Multimodal representation and learning

ir. Shah Nawaz
University of Insubria - Varese, Italy

Outline

- Motivation
- Cross modal retrieval - text and image
  - Semantic text encoding
- Cross modal verification - audio and image
  - Test protocols
- Future research directions

Motivation

- Deep learning has remarkably improved the state-of-the-art in speech recognition, visual object recognition, object detection, and text processing
- Majority of these techniques focused on unimodality
- Real-world scenario presents data in a multimodal fashion
  - We see objects, listen sounds, feel texture, smell odors, and taste flavor
- Therefore, it is important to perform multimodal learning to understand the web and the world around us

Multimodal Examples

(a) Useful accessory for those who ride a **bike**. Size 46-52.

(b) The First **Bike** Pink Arrow dedicated to little girls.

(c) Telescopic **ladder** to partial or total opening. Ideal for any external intervention.

(d) Custom multifunction dynamic construction **scaffolding**, simple for decoration.

- In the top row, an example of ambiguous text descriptions that can be disambiguated with the analysis of the accompanying images
- In the bottom row, an examples of ambiguous images that can be disambiguated with the analysis of the associated text descriptions

Ignazio Gallo, Alessandro Calefati, **Shah Nawaz**. “Multimodal Classification Fusion in Real-World Scenarios”. ICDAR 2017

Multimodal Examples

- An example of multimodal tweets. In this tweet, "**Rocky**" is the name of the dog

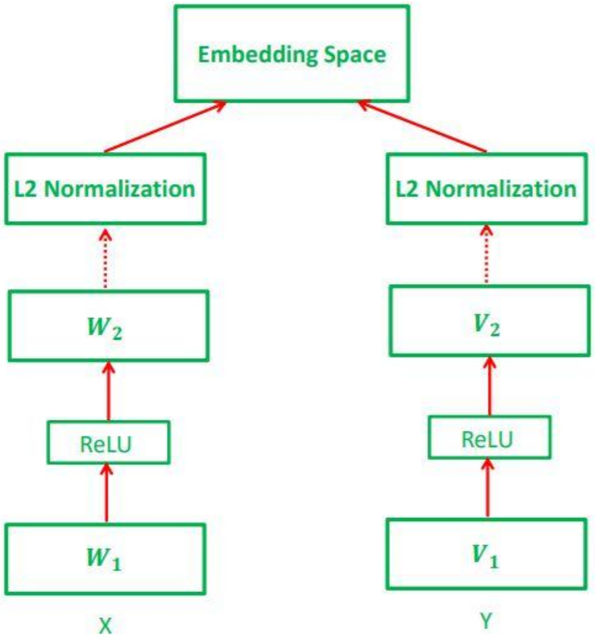
Representation - Feature Extraction

- Traditional representation or feature extraction
  - Histogram of oriented gradients
  - Scale-invariant feature transform
- Convolutional Neural Network
  - Produce state-of-the-art features

Multimodal Applications

- Numerous applications
  - **Cross-modal retrieval**
  - **Cross-modal verification**
  - **Classification**
  - Visual question answer
  - Semantic relatedness
  - Multimodal named entity recognition

Cross Modal Retrieval (Image-Text)

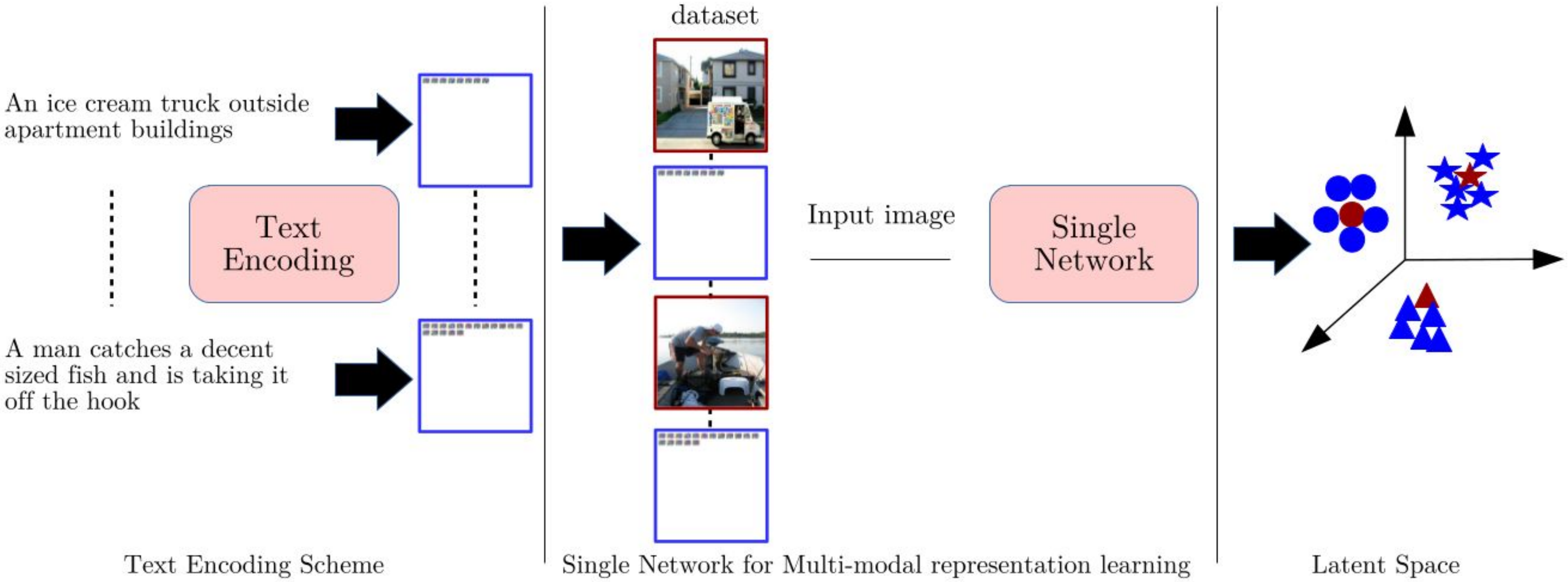


Two Branch Network

- Two modalities are often encoded separately
  - Image branch
  - Text branch
- A loss function minimizes the gap between image and text descriptions
- Evaluation on Recall-at-K(R@K)

Liwei Wang, Yin Li and Svetlana Lazebnik. “Learning Deep Structure-Preserving Image-Text Embeddings”. CVPR 2016

Cross Modal Retrieval (Image-Text)



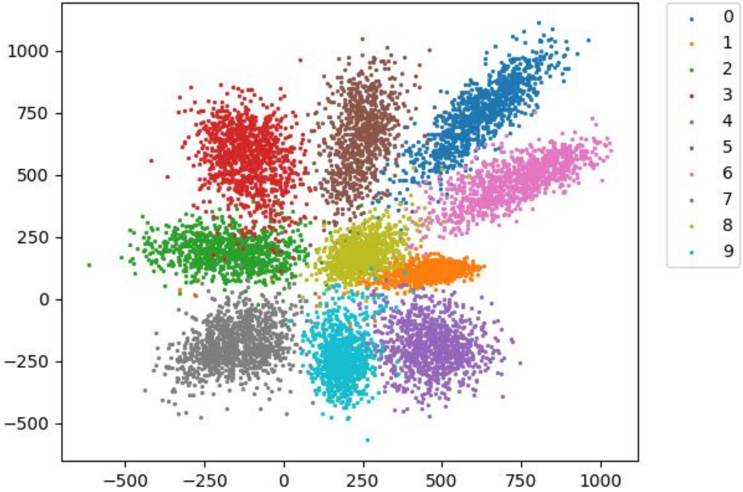
Shah Nawaz, Muhammad Kamran Janjua, Alessandro Calefati, Ignazio Gallo, Arif Mahmood and Faisal Shafait. "Do Cross Modal Systems Leverage Semantic Relationships?". **Submitted:** ACMM- 2019

Cross Modal Retrieval - Mod Center Loss

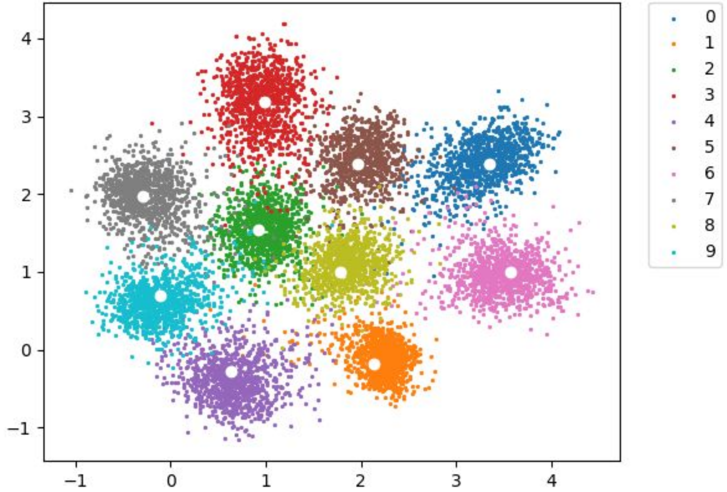
$$\mathcal{L}_S = -\sum_{i=1}^{m} \log \frac{e^{W_{y_i}^T \boldsymbol{x}_i + b_{y_i}}}{\sum_{j=1}^{n} e^{W_j^T \boldsymbol{x}_i + b_j}}$$

$$-\sum_{i=1}^{m} \log \frac{e^{W_{y_i}^T \boldsymbol{x}_i + b_{y_i}}}{\sum_{j=1}^{n} e^{W_j^T \boldsymbol{x}_i + b_j}} + \frac{\lambda}{2} \sum_{i=1}^{m} \|\boldsymbol{x}_i - \boldsymbol{c}_{y_i}\|_2^2$$

Softmax Loss



Softmax Loss + Center Loss



Cross Modal Retrieval (Image-Text)

	MSCOCO						Flickr30K					
Model	Image-to-Text			Text-to-Image			Image-to-Text			Text-to-Image		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
DVSA (2015)	38.4	69.9	80.5	27.4	60.2	74.8	–	–	–	–	–	–
HM-LSTM (2017)	43.9	-	87.8	36.1	-	86.7	–	–	–	–	–	–
m-RNN-vgg (2015)	41.0	73.0	83.5	29.0	42.2	77.0	35.4	63.8	73.7	22.8	50.7	63.1
Order-embedding (2015)	46.7	-	88.9	37.9	-	85.9	–	–	–	–	–	–
m-CNN(ensemble) (2015)	42.8	73.1	84.1	32.6	68.6	82.8	33.6	64.1	74.9	26.2	56.3	69.6
Str. Pres. (2016)	50.1	79.7	89.2	39.6	75.2	86.9	40.3	68.9	79.9	29.7	60.1	72.1
Two-Way (2017)	–	–	–	–	–	–	49.8	67.5	-	36.0	55.6	-
TextCNN (2016)	13.6	39.6	54.6	10.3	35.5	55.5	–	–	–	–	–	–
FV-HGLMM (2016)	14.3	40.5	55.8	12.7	39.0	57.2	–	–	–	–	–	–
Our Work (*cfg-std*)	13.2	30.4	41.9	12.2	33.0	46.7	10.5	26.2	36.8	8.2	22.82	32.0
Our Work (*cfg-2*)	13.0	32.9	46.0	12.94	36.62	49.94	21.2	40.7	50.4	19.36	38.12	47.34
Our Work (*cfg-3*)	32.9	58.5	70.2	26.22	56.7	69.96	–	–	–	–	–	–

Shah Nawaz, Muhammad Kamran Janjua, Alessandro Calefati, Ignazio Gallo, Arif Mahmood and Faisal Shafait. “Do Cross Modal Systems Leverage Semantic Relationships?”. **Submitted:** ACMM- 2019

Cross Modal Retrieval (Image-Text)

- R@K inadanacy

Query Image	Retrieved Text	λ
	– A man riding a skateboard at a skatepark	0.82
	– A man riding a skateboard down a ramp	0.81
	– A person riding a skateboard at a skatepark	0.78
	– A skateboarder trying to get up a ramp	0.78
	– A man riding a skateboard over an obstacle	0.77

Query Text: A man is playing tennis on the tennis court.

0.82 0.75 0.69 0.68 0.64

Shah Nawaz, Muhammad Kamran Janjua, Alessandro Calefati, Ignazio Gallo, Arif Mahmood and Faisal Shafait. "Do Cross Modal Systems Leverage Semantic Relationships?". **Submitted:** ACMM- 2019

Cross Modal Retrieval (Image-Text)

- Similarity Index

$$\lambda(X_l, Y_m) = \frac{\sum_{i=1}^{n} x_i^l y_i^m}{\sqrt{\sum_{i=1}^{n} (x_i^l)^2} \sqrt{\sum_{i=1}^{n} (y_i^m)^2}}; \forall X_l, Y_m$$

$X_l = (x_1^l, x_2^l, x_3^l, ..., x_n^l) \in \mathcal{R}^n$ belonging to l^{th} class of images

$Y_m = (y_1^m, y_2^m, y_3^m, ..., y_n^m) \in \mathcal{R}^n$ belonging to m^{th} class of text descriptions

- Semantic Map

$$\lambda@K = \frac{1}{N \cdot K} \sum_{l=1}^{c} \sum_{m=1}^{K} \lambda(X_l, Y_m)$$

Shah Nawaz, Muhammad Kamran Janjua, Alessandro Calefati, Ignazio Gallo, Arif Mahmood and Faisal Shafait. "Do Cross Modal Systems Leverage Semantic Relationships?". **Submitted:** ACMM- 2019

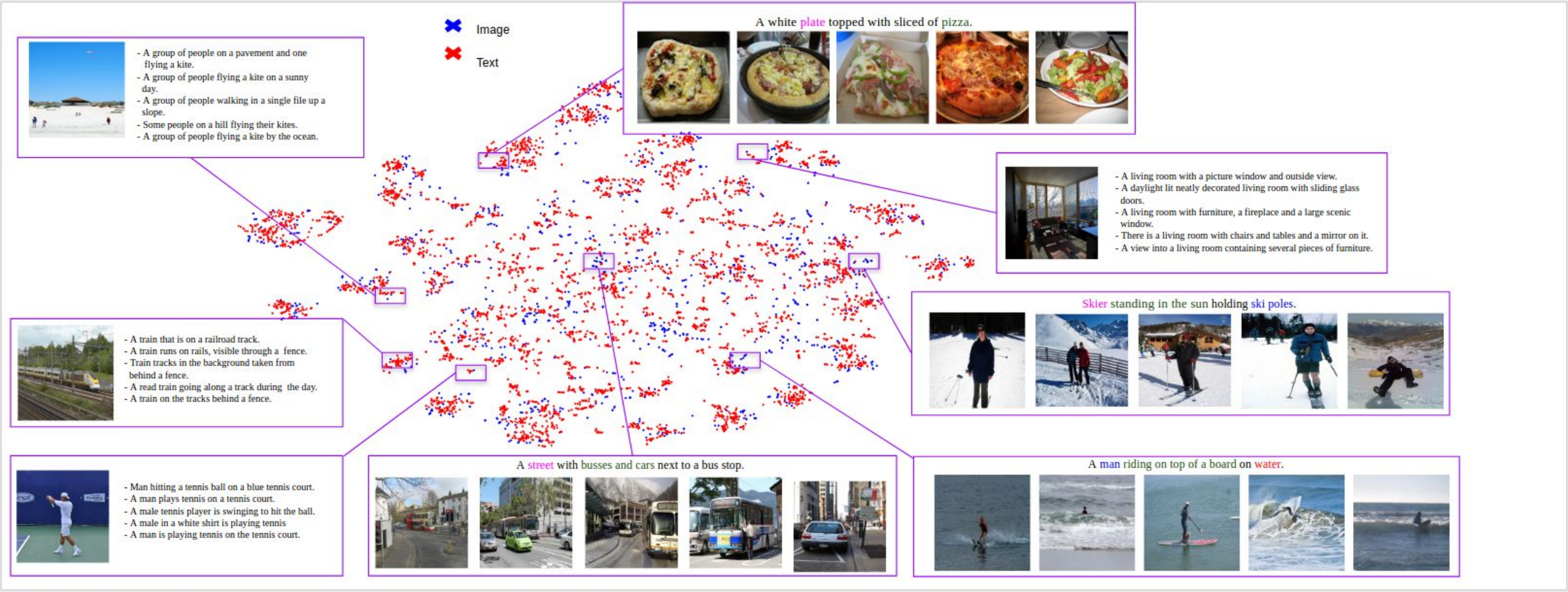
Cross Modal Retrieval (Image-Text)

	MSCOCO						Flickr30K					
Model	Image-to-Text			Text-to-Image			Image-to-Text			Text-to-Image		
	$\lambda@1$	$\lambda@5$	$\lambda@10$	$\lambda@1$	$\lambda@5$	$\lambda@10$	$\lambda@1$	$\lambda@5$	$\lambda@10$	$\lambda@1$	$\lambda@5$	$\lambda@10$
Str. Pres.	67.24	64.63	62.74	64.07	59.29	56.30	62.30	59.59	57.79	59.05	54.60	51.97
Our Work	68.67	65.25	62.86	66.70	59.42	54.45	49.13	45.52	43.33	43.97	39.22	36.43

Model	Image-to-Text			Text-to-Image		
	$\lambda@1$	$\lambda@5$	$\lambda@10$	$\lambda@1$	$\lambda@5$	$\lambda@10$
Str. Pres.	60.68	58.16	56.54	57.53	53.66	51.28
Our Work	**67.57**	**64.17**	**61.81**	**65.42**	**58.36**	**53.55**

Shah Nawaz, Muhammad Kamran Janjua, Alessandro Calefati, Ignazio Gallo, Arif Mahmood and Faisal Shafait. “Do Cross Modal Systems Leverage Semantic Relationships?”. **Submitted:** ACMM- 2019

Cross Modal Retrieval (Image-Text)



Shah Nawaz, Muhammad Kamran Janjua, Alessandro Calefati, Ignazio Gallo, Arif Mahmood and Faisal Shafait. “Do Cross Modal Systems Leverage Semantic Relationships?”. **Submitted:** ACMM- 2019

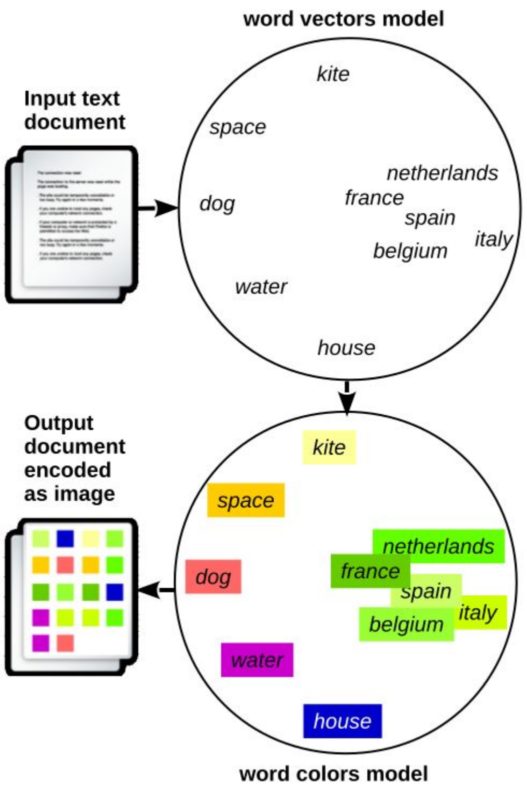
Semantically Reasonable Results

Human Experts	Image-to-Sentence
Expert # 1	85.40
Expert # 2	85.00
Expert # 3	83.40
$R@K$	84.6

Shah Nawaz, Muhammad Kamran Janjua, Alessandro Calefati, Ignazio Gallo, Arif Mahmood and Faisal Shafait. "Do Cross Modal Systems Leverage Semantic Relationships?". **Submitted:** ACMM- 2019

Semantic Text Encoding

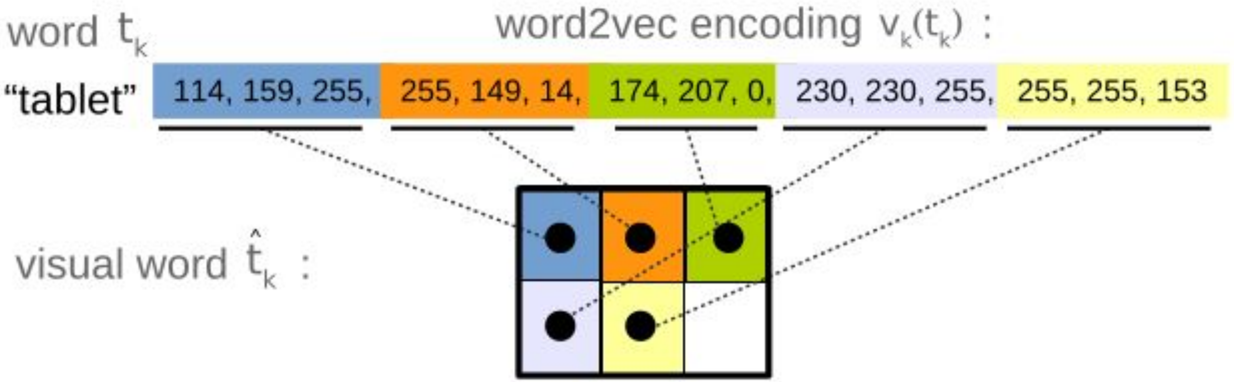
- ❑ Transform text embedding into an encoded image
- ❑ Based on **Word2Vec word embedding**
- ❑ Words that occur in **similar contexts** have **similar word embeddings**
- ❑ Intuitively, **similar words** will have **similar color encodings**



Ignazio Gallo, **Shah Nawaz**, Alessandro Calefati. "Semantic Text Encoding for Text Classification using Convolutional Neural Networks". ICDAR 2017.
Shah Nawaz, Alessandro Calefati, Muhammad Kamran Janjua, Ignazio Gallo. "Visual Word Embedding for Text Classification". **Submitted:** MVA 2019.

Semantic Text Encoding

- Toy example



Ignazio Gallo, **Shah Nawaz**, Alessandro Calefati. "Semantic Text Encoding for Text Classification using Convolutional Neural Networks". ICDAR 2017.
Shah Nawaz, Alessandro Calefati, Muhammad Kamran Janjua, Ignazio Gallo. "Visual Word Embedding for Text Classification". **Submitted:** MVA 2019.

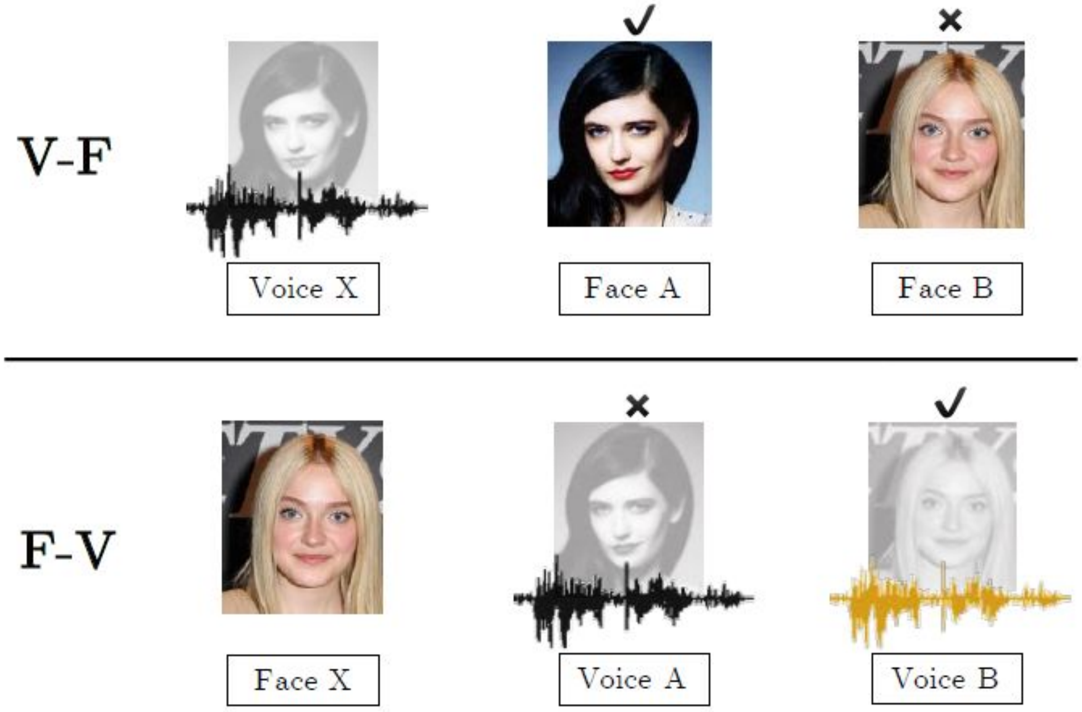
Semantic Text Encoding

- Comparison with state-of-the-art text classification methods

Model	AG	Sogou	DBP.	Yelp P.	Yelp F.	Yah. A.	Amz. F.	Amz. P.
Xiao *et al.*	8.64	4.83	1.43	5.51	38.18	28.26	40.77	5.87
Zhang *et al.*	**7.64**	**2.81**	1.31	4.36	37.95	28.80	40.43	4.93
Conneau *et al.*	8.67	3.18	1.29	**4.28**	**35.28**	26.57	**37.00**	4.28
Encoding scheme + AlexNet	9.19	8.02	1.36	11.55	49.00	25.00	43.75	3.12
Encoding scheme + GoogleNet	7.98	6.12	**1.07**	9.55	43.55	**24.10**	40.35	**3.01**

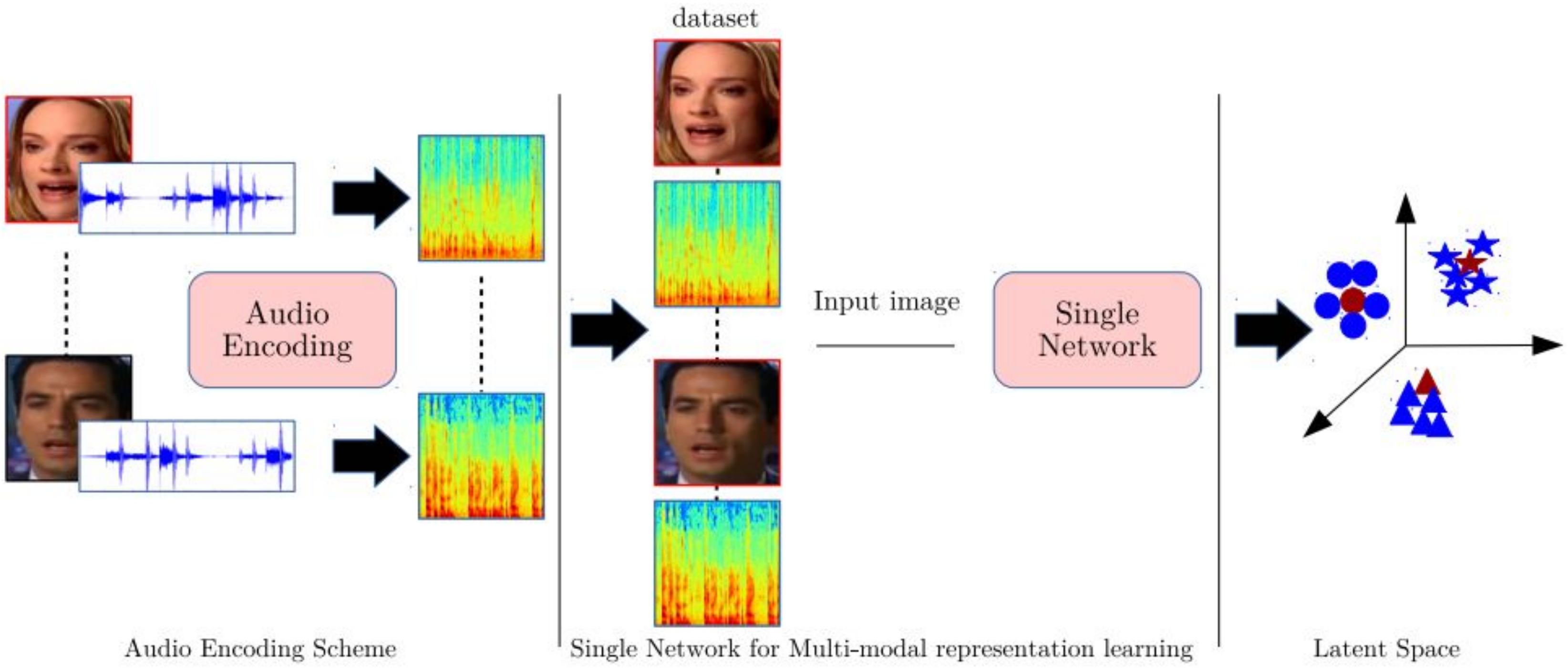
Ignazio Gallo, **Shah Nawaz**, Alessandro Calefati. "Semantic Text Encoding for Text Classification using Convolutional Neural Networks". ICDAR 2017.
Shah Nawaz, Alessandro Calefati, Muhammad Kamran Janjua, Ignazio Gallo. “Visual Word Embedding for Text Classification”. **Submitted:** MVA 2019.

Cross Modal (Audio - Image)



Arsha Nagrani, Samuel Albanie, Andrew Zisserman. “Seeing Voices and Hearing Faces: Cross-modal biometric matching”. CVPR 2018.

Cross Modal (Audio - Image)



Shah Nawaz, Muhammad Kamran Janjua, Ignazio Gallo, Arif Mahmood and Alessandro Calefati, "Deep Latent Space for Cross Modal Mappings of Audio and Visual Signals". **Submitted:** BMVC 2019.

Cross Modal (Audio - Image)

- Two test protocols
  - Seen - Heard
  - Unseen - Unheard
- Verification
  - Gender, Age and Nationality
- Force Matching

Shah Nawaz, Muhammad Kamran Janjua, Ignazio Gallo, Arif Mahmood and Alessandro Calefati, "Deep Latent Space for Cross Modal Mappings of Audio and Visual Signals". **Submitted:** BMVC 2019.

Cross Modal (Audio - Image)

- Cross Modal Verification

	AUC %	EER %
	Seen-Heard	
Learnable Pins	73.8	34.1
SSNet	**89.2**	**19.0**
	Un-seen-Un-heard	
Learnable Pins	63.5	39.2
SSNet	**71.8**	**34.2**

Shah Nawaz, Muhammad Kamran Janjua, Ignazio Gallo, Arif Mahmood and Alessandro Calefati, "Deep Latent Space for Cross Modal Mappings of Audio and Visual Signals". **Submitted:** BMVC 2019.

Cross Modal (Audio - Image)

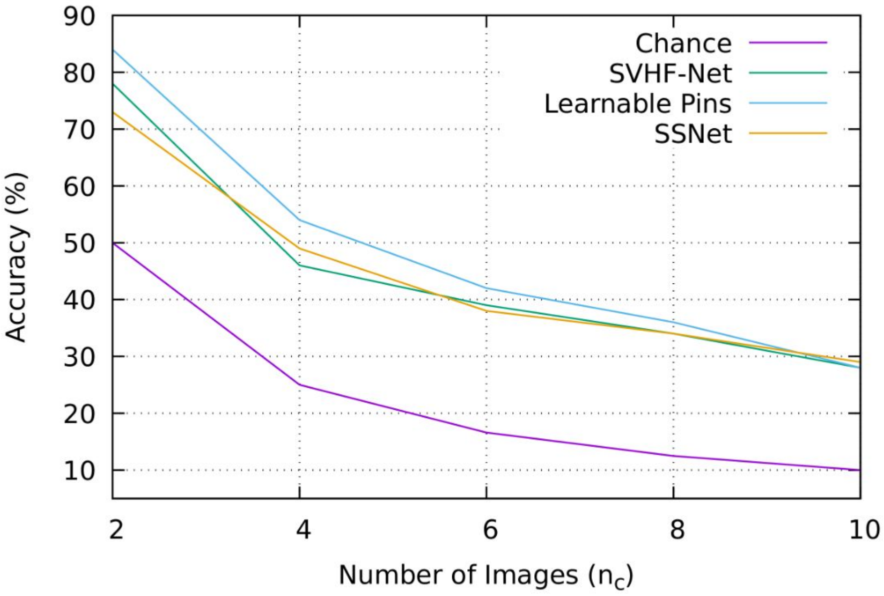
- Demographic Criteria - Gender, Age, Nationality

Demographic Criteria	Random	G	N	A	GNA
	Seen-Heard (AUC %)				
Learnable Pins-[Scratch]	73.8	-	-	-	-
Learnable Pin-[Pre-train]	87.0	74.2	85.9	86.6	74.0
SSNet-[Scratch]	**89.2**	**82.7**	**88.6**	**89.2**	**82.4**
	Unseen-Unheard (AUC %)				
Learnable Pins-[Scratch]	63.5	-	-	-	-
Learnable Pins-[Pre-train]	**78.5**	61.1	**77.2**	**74.9**	**58.8**
SSNet-[Scratch]	71.8	**61.9**	51.9	69.5	52.1

Shah Nawaz, Muhammad Kamran Janjua, Ignazio Gallo, Arif Mahmood and Alessandro Calefati, "Deep Latent Space for Cross Modal Mappings of Audio and Visual Signals". **Submitted:** BMVC 2019.

Cross Modal (Audio - Image)

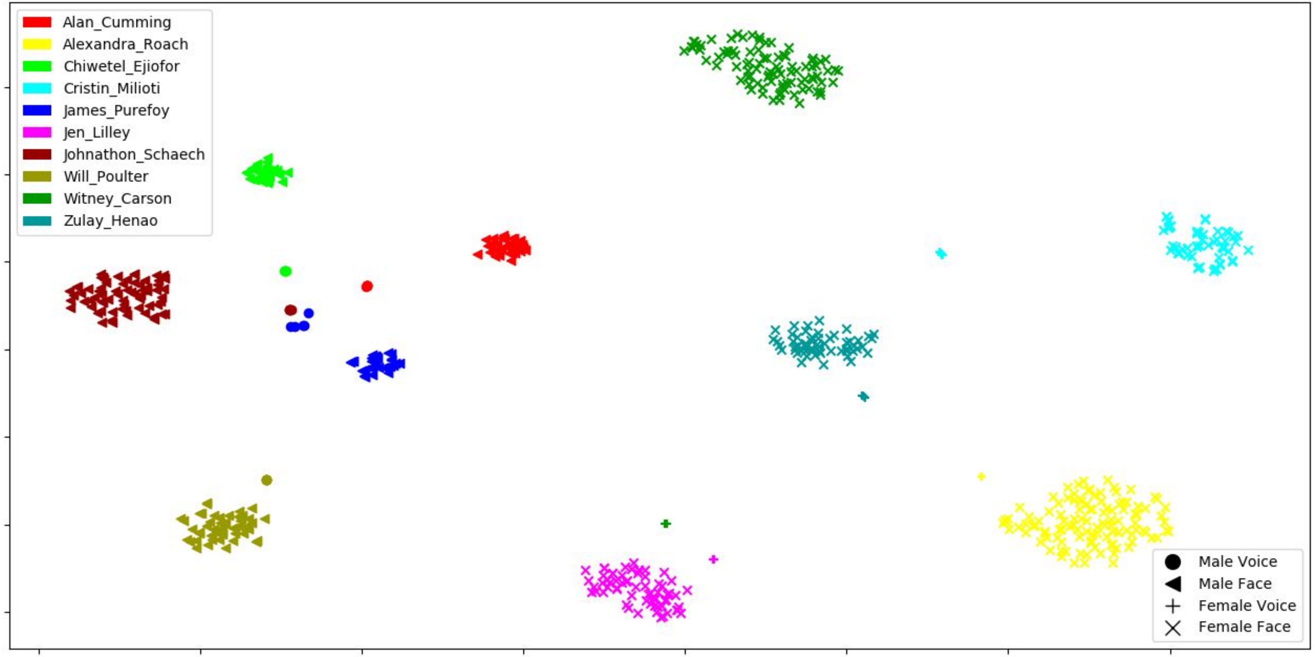
- Force matching



Shah Nawaz, Muhammad Kamran Janjua, Ignazio Gallo, Arif Mahmood and Alessandro Calefati, "Deep Latent Space for Cross Modal Mappings of Audio and Visual Signals". **Submitted:** BMVC 2019.

Cross Modal (Audio - Image)

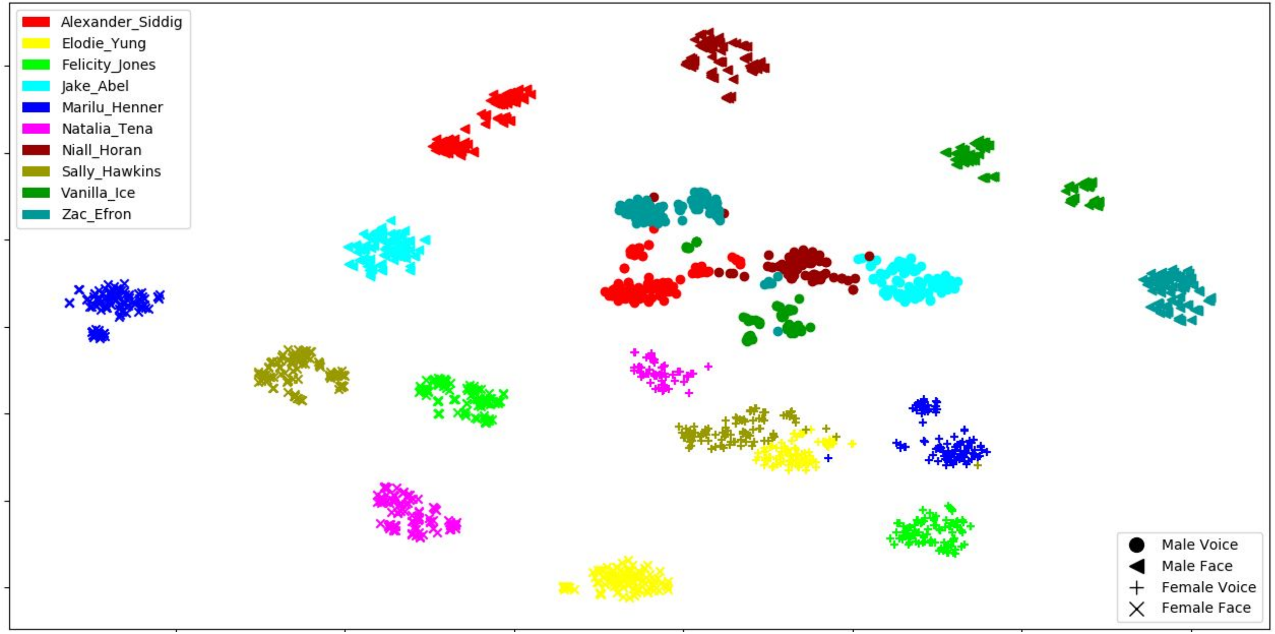
- T-sne Visualization - Seen - Heard



Shah Nawaz, Muhammad Kamran Janjua, Ignazio Gallo, Arif Mahmood and Alessandro Calefati, “Deep Latent Space for Cross Modal Mappings of Audio and Visual Signals”. **Submitted:** BMVC 2019.

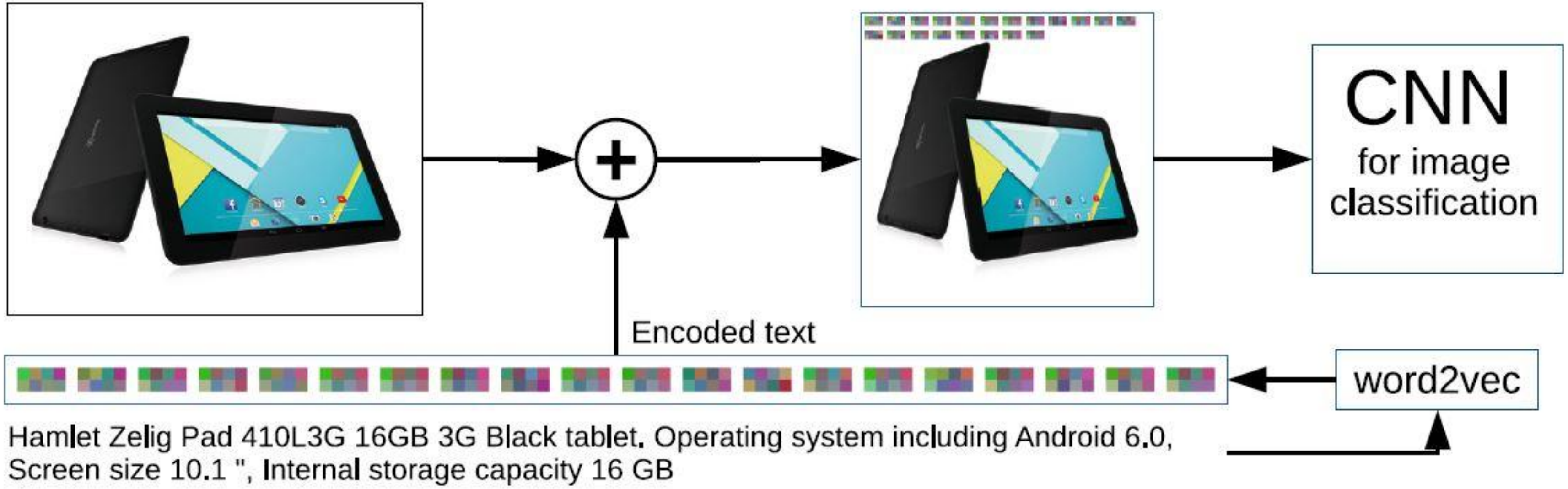
Cross Modal (Audio - Image)

❑ T-sne Visualization - Unseen - Unheard



Shah Nawaz, Muhammad Kamran Janjua, Ignazio Gallo, Arif Mahmood and Alessandro Calefati, "Deep Latent Space for Cross Modal Mappings of Audio and Visual Signals". **Submitted:** BMVC 2019.

Fused Representation for Classification



Shah Nawaz, Alessandro Calefati, Muhammad Kamran Janjua, Muhammad Umer An-waar, Ignazio Gallo. Learning Fused Representations for Large Scale Multimodal Classification. Accepted: IEEE Sensors Letters 2018.

Fused Representation for Classification

Image | Encoded Text | Fused representation

- ❑ Image only - 77.42% **Baby**, 11.16% **Home and Kitchen**, ...
- ❑ Fused representation - 100 **Baby**

Shah Nawaz, Alessandro Calefati, Muhammad Kamran Janjua, Muhammad Umer An-waar, Ignazio Gallo. Learning Fused Representations for Large Scale Multimodal Classification. Accepted: IEEE Sensors Letters 2018.

Fused Representation for Classification

	Model	Ferramenta	UPMC Food-101	Amazon Product Data
	Wang et al	–	85.10	–
Previous work	Kiela et al	–	**90.8±0.1**	–
	Gallo et al	94.42	60.63	–
Baseline	Image	92.47	55.65	51.42
	Text	84.50	56.75	64.37
Ours	Proposed	**95.87**	85.69	**78.26**

Shah Nawaz, Alessandro Calefati, Muhammad Kamran Janjua, Muhammad Umer An-waar, Ignazio Gallo. Learning Fused Representations for Large Scale Multimodal Classification. Accepted: IEEE Sensors Letters 2018.

Current Research Directions

- Multimodal Named Entity Recognition

My daughter got 1 place in [Apple valley **LOC**] Tags gymnastics

[Apple **ORG**] 's latest [iOS **OTHERS**] update is bad for advertisers

Questions

Multimodal Representation

- ❑ Unimodal representation
  - ❑ Images, text, audio etc.
- ❑ Joint representation
- ❑ Challenges
  - ❑ Different representation of each modality
  - ❑ Different correlation structure
- ❑ State-of-the-art representation