IRISA @ TRECVID2017

Beyond Crossmodal and Multimodal Models

Task: Video Hyperlinking

Mikail Demirdelen, Mateusz Budnik, Gabriel Sargent, Rémi Bois, Guillaume Gravier

IRISA, Université de Rennes 1, CNRS

Table of contents

Introduction

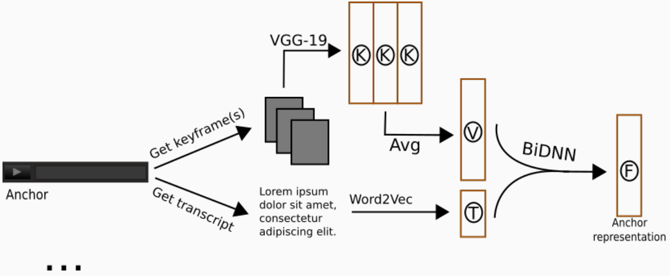
A crossmodal system

In 2016, IRISA used a crossmodal system[1]:

- Segmentation step

  → Get segments from whole videos

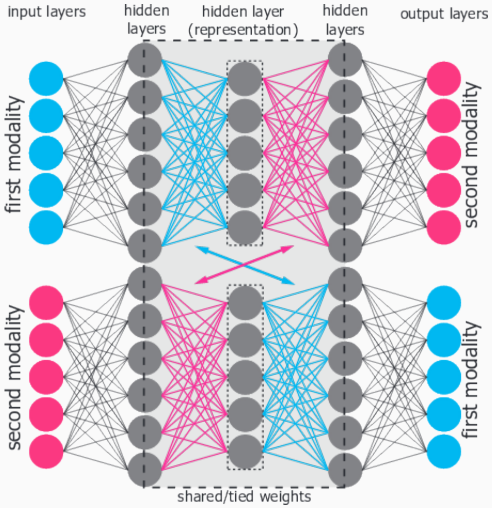
- Segments/anchors embedding step:



- Comparing and ranking step

  → For each anchor, compare and rank each segment

The BiDNN



This system had the best score on P@5

→ Go further with this approach?

Segmentation

Motivation

In 2016, we had around **300,000 segments**
$\rightarrow$ Limited number of segments
$\rightarrow$ Problems with the overlap

Create more segments!

Some constraints:
$\rightarrow$ The segment should not cut the speech
$\rightarrow$ They must last between 10 and 120 seconds

The method

With a constraint programming framework:

- Keep all the segments that last between 50 and 60 seconds without cutting the speech
- When there we none, expand the duration between 10 and 120 seconds

1.1 million new segments $\rightarrow$ **1.4 million segments** in total (around 4 times more)

Representations

Motivation

Our model greatly depends on the quality of the representation of each modality
$\rightarrow$ Can we improve them?

Development set: each triplet (anchor, target, matching) submitted last year

We extracted/recovered:

- For each anchor, its transcript and one or more keyframes
- For each target, its transcript and one keyframe

Visual Representation

Embedding of the keyframes using different pre-trained CNNs (VGG-19[7], ResNet[2], ResNext[9] and Inception[8])

When multiples keyframes, there was an additional step of ***keyframe representation fusion***:

- Single: Using a single keyframe and discarding the rest
- Avg: The embedding is the average of all of the keyframes embeddings
- Max: Each feature of the embedding is the maximum of all keyframes corresponding feature

Visual Representation

	Single		Average		Max	
Models	P@5	P@10	P@5	P@10	P@5	P@10
VGG19	41.60	41.27	43.40	41.60	42.60	41.03
Inception	40.40	41.83	41.00	41.39	42.60	41.73
ResNext-101	41.00	39.37	41.40	40.10	41.80	39.90
ResNet-200	43.80	41.57	47.20	44.37	**47.60**	**44.87**
ResNet-152	44.40	41.37	45.60	41.67	45.20	40.40

$\rightarrow$ We chose to use a *ResNet-200* network and a *Max* keyframe representation fusion method

Textual Representation

Same experiments with transcripts:

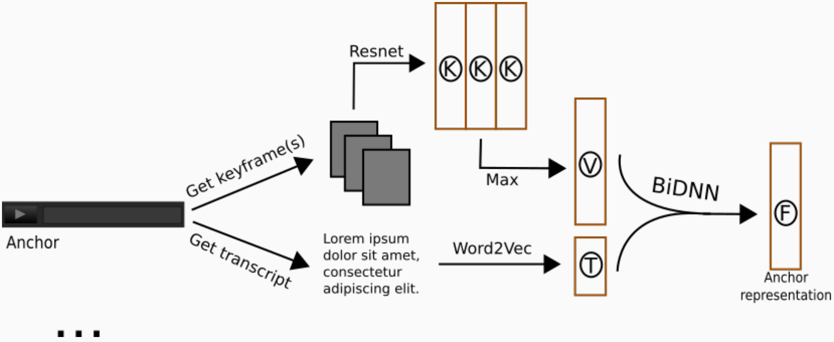
Models	P@5	MAP
Average Word2Vec[5]	**44.2**	**45.3**
Doc2Vec[4]	38.4	39.4
Skip-Thought[3]	40.2	41.6

→ We chose to keep *Word2Vec*.

Runs description

BiDNNFull - Crossmodal Bidirectional Joint Learning

A bidirectional deep neural network (BiDNN) was trained with ResNet as a visual descriptor and a Word2Vec as a textual descriptor:



→ *BiDNNFull* is our baseline for testing other improvements to the system.

BiDNNFilter - BiDNN with metadata filter


```xml
<video>
  <title><![CDATA[iPhone 3G Street Interview Episode 2 - What do you do with your
  iPhone ?]]></title>
  <description><![CDATA[<p><strong>In Episode 2 of our iPhone 3G street
  interviews. People were asked as they were walking out of a SF Bay Area Apple
  Store about what they do with their iPhones. And we just let the camera roll. </
  strong></p><p>]]></description>
  <explicit>false</explicit>
  <duration>122</duration>
  <url>http://blip.tv/file/1059784</url>
  <license>
    <type>Creative Commons Attribution-NonCommercial-NoDerivs 2.0</type>
    <id>3</id>
  </license>
  <tags>
    <string>iphone</string>
    <string>3g</string>
    <string>gossip</string>
    <string>interviews</string>
    <string>apple</string>
    <string>google</string>
    <string>teens</string>
    <string>blogger</string>
  </tags>
  <uploader>
    <uid>219192</uid>
    <login>1801Media</login>
  </uploader>
```


Description

License

Tags

Uploader

We chose to keep the *list of tags* as a filter to compare anchors and targets that **share at least one tag in common**.

BiDNNFilter - BiDNN with metadata filter

However:

- 77% of videos have tags
- They have a mean number of tags of 4.71

Too restrictive?

Use the text of the descriptions:

- Selection of only verbs, nouns and adjectives
- Lemmatization
- Exclusion of stopwords and hapaxes

→ *BiDNNFilter* is the same as *BiDNNFull* but with the addition of the **list of keywords—tags and description—used as a filter**.

BiDNNPinv - Multimodal model with pseudo-inverse

Some issues about the keyframe representation fusion method:
$\rightarrow$ Basic treatment of information contained in multiple keyframes

We use the Moore-Penrose pseudo-inverse:

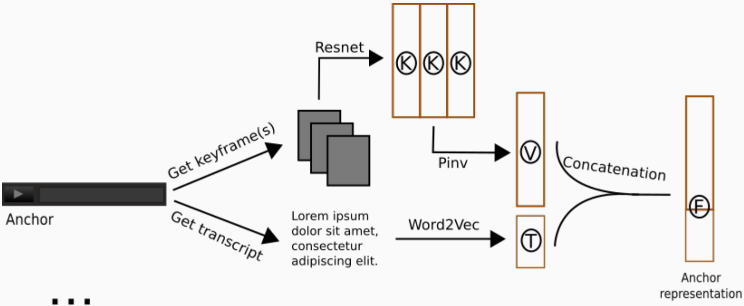
- **Capures a notion of movement** between multiple keyframes
- **Deals with different variations** found across all keyframes.
- It can improve the search quality[6].

$\rightarrow$ *BiDNNPinv* is the same as *BiDNNFull* where the Max function is replaced by the pseudo-inverse.

NoBiDNNPinv - Concatenation with pseudo-inverse

Quantify the usefulness of the BiDNN in this system

We replaced the BiDNN by a L2-normalization followed by a concatenation:



→ *NoBiDNNPinv*'s embedding pipeline is described by the picture.

Results

Results

Runs	MAP	MAISP	P@5	P@10	P@20
BiDNNFull	13.34	10.14	68.80	71.20	42.40
BiDNNFilter	10.81	8.43	**76.00**	**74.40**	38.00
BiDNNPinv	**15.29**	**11.52**	75.20	**74.40**	**43.40**
noBiDNNPinv	12.46	10.16	72.80	73.20	39.60

- *BiDNNFilter* obtained the best P@5 and P@10 showing the interest of **the filter to increase precision**.
- *BiDNNPinv* obtained the best MAP, MAISP and P@20 showing the **pseudo-inverse gives more precision stability**.
- The score difference between *BiDNNPinv* and *noBiDNNPinv* confirms the **relevance of the crossmodal model**.

Conclusion

Conclusion

Adding a filter increases the precision

The pseudo-inverse succeeds at capturing relevant information on multiple keyframes

We can think of future interesting developments:

- Combine both the filter and the pseudo-inverse
- Incorporate the metadata within the neural network, using it as a third modality
- Use the pseudo-inverse on both anchors and targets

Thank you for your attention!

References I

R. Bois, V. Vukotić, R. Sicre, C. Raymond, G. Gravier, and P. Sébillot.
Irisa at trecvid2016: Crossmodality, multimodality and monomodality for video hyperlinking.
In *Working Notes of the TRECVid 2016 Workshop*, 2016.

K. He, X. Zhang, S. Ren, and J. Sun.
Deep residual learning for image recognition.
In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

R. Kiros, Y. Zhu, R. R. Salakhutdinov, R. Zemel, R. Urtasun, A. Torralba, and S. Fidler.
Skip-thought vectors.
In *Advances in neural information processing systems*, pages 3294–3302, 2015.

References II

- Q. Le and T. Mikolov.
  Distributed representations of sentences and documents.
  In *Proceedings of the 31st International Conference on Machine Learning*, pages 1188–1196, 2014.
- T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean.
  Distributed representations of words and phrases and their compositionality.
  In *Advances in neural information processing systems*, pages 3111–3119, 2013.
- R. Sicre and H. Jégou.
  Memory vectors for particular object retrieval with multiple queries.
  In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, pages 479–482. ACM, 2015.

References III

K. Simonyan and A. Zisserman.
Very deep convolutional networks for large-scale image recognition.
arXiv preprint arXiv:1409.1556, 2014.

C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich.
Going deeper with convolutions.
In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He.
Aggregated residual transformations for deep neural networks.
arXiv preprint arXiv:1611.05431, 2016.

Some good/bad cases

BiDNNFilter:

Good cases

- anchor_131: good description + tags
- anchor_132&137: good description with no tags

Bad cases

- anchor_124: very general tags → not better than BiDNNFull
- anchor_126: only three tags that do not describe the video (grit, grittv, laura_flanders)
- anchor_141: no tags and a very long description (709 words)

BiDNNPinv:

Good cases

- anchor_141: an anchor with a lot of keyframes?

The **bad cases** are hard to identify

Moore-Penrose pseudo-inverse

Moore-Penrose pseudo-inverse

Given a set of anchor vectors represented as columns in a $d \times n$ matrix $X = [x_1, ..., x_n]$ where $x_i \in R^d$:

$$m(X) = X(X^T X)^{-1}\mathbf{1}_n \tag{1}$$

where $\mathbf{1}_n$ is a n dimensional vector with all values set to 1.