FRIEDRICH-ALEXANDER-
UNIVERSITÄT
ERLANGEN-NÜRNBERG
SCHOOL OF ENGINEERING

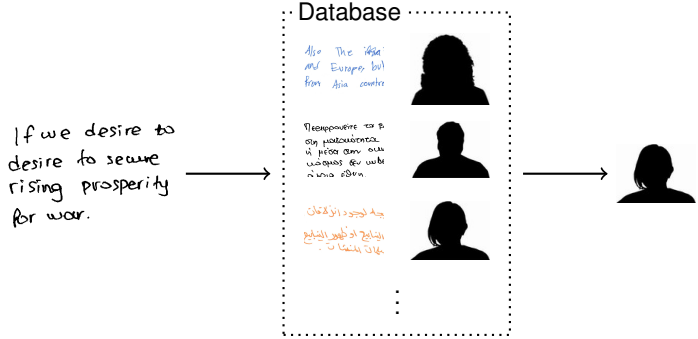
Encoding CNN Activations for Writer Recognition

Vincent Christlein, Andreas Maier
Pattern Recognition Lab, Friedrich-Alexander University of Erlangen-Nuremberg
April 27th, 2018

FRIEDRICH-ALEXANDER-
UNIVERSITÄT
ERLANGEN-NÜRNBERG
SCHOOL OF ENGINEERING

Writer Recognition

Writer Identification vs. Writer Retrieval



Writer Identification

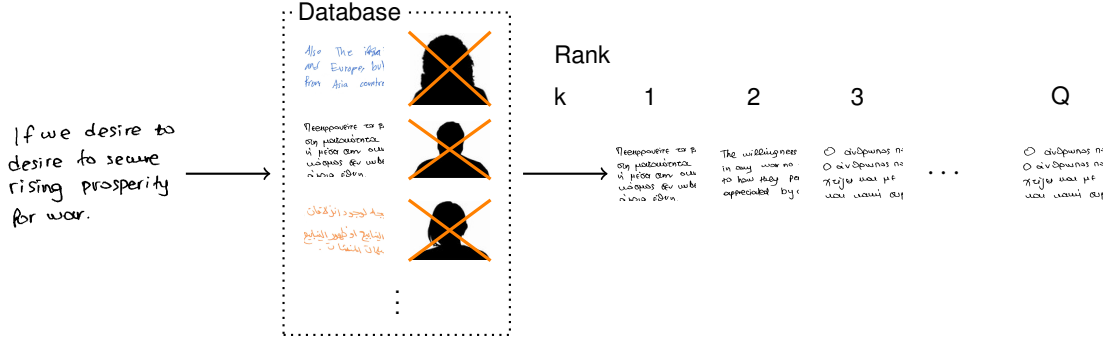
Given:

- Query document
- Documents of known writers

Wanted:

- Writer-ID

Writer Identification vs. Writer Retrieval



Writer Retrieval

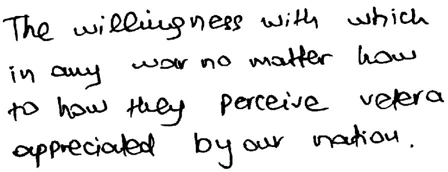
Given:

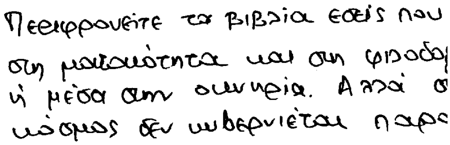
- Query document
- Documents of (possibly unknown) writers

Wanted:

- Most similar documents

Contemporary Datasets





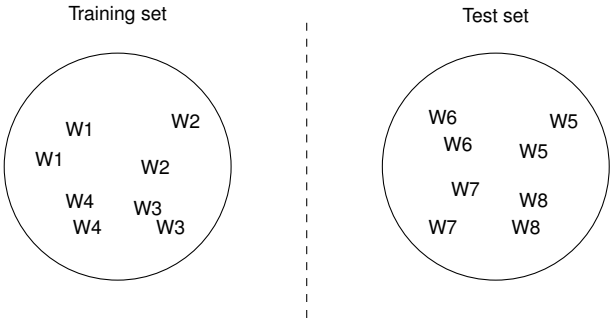
ICDAR13 benchmark dataset[1]

- 4 documents per writer (2 English, 2 Greek)
- Train: 100 writers
- Test: 250 writers

Other datasets: CVL (English, German), KHATT (Arabic), IAM (English)

[1] G. Louloudis, B. Gatos, N. Stamatopoulos, *et al.*, "ICDAR 2013 Competition on Writer Identification", in *ICDAR*, Washington DC, NY, Aug. 2013, pp. 1397–1401.

Writer-Independent Datasets

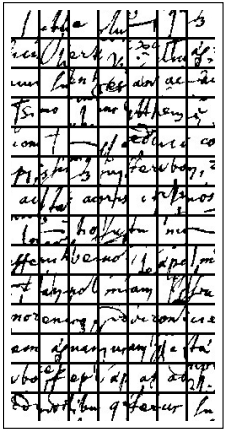


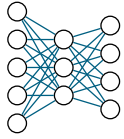
Training and test sets are independent

$\Rightarrow$ No training for a specific writer possible!

Typical Methodology For Deep Learning Feature Extraction

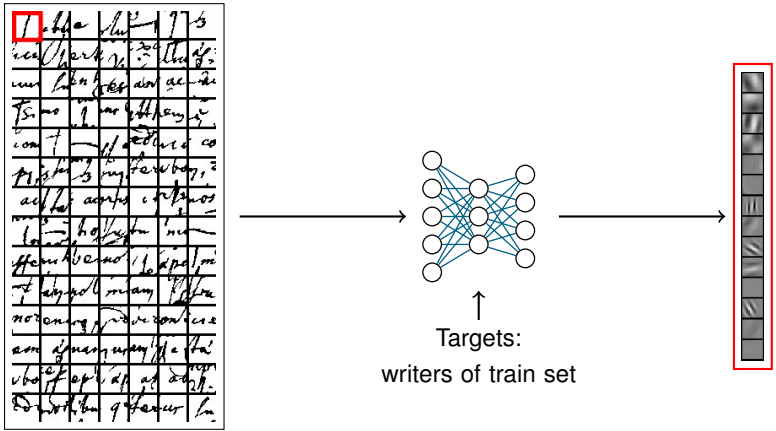
CNN Activation Features



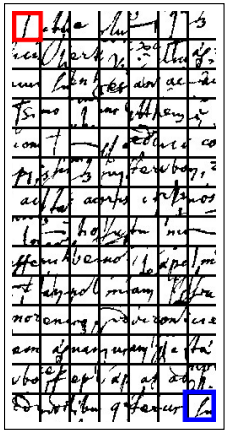


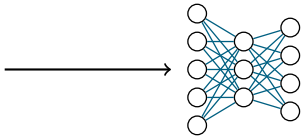
Targets:
writers of train set

CNN Activation Features



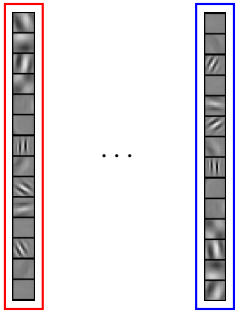
CNN Activation Features



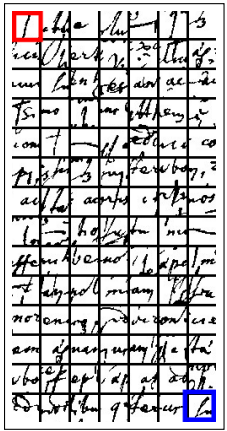


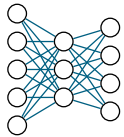
Targets:
writers of train set

CNN activation features



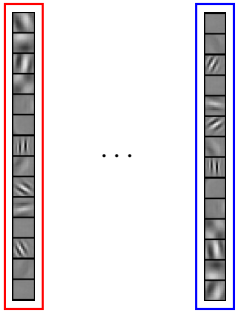
CNN Activation Features





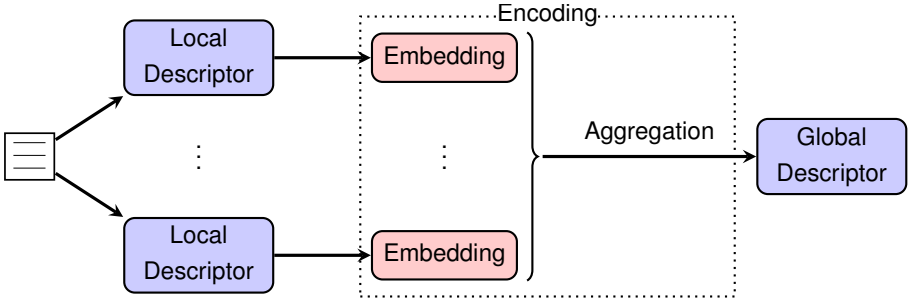
Targets:
writers of train set

CNN activation features



Encoding

Encoding



- Embedding: Fisher vectors, GMM supervectors, VLAD, triangulation embedding[2]
- Aggregation: Sum pooling, democratic aggregation, generalized max-pooling[3]

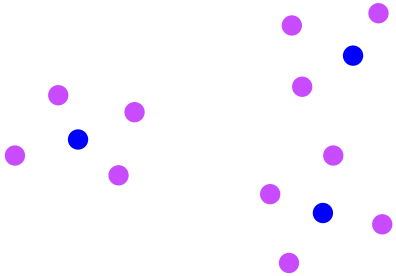
[2] H. Jégou and A. Zisserman, "Triangulation Embedding and Democratic Aggregation for Image Search", in *CVPR*, Columbus, Jun. 2014, pp. 3310–3317.

[3] N. Murray, H. Jegou, F. Perronnin, *et al.*, "Interferences in Match Kernels", *TPAMI*, vol. 39, no. 9, pp. 1797–1810, Oct. 2016. arXiv: 1611.08194.

VLAD[4]

$$\mathcal{X} = \{\boldsymbol{x}_i \in \mathbb{R}^D, i = 1, \ldots, T\}, \ \mathcal{D} = \{\boldsymbol{\mu}_k \in \mathbb{R}^D, k = 1, \ldots, K\}$$

VLAD

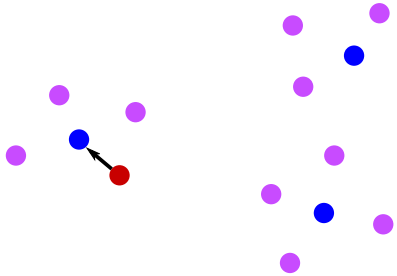


[4] H. Jégou, F. Perronnin, M. Douze, *et al.*, "Aggregating Local Image Descriptors into Compact Codes.", *PAMI*, vol. 34, no. 9, pp. 1704–1716, Sep. 2012.

VLAD[4]

$$\mathcal{X} = \{\boldsymbol{x}_i \in \mathbb{R}^D, i = 1, \ldots, T\},\ \mathcal{D} = \{\boldsymbol{\mu}_k \in \mathbb{R}^D, k = 1, \ldots, K\}$$

VLAD

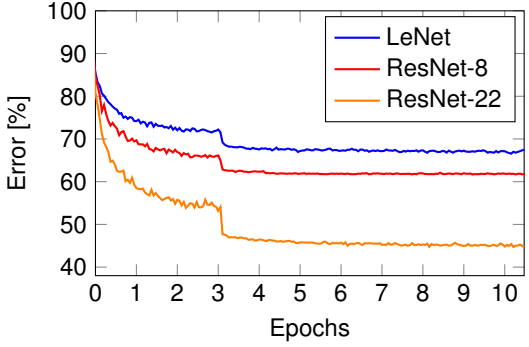


$$\phi_{\text{VLAD},k}(\boldsymbol{x}) = \text{NN}(\boldsymbol{x}, \boldsymbol{\mu}_k)(\boldsymbol{x} - \boldsymbol{\mu}_k)$$

[4] H. Jégou, F. Perronnin, M. Douze, *et al.*, "Aggregating Local Image Descriptors into Compact Codes.", *PAMI*, vol. 34, no. 9, pp. 1704–1716, Sep. 2012.

CNN Training

- Test error on independent validation set



- SGD w. Nesterov momentum 0.9, weight decay 10^{-4}, learning rate schedule

- Test with VLAD ($K = 100$, power normalization)

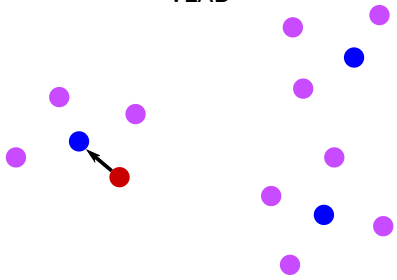
Method	mAP
LeNet	86.75
ResNet-8	88.39
ResNet-22	**89.86**

VLAD vs. Triangulation Embedding

VLAD vs. Triangulation Embedding

$$\mathcal{X} = \{\boldsymbol{x}_i \in \mathbb{R}^D, i = 1, \ldots, T\},\ \mathcal{D} = \{\boldsymbol{\mu}_k \in \mathbb{R}^D, k = 1, \ldots, K\}$$

VLAD

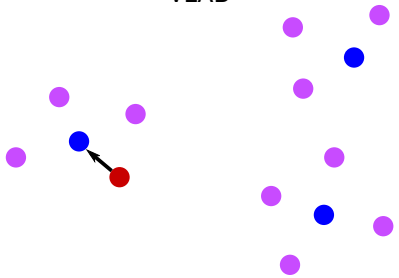


$$\phi_{\text{VLAD},k}(\boldsymbol{x}) = \text{NN}(\boldsymbol{x}, \boldsymbol{\mu}_k)(\boldsymbol{x} - \boldsymbol{\mu}_k)$$

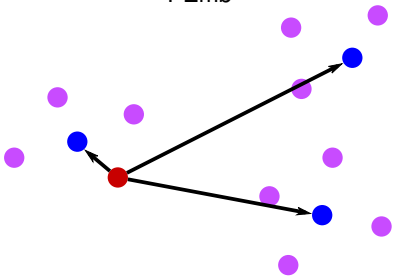
VLAD vs. Triangulation Embedding

$$\mathcal{X} = \{\boldsymbol{x}_i \in \mathbb{R}^D, i = 1, \ldots, T\}, \; \mathcal{D} = \{\boldsymbol{\mu}_k \in \mathbb{R}^D, k = 1, \ldots, K\}$$

VLAD



T-Emb



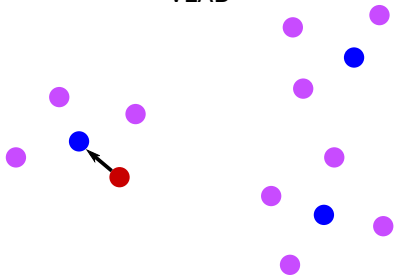
$$\phi_{\text{VLAD},k}(\boldsymbol{x}) = \text{NN}(\boldsymbol{x}, \boldsymbol{\mu}_k)(\boldsymbol{x} - \boldsymbol{\mu}_k)$$

$$\phi_{\text{T-Emb},k}(\boldsymbol{x}) = \frac{\boldsymbol{x} - \boldsymbol{\mu}_k}{\|\boldsymbol{x} - \boldsymbol{\mu}_k\|_2} \text{ [2]}$$

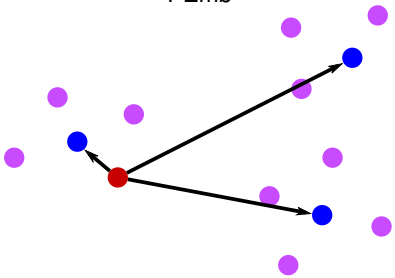
VLAD vs. Triangulation Embedding

$$\mathcal{X} = \{\boldsymbol{x}_i \in \mathbb{R}^D, i = 1, \ldots, T\}, \; \mathcal{D} = \{\boldsymbol{\mu}_k \in \mathbb{R}^D, k = 1, \ldots, K\}$$

VLAD



T-Emb



$$\phi_{\text{VLAD},k}(\boldsymbol{x}) = \text{NN}(\boldsymbol{x}, \boldsymbol{\mu}_k)(\boldsymbol{x} - \boldsymbol{\mu}_k)$$

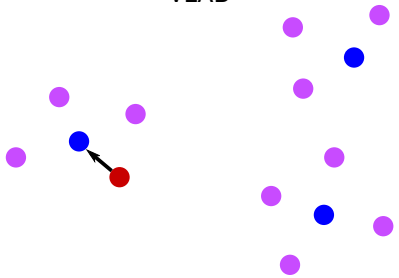
$$\phi_{\text{T-Emb},k}(\boldsymbol{x}) = \frac{\boldsymbol{x} - \boldsymbol{\mu}_k}{\|\boldsymbol{x} - \boldsymbol{\mu}_k\|_2} \text{ [2]}$$

Q: PCA whitening transformation

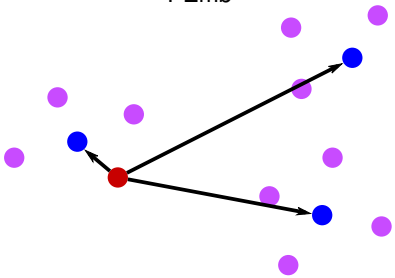
VLAD vs. Triangulation Embedding

$$\mathcal{X} = \{\boldsymbol{x}_i \in \mathbb{R}^D, i = 1, \ldots, T\},\ \mathcal{D} = \{\boldsymbol{\mu}_k \in \mathbb{R}^D, k = 1, \ldots, K\}$$

VLAD



T-Emb



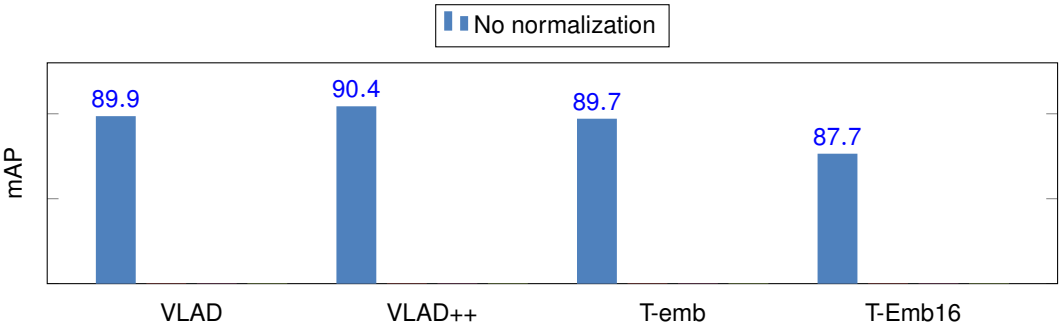
$$\phi_{\text{VLAD},k}(\boldsymbol{x}) = \text{NN}(\boldsymbol{x}, \boldsymbol{\mu}_k)(\boldsymbol{x} - \boldsymbol{\mu}_k)$$

$$\phi_{\text{VLAD++},k}(\boldsymbol{x}) = Q\,\text{NN}(\boldsymbol{x}, \boldsymbol{\mu}_k)(\boldsymbol{x})\frac{\boldsymbol{x} - \boldsymbol{\mu}_k}{\|\boldsymbol{x} - \boldsymbol{\mu}_k\|}\,(\approx[5])$$

$$\phi_{\text{T-Emb},k}(\boldsymbol{x}) = Q\frac{\boldsymbol{x} - \boldsymbol{\mu}_k}{\|\boldsymbol{x} - \boldsymbol{\mu}_k\|_2}\ [2]$$

Q: PCA whitening transformation

VLAD vs. Triangulation Embedding

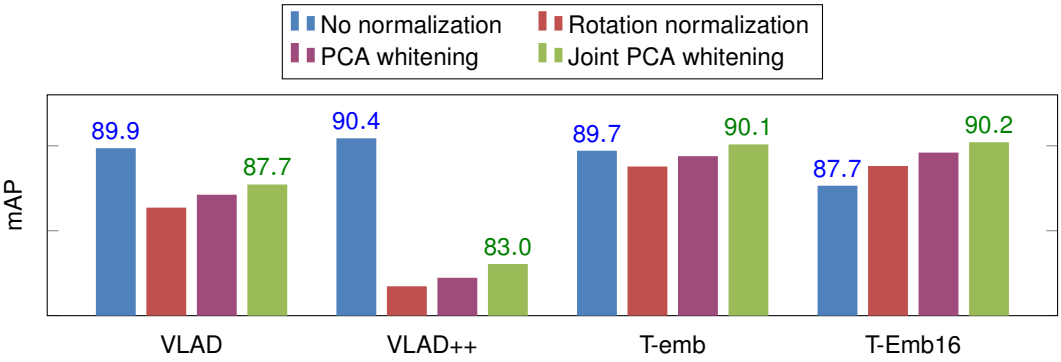


- T-Embedding not better than VLAD++

➔ Benefit from PCA whitening and residual normalization

➔ Not from triangulation

VLAD vs. Triangulation Embedding



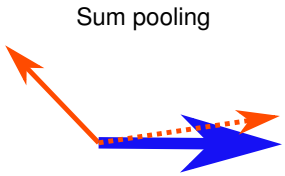
- T-Embedding not better than VLAD++

➔ Benefit from PCA whitening and residual normalization

➔ Not from triangulation

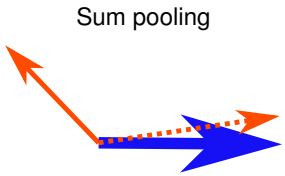
Sum Pooling vs. Generalized Max Pooling

Visual Burstiness



- Unrelated descriptors produce interference
- Frequent descriptors dominate similarity

Visual Burstiness



- Unrelated descriptors produce interference
- Frequent descriptors dominate similarity

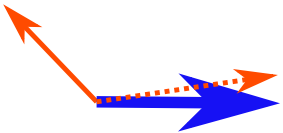
➔ Choose better embedding

➔ Normalize encoding

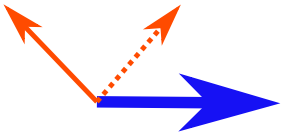
  - Power normalization
  - Intra normalization
  - ...

Visual Burstiness

Sum pooling



Generalized max pooling [3]



- Unrelated descriptors produce interference
- Frequent descriptors dominate similarity

➔ Choose better embedding

➔ Normalize encoding

- Power normalization
- Intra normalization
- ...

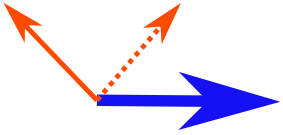
➔ Balance pooling

Generalized Max Pooling

- Seek encoding $\boldsymbol{\xi}$ which weights each embedding ϕ

$$\boldsymbol{\xi} = \sum_{\boldsymbol{x} \in \mathcal{X}} \alpha(\boldsymbol{x}) \phi(\boldsymbol{x}) = \boldsymbol{\Phi} \boldsymbol{\alpha}$$

Generalized max pooling [3]



Generalized Max Pooling

- Seek encoding $\boldsymbol{\xi}$ which weights each embedding ϕ

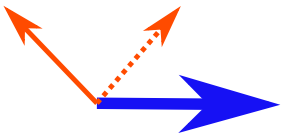
$$\boldsymbol{\xi} = \sum_{\boldsymbol{x} \in \mathcal{X}} \alpha(\boldsymbol{x}) \phi(\boldsymbol{x}) = \boldsymbol{\Phi} \boldsymbol{\alpha}$$

- Max pooling: equally similar to frequent and rare patches

➔ Enforce similarity between any patch encoding and aggregated representation to be constant

$$\boldsymbol{\Phi}^\top \boldsymbol{\xi}_{\text{gmp}} = \mathbf{1}_n,$$

Generalized max pooling [3]



Generalized Max Pooling

- Seek encoding $\boldsymbol{\xi}$ which weights each embedding ϕ

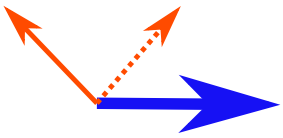
$$\boldsymbol{\xi} = \sum_{\boldsymbol{x} \in \mathcal{X}} \alpha(\boldsymbol{x}) \phi(\boldsymbol{x}) = \boldsymbol{\Phi} \boldsymbol{\alpha}$$

- Max pooling: equally similar to frequent and rare patches

➔ Enforce similarity between any patch encoding and aggregated representation to be constant

$$\boldsymbol{\Phi}^\top \boldsymbol{\xi}_{\text{gmp}} = \mathbf{1}_n,$$

Generalized max pooling [3]

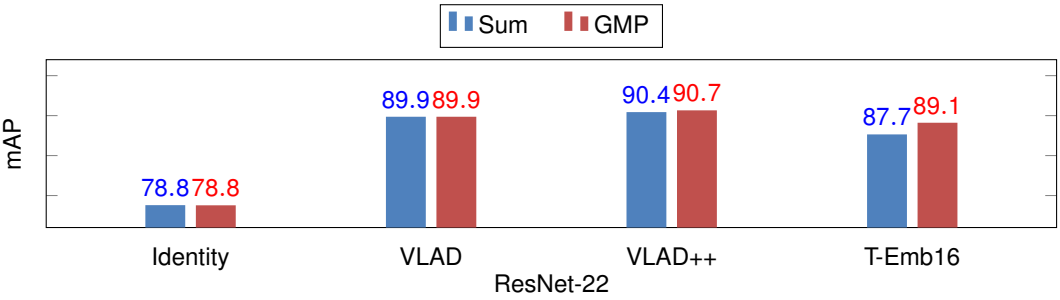


➔ Optimization problem can be cast as a ridge regression problem

$$\boldsymbol{\xi}_{\text{gmp}} = \underset{\boldsymbol{\xi}}{\operatorname{argmin}} \|\boldsymbol{\Phi}^\top \boldsymbol{\xi} - \mathbf{1}_n\|^2 + \lambda \|\boldsymbol{\xi}\|^2 \ ,$$

$\lambda \to 0$: max pooling
$\lambda \to \infty$: sum pooling

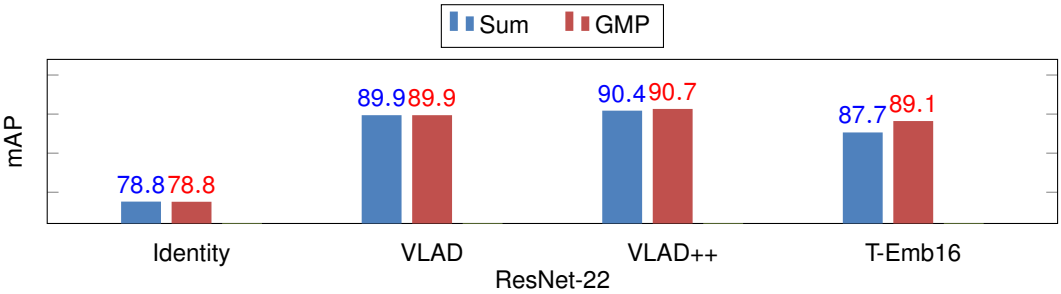
Sum Pooling vs. Generalized Max Pooling



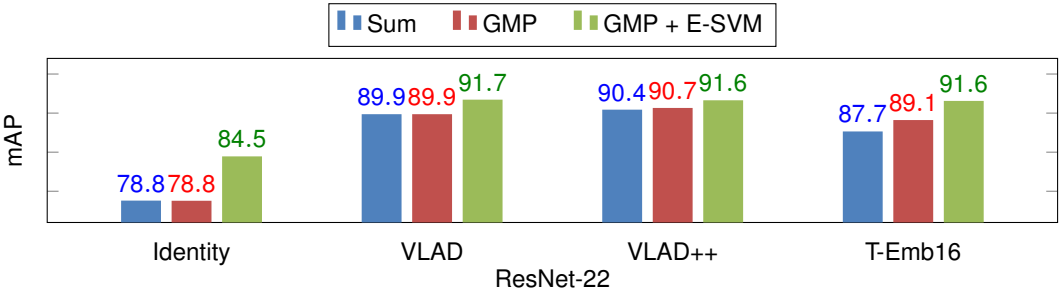
- GMP gives only slight improvements

Further Impovements

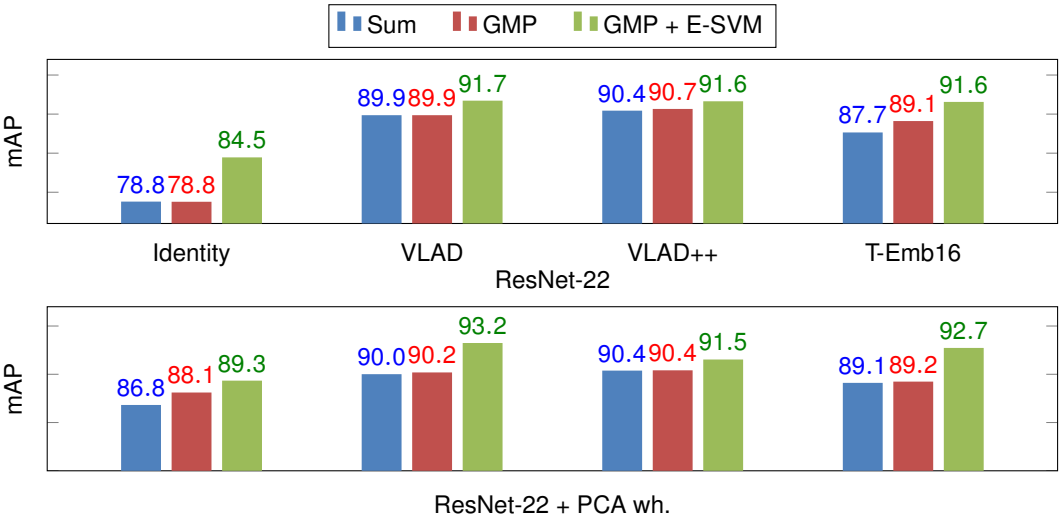
Exemplar SVMs + (local descriptor) PCA whitening



Exemplar SVMs + (local descriptor) PCA whitening



Exemplar SVMs + (local descriptor) PCA whitening



Comparison with State of the Art

Comparison with State of the Art

Method	Top-1	H-2	H-3	S-5	S-10	mAP
Fiel '15	96.8	42.3	23.1	98.9	99.4	
Christlein et al. '15	99.4	81.0	61.8	99.6	99.7	88.0
Tang & Wu '16	99.0	84.4	68.1	99.2	99.6	
Christlein et al. '17	**99.7**	84.8	63.5	**99.8**	99.8	89.4
Mohammed et al. '17	97.9					
VLAD + GMP + E-SVM	99.6	**89.8**	**77.0**	**99.8**	**99.9**	**93.2** $\pm$0.14

(a) ICDAR'13

Comparison with State of the Art

Method	Top-1	H-2	H-3	H-4	S-5	S-10	mAP
Tang & Wu '16	**99.7**	**99.0**	**97.9**	93.0	**99.8**	**100**	–
Christlein '17	99.2	98.4	97.1	93.6	99.6	99.7	98.0
VLAD + GMP + E-SVM	99.5	**99.0**	97.7	**94.5**	99.6	99.8	**98.4**

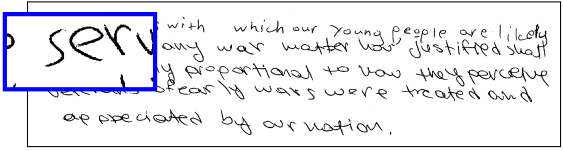
(a) CVL

Method	Top-1	H-2	H-3	S-5	S-10	mAP
Christlein '17	99.5	96.5	92.5	99.5	99.5	97.2
VLAD + GMP + E-SVM	**99.6**	**97.6**	**94.5**	**99.7**	**99.7**	**98.0**

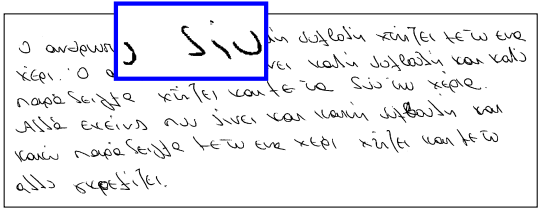
(b) KHATT

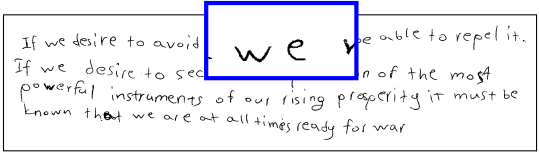
Failures

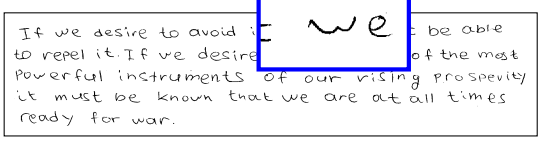
Query

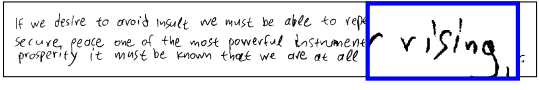


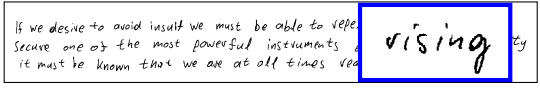
Top-1





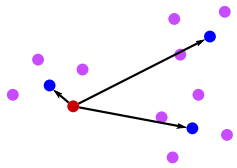


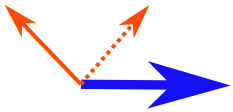




Conclusion

Conclusion

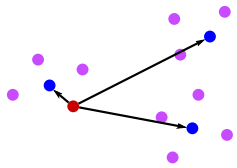


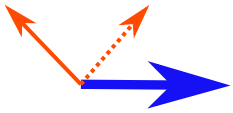


Summary

- Investigated two popular encoding techniques
  - ➔ T-Embedding and VLAD perform similarly
- Investigated generalized max pooling, PCA whitening and the combination with E-SVMs
- Improved state of the art in writer recognition (ICDAR'13,KHATT)

Conclusion





Summary

- Investigated two popular encoding techniques
  - ➔ T-Embedding and VLAD perform similarly
- Investigated generalized max pooling, PCA whitening and the combination with E-SVMs
- Improved state of the art in writer recognition (ICDAR'13,KHATT)

Outlook

- Try activations from other layers
- Incorporate text detection into the pipeline

Questions?

References

References I

[1] G. Louloudis, B. Gatos, N. Stamatopoulos, and A. Papandreou, “ICDAR 2013 Competition on Writer Identification”, in *ICDAR*, Washington DC, NY, Aug. 2013, pp. 1397–1401.

[2] H. Jégou and A. Zisserman, “Triangulation Embedding and Democratic Aggregation for Image Search”, in *CVPR*, Columbus, Jun. 2014, pp. 3310–3317.

[3] N. Murray, H. Jegou, F. Perronnin, and A. Zisserman, “Interferences in Match Kernels”, *TPAMI*, vol. 39, no. 9, pp. 1797–1810, Oct. 2016. arXiv: 1611.08194.

[4] H. Jégou, F. Perronnin, M. Douze, J. Sánchez, P. Pérez, and C. Schmid, “Aggregating Local Image Descriptors into Compact Codes.”, *PAMI*, vol. 34, no. 9, pp. 1704–1716, Sep. 2012.

[5] J. Delhumeau, P.-H. Gosselin, H. Jégou, and P. Pérez, “Revisiting the VLAD Image Representation”, in *21st ACM International Conference on Multimedia - MM ’13*, Barcelona: ACM, Oct. 2013, pp. 653–656.