On Identification and Retrieval of Near-Duplicate Biological Images: a New Dataset and Protocol

T.E. Koker *, S.S. Chintapalli*, S. Wang, B.A. Talbot, D. Wainstock, M. Ciccone, and M.C. Walsh
Harvard Medical School

INTRODUCTION

- Duplicative image data reporting in the biomedical literature is more prevalent [1] than previously realized and raises questions concerning data integrity, quality, and reliability.
- Often image data duplication and/or manipulation are detected in an ad hoc manner by fellow scientists or by editorial staff during a manuscript review process.
- At the rate at which the scientific literature is expanding, it is not feasible for all cases of image manipulation to be detected by peer review. Thus, there is a continued need for more systematic and automated tools to detect potential duplication.

We introduce **BINDER - Bio-Image Near-Duplicate Examples Repository**, a novel dataset to help researchers develop, train, and test models to detect same-source biomedical images. We further use the dataset to demonstrate how novel adaptations of existing image retrieval and metric learning models (deep neural networks) can be applied to achieve high accuracy inference results, creating a baseline for future work. In aggregate, we present a supervised protocol for near-duplicate image identification and retrieval without any "real-world" training example.

BINDER: Bio-Image Near-Duplicate Examples Repository

Limitations –

- Difficulty obtaining published images (flagged for duplication) at a volume necessary to train high-capacity neural networks.
- Making the dataset open-access would require author/publisher's consent.

Our Solution –

- Synthetic manipulations (ref. Table. I) identified from a small real-world examples (RWE – 126 pairs) set were applied to a large collection of biological (microscopy) images gathered from open-source public repositories‡ encompassing 17 classes/categories of cell types/model organisms.
- The microscopy images were split class-wise into training, test, and validation sets (17 categories split into 9, 4, and 4 classes respectively).

BINDER contains 7,490 unique image patches for model **training**, 1,821 same-size patch duplicates for **validation** and **testing**, and 868 different-size image/patch pairs for image-retrieval validation and testing.

TABLE I
SYNTHETIC MANIPULATIONS

Manipulation	Procedure
Vertical Flip	$P = 0.5$
Horizontal Flip	$P = 0.5$
Color Invert	$P = 0.1$
Perspective	Corners perturbed $U(-20, 20)$ pixels
Scale	$U(0.75, 1.25)$
Rotation	$U(-20, 20)$ degrees
x and y Translation	$U(-10, 10)$ pixels each
Gamma Correction	$U(0.5, 1.5)$
Brightness	$U(0.9, 1.1)^1$
Contrast	$U(0.5, 1.5)^1$
JPEG Compression	$P = 0.1$, quality $= 50$

P signifies probability of manipulation being performed, U represents the range of random uniform distribution in which manipulation will be applied.

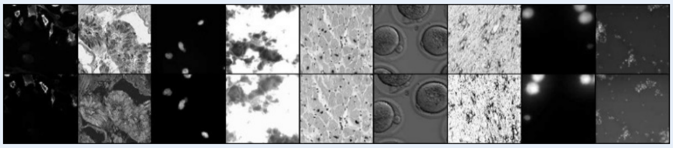

Fig. 1. Examples of duplicate pairs (9 categories - training). Row 1 - original patches; Row 2 - their respective manipulations.

DISCUSSIONS

- ❑ Our primary contribution is towards creating an assessment platform and benchmark dataset – BINDER - to identify and metrically assess image duplications in biomedical literature.
- ❑ A tool like this can be extremely useful for our community in the surveillance and maintenance of research data integrity, but it needs to be used with caution.
- ❑ Instances of duplication within the literature could be genuine errors. Each scenario of potential image duplication/manipulation requires thoughtful discussion and review with authors and stakeholders invested in the recording or reporting of associated research data.

MODELS & EVALUATION

- Three Siamese Neural Network based models were trained and tested for Near-Duplicate Image Retrieval (ref Figure 2). On a high level, the model outputs a global descriptor (an embedding) for each query and potential duplicate image. And the duplication is confirmed if the Euclidean distance between their embeddings is less than a threshold, *t*.
- The three models [2, 3] are trained to reduce triplet loss with hard negative mining, to bring duplicates closer and push non-duplicates farther apart.
- The models are pretrained on COCO and fine-tuned on BINDER. To quantify the models' ability to retrieve duplicate image pairs – area under the ROC curve (AUC) is used.
- The models are evaluated on 3 datasets – BINDER test set, RWE set, and MFND IND set [4].

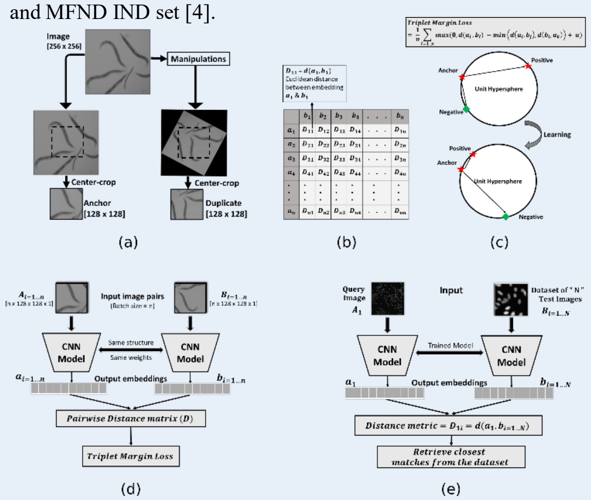
Image
[256 x 256]
Manipulations
Center-crop
Anchor
[128 x 128]
Center-crop
Duplicate
[128 x 128]
(a)
$D_{11} = d(a_1, b_1)$
Euclidean distance between embedding a_1 & b_1
b_1 b_2 b_3 b_4 . . . b_n
a_1 D_{11} D_{12} D_{13} D_{14} . . . D_{1n}
a_2 D_{21} D_{22} D_{23} D_{24} . . . D_{2n}
a_3 D_{31} D_{32} D_{33} D_{34} . . . D_{3n}
a_4 D_{41} D_{42} D_{43} D_{44} . . . D_{4n}
a_n D_{n1} D_{n2} D_{n3} D_{n4} . . . D_{nn}
(b)
Triplet Margin Loss
$= \frac{1}{n}\sum_{i=1..n} max(0, d(a_i, b_i) - min(d(a_i, b_j), d(b_i, a_k)) + \alpha)$
Anchor
Positive
Unit Hypersphere
Negative
Learning
Positive
Anchor
Unit Hypersphere
Negative
(c)
$A_{i=1...n}$
[n x 128 x 128 x 1]
Input image pairs
(Batch size = n)
$B_{i=1...n}$
[n x 128 x 128 x 1]
CNN Model
Same structure
Same weights
CNN Model
$a_{i=1...n}$
Output embeddings
$b_{i=1...n}$
Pairwise Distance matrix (D)
Triplet Margin Loss
(d)
Query Image A_1
Input
Dataset of " N " Test Images $B_{i=1...N}$
CNN Model
Trained Model
CNN Model
a_1
Output embeddings
$b_{i=1...N}$
Distance metric = $D_{1i} = d(a_1, b_{i=1...N})$
Retrieve closest matches from the dataset
(e)

Fig. 2. (a) Synthetic image manipulation for duplicate pair generation. Random manipulations (see Table I) are applied to generate duplicate images. (b) Pairwise distance matrix as used during training. (c) Triplet margin loss minimizes the distance between positive pairs and maximizes the distance between negative pairs during training. (d) Model training diagram. (e) Model testing/deploying diagram.

RESULTS

VGG19 + GeM model outperforms the other models (see purple lines on plots). In addition, we observe that fine-tuning on the field specific dataset (BINDER) yields improved results.

TABLE II
MODEL BENCHMARKS

Model	Fine-tune	BINDER Test	RWE	MFND IND.
Autoencoder	None	0.25/0.92	0.53/0.95	0.935/0.999
Autoencoder	Bio	0.25/0.93	0.55/0.96	
RN50+GeM	None	0.54/0.99	0.70/0.98	0.985/1.000
RN50+GeM	Bio	0.55/0.99	0.68/0.98	
VGG19+GeM	None	0.58/0.99	0.69/0.99	0.988/1.000
VGG19+GeM	Bio	**0.63**/0.99	0.69/0.98	

Area under the ROC curve for each dataset tested with [hardest/random] negative sampling. Each result is the median of 5 runs.

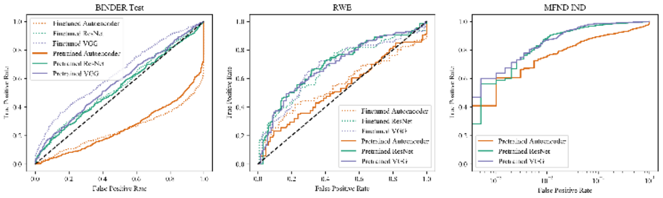

Fig. 3. ROC curve for one run on each dataset, using hard negative mining. Log scaling is used for MFND IND. Threshold is sampled in [0, 2].

REFERENCES

[1] Bik, Elisabeth et.al., "*The Prevalence of Inappropriate Image Duplication in Biomedical Research Publications*," DOI: 10.1128/mBio.00809-16.

[2] F. Radenović et.al., "*Fine-Tuning CNN Image Retrieval with No Human Annotation*," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 7, pp. 1655-1668, 1 July 2019 , DOI: 10.1109/TPAMI.2018.2846566.

[3] L. V. Utkin et. al., "*A Siamese Autoencoder Preserving Distances for Anomaly Detection in Multi-robot Systems*," *2017 ICCAIRO*, Prague, 2017, pp. 39-44, DOI: 10.1109/ICCAIRO.2017.17.

[4] Connor, Richard et. al., "*Identification of MIR-Flickr Near-duplicate Images - A Benchmark Collection for Near-duplicate Detection*," DOI: 10.5220/0005359705650571.

‡ NYU Mouse Embryo Tracking Database (METD), Broad Bioimage Benchmark Collection (BBBC), Adiposoft Image Dataset (AID), and Open Microscopy Image Data Resource (IDR).

* Equal authorship contribution