



FAU

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



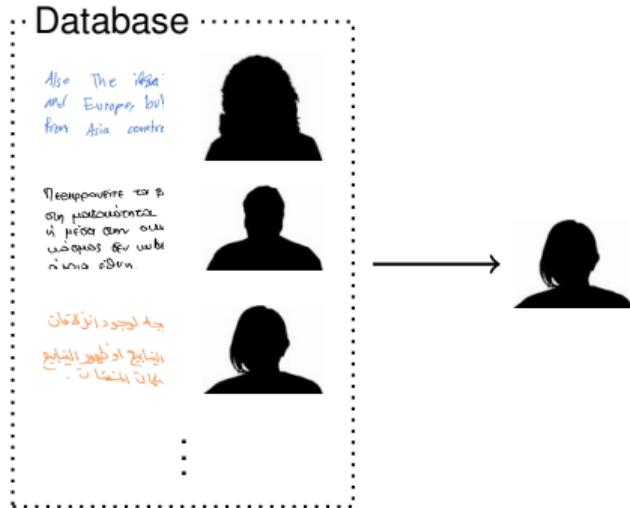


Writer Recognition



Writer Identification vs. Writer Retrieval

If we desire to
desire to secure
rising prosperity
for war.



Writer Identification

Given:

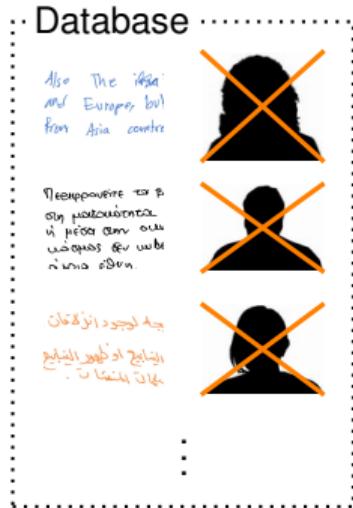
- Query document
 - Documents of known writers

Wanted:

- Writer-ID

Writer Identification vs. Writer Retrieval

If we desire to
desire to secure
rising prosperity
for war.



Rank

k

1

2

3

Q



Unappropriate for the publication in press own own unappropriate for war in Asia.

The willingness in any war no to how they are appreciated by

○ unappropriate
○ unappropriate no
○ unappropriate yes pt
○ unappropriate yes qd

⋮

○ unappropriate no
○ unappropriate no
○ unappropriate yes pt
○ unappropriate yes qd

Writer Retrieval

Given:

- Query document
- Documents of (possibly unknown) writers

Wanted:

- Most similar documents

Contemporary Datasets

The willingness with which
in any war no matter how
to how they perceive veterans
appreciated by our nation.

Περιποντή τα βίβλια εστιά να
σημαντικότερη ως οι γραδού
η μέσα σε αυτήν. Απότι ο
νέος δεν περιέχει νέα

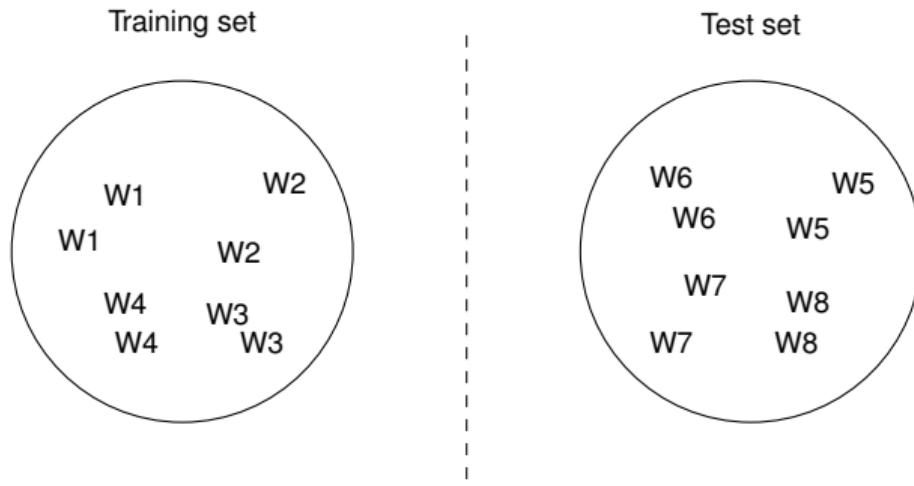
ICDAR13 benchmark dataset¹

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

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

¹G. 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
⇒ No training for a specific writer possible!



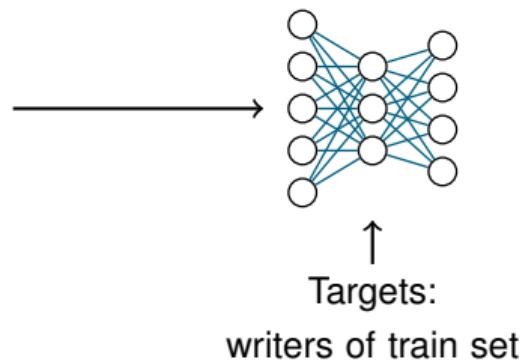
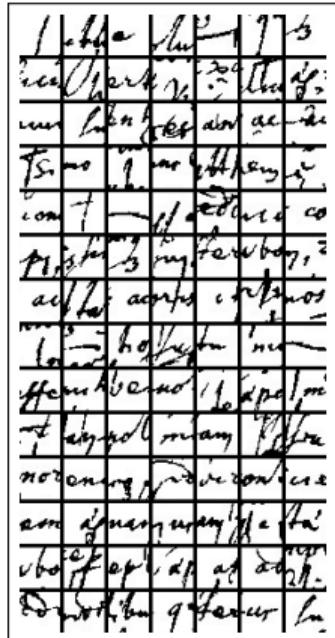
FAU

FRIEDRICH-ALEXANDER-
UNIVERSITÄT
ERLANGEN-NÜRNBERG
SCHOOL OF ENGINEERING

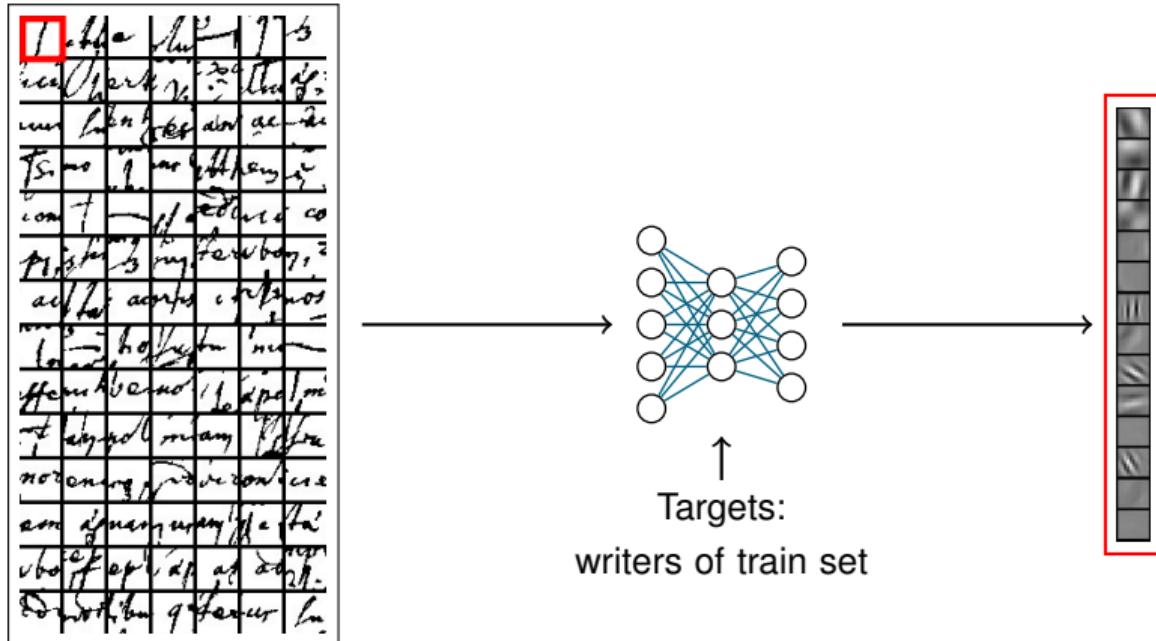
Typical Methodology For Deep Learning Feature Extraction



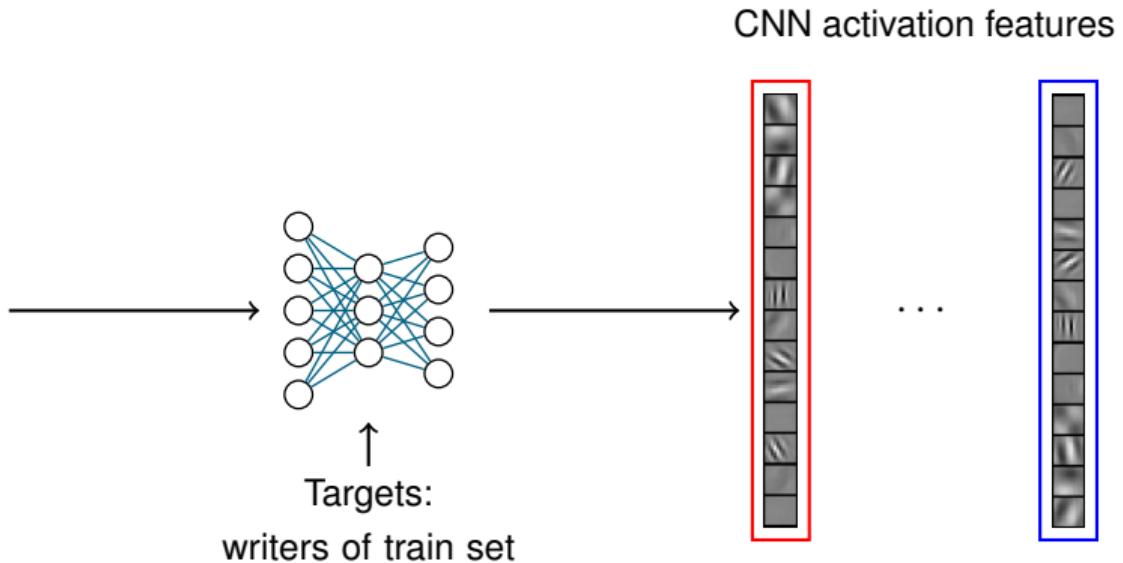
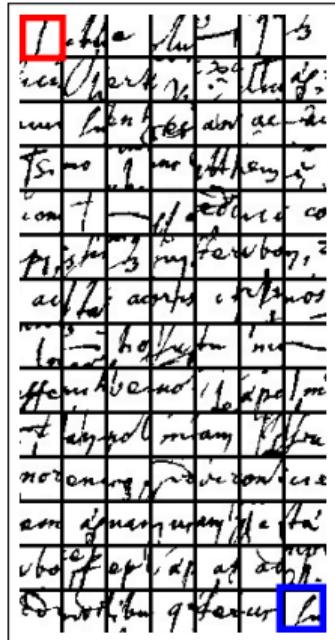
CNN Activation Features



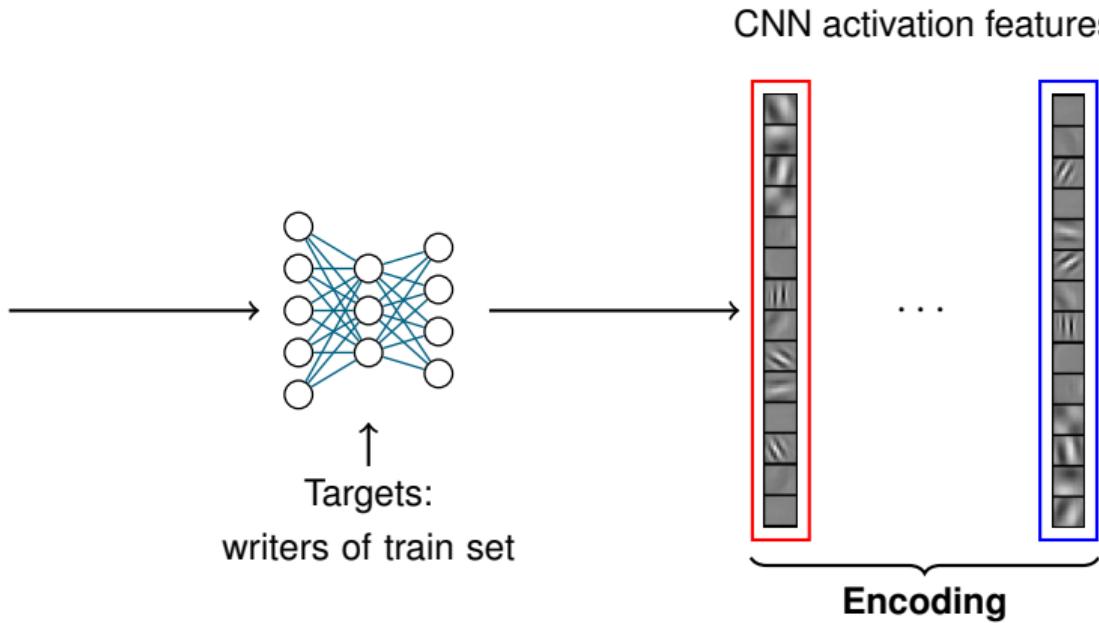
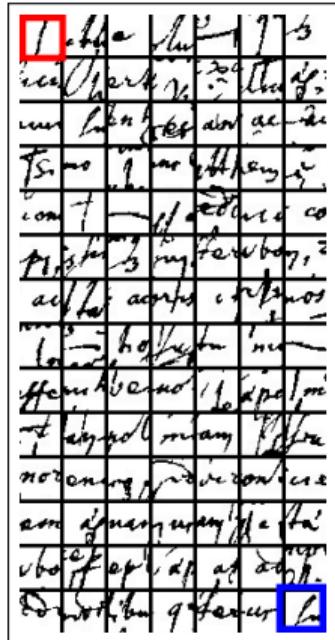
CNN Activation Features



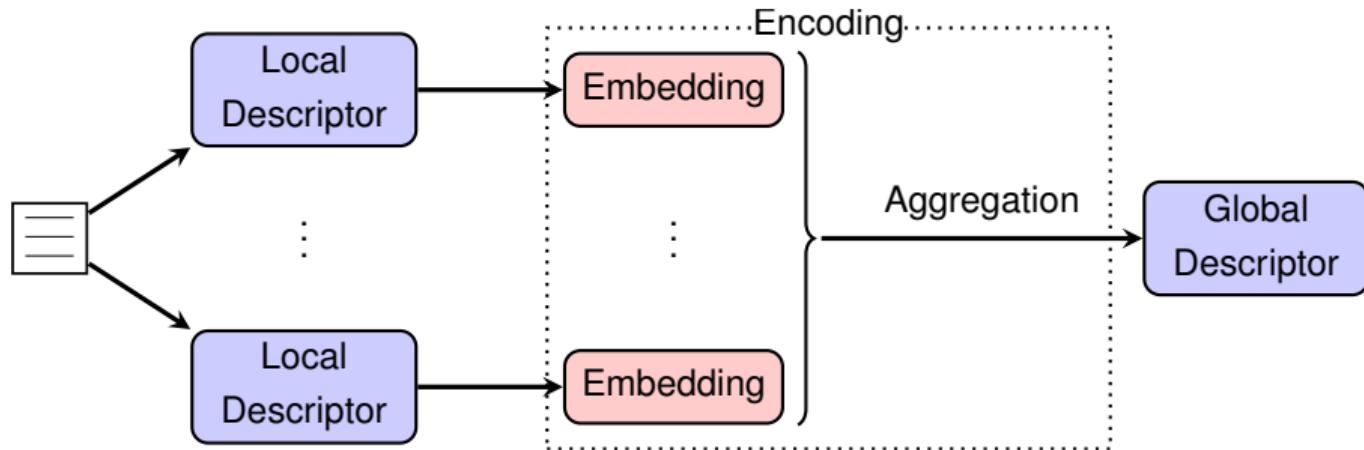
CNN Activation Features



CNN Activation Features



Encoding



- Embedding: Fisher vectors, GMM supervectors, VLAD, triangulation embedding²
- Aggregation: Sum pooling, democratic aggregation, generalized max-pooling³

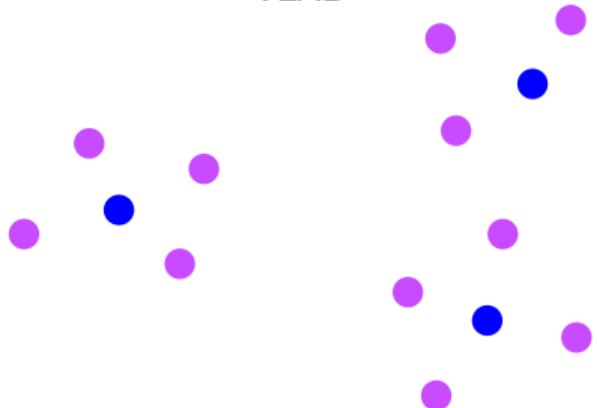
²H. Jegou and A. Zisserman, "Triangulation Embedding and Democratic Aggregation for Image Search", in *CVPR*, Columbus, Jun. 2014, pp. 3310–3317.

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

VLAD⁴

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

VLAD

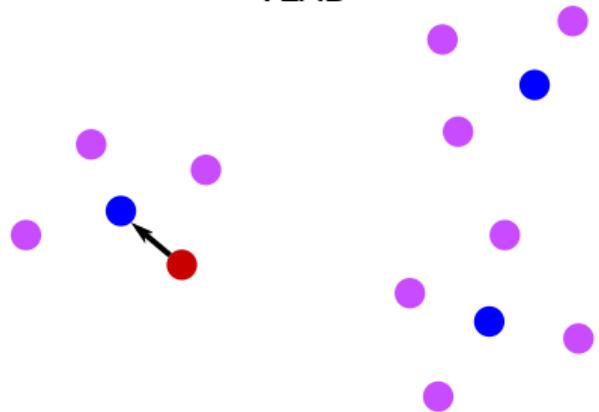


⁴H. Jégou, F. Perronnin, M. Douze, et al., "Aggregating Local Image Descriptors into Compact Codes.", *PAMI*, vol. 34, no. 9, pp. 1704–1716, Sep. 2012.

VLAD⁴

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

VLAD

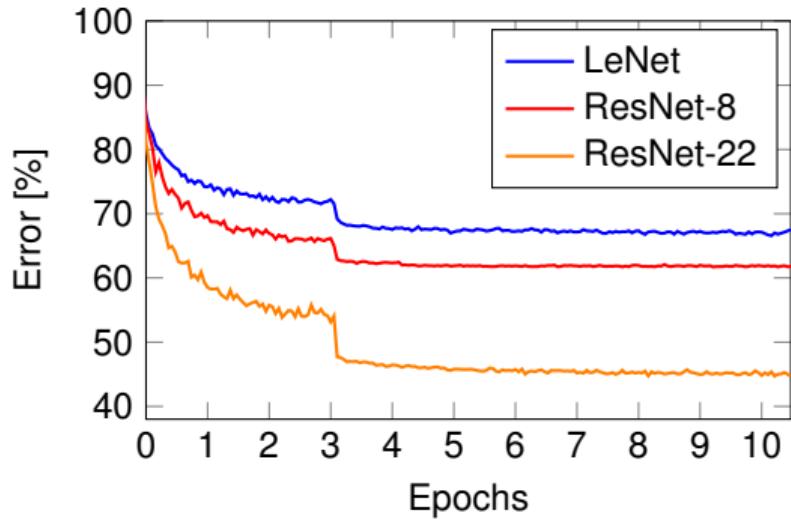


$$\phi_{\text{VLAD},k}(\mathbf{x}) = \text{NN}(\mathbf{x}, \boldsymbol{\mu}_k)(\mathbf{x} - \boldsymbol{\mu}_k)$$

⁴H. 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



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

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

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



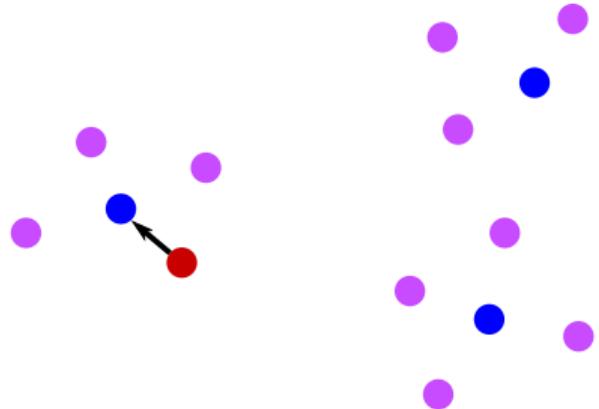
VLAD vs. Triangulation Embedding



VLAD vs. Triangulation Embedding

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

VLAD

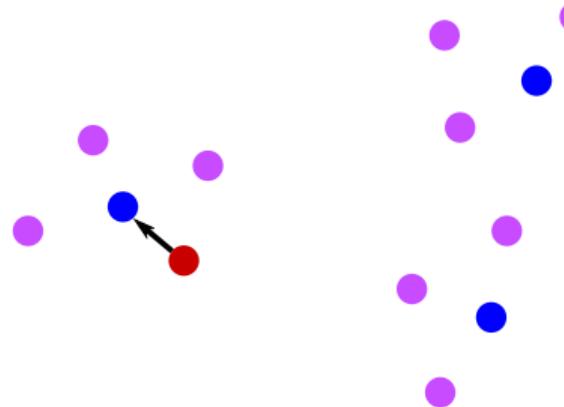


$$\phi_{\text{VLAD},k}(\mathbf{x}) = \text{NN}(\mathbf{x}, \boldsymbol{\mu}_k)(\mathbf{x} - \boldsymbol{\mu}_k)$$

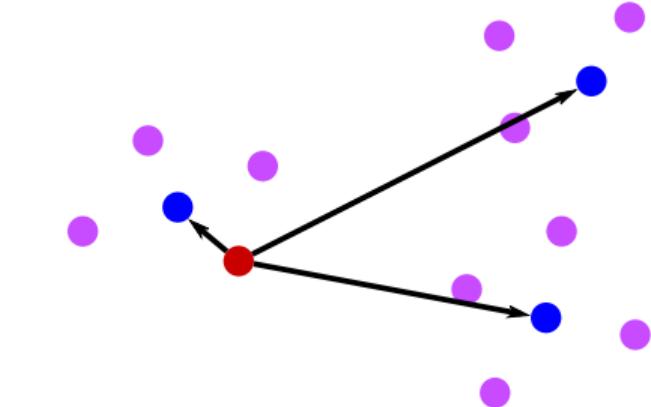
VLAD vs. Triangulation Embedding

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

VLAD



T-Emb



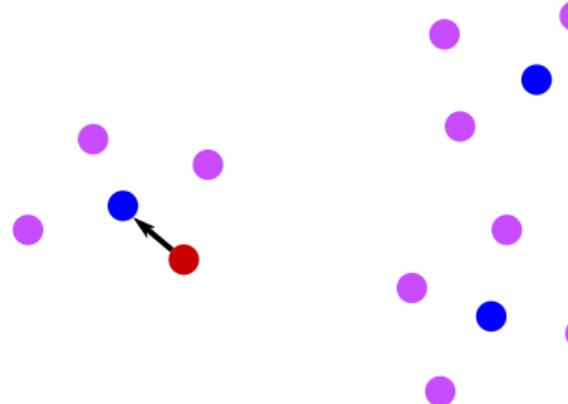
$$\phi_{\text{VLAD},k}(\mathbf{x}) = \text{NN}(\mathbf{x}, \boldsymbol{\mu}_k)(\mathbf{x} - \boldsymbol{\mu}_k)$$

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

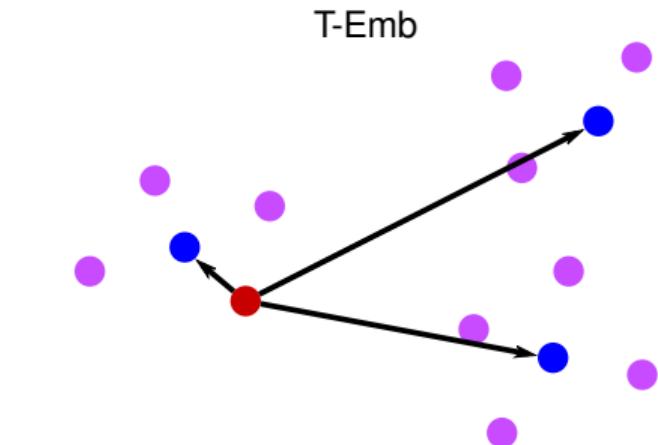
VLAD vs. Triangulation Embedding

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

VLAD



T-Emb



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

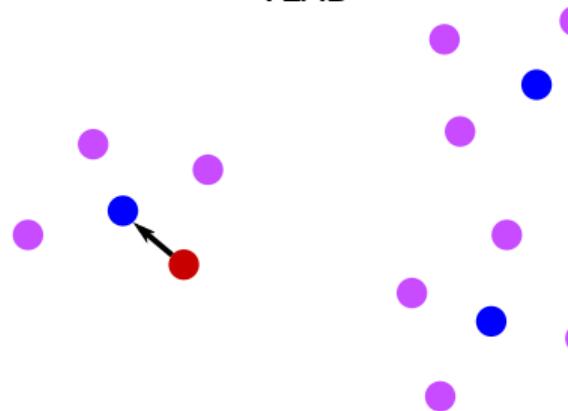
$$\phi_{\text{T-Emb},k}(\mathbf{x}) = \frac{\mathbf{x} - \boldsymbol{\mu}_k}{\|\mathbf{x} - \boldsymbol{\mu}_k\|_2} [2]$$

Q: PCA whitening transformation

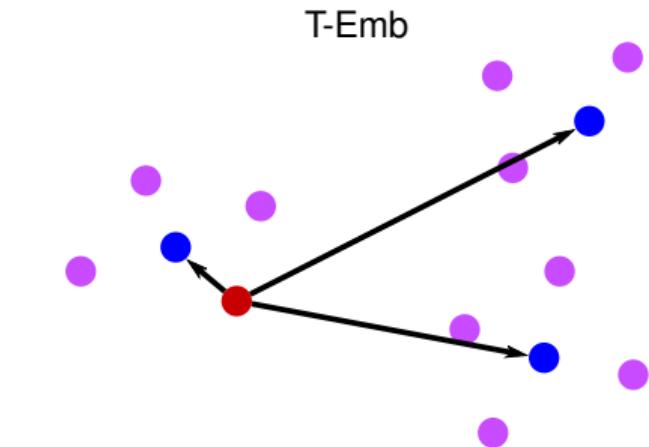
VLAD vs. Triangulation Embedding

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

VLAD



T-Emb



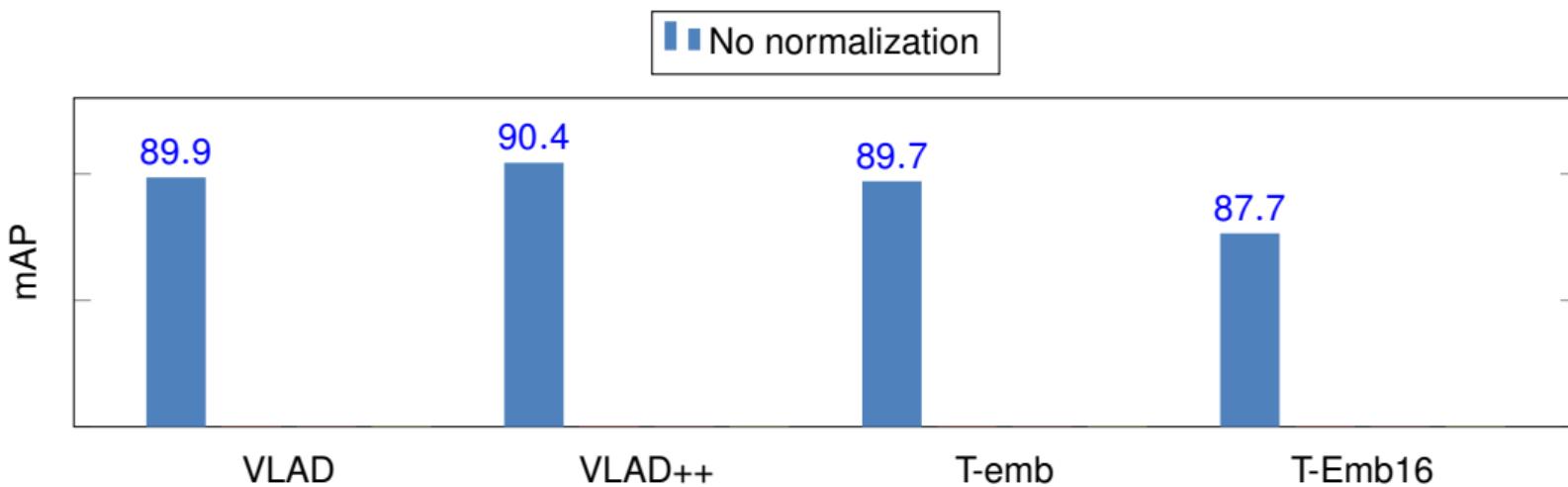
$$\phi_{\text{VLAD},k}(\mathbf{x}) = \text{NN}(\mathbf{x}, \boldsymbol{\mu}_k)(\mathbf{x} - \boldsymbol{\mu}_k)$$

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

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

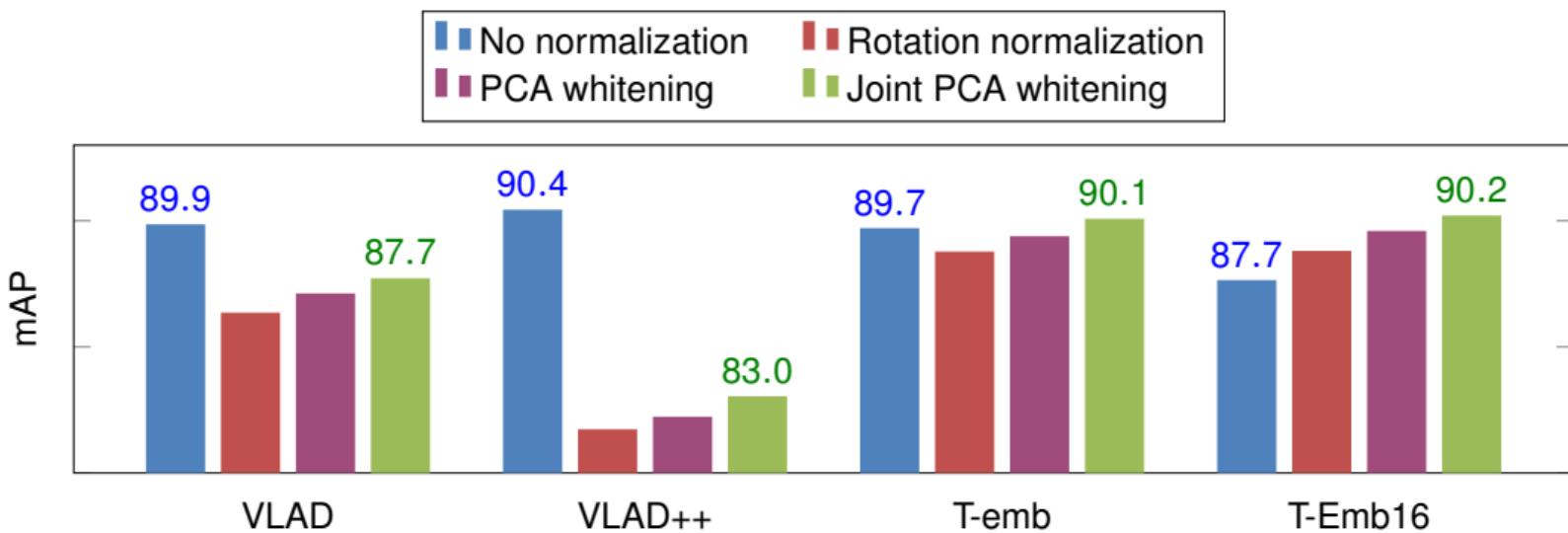
Q: PCA whitening transformation

VLAD vs. Triangulation Embedding



- T-Embedding not better than VLAD++
 - Benefit from PCA whitening and residual normalization
 - Not from triangulation

VLAD vs. Triangulation Embedding



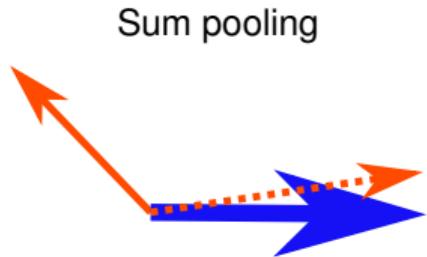
- T-Embedding not better than VLAD++
- Benefit from PCA whitening and residual normalization
- Not from triangulation



Sum Pooling vs. Generalized Max Pooling

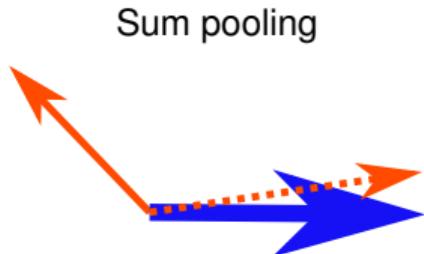


Visual Burstiness



- Unrelated descriptors produce interference
- Frequent descriptors dominate similarity

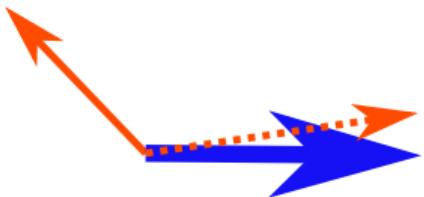
Visual Burstiness



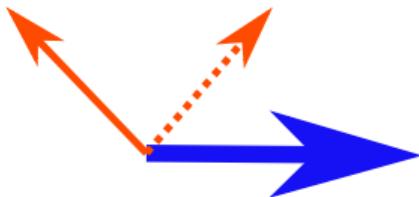
- Unrelated descriptors produce interference
- Frequent descriptors dominate similarity
- Choose better embedding
- Normalize encoding
 - Power normalization
 - Intra normalization
 - ...

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
 - ...

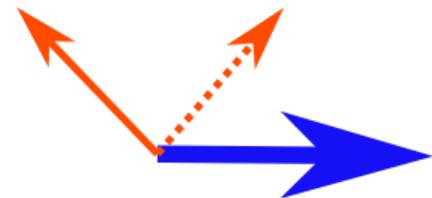
→ Balance pooling

Generalized Max Pooling

- Seek encoding ξ which weights each embedding ϕ

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

Generalized max pooling [3]



Generalized Max Pooling

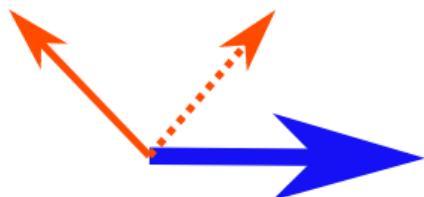
- Seek encoding ξ which weights each embedding ϕ

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

- Max pooling: equally similar to frequent and rare patches
- Enforce similarity between any patch encoding and aggregated representation to be constant

$$\Phi^\top \xi_{\text{gmp}} = \mathbf{1}_n,$$

Generalized max pooling [3]



Generalized Max Pooling

- Seek encoding ξ which weights each embedding ϕ

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

- Max pooling: equally similar to frequent and rare patches
- Enforce similarity between any patch encoding and aggregated representation to be constant

$$\Phi^\top \xi_{\text{gmp}} = \mathbf{1}_n,$$

Generalized max pooling [3]

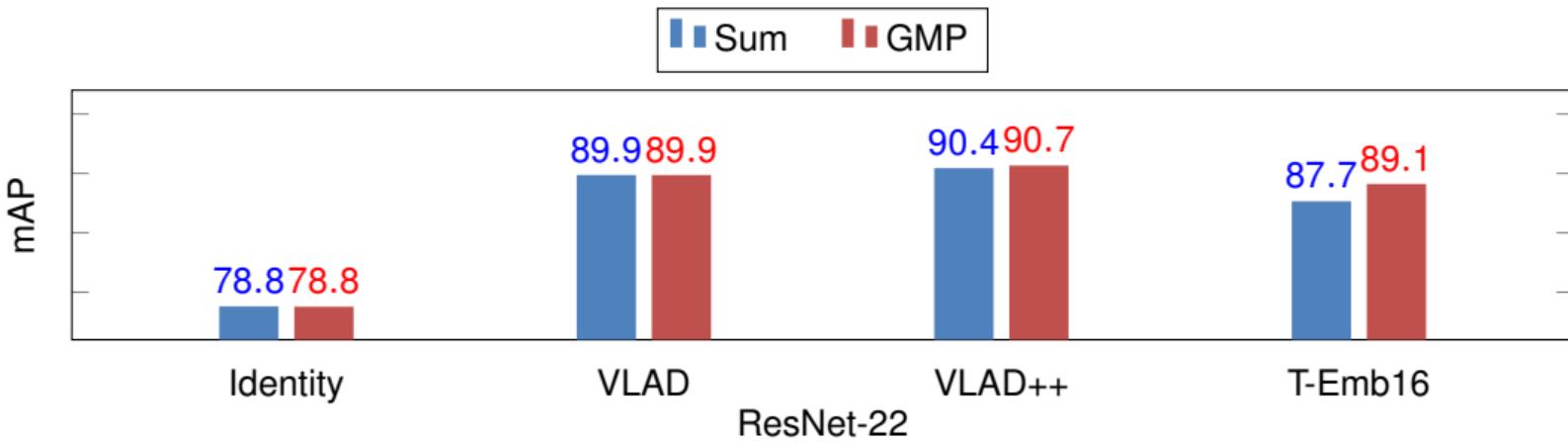


→ Optimization problem can be cast as a ridge regression problem

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

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

Sum Pooling vs. Generalized Max Pooling



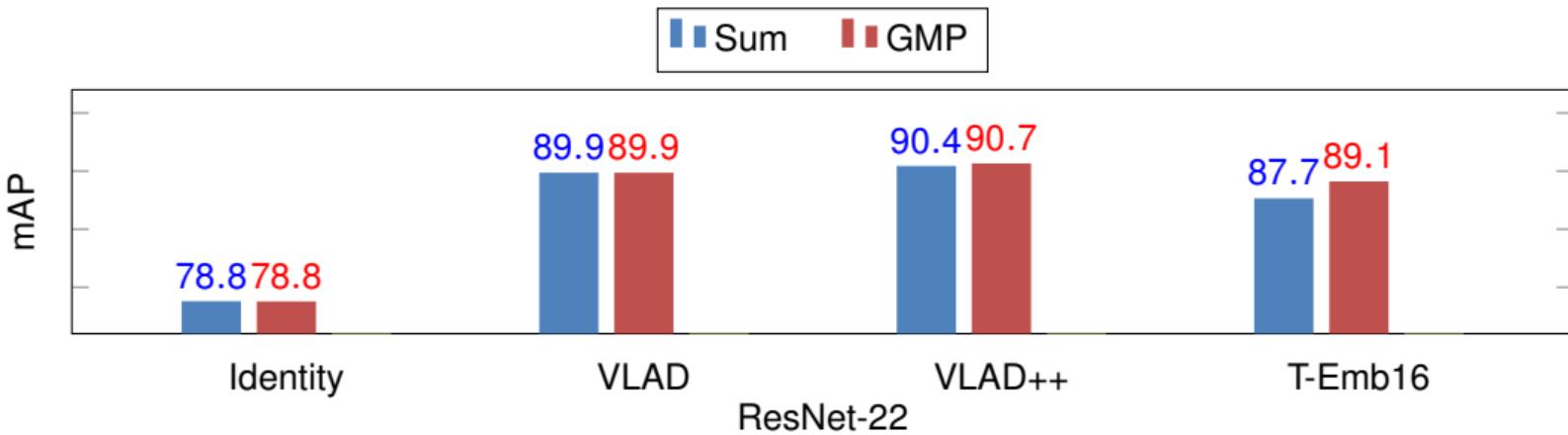
- GMP gives only slight improvements



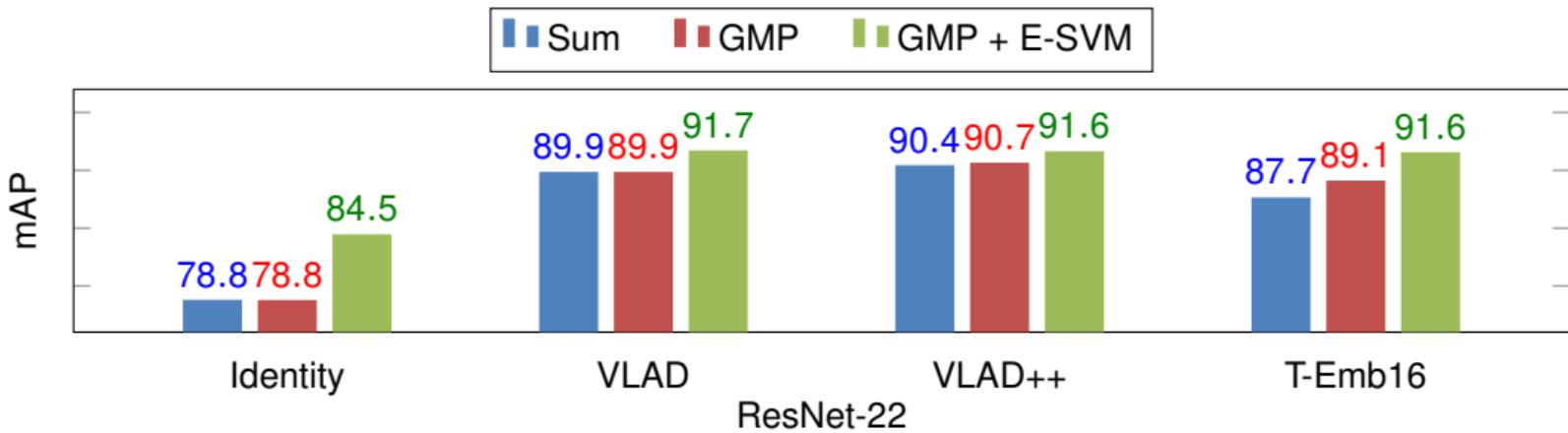
Further Improvements



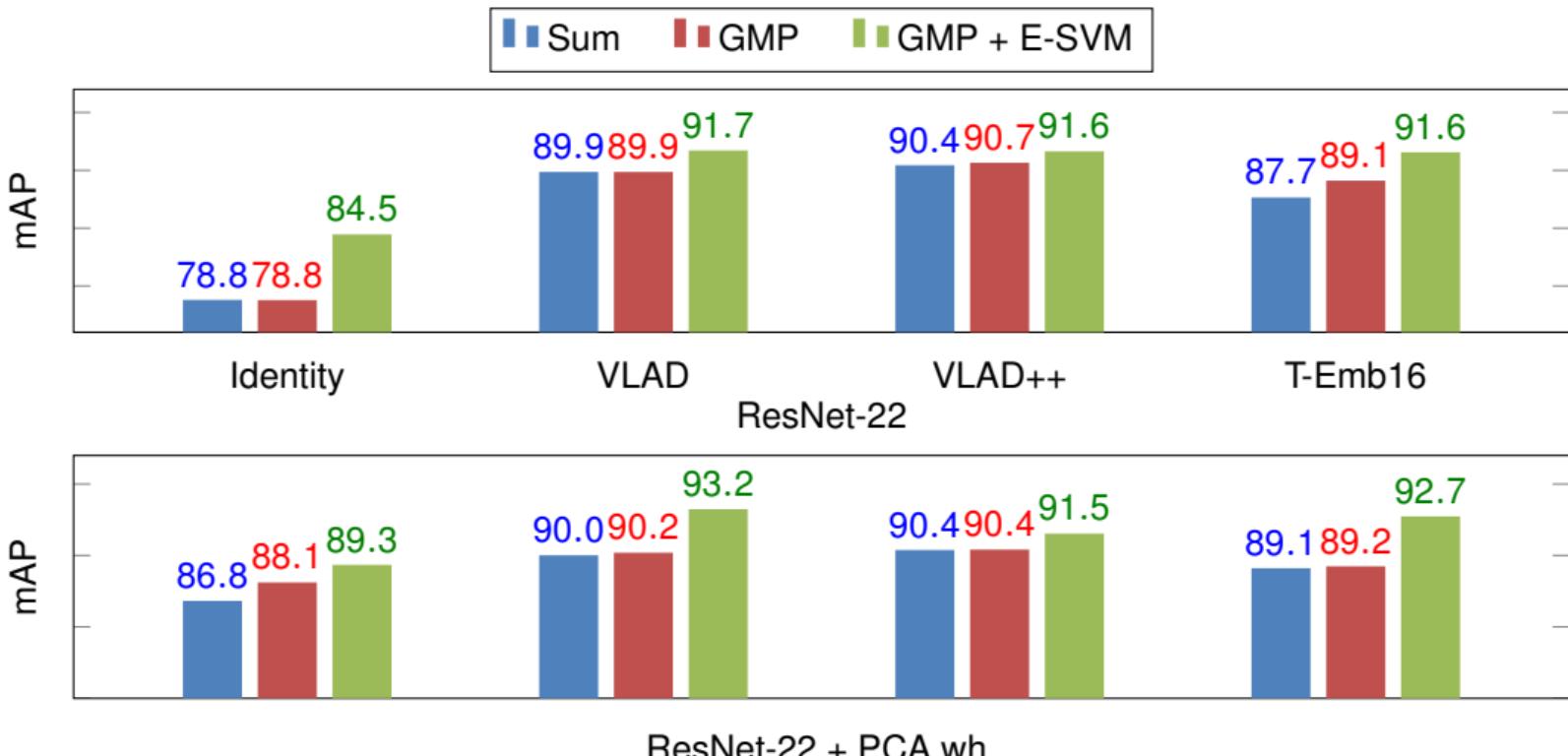
Exemplar SVMs + (local descriptor) PCA whitening



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 ±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

P SERV with which our young people are likely any war whether how justified shall y proportional to how they perceive demands of early wars were treated and are expected by our nation.

Top-1

J andwa
kelp. O e
napo seite xin[er] wante ie Sisiu xape.
Alle exius our sive van vanin w[er]bun van
van napo seite few we xepi xin[er] wante
allo expe[te].

If we desire to avoid we are able to repel it.
If we desire to see one of the most powerful instruments of our rising prosperity it must be known that we are at all times ready for war

If we desire to avoid it to repel it. If we desire powerful instruments of our rising prosperity it must be known that we are at all times ready for war.

If we desire to avoid insult we must be able to repel secure peace one of the most powerful instruments of prosperity it must be known that we are at all times ready for war

If we desire to avoid insult we must be able to repel secure one of the most powerful instruments of prosperity it must be known that we are at all times ready for war



Conclusion



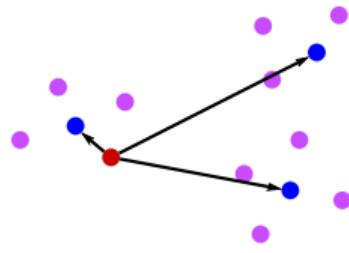
Conclusion



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

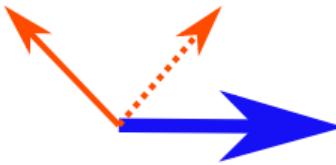


Summary

- 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.