

Bidirectional Joint Representation Learning with Symmetrical Deep Neural Networks for Multimodal and Crossmodal Applications

Vedran Vukotić^{1,2}, Christian Raymond^{1,2}, Guillaume Gravier^{1,3}
 vedran.vukotic@irisa.fr christian.raymond@irisa.fr guillaume.gravier@irisa.fr

Problem

- given multimodal data in a continuous representation space

Goal:

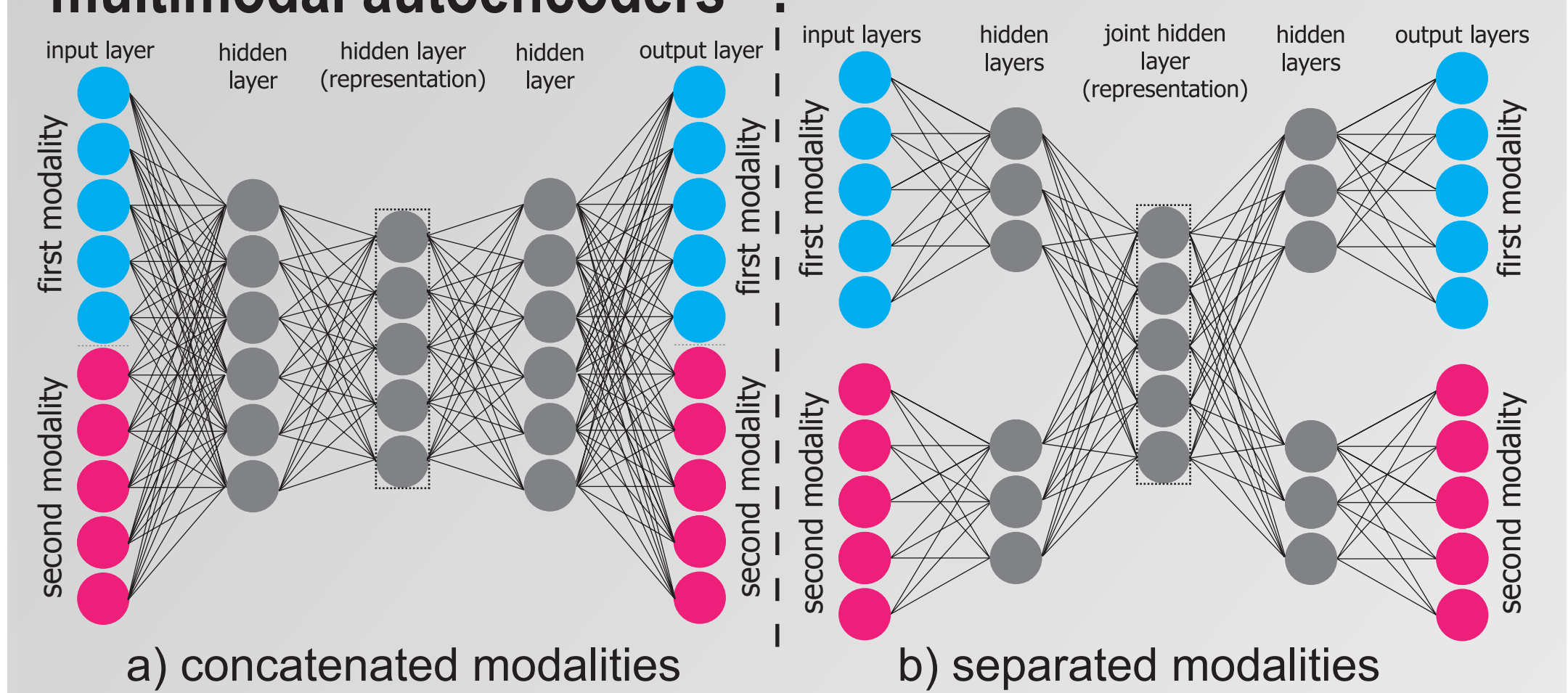
- perform retrieval, ranking, classification, etc.

Means:

- crossmodal translation
- early fusion / multimodal embedding

Common Approaches

• multimodal autoencoders^{1,2}:

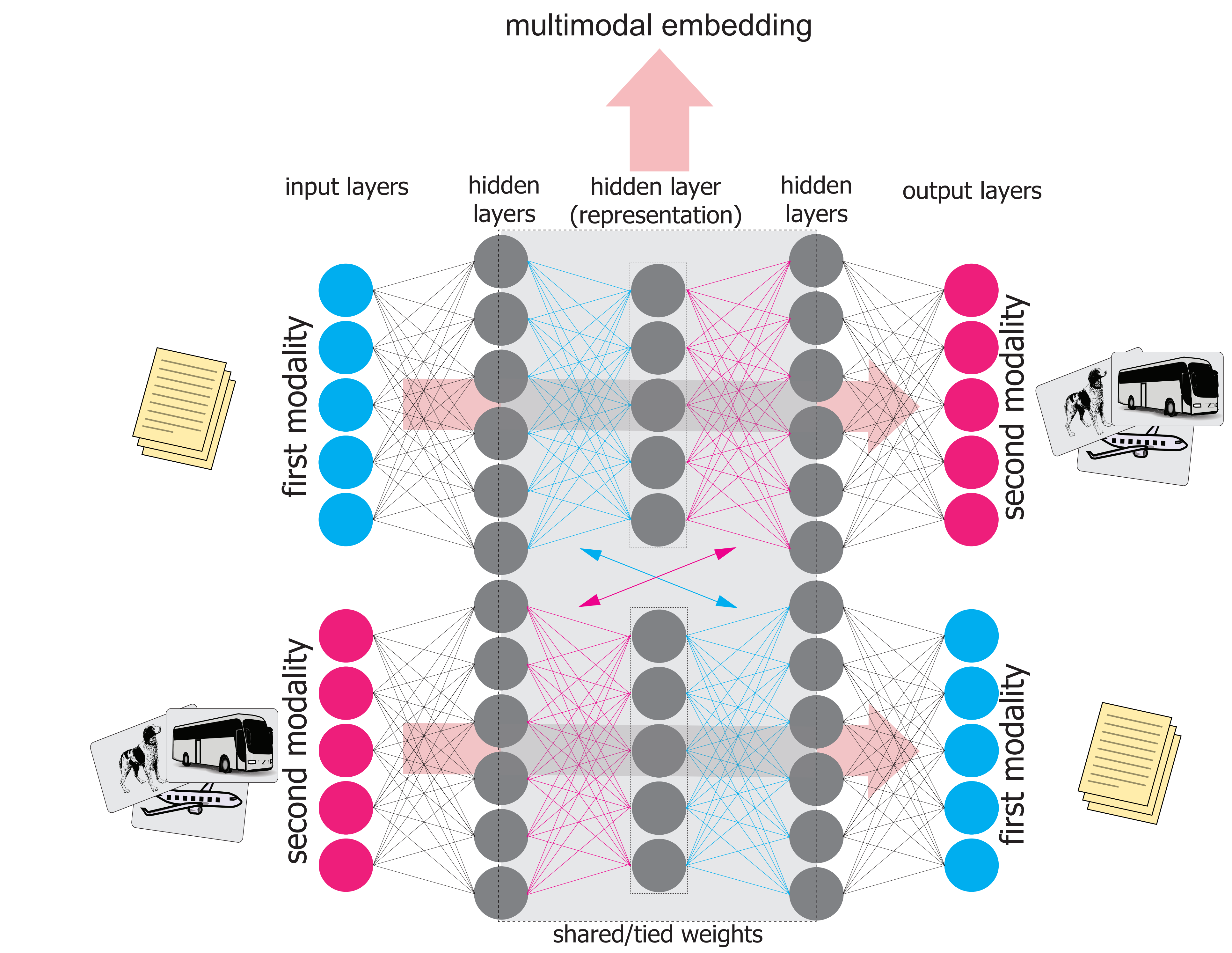


a) concatenated modalities b) separated modalities

- additional improvement with:
 - superposition of noise to input
 - sporadic removal of one input modality
 - dropout

Downsides

- both inputs influence the same layer (directly or through other hidden layers) - mixed influence
- autoencoders need to learn to reconstruct the same output both when one modality is marked missing (e.g. 000...0000) and when both modalities are presented as input
- primarily made for multimodal embedding; crossmodal translation is a secondary function



Evaluation & Dataset Representation

Evaluation:

- video hyperlinking task
- ranking⁴ relevant video segments by similarity to referent video segment
- dataset from TRECVID 2015 - relevance judged by AMT
- two modalities used: automatic speech transcripts and detected visual concepts provided by KU Leuven

Automatic Transcripts:

- averaged³ Word2Vec representation of each word appearing in the video segment

Visual Concepts:

- averaged Word2Vec representation of all Leuven visual concepts appearing in the video segment (sorted)

Dataset Statistics:

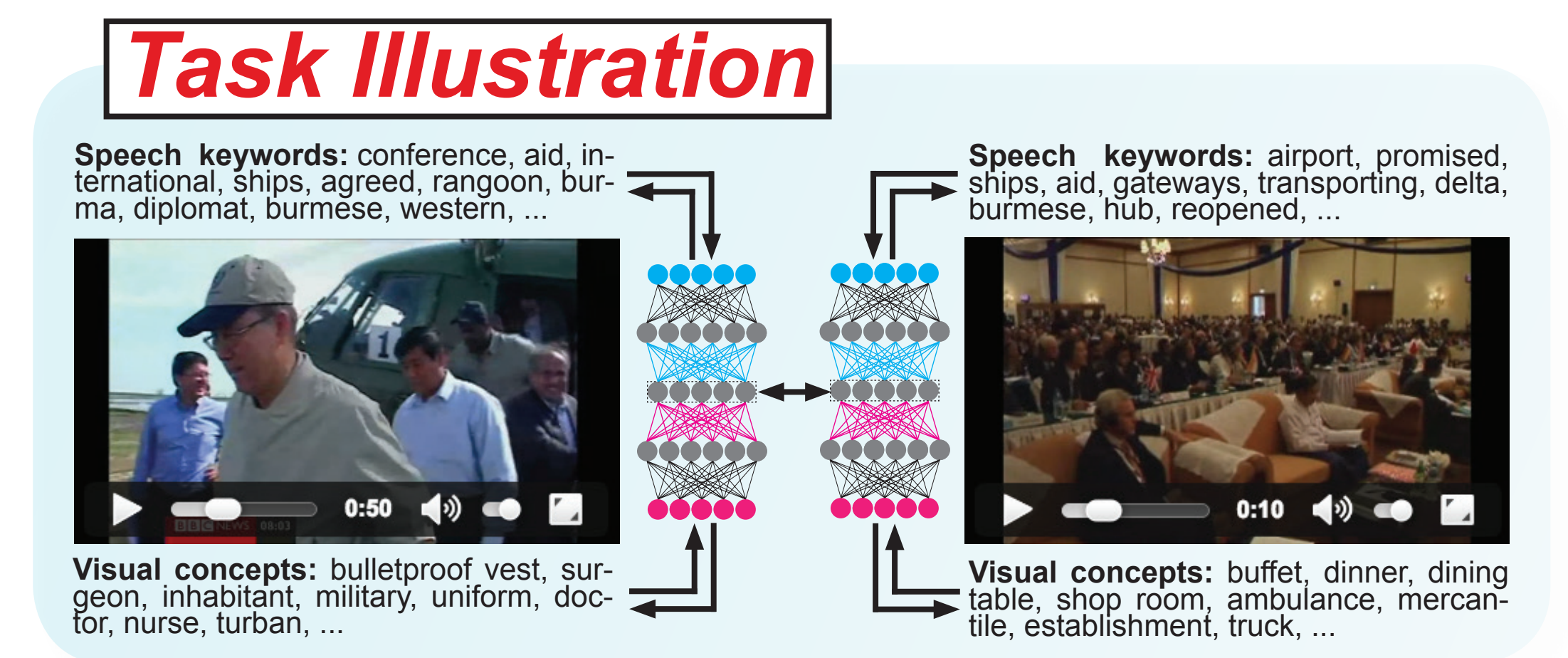
- 30 referent video segments (anchors), 10,809 video segments to match (targets) and a ground truth with 12,340 anchor-target pairs

Idea

- use DNNs with identical architectures to translate from one modality to the other and conversely
 - direct crossmodal translation
 - if one modality is missing, only the other is used (no zeroed inputs)
- enforced symmetry by tying weights in the central part
 - equivalent to training a DNN to minimize reconstruction errors in both directions
 - creates symmetrical mappings and a **joint multimodal representation space** in the central hidden layer

Results

Method	P@10	σ
Baseline		
Only transcripts	58.67	-
Only visual concepts	50.00	-
Linear combination	61.32	3.1
Autoencoders		
Embedded tran. and v. c. with AE a	57.40	1.24
Embedded tran. and v. c. with AE b	59.60	0.65
Bidirectional DNNs with tied weights		
Embedded transcripts	70.43	0.46
Embedded visual concepts	54.92	0.99
Embedded transcripts and visual c.	73.74	0.82
Expanded transcripts	58.16	0.24
Expanded visual concepts	55.75	0.13
Expanded transcripts and visual c.	62.35	0.25



References

- H. Lu, Y. Liou, H. Lee, and L. Lee. Semantic retrieval of personal photos using a deep autoencoder fusing visual features with speech annotations represented as word/paragraph vectors. In *Annual Conf. of the Intl. Speech Communication Association*, 2015.
- J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng. Multimodal deep learning. In *Intl. Conf. on Machine Learning*, 2011.
- M. Campr and K. Jezek. Comparing semantic models for evaluating automatic document summarization. In *Text, Speech, and Dialogue*, 2015.
- J. Weston, S. Bengio, and N. Usunier. Large scale image annotation: learning to rank with joint word-image embeddings. *Machine learning*, 81(1):21–35, 2010.