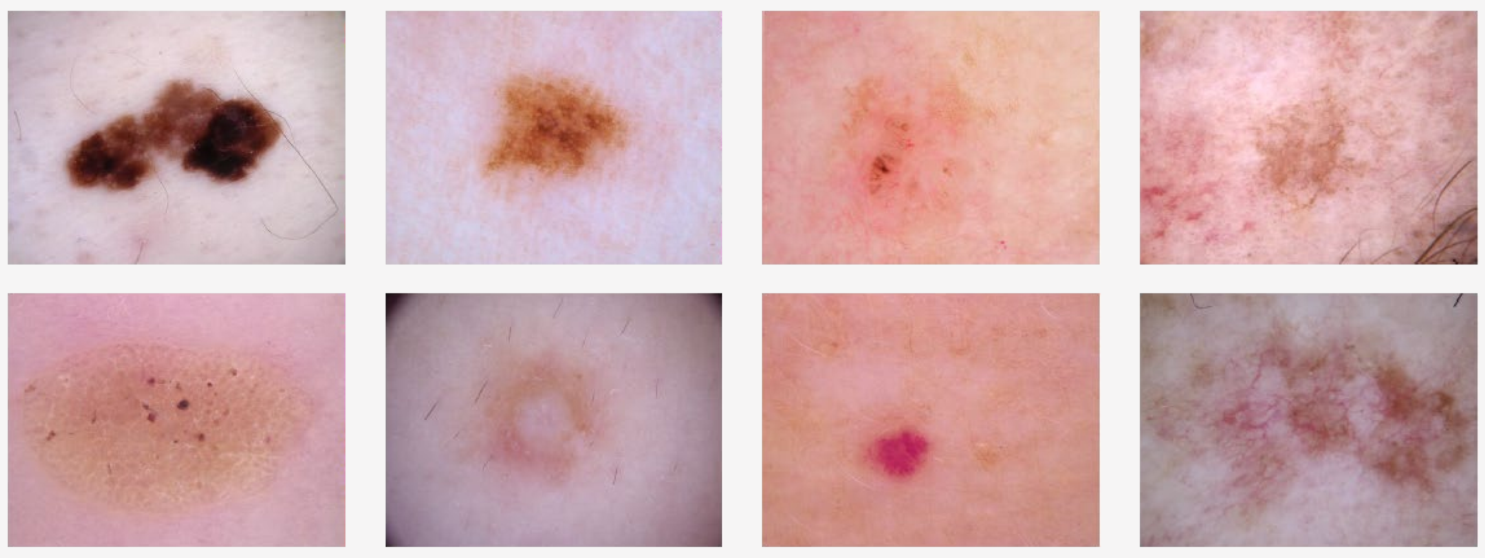
Supporting Skin Lesion Diagnosis with Content-Based Image Retrieval

Stefano Allegretti, Federico Bolelli, Federico Pollastri, Sabrina Longhitano, Giovanni Pellacani, and Costantino Grana
Università degli Studi di Modena e Reggio Emilia, Italy

Skin Cancer

- Skin cancer is one of the most common forms of human cancer worldwide
- If caught early, it is usually curable
- Distinguishing skin cancer from other kinds of skin lesion is a difficult task



Goal— Realize a skin lesion retrieval system, which exploits features extracted by a CNN to gather images from a classified dataset that are similar to a new lesion, in order to assist dermatologists in the diagnosis process.

Computer Aided Diagnosis with CNN

- Convolutional Neural Networks have been widely employed for **skin lesion classification**[1]
- Classification CNNs have pros and cons:

Good classification accuracy, comparable to expert dermatologists

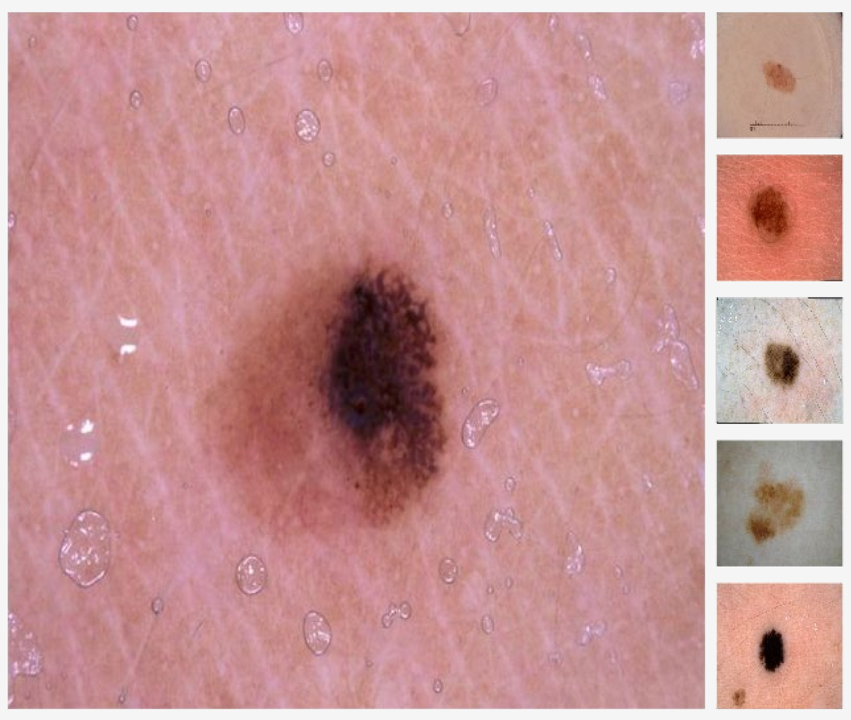
Low interpretability: Scarce diagnostic aid for physicians

- Dermatologists are always responsible for the final diagnosis of skin lesions. Can they blindly trust automatic classifiers?
- How can interpretability be improved?

[1] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, 2017.

Content-Based Image Retrieval (CBIR)

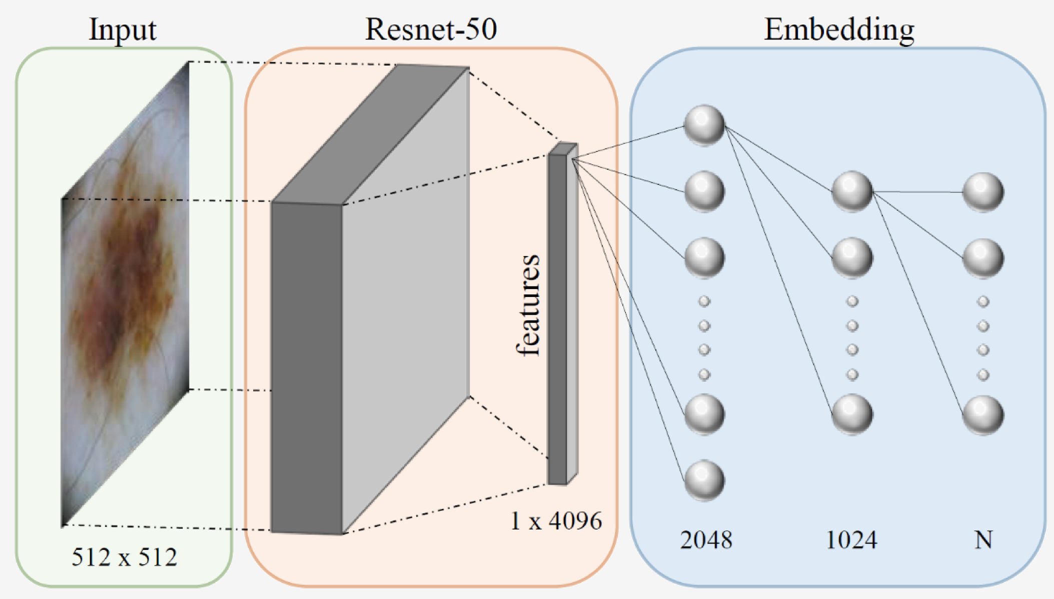
- Given a new lesion, retrieve **similar cases** from a labeled database
- How to define image similarity?
- Past works:
  - Euclidean or Bhattacharyya distance between handcrafted features
  - Hamming distance between hash codes, computed with a modified classification CNN (AlexNet)[1]



[1] X. Pu, Y. Li, H. Qiu, and Y. Sun, "Deep Semantics-Preserving Hashing Based Skin Lesion Image Retrieval," in *Advances in Neural Networks - ISNN 2017.* Springer, 2017, pp. 282–289.

Proposed CBIR system

- ResNet Feature Extractor: ResNet, except for the last FC layer
- Embedding Network: 2 FC layers, built on ResNet-extracted features



- **Cosine similarity** between image embeddings
- Embedding training with **triplet loss** function
- **ISIC dataset -** 20K images, 8 classes

Experimental Results

QUANTITATIVE

4 variations of the proposed model:

- ResNet FE trained for **classification, with cross-entropy loss**
- ResNet FE + EmbNet, **end-to-end** trained with **triplet loss**
- ResNet FE pretrained for classification + EmbNet, **only** EmbNet trained with **triplet loss**
- ResNet FE pretrained for classification + EmbNet, **end-to-end** trained with **triplet loss**

Average precision at k measured for every model analyzed, for three values of k, 1, 5, and 10. Central columns report average values separated for each class, and the last column reports the balanced average. Novel proposals are identified by *.

Model	Cut-Off k	Per class P@k								AP@k
		MEL	NV	BCC	AK	BKL	DF	VASC	SCC	
Hash-AP [45]	-	0.4786	0.6111	0.5730	0.1896	0.1984	0.1505	0.3842	0.1375	0.3404
Hash-AP ResNet*	-	0.8176	0.7558	0.8509	0.7417	0.6256	0.7604	0.8271	0.6851	0.7580
Classification*	1	0.7840	0.9369	0.9347	0.7400	0.8300	0.7733	0.8667	0.7133	0.8224
	5	0.7262	0.9111	0.9029	0.7190	0.7724	0.7333	0.8373	0.6853	0.7859
	10	0.7040	0.9038	0.8957	0.7160	0.7470	0.7213	0.8307	0.6787	0.7746
Embedding End-to-End*	1	0.7400	0.9018	0.9133	0.6600	0.7520	0.7333	0.8267	0.7000	0.7784
	5	0.7314	0.8923	0.9005	0.6820	0.7576	0.7387	0.8240	0.7027	0.7786
	10	0.7322	0.8905	0.8993	0.6855	0.7572	0.7440	0.8253	0.7093	0.7804
Class & Embedding*	1	0.7490	0.9347	0.8973	0.7150	0.7700	0.8400	0.9067	0.7133	0.8157
	5	0.7542	0.9347	0.9013	0.7170	0.7768	0.8400	0.9013	0.7093	0.8168
	10	0.7531	0.9331	0.9032	0.7170	0.7814	0.8453	0.9040	0.7147	0.8190
Class & Embedding End-to-End*	1	0.7560	0.9022	0.9027	0.6600	0.7600	0.7733	0.8267	0.7133	0.7867
	5	0.7458	0.9030	0.8992	0.6770	0.7588	0.7707	0.8293	0.7213	0.7881
	10	0.7436	0.9012	0.9009	0.6830	0.7668	0.7760	0.8373	0.7273	0.7920

QUALITATIVE

- Dermatologists classified 100 lesions two times
  - **Task 1 –** without aid
  - **Task 2 –** with the 5 most similar labeled images, retrieved by the best performing model
- Average results are:
  - **Task 1 –** 67.4%
  - **Task 2 –** 76.6%
- Mean classification improvement of 9.2%

	Task 1	Task 2
Dermatologist #1	75%	79%
Dermatologist #2	64%	80%
Dermatologist #3	69%	71%
Dermatologist #4	68%	82%
Dermatologist #5	61%	71%