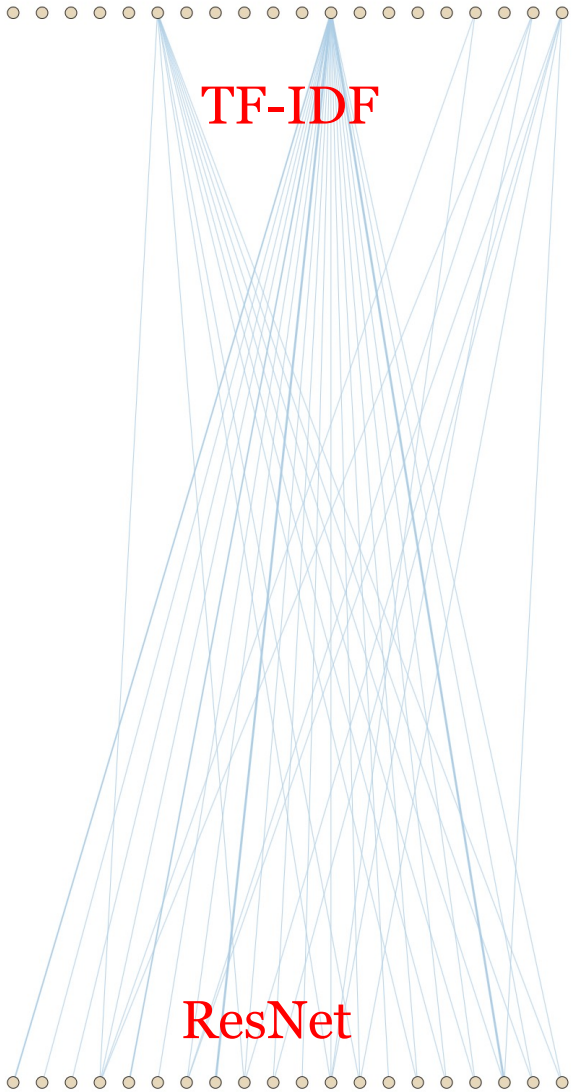
Results – weights

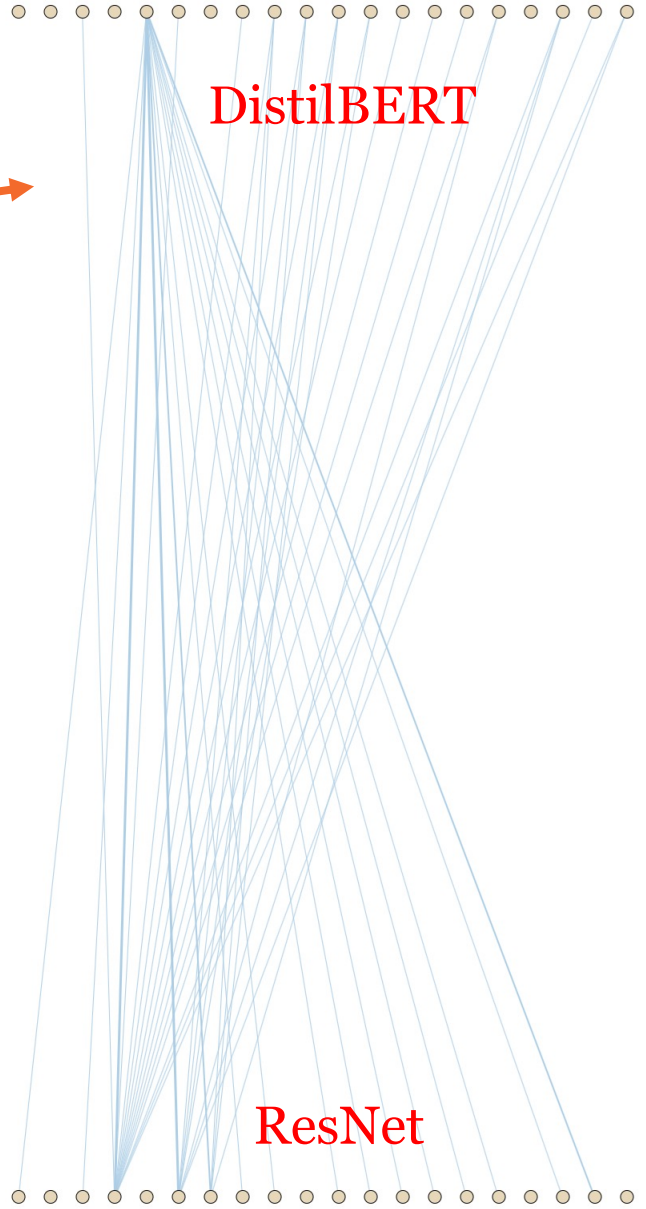




Colour-adjusted for better visibility; see original in the report.

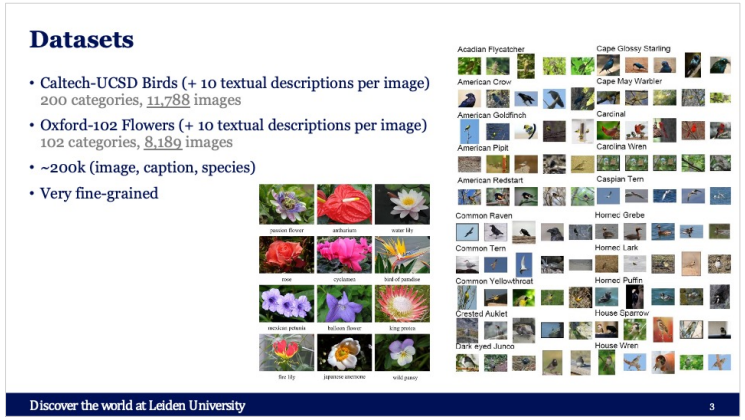
Results – top weights

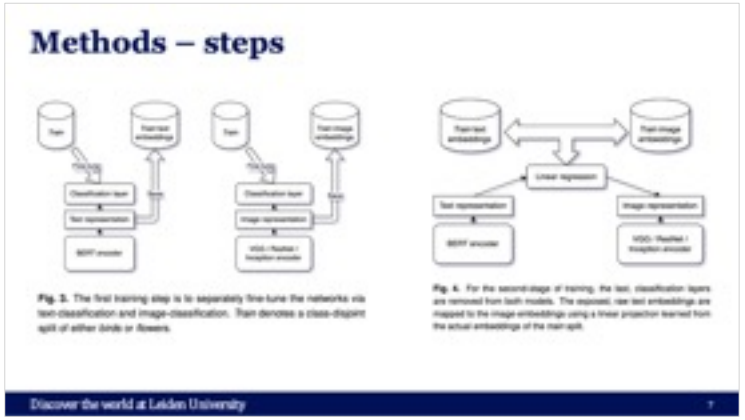


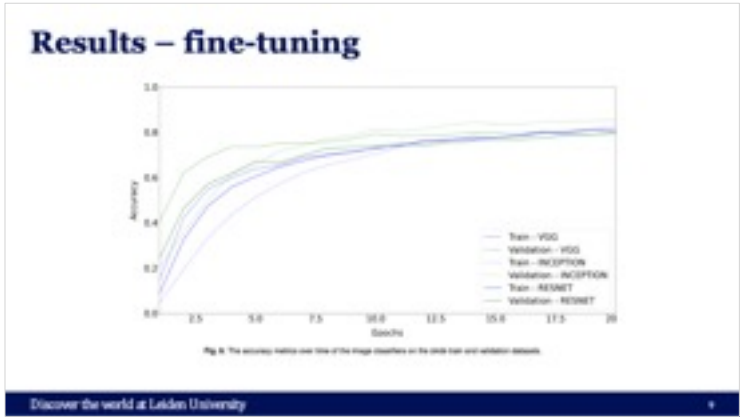


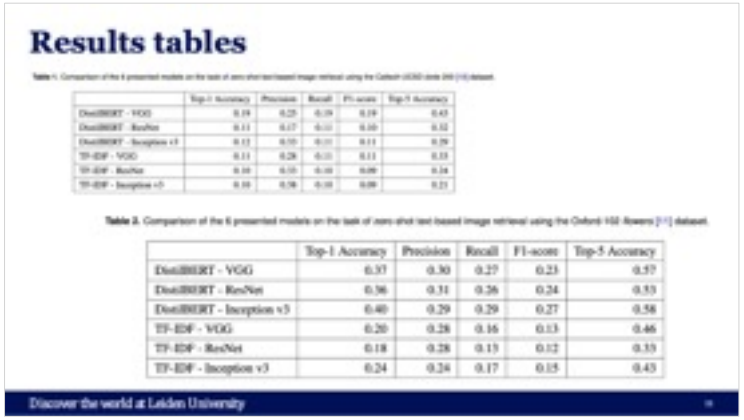
Conclusion

- **6 architectures** were compared
- On highly domain-specific **fine-grained** classes, in a **zero-shot** setup
- Fine-tuned on different tasks then combined together
- **It worked**
- **There is similarity in the embedding spaces**
- TF-IDF << **DistilBERT**
- No absolute winner for image processing architectures
- **Worse** than Reed et al.'s **but** more **general**

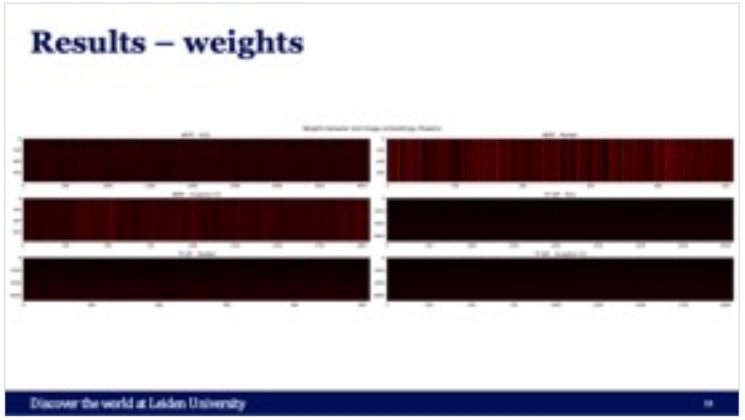


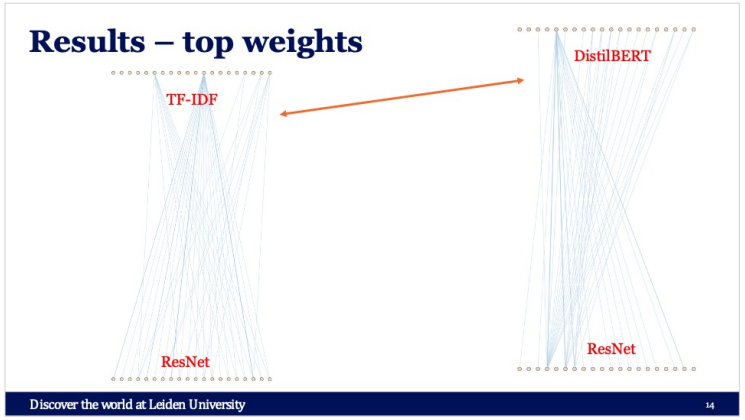






	Top-1 Accuracy	Precision	Recall	F1-score	Top-5 Accuracy
DistilBERT - VGG	0.37	0.30	0.27	0.23	0.57
DistilBERT - ResNet	0.36	0.31	0.26	0.24	0.53
DistilBERT - Inception v3	0.40	0.29	0.29	0.27	0.58
TF-IDF - VGG	0.20	0.28	0.16	0.13	0.46
TF-IDF - ResNet	0.18	0.28	0.15	0.12	0.35
TF-IDF - Inception v3	0.24	0.24	0.17	0.15	0.43





https://github.com/schmelczer/mir-final

Universiteit Leiden
The Netherlands

Thank you!