Deep Multimodal and Cross-modal Embeddings

DSPT #80 Webinar

David Semedo, Ph.D.   df.semedo@campus.fct.unl.pt

Universidade NOVA de Lisboa, Portugal

NOVALINCS LABORATORY FOR COMPUTER SCIENCE AND INFORMATICS

FACULDADE DE CIÊNCIAS E TECNOLOGIA UNIVERSIDADE NOVA DE LISBOA

About me and my research

- Researcher at the Web and Media Search group, from NOVA LINCS since 2015.
- Ph.D. in Deep Learning for Multimedia Understanding.
- Topic:
  - *"Bridging Vision and Language over Time with Neural Cross-modal Embeddings"*.
- Main interests: multimodal machine learning, at the intersection of CV and NLP, neural networks and data mining.

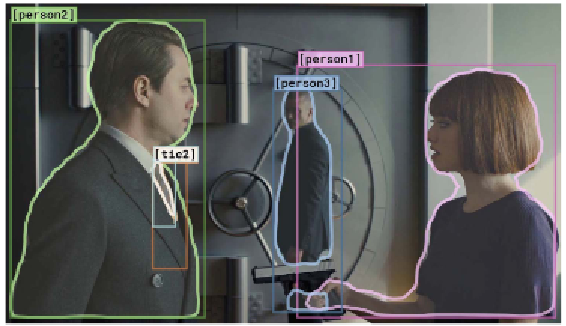
Webinar outline

- Motivation
- Learning deep multimodal and cross-modal embeddings
- State-of-the-art and Applications
- Final remarks

The world is multimodal!

Vision and Language

Visual Commonsense Reasoning: Answer a question about an image and provide a rationale justifying the answer.



Why is [person1] pointing a gun at [person2]?

a) [person1] wants to kill [person2].(1%)

b) [person1] and [person3] are robbing the bank and [person2] is the bank manager. (71%)

c) [person2] has done something to upset [person1]. (18%)

d) Because [person2] is [person1]'s daughter. [person1] wants to protect [person2]. (8%)

b) is right because...

a) [person1] is chasing [person1] and [person3] because they just robbed a bank. (33%)

b) Robbers will sometimes hold their gun in the air to get everyone's attention. (5%)

c) The vault in the background is similar to a bank vault. [person3] is waiting by the vault for someone to open it. (49%)

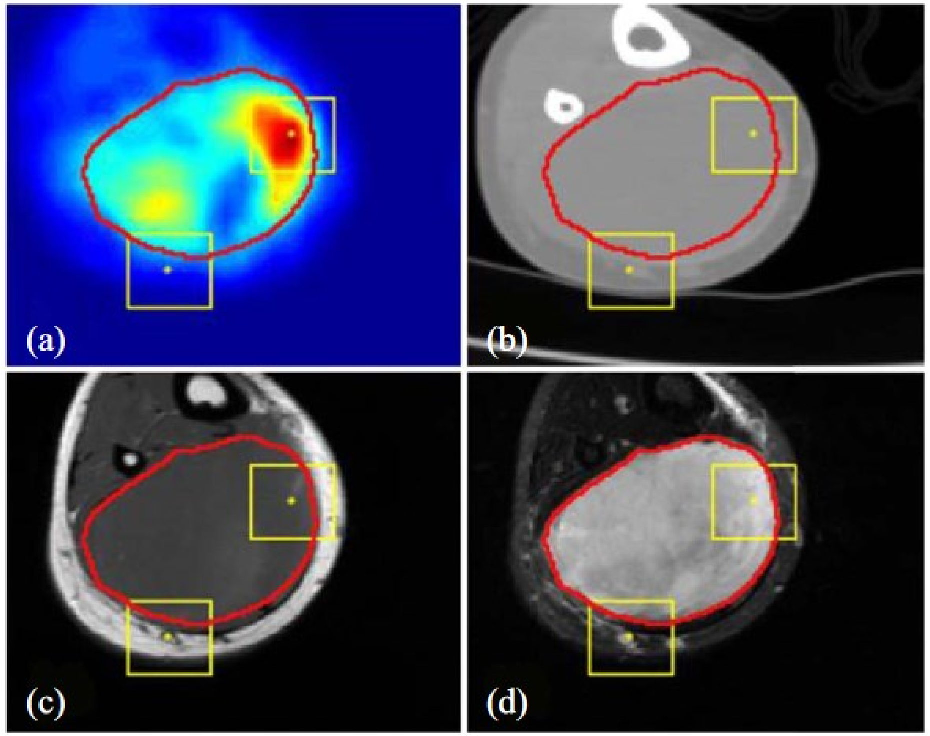
d) A room with barred windows and a counter usually resembles a bank. (11%)

Zellers, Rowan et al. **From Recognition to Cognition: Visual Commonsense Reasoning**. CVPR 2019.

The world is multimodal!

Different diagnostic exams

- Positron Emission Tomography (PET)
- Computer Tomography (CT)
- Magnetic Resonance Image (MRI)



Tumor segmentation

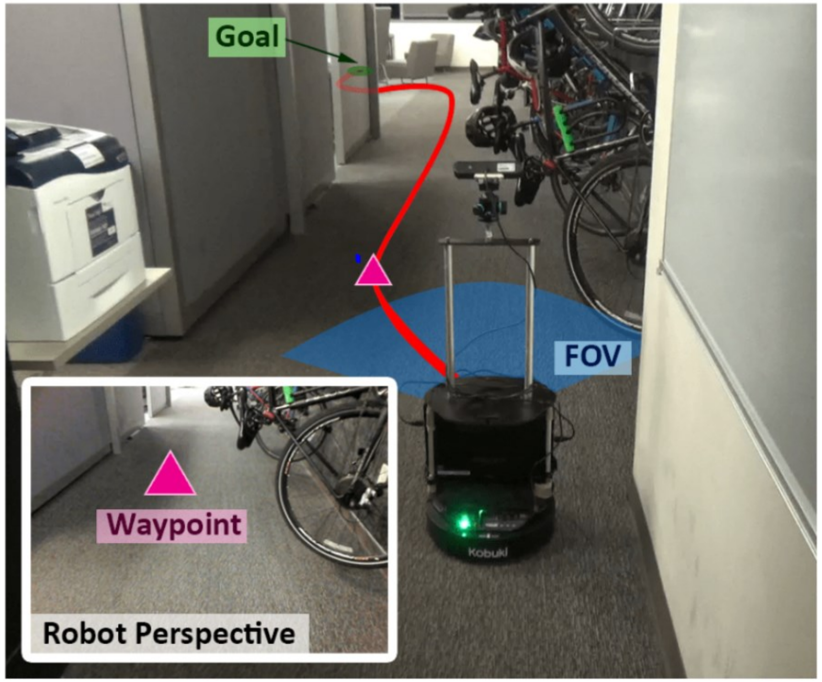
Z. Guo, et.al. **Deep Learning-Based Image Segmentation on Multimodal Medical Imaging**, *IEEE Transactions on Radiation and Plasma Medical Sciences*, 2019

NOVALINCS LABORATORY FOR COMPUTER SCIENCE AND INFORMATICS
FACULDADE DE CIÊNCIAS E TECNOLOGIA UNIVERSIDADE NOVA DE LISBOA

The world is multimodal!

Autonomous Navigation

- Vision (e.g. RGB, Infrared Cameras)
- Sensors (e.g. proximity, depth)
- Speech
- ...

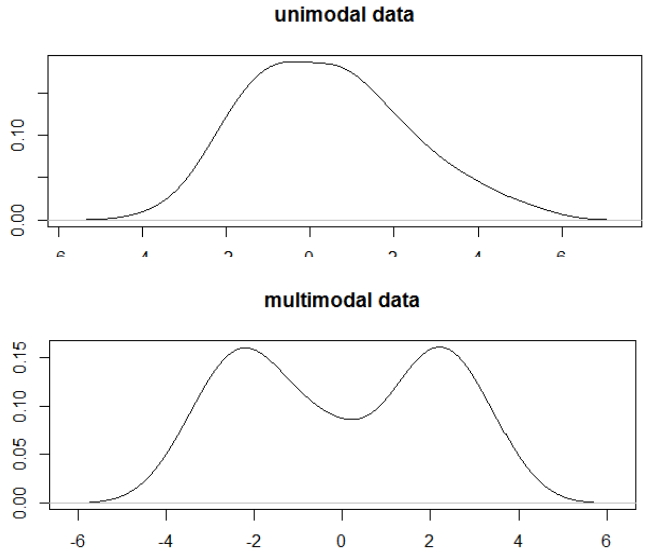


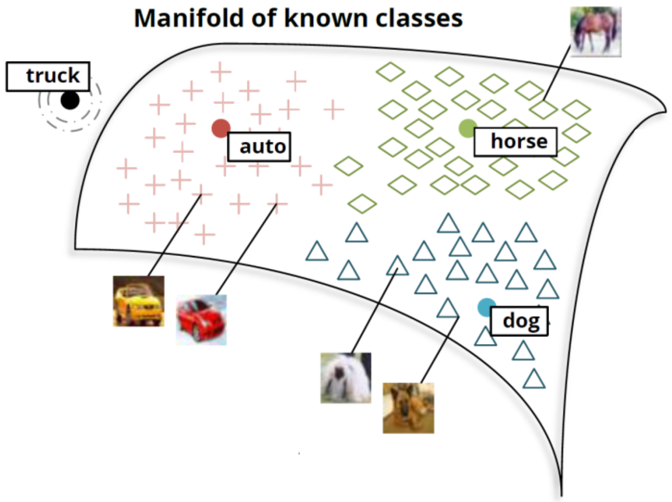
Bansal, Somil et al. **Combining Optimal Control and Learning for Visual Navigation in Novel Environments,** ArXiv abs/1903.02531, 2019.

Motivation - Definitions

Multimodal: “Having or involving several modes, modalities, or maxima”

Multimodal Embedding: Vector representation of a given document, in a multimodal space





Multiple modalities in ML pipelines

Why would we want to bridge two (or more) modalities?

✓ Combining information from multiple information sources!

Description:

“Deep Time light show at Edinburgh Festival, 2016”

NOVALINCS LABORATORY FOR COMPUTER SCIENCE AND INFORMATICS
FCT FACULDADE DE CIÊNCIAS E TECNOLOGIA UNIVERSIDADE NOVA DE LISBOA

Structured vs. Unstructured data

Tabular data – After vectorizing, each dimension potentially discriminates different samples.

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

Images (RGB - 3-channel width x height matrix)

Why not just concatenate representations?

Heterogeneous Representations

Textual Data

Deep Time light Show

Bag of Words, Word/Sentence Embeddings, ...

Visual Data

Colors (HoC), Gradients (HoG), Keypoints (SIFT, SURF), CNN-based, etc..

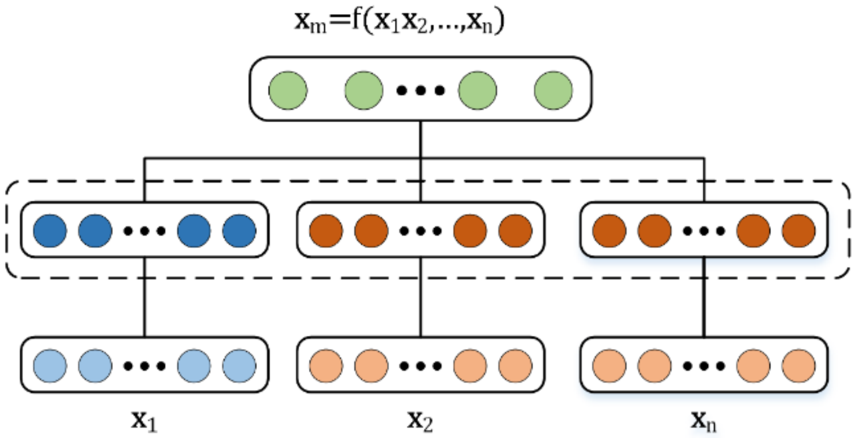
Early fusion (e.g. concatenation)

- Representations need to be well-aligned;
- Single model for both modalities;
- Limited capability to unveil complex correlations between modalities.

Multimodal Embeddings – Joint Representations

- Merges information from all input modalities in a single embedding after non-linear transformations (Late fusion);

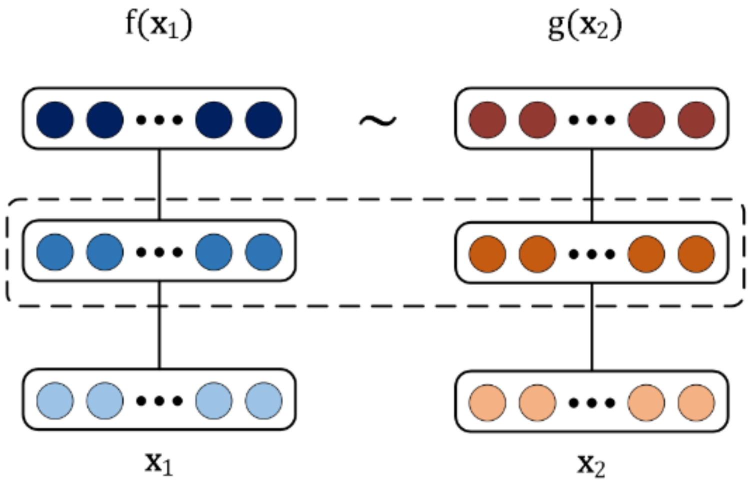
- Multimodal data present at training and inference time.



T. Baltrušaitis, C. Ahuja and L. Morency, **Multimodal Machine Learning: A Survey and Taxonomy,** in *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 2019.*

Multimodal Embeddings – Coordinated Representations

- Learn two separate spaces, that are constrained to be aligned:
    - Minimize distance metric, maximize correlation, etc..
- Independent projection functions - copes with missing modalities!



T. Baltrušaitis, C. Ahuja and L. Morency, **Multimodal Machine Learning: A Survey and Taxonomy,** in *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 2019.*

Linear Cross-modal Embeddings

Given a projection basis A for texts and B for images, project both modalities to correlated subspaces.

Data View 2

X_i

Deep Time light show

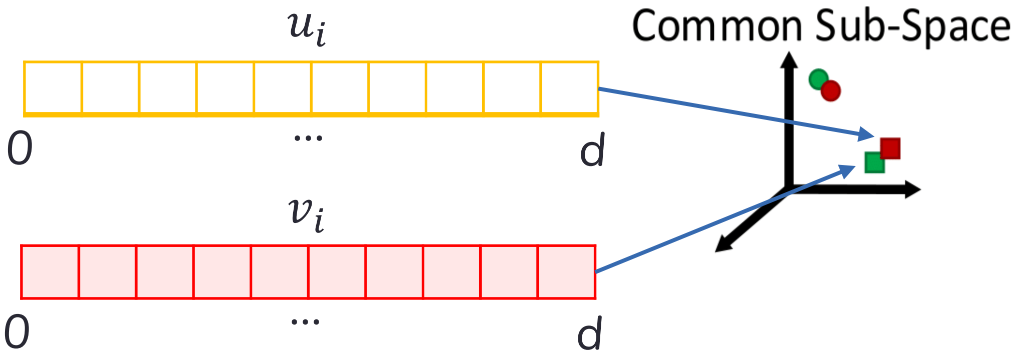
$$\begin{bmatrix} \vdots & \vdots & \vdots \\ a_{11} & & a_{1N} \\ \vdots & \vdots & \vdots \end{bmatrix}^T \cdot X_i = u_i$$

Data View 1

Y_i

$$\begin{bmatrix} \vdots & \vdots & \vdots \\ b_{11} & & b_{1N} \\ \vdots & \vdots & \vdots \end{bmatrix}^T \cdot Y_i = v_i$$

Projections of the document



Canonical Correlation Analysis

CCA objective (maximize):

$$\rho = \frac{E[uv]}{\sigma_u \sigma_v} = \frac{E[uv]}{\sqrt{E[u^2]E[v^2]}}$$

Rasiwasia , Nikhil et. al. **A new approach to cross-modal multimedia retrieval**. ACM Multimedia (MM ‘14).

Going beyond linear transformations

Do we know a *priori* *how* any two modalities are correlated?

We may have a hint, but not exactly
how to mathematically express that hint.

State-of-the-art:

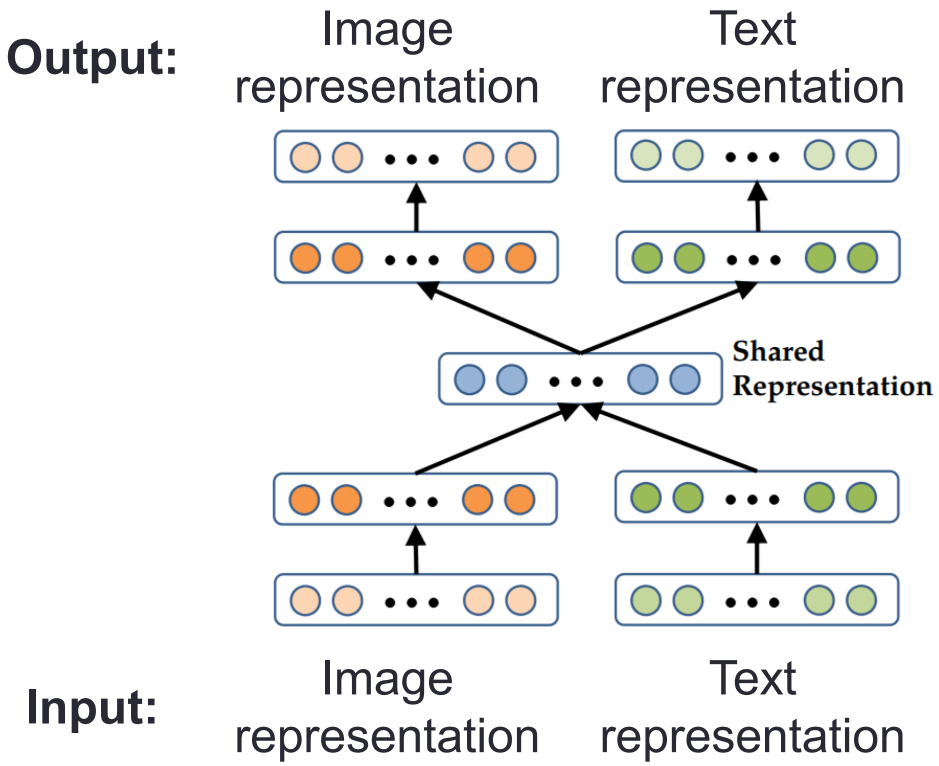
✓ Use neural networks to learn non-linear projection functions.

Universal Approximation Theorem:
We just need enough width (neurons) or depth (hidden layers)!

NOVALINCS LABORATORY FOR COMPUTER SCIENCE AND INFORMATICS
FCT FACULDADE DE CIÊNCIAS E TECNOLOGIA UNIVERSIDADE NOVA DE LISBOA

Intuitive explanation: http://neuralnetworksanddeeplearning.com/chap4.html, by Michael Nielsen.
Kidger, Patrick and Terry Lyons. **Universal Approximation with Deep Narrow Networks**, COLT 2020.

Multimodal Autoencoder



Loss: l_2 reconstruction error.

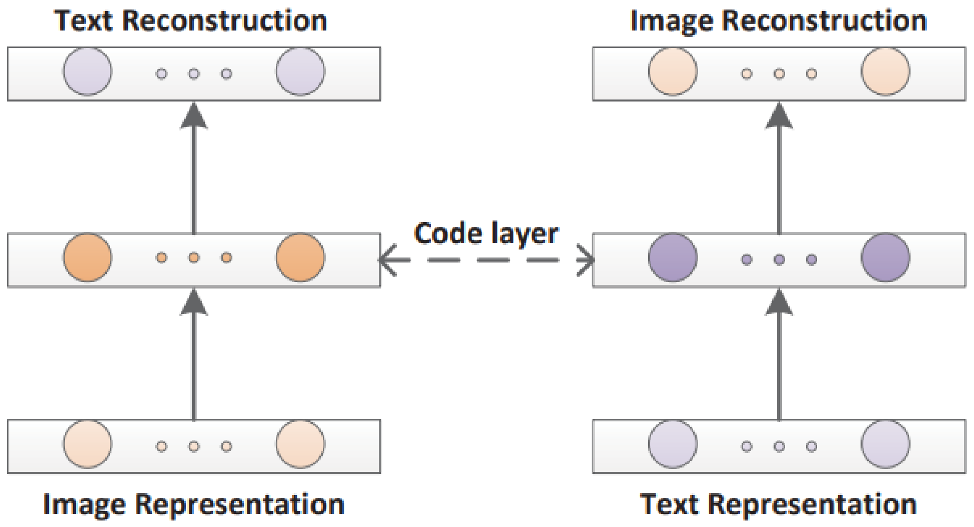
What can we do with such architecture/embedding?

- Extend it to more than 2 modalities;
- Multimodal queries;
- Use as features on a classifier.

Not a cross-modal embedding yet

Jiquan Ngiam, et.al. **Multimodal deep learning**. ICML 2011.

Correspondence Autoencoder



Loss: l_2 reconstruction error.

Same input reconstruction principle as in an Autoencoder, but ...

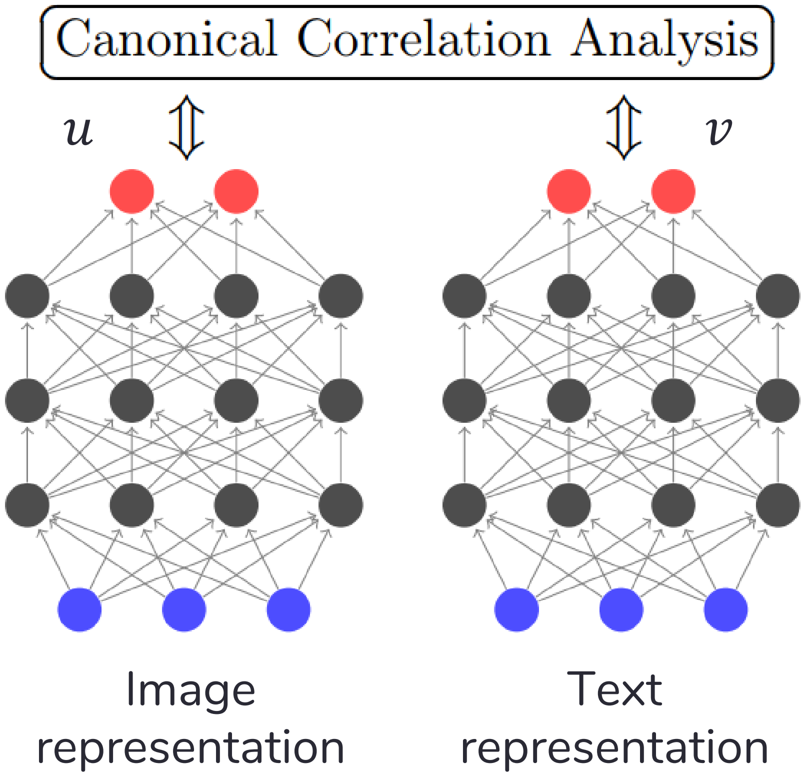
Weights of the Intermediate layer are shared.

This means that each modality network can be used independently!

Feng, Fangxiang, et.al. **Cross-modal Retrieval with Correspondence Autoencoder**. ACM Multimedia (MM '14).

Deep Canonical Correlation Analysis

Neural network projection function, with the CCA objective as loss function.



Learns a representation by **maximizing correlation.**

$$\rho = \frac{E[uv]}{\sigma_u \sigma_v} = \frac{E[uv]}{\sqrt{E[u^2]E[v^2]}}$$

Unsupervised approach, but it's quite an effective architecture!

Yan, Fei and Krystian Mikolajczyk. **Deep correlation for matching images and text.** CVPR 2015.

Going beyond pairwise-correlations

State-of-the-art:

✓ Use neural networks to learn **non-linear projection functions**.

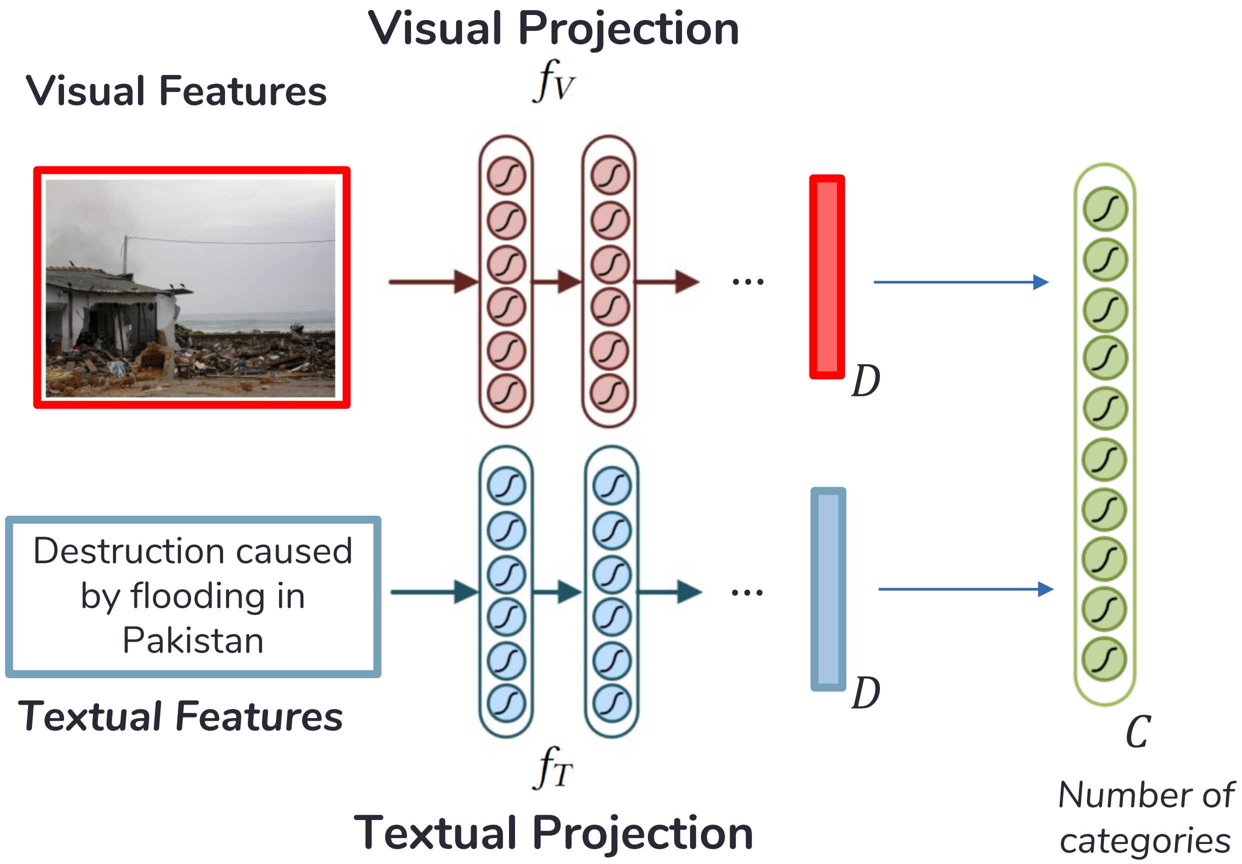
✓ **Supervised**: Consider category/class information to help structuring data.

Same class

Lights at the Castle!

Deep Time light show

Supervised Cross-modal Embeddings – Classifier as proxy



Multi-class setting:
softmax layer

$$\sigma(\vec{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$$

Loss: categorical cross-entropy

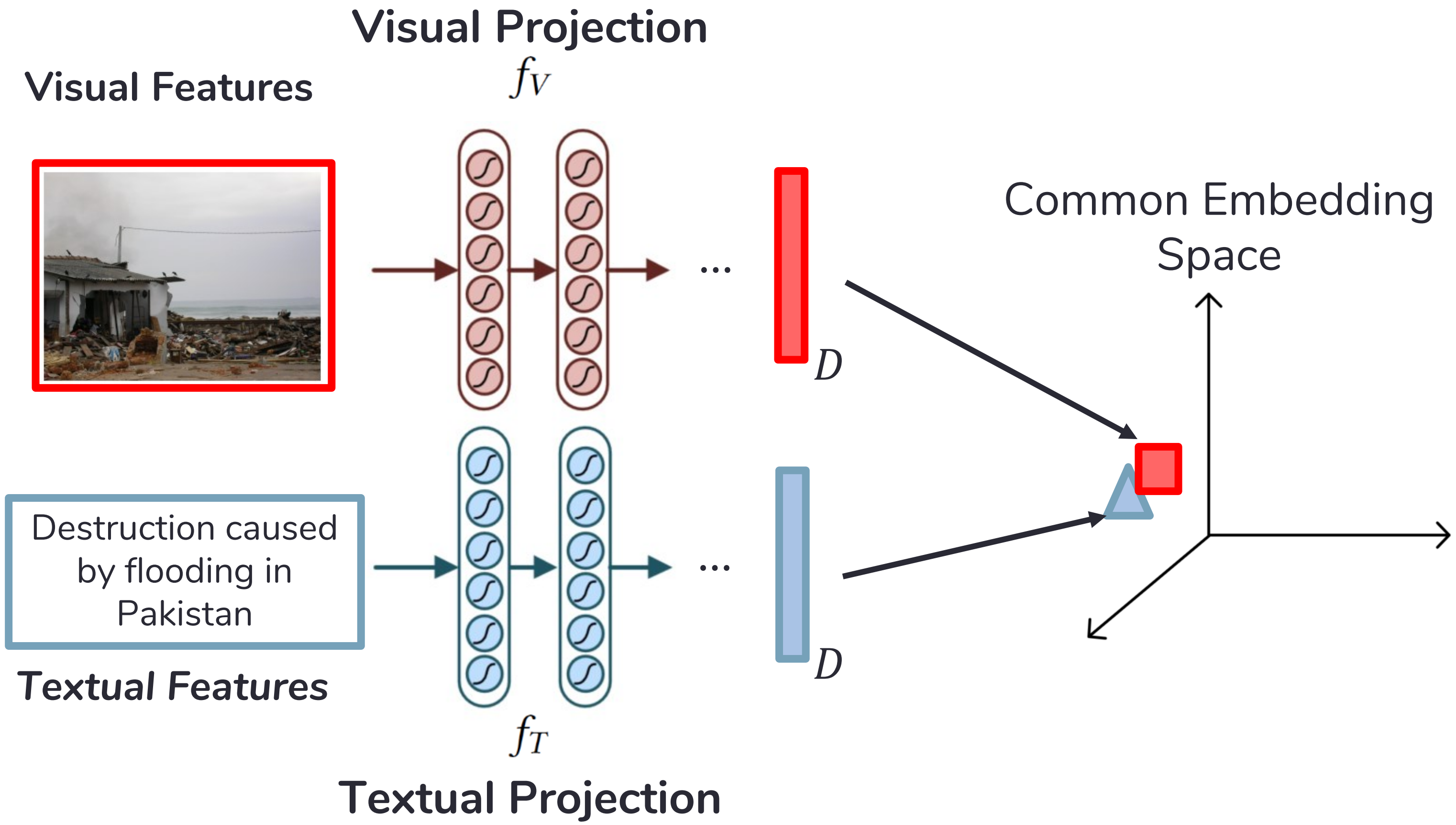
Multi-label setting:
logits layer (multiple regression functions)

Loss: binary cross-entropy

$$[y_n \cdot \log \sigma(x_n) + (1 - y_n) \cdot \log(1 - \sigma(x_n))]$$

sigmoid

Supervised Cross-modal Embeddings – Metric Learning



Loss: ranking loss

The loss directly imposes the structure we want on the embedding space.

Example of a Ranking loss - Triplet Loss

Directly enforces the desired embedding structure.

Same semantic class

anchor

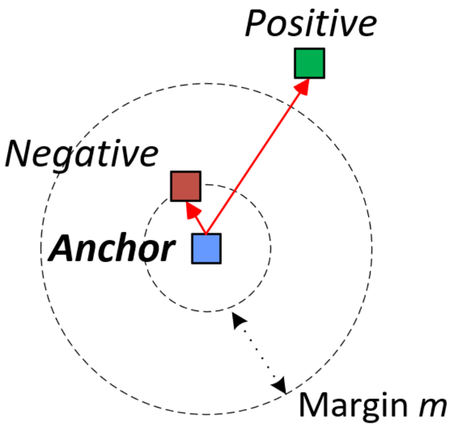
positive

Destruction caused by flooding in Pakistan

negative

People Dancing in the streets of Lisbon

Triplet loss: $s(\mathbf{x}_*^a, \mathbf{x}_*^p) > s(\mathbf{x}_*^a, \mathbf{x}_*^n) + m$



✗

✓

Florian Schroff, et.al. **FaceNet: A unified embedding for face recognition and clustering**. *CVPR, 2015.*

Example of a Ranking loss - Triplet Loss

$$s(\mathbf{x}^a_*, \mathbf{x}^p_*) > s(\mathbf{x}^a_*, \mathbf{x}^n_*) + \boxed{m}$$

Constraint is formulated as
a **differentiable function**:

$$loss(\mathbf{x}^a_*, \mathbf{x}^p_*, \mathbf{x}^n_*) = max(0, m - s(\mathbf{x}^a_*, \mathbf{x}^p_*) + s(\mathbf{x}^a_*, \mathbf{x}^n_*))$$

Then, **sample several triplets** and keep enforcing the constraint over those triplets:

Triplet 1:

anchor

positive

Destruction caused by flooding in Pakistan

negative

People Dancing in the streets of Lisbon

...

Triplet N:

anchor

"Deep Time light show at Edinburgh Festival, 2016"

positive

negative

$$loss = \underbrace{loss(\mathbf{x}^a_*, \mathbf{x}^p_*, \mathbf{x}^n_*)}_{\text{Triplet 1}} + \ldots + \underbrace{loss(\mathbf{x}^a_*, \mathbf{x}^p_*, \mathbf{x}^n_*)}_{\text{Triplet N}}$$

Adaptive Triplet Loss: scheduled optimization approach – increases expressiveness (ref. below).

Florian Schroff, et.al. **FaceNet: A unified embedding for face recognition and clustering**. CVPR, 2015.
David Semedo and Joao Magalhães. **Cross-Modal Subspace Learning with Scheduled Adaptive Margin Constraints**. ACM Multimedia (MM '19).

Some results on cross-modal retrieval

Results (mean Average Precision) on 3 benchmark datasets.

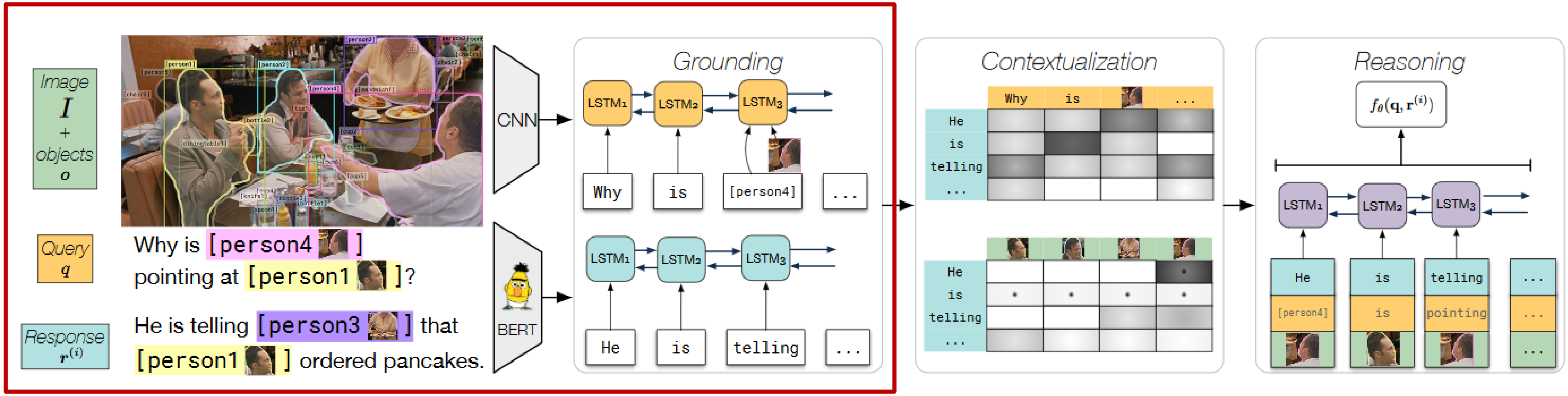
	Method	Pascal Sentences			NUS-WIDE-10k			Wikipedia		
		$I \mapsto T$	$T \mapsto I$	Avg	$I \mapsto T$	$T \mapsto I$	Avg	$I \mapsto T$	$T \mapsto I$	Avg
Linear	CCA	0.203	0.208	0.206	0.167	0.181	0.174	0.298	0.273	0.286
	CFA	0.476	0.470	0.473	0.406	0.435	0.421	0.319	0.316	0.318
	KCCA	0.488	0.446	0.467	0.351	0.356	0.354	0.438	0.389	0.414
	LGCFL	0.539	0.503	0.521	0.453	0.485	0.469	0.466	0.431	0.449
	JRL	0.563	0.505	0.534	0.466	0.499	0.483	0.479	0.428	0.454
Correspondence Autoencoder	Corr-AE	0.532	0.521	0.527	0.441	0.494	0.468	0.442	0.429	0.436
Deep CCA	DCCA	0.568	0.509	0.539	0.452	0.465	0.459	0.445	0.399	0.422
	CMDN	0.544	0.526	0.535	0.492	0.542	0.517	0.487	0.427	0.457
Classifier-based	Deep-SM	0.560	0.539	0.550	0.497	0.478	0.488	0.478	0.422	0.450
	ACMR	0.538	0.544	0.541	0.519	0.542	0.531	0.468	0.412	0.440
	CCL	0.576	0.561	0.569	0.481	0.520	0.501	0.505	**0.457**	0.481
Adaptive Triplet loss	SAM	**0.637**	**0.643**	**0.640**	**0.563**	**0.594**	**0.579**	**0.518**	**0.457**	**0.487**

David Semedo and Joao Magalhães. **Cross-Modal Subspace Learning with Scheduled Adaptive Margin Constraints**. ACM Multimedia (MM '19).

State-of-the-art and Applications

Revisiting Visual Commonsense Reasoning: Answer a question about an image and provide a rationale justifying the answer.

Model Architecture:



Zellers, Rowan et al. **From Recognition to Cognition: Visual Commonsense Reasoning**. CVPR 2019.

Diachronic Cross-modal Embeddings

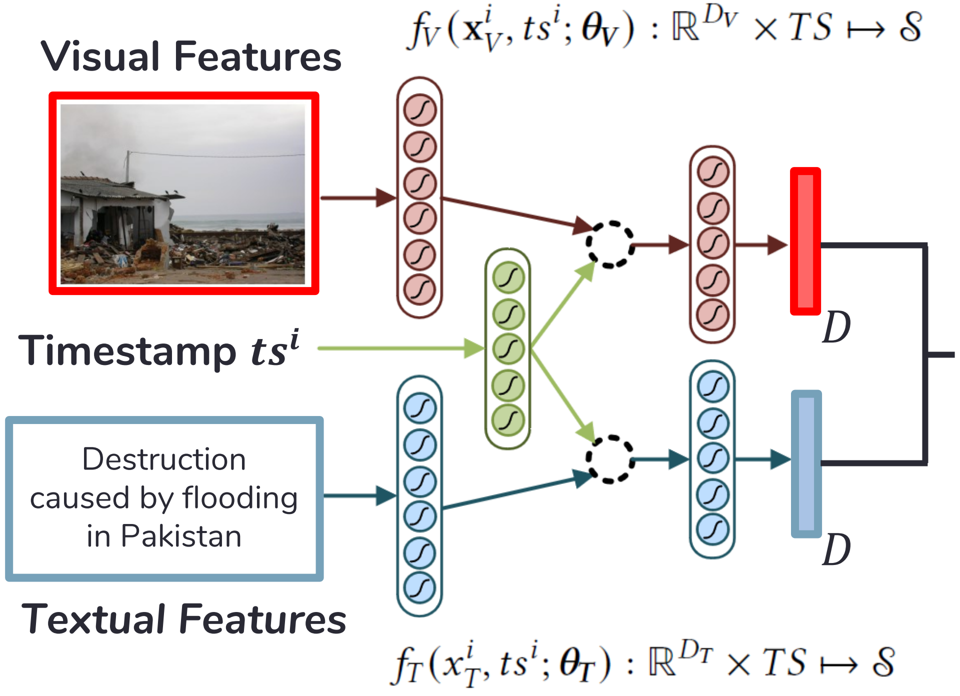
Adding the time dimension:

- Modeling the evolution of vision and language interactions;
- Learn the embedding from a large temporal span dataset (20 years).

David Semedo and Joao Magalhaes. **Diachronic Cross-modal Embeddings**. ACM Multimedia (MM '19).

Diachronic Cross-modal Embeddings

Modeling the evolution of vision and language interactions



- ✓ **Conditions** representations on time (*continuous input*);
- ✓ **Jointly learned**.

$$\mathcal{L}(x_*^a, x_*^p, x_*^n; \theta) = \mathcal{L}_{inter}(x_*^a, x_*^n; \theta) + \mathcal{L}_{intra}(x_*^a, x_*^p; \theta)$$

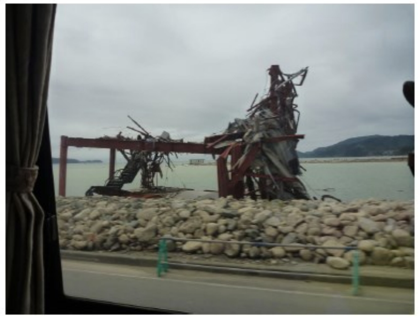
Triplet loss

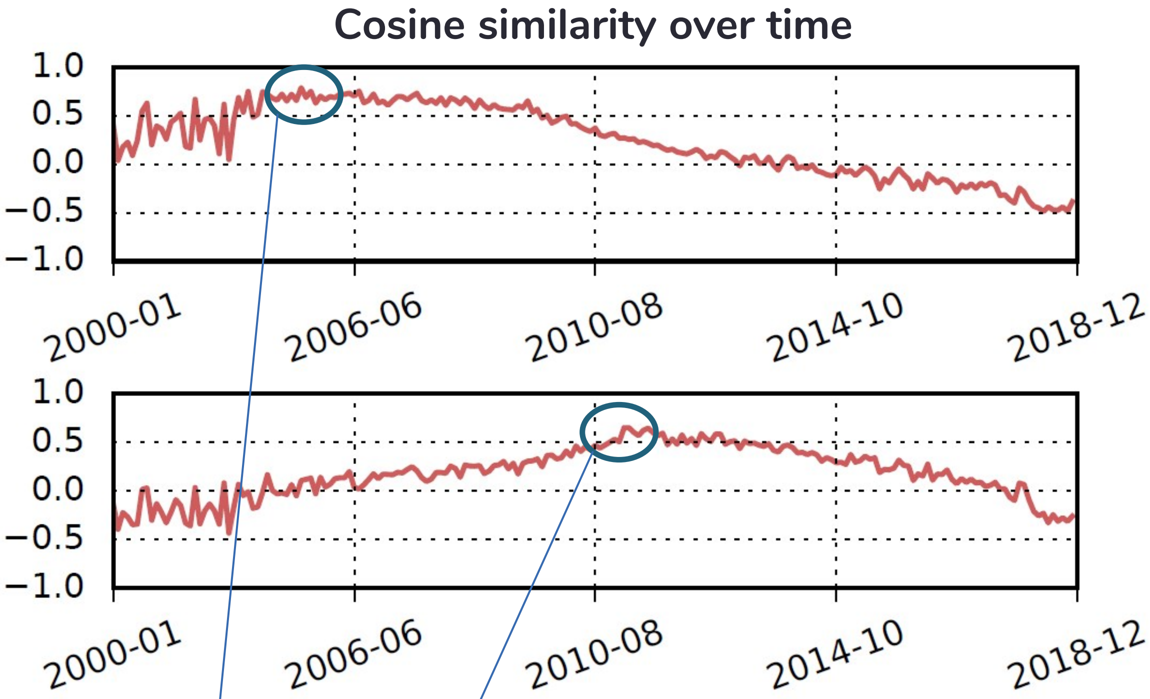
Temporal Alignment to preserve data original timeline

Analysis of Semantic Dispersion over Time on 20 years of data

Tsunami
Indonesia

Tsunami
Japan

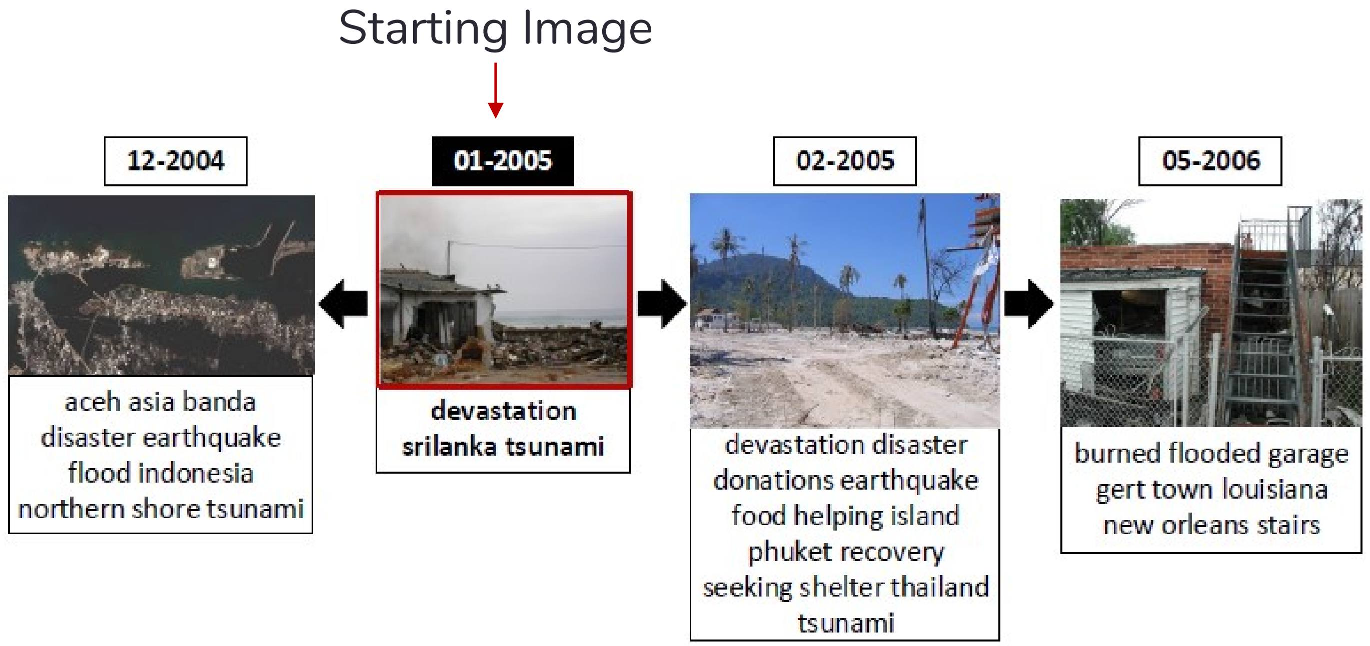




Corresponds to the timestamps
of the images

Cross-modal Evolution – Summarizing Trajectories

Automatically generate summaries given a single Image/Text,
based on temporal trajectories

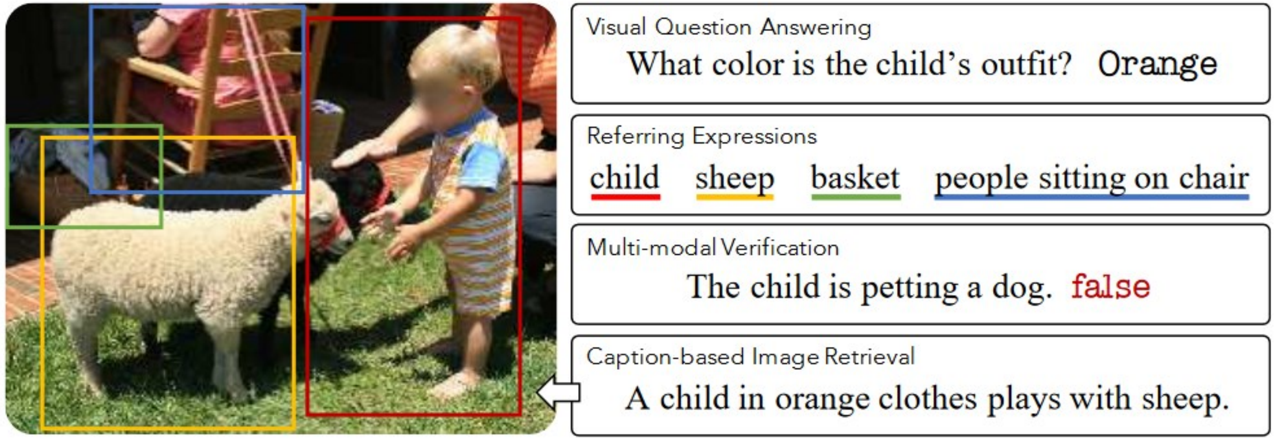


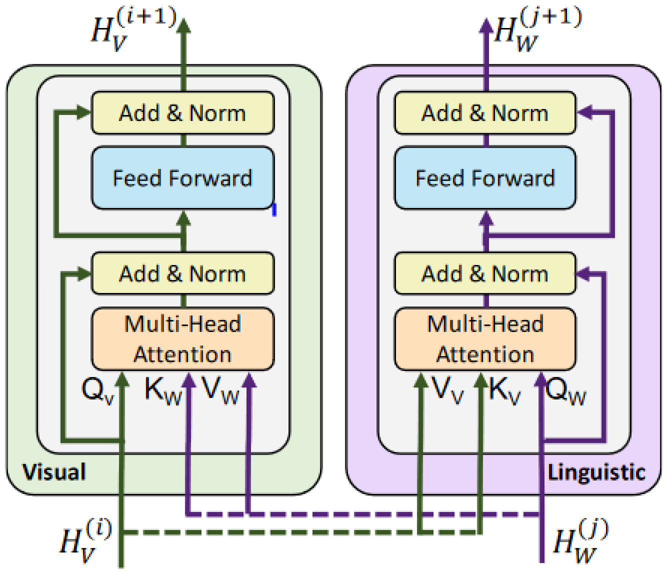
David Semedo and Joao Magalhaes. **Diachronic Cross-modal Embeddings**. ACM Multimedia (MM '19).

State-of-the-art and Applications

12-in-1: Multi-Task Vision and Language Representation Learning:

- ✓ Based on ViLBERT - BERT extension to Jointly Represent Images and Text.





Co-attention transformer layer

Information exchange between modalities.

Research direction: Multi-task learning + several datasets.

Lu, Jiasen et al. **ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks**. NeurIPS 2019.
Lu, Jiasen et al. **12-in-1: Multi-Task Vision and Language Representation Learning**. CVPR 2020.

Summary and Key Takeaways

Multimodal and Cross-modal embeddings can effectively combine information from multiple modalities.

- We've covered the basic building blocks to learn effective multimodal embeddings.
- Feed ML models to address down-stream tasks.

Further reading and pointers

- Awesome multimodal ML list (Link);
- Multimodal Machine Learning tutorial (Link);
- *Nice survey: Baltrušaitis, Tadas et al.* **Multimodal Machine Learning: A Survey and Taxonomy**, PAMI 2019;
- Some relevant conferences: ACM MM, CVPR, NeurIPS, ICCV, ICML, …
- Deep Learning book, Ian Goodfellow, Yoshua Bengio and Aaron Courville (Link).

Thank you!

David Semedo, Ph.D.   df.semedo@campus.fct.unl.pt

Universidade NOVA de Lisboa, Portugal

NOVALINCS
LABORATORY FOR COMPUTER SCIENCE AND INFORMATICS

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA