A Multi-modal Neural Embeddings Approach for Detecting Mobile Counterfeit Apps

Jathushan Rajasegaran, Naveen Karunanayake, Ashanie Gunathillake,
Suranga Seneviratne and Guillaume Jourjon

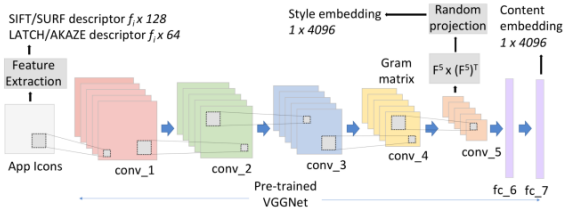
Motivation

- Counterfeit apps impersonate existing popular apps in attempts to misguide users.
- Reasons behind app impersonations include:
  - Harvesting user credentials
  - Increased advertising revenue
  - Spreading malware

Angry Birds Angry Purrs Temple Run Temple Piggy

Approach

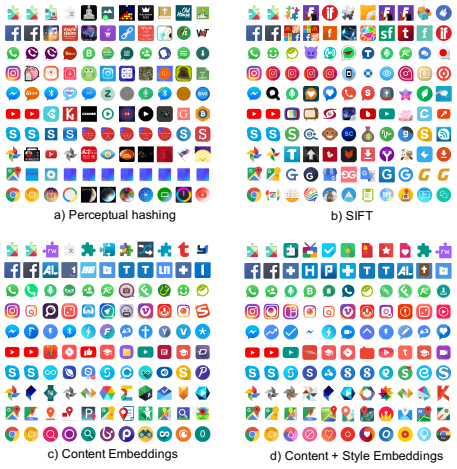
- Multi-modal embeddings for app similarity
  - Content embeddings (VGG19 fc_7)
  - Style embeddings (Gram matrix of conv_5)
  - Very spare random projection to reduce size
  - Paragraph vectors for text embeddings



Neural embeddings (Cosine distance)		
Content (C_{cos})	4,096	$1 - \frac{X_i^{cont} \cdot X_t^{cont}}{\|\|X_i^{cont}\|\|_2 \|\|X_t^{cont}\|\|_2}$
Style (S_{cos})	4,096	$1 - \frac{X_i^{style} \cdot X_t^{style}}{\|\|X_i^{style}\|\|_2 \|\|X_t^{style}\|\|_2}$
Text (T_{cos})	100	$1 - \frac{X_i^{text} \cdot X_t^{text}}{\|\|X_i^{text}\|\|_2 \|\|X_t^{text}\|\|_2}$
Content+Style	8,192	$\alpha\ C_{cos} + \beta\ S_{cos}$
Content+Style+Text	8,292	$\alpha\ C_{cos} + \beta\ S_{cos} + \gamma\ T_{cos}$

Table 1: Neural embeddings and their sizes

- Evaluate the performance on **standard image retrieval datasets**: UKBench and Holidays as well as a manually labelled app icon dataset.
- Compare performance with **hashing-based methods** (e.g. average, difference, perceptual) and **feature-based methods** (e.g. SIFT, SURF).



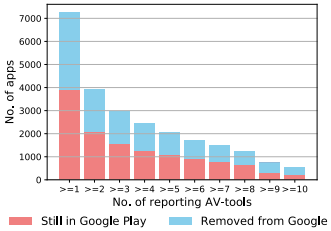
Retrieved icons for top-10 apps in Google Play Store

		Difference	Perceptual	SIFT	SURF	C_{cos}	S_{cos}	$C_{cos} + \beta S_{cos}$	$C_{cos} + \beta S_{cos} + \gamma T_{cos}$
Holidays	5-NN	22.68	21.60	33.00	31.12	46.36	46.72	**47.92**	N/A
	10-NN	11.74	10.96	17.58	16.66	25.28	25.24	**25.92**	N/A
	15-NN	8.08	7.36	12.15	11.55	17.47	17.25	**17.89**	N/A
	20-NN	6.19	5.58	9.34	8.91	13.31	13.13	**13.57**	N/A
UKBench	5-NN	22.44	21.63	55.27	52.97	**70.22**	65.01	70.06	N/A
	10-NN	11.70	10.97	28.99	27.93	**36.90**	33.86	36.62	N/A
	15-NN	7.95	7.38	19.82	19.12	**25.03**	22.95	24.79	N/A
	20-NN	6.08	5.59	15.11	14.60	**18.97**	17.40	18.75	N/A
Labelled	5-NN	48.41	47.62	48.92	47.67	56.43	60.42	62.23	**64.76**
	10-NN	28.10	27.42	26.79	27.05	33.69	35.39	36.04	**38.47**
	15-NN	19.92	19.45	18.86	19.00	24.05	25.25	25.57	**27.19**
	20-NN	15.56	15.24	14.57	14.69	18.69	19.66	19.86	**21.09**
All	5-NN	38.01	37.07	38.23	39.13	45.51	50.72	50.91	**55.96**
	10-NN	21.53	20.79	21.82	22.10	26.08	29.57	29.81	**32.99**
	15-NN	15.30	14.74	15.31	15.52	18.30	20.90	21.12	**23.46**
	20-NN	11.89	11.40	11.87	11.97	14.07	16.14	16.31	**18.23**

Table 2: Precision@k (NN* - Nearest Neighbors)

Results

- We next do a 10-NN search on the top-10,000 apps in the Google Play Store and check retrieved apps for:
  - Malware inclusion (VirusTotal)
  - Additional ad library inclusion
  - Requesting extra dangerous permissions



Original app	Similar app	AV-tools	Downloads (Original)	Downloads (Similar)
Clean Master	Ram Booster*	12	500 million - 1 billion	500 - 1,000
Temple Run	Endless Run*	12	100 million - 500 million	5,000 - 10,000
Temple Run 2	Temple Theft Run*	12	500 million - 1 billion	500, 000 - 1 million
Hill Climb Racing	Offroad Racing: Mountain Climb	9	100 million - 500 million	1 million - 5 million
Flow Free	Colored Pipes	8	100 million - 500 million	1 million - 5 million
Parallel Space	Double Account*	17	50 million - 100 million	100, 000 - 500, 000

* The app is currently not available in Google Play Store

Acknowledgement

This project is partially funded by the Google Faculty Awards 2017, NSW Cyber Security Network's Pilot Grant Program 2018, and the DSTG Next Generation Technologies Program.