Multimodal and Crossmodal Representation Learning from Textual and Visual Features with Bidirectional Deep Neural Networks for Video Hyperlinking

Vedran Vukotić, Christian Raymond, Guillaume Gravier

INRIA/IRISA, Rennes, France

iV&L-MM'16

October 16th 2016, Amsterdam, NL

Outline:

- short introduction on video hyperlinking
- overview of single-modal representations and related NN models
- crossmodal and multimodal methods:
  - score fusion
  - multimodal autoencoders
  - (novel) bidirectional DNNs
- comparative results
- conclusions & future work

Introduction

Video Hyperlinking

"creating hyperlinks within video data based on content analysis and comparison, where links might reflect various types of relations between the source (i.e., anchor) and target fragments of the link"[a]

[a] Anca-Roxana Simon. "Semantic structuring of video collections from speech: segmentation and hyperlinking". PhD thesis. Rennes: Univ. Rennes 1, 2015.

Introduction

Video Hyperlinking

"creating hyperlinks within video data based on content analysis and comparison, where links might reflect various types of relations between the source (i.e., anchor) and target fragments of the link"[a]

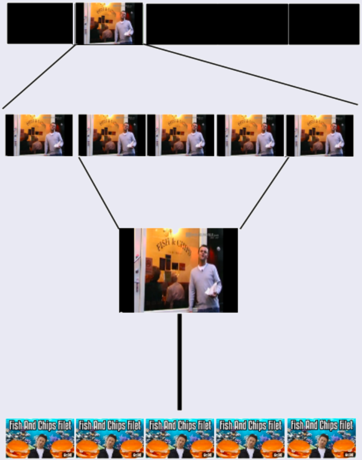
[a] Anca-Roxana Simon. "Semantic structuring of video collections from speech: segmentation and hyperlinking". PhD thesis. Rennes: Univ. Rennes 1, 2015.

Video Hyperlinking - Notions

- **anchor** - currently viewing video segment for which the viewer is asking for references
- **target** - hyperlinked video segment that is somehow relevant for the anchor

Video Hyperlinking

Illustration



Modalities and Representations

Two Streams

- **video stream**:
  - sequence of frames:
    - keyframes (intraframes)
    - other frames (interframes) ✗
  - possible representations:
    - higher level concepts
    - higher level embeddings
    - lower level embeddings

Modalities and Representations

Two Streams

- **video stream**:
  - sequence of frames:
    - keyframes (intraframes)
    - other frames (interframes) ✗
  - possible representations:
    - higher level concepts
    - higher level embeddings
    - lower level embeddings
- **audio stream - speech:**
  - subtitles ✗
  - automatic transcripts

Modalities and Representations

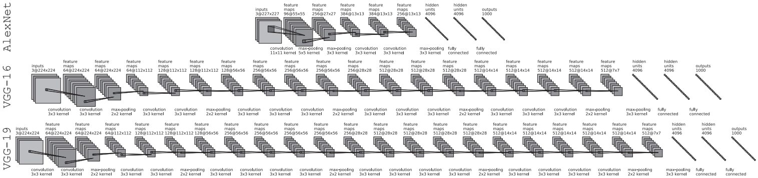
Two Streams

- **video stream**:
  - sequence of frames:
    - keyframes (intraframes)
    - other frames (interframes) ✗
  - possible representations:
    - higher level concepts
    - higher level embeddings
    - lower level embeddings
- **audio stream - speech:**
  - subtitles ✗
  - automatic transcripts

Representing Data:

- discrete representation spaces ✗
- continuous representation spaces ✓

Visual Representations - Convolutional Neural Networks (CNNs)



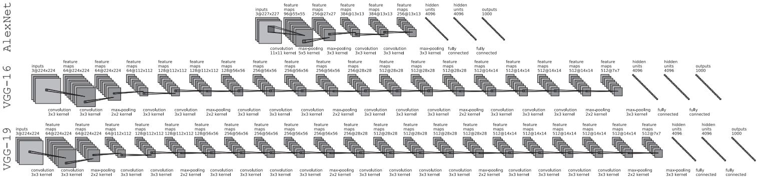
CNN Architecures [1]

three common architectures:

- AlexNet
- VGG-16
- VGG-19

[1] https://github.com/heuritech/convnets-keras

Visual Representations - Convolutional Neural Networks (CNNs)



CNN Architecures [1]

three common architectures:

- AlexNet
- VGG-16
- VGG-19

provide:

- ImageNet concepts
- high level descriptors
- low level descriptors

[1] https://github.com/heuritech/convnets-keras

Speech Representations

Speech as Textual Information

two common continuous representations:

- Word2Vec
  - state-of-the art results
  - designed to represent words (word embeddings can be aggregated)
  - two main models: **skip-gram** and CBOW (continuous bag of words)

Speech Representations

Speech as Textual Information

two common continuous representations:

- Word2Vec
  - state-of-the art results
  - designed to represent words (word embeddings can be aggregated)
  - two main models: **skip-gram** and CBOW (continuous bag of words)
- paragraph vectors[a]
  - made to represent multiple words (paragraphs, blocks, documents, ...)
  - two models: PV-DM (distributed memory) and PV-DBOW (distributed bag of words)

[a] Quoc V Le and Tomas Mikolov. "Distributed Representations of Sentences and Documents." In: *ICML*. vol. 14. 2014, pp. 1188–1196.

Dataset

MediaEval 2014

- part of the search and hyperlinking task in 2014 (now in TRECVID)
- automatic transcripts, subtitles and KU Leuven visual concepts[a] are offered
- 30 anchors provided - submissions judged post-hoc on Amazon mechanical turk

Dataset

MediaEval 2014

- part of the search and hyperlinking task in 2014 (now in TRECVID)
- automatic transcripts, subtitles and KU Leuven visual concepts[a] are offered
- 30 anchors provided - submissions judged post-hoc on Amazon mechanical turk
- a groundtruth is formed after the challenge
  - 30 anchors and 10 809 targets
  - 12 340 anchor-target pairs (related and unrelated)
  - 371 664 keyframes ($\approx$ 34.3 keyframes per video segment)

[a] Tatiana Tommasi, Tinne Tuytelaars, and Barbara Caputo. "A Testbed for Cross-Dataset Analysis". In: *CoRR* abs/1402.5923 (2014).

Single-modal Representations: Visual Representations

Precision at 10 (%) - Single Modality

Representation	Aggregation	P@10 (%)
KU Leuven v. c., W2V	**average**	**50.00**
KU Leuven v. c., W2V	Fisher	47.80
KU Leuven v. c., PV-DM	-	45.33
KU Leuven v. c., PV-DBOW	-	48.33
AlexNet	average	63.00
AlexNet	**Fisher**	**65.00**
VGG-16	**average**	**70.67**
VGG-16	Fisher	64.67
VGG-19	**average**	**68.67**
VGG-19	Fisher	66.00

Single-modal Representations: Representations of Transcripts

Precision at 10 (%) - Single Modality

Representation	Aggregation	P@10 (%)
Word2Vec	**average**	**58.67**
Word2Vec	Fisher	54.00
PV-DM	-	45.00
PV-DBOW	-	41.67

Using Multiple Modalities

Approaches:

- multimodal fusion
  - score fusion
  - early fusion
  - late fusion
- crossmodal translation

Using Multiple Modalities

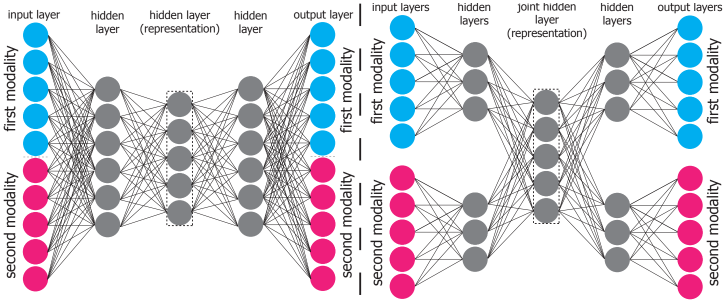
Approaches:

- multimodal fusion
  - score fusion
  - early fusion
  - late fusion
- crossmodal translation

Popular Methods:

- (generative) statistical modeling
- multimodal/crossmodal autoencoders

Multimodal Autoencoders



Features:

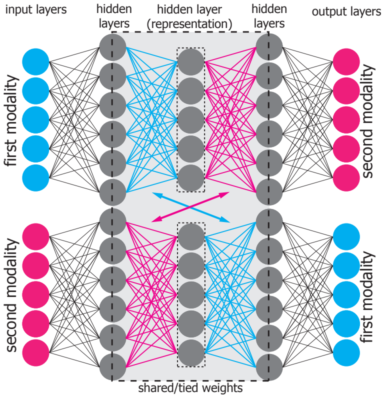
- multimodal fusion
  - early fusion
  - late fusion
- crossmodal translation

Multimodal Autoencoders - Downsides

Downsides:

- both modalities influence the same central layer
- even when translating, one modality is mixed with the other or with a zeroed input
- need to reconstruct both the same both when both modalities are present and when one is zeroed
- primarily made for multimodal embedding, crossmodal translation is secondary

Bidirectional (Symmetrical) Deep Neural Networks



$$\mathbf{h}_1^{(1)} = f(\mathbf{W}_1^{(1)} \times \mathbf{x}_{\mathbf{m}_1} + \mathbf{b}_1^{(1)})$$
$$\mathbf{h}_2^{(1)} = f(\mathbf{W}_2^{(1)} \times \mathbf{x}_{\mathbf{m}_2} + \mathbf{b}_2^{(1)})$$

$$\mathbf{h}_1^{(2)} = f(\mathbf{W}^{(2)} \times \mathbf{h}_1^{(1)} + \mathbf{b}_1^{(2)})$$
$$\mathbf{h}_2^{(2)} = f(\mathbf{W}^{(3)^T} \times \mathbf{h}_2^{(1)} + \mathbf{b}_2^{(2)})$$

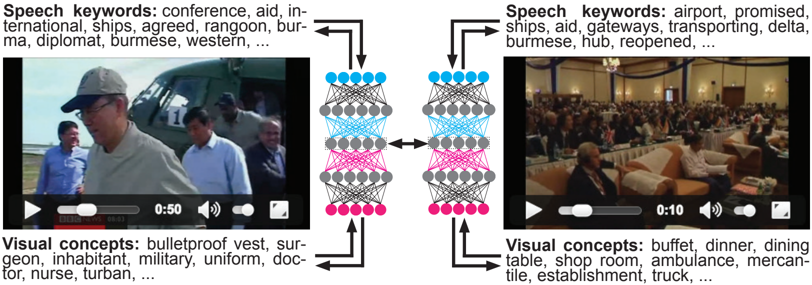
$$\mathbf{h}_1^{(3)} = f(\mathbf{W}^{(3)} \times \mathbf{h}_1^{(2)} + \mathbf{b}_1^{(3)})$$
$$\mathbf{h}_2^{(3)} = f(\mathbf{W}^{(2)^T} \times \mathbf{h}_2^{(2)} + \mathbf{b}_2^{(3)})$$

$$\mathbf{o}_1 = f(\mathbf{W}_1^{(4)} \times \mathbf{h}_1^{(3)} + \mathbf{b}_1^{(4)})$$
$$\mathbf{o}_2 = f(\mathbf{W}_2^{(4)} \times \mathbf{h}_2^{(3)} + \mathbf{b}_2^{(4)})$$

Main Idea:

- perform bidirectional crossmodal translation primarily
- enforce symmetry (tie weights) in the middle part
- use concatenated central layers as representation

Video Hyperlinking with BiDNNs



Simple Multimodal Approaches

Concatenation

Modalities	P@10 (%)	σ (%)
transcripts, visual concepts	58.00	-
transcripts, AlexNet	70.00	-
transcripts, VGG-16	**75.33**	-
transcripts, VGG-19	74.33	-

Linear Combination of Scores

Modalities	P@10 (%)	σ (%)
transcripts, visual concepts	61.32	3.10
transcripts, AlexNet	67.38	2.66
transcripts, VGG-16	**71.86**	**4.11**
transcripts, VGG-19	71.78	3.90

Multimodal Autoencoders

Multimodal Autoencoders (Separate Branches)

Modalities	Method	P@10 (%)	σ (%)	imp.
Simple Multimodal Approaches				
transcripts, v.c.	lin. comb.	61.32	3.10	-
transcripts, AlexNet	lin. comb.	67.38	2.66	-
transcripts, VGG-16	lin. comb.	71.86	4.11	-
transcripts, VGG-19	lin. comb.	71.78	3.90	-
Multimodal Autoencoders				
transcripts, visual concepts		59.60	0.65	✗
transcripts, AlexNet		69.87	1.64	✓
transcripts, VGG-16		74.53	1.52	✓
transcripts, VGG-19		75.73	1.79	✓

BiDNN Multimodal Embedding

Multimodal Embedding:

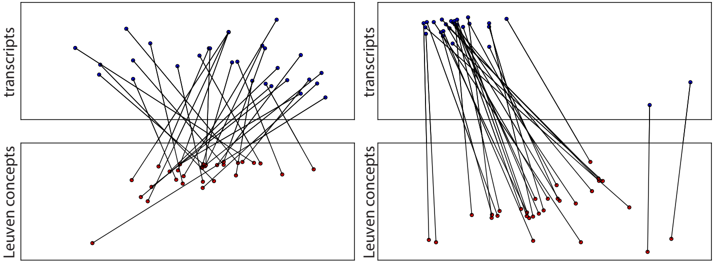
Modalities	Method	P@10 (%)	σ (%)	imp.
Multimodal Autoencoders				
transcripts, visual concepts		59.60	0.65	-
transcripts, AlexNet		69.87	1.64	-
transcripts, VGG-16		74.53	1.52	-
transcripts, VGG-19		75.73	1.79	-
BiDNN Multimodal Embedding				
transcripts, visual concepts		73.74	0.46	✓
transcripts, AlexNet		73.41	1.08	✓
transcripts, VGG-16		76.33	1.60	✓
transcripts, VGG-19		**80.00**	**0.80**	✓

BiDNN Single-modality Embedding

Embedding One Modality + Crossmodal Training:

Modalitiy	Method	P@10 (%)	σ (%)	imp.
Original Single-modal Embedding				
transcripts		58.00	-	-
visual concepts		50.00	-	-
AlexNet		65.00	-	-
VGG-16		70.67	-	-
VGG-19		68.67	-	-
BiDNN Single-modality Embedding				
transcripts		66.78	1.05	✓
visual concepts		54.92	0.99	✓
AlexNet		66.33	0.58	✓
VGG-16		68.70	1.98	✗
VGG-19		70.81	1.08	✓

BiDNN Embedding Space



a) original representation spaces b) joint representation space

BiDNN Query Expansion

Crossmodal Translation - Query Expansion:

Modalities	Method	P@10 (%)	σ (%)	imp.
Simple Multimodal Approaches				
transcripts, v.c.	concat	58.00	-	-
transcripts, AlexNet	concat	70.00	-	-
transcripts, VGG-16	concat	75.33	-	-
transcripts, VGG-19	concat	74.33	-	-
BiDNN query expansion				
transcripts, visual concepts		62.35	0.25	✓
transcripts, AlexNet		70.11	1.25	✓
transcripts, VGG-16		75.33	0.10	✗
transcripts, VGG-19		74.33	0.10	✗

Conclusion

Single-modal Takeaways:

- averaged word2vecs on transcripts seem to work best
- higher-level CNN embeddings work better than visual concepts
- deeper architecture are better (VGG-x are deep enough)
- embedding even single modalities can help

Conclusion

Single-modal Takeaways:

- averaged word2vecs on transcripts seem to work best
- higher-level CNN embeddings work better than visual concepts
- deeper architecture are better (VGG-x are deep enough)
- embedding even single modalities can help

Multimodal / Crossmodal Conclusions:

- classical multimodal autoencoders are great but have few downsides
- BiDNNs seem to tackle those downsides and yield an improved representation space
- focusing on crossmodal translation > focusing on multimodal embedding ($\Longrightarrow$ both are improved!)

Future Work

Plan:

- test performance outside video-hyperlinking (e.g. sentence-image matching on flickr8k and flickr30k)
- add and evaluate batch normalization
- test if introducing sparsity would further improve the representation space
- potentially:
  - check if joint the costs of end-to-end learning would pay out
  - check if generative models (e.g. variational autoencoders) would help

Thank you! Questions?