Towards a Unified Multi-modal Network Data Representation

Zied Ben Houidi, Raphael Azorin, Massimo Gallo,
Alessandro Finamore, Dario Rossi

Correspondenc
zied.ben.houidi@huawei.com
raphael.azorin@huawei.com

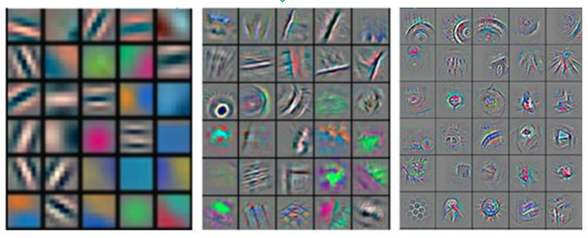
Accepted at ACM HotNets'2022
https://doi.org/10.1145/3563766.3564108

Motivation

Learning good data representations is paramount to Deep Learning's success

Quantities **in Computer Vision**

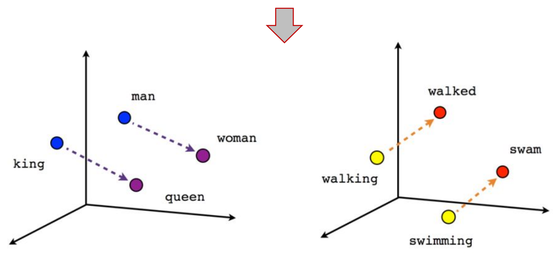
Encoder (e.g., CNN)



Low-level Mid-level High-level

Entities **in Natural Language Processing**

Encoder (e.g., Word2Vec)



Male-Female Verb tense

Entities (IPs, BGP, …) and Quantities (size, IAT, …)
in Networking

What is the best representation learning strategy?

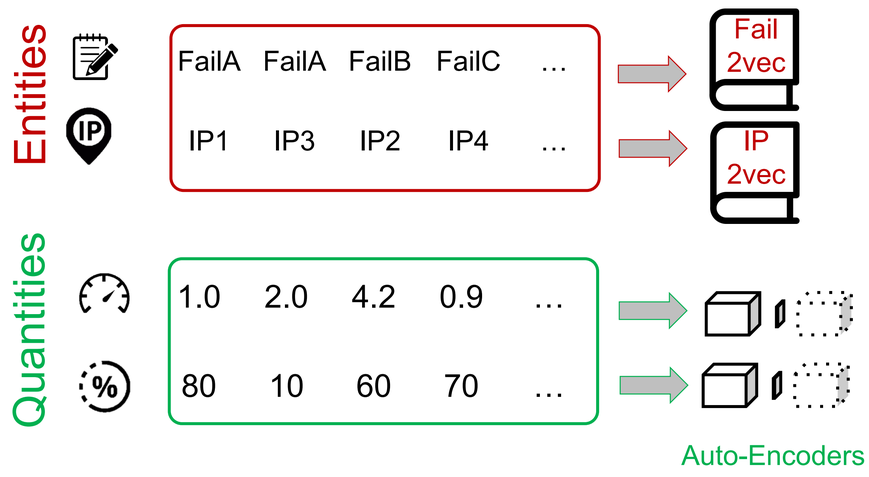
Methodology

Exploit the systematic bi-modality of network data

Pipeline

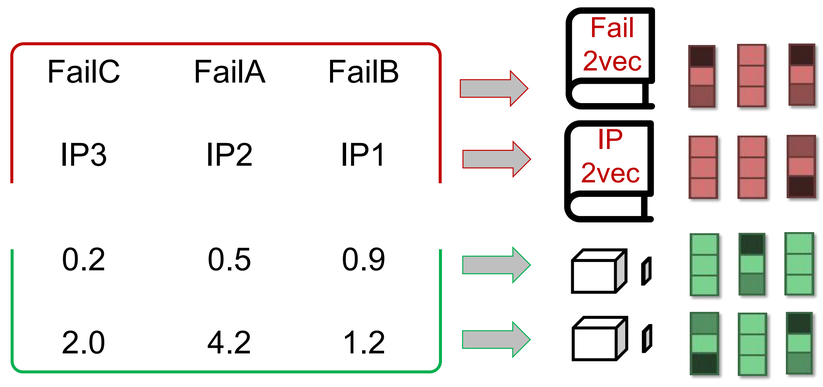
1. Pre-training

Leverage huge amounts of unlabeled data using Word2Vec-like embeddings



2. Sample selection

Define the input sample for the use case (e.g., flows)



3. Training

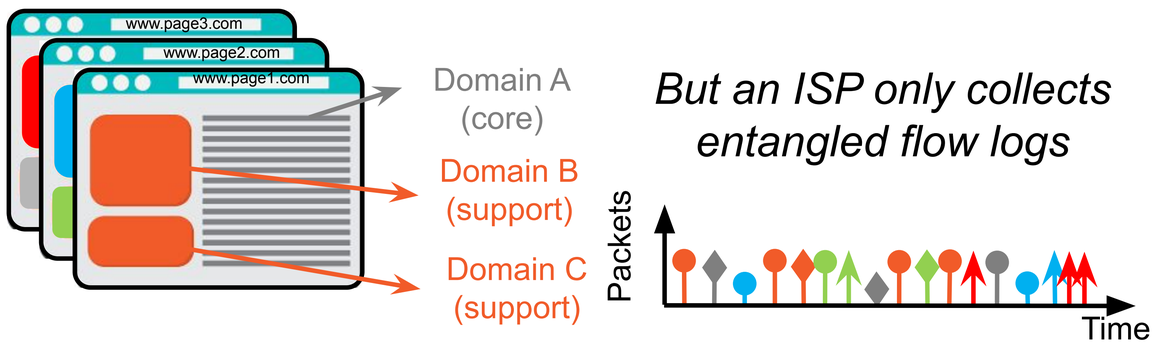
Using embedded samples, train an ML model for the downstream task

4. Inference

Predict and keep the model up-to-date

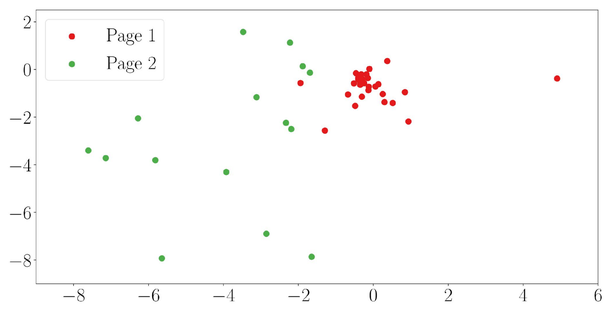
Use case 1: Clickstream identification

In web traffic, a single page downloads tens of objects, from tens of different locations.



But an ISP only collects entangled flow logs

***(i)** which entry belongs to which page?*



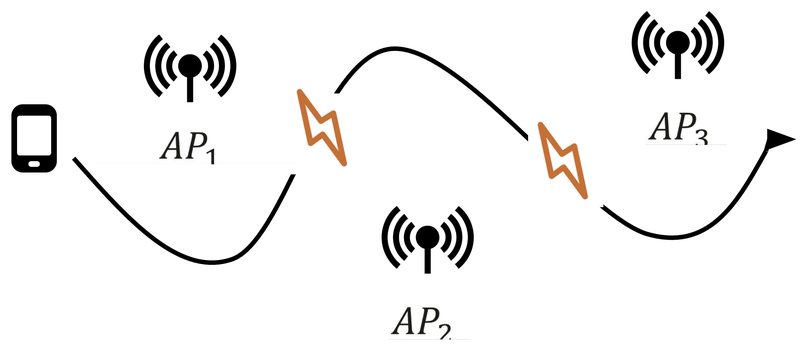
Easy separation of domains in embedded space

***(ii)** which flow is core and which is support domain?*

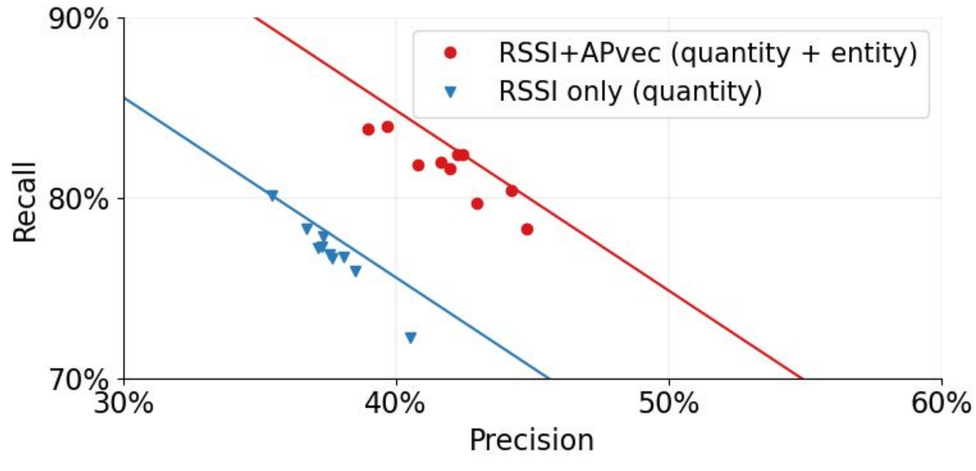
Approach	Precision			Recall		
	Q1	Q2	Q3	Q1	Q2	Q3
Quantities	66%	75%	100%	50%	60%	75%
Domain2Vec	75%	100%	100%	33%	50%	66%
Domain2Vec + Char emb.	80%	100%	100%	58%	77%	87%

Use case 2: WLAN movement prediction

A Wireless LAN deployment typically involves several access points (AP).



One problem is to predict whether a terminal is going to move from its AP, or stay (binary classification).



[1] Visualizing and Understanding Convolutional Networks [Zeiler & Fergus, 2013]
[2] Image from Word2Vec Explained [Gilyadov, 2017]