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

Writer Identification

Given:
- Query document
- Documents of known writers

Wanted:
- Writer-ID

If we desire to secure rising prosperity for war.

Source: ICDAR 17 dataset, QUWI15 dataset, freepik.com

Vincent Christlein | Encoding CNN Activations for Writer Recognition
April 27th, 2018
Writer Identification vs. Writer Retrieval

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

**Database**

**Rank**

1 2 3 4

**Q**

**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.

ICDAR13 benchmark dataset

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

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

---

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
⇒ 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

Targets: writers of train set
CNN Activation Features

 Targets:
 writers of train set

CNN activation features
CNN Activation Features

Targets: writers of train set

Encoding

CNN activation features
**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

\[ \chi = \{ x_i \in \mathbb{R}^D, i = 1, \ldots, T \}, \quad \mathcal{D} = \{ \mu_k \in \mathbb{R}^D, k = 1, \ldots, K \} \]

\[ \text{VLAD} \]

---

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.
\[ \mathcal{X} = \{ x_i \in \mathbb{R}^D, i = 1, \ldots, T \}, \quad \mathcal{D} = \{ \mu_k \in \mathbb{R}^D, k = 1, \ldots, K \} \]

\[ \phi_{\text{VLAD}, k}(x) = \text{NN}(x, \mu_k)(x - \mu_k) \]

---

*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

\[ \mathcal{X} = \{x_i \in \mathbb{R}^D, i = 1, \ldots, T\}, \quad \mathcal{D} = \{\mu_k \in \mathbb{R}^D, k = 1, \ldots, K\} \]

\[ \phi_{VLAD,k}(x) = \text{NN}(x, \mu_k)(x - \mu_k) \]
**VLAD vs. Triangulation Embedding**

\[ \mathcal{X} = \{ \mathbf{x}_i \in \mathbb{R}^D, i = 1, \ldots, T \}, \quad \mathcal{D} = \{ \mu_k \in \mathbb{R}^D, k = 1, \ldots, K \} \]

**VLAD**

\[ \phi_{\text{VLAD},k}(\mathbf{x}) = \text{NN}(\mathbf{x}, \mu_k)(\mathbf{x} - \mu_k) \]

**T-Emb**

\[ \phi_{\text{T-Emb},k}(\mathbf{x}) = \frac{\mathbf{x} - \mu_k}{\| \mathbf{x} - \mu_k \|_2} \quad [2] \]

\(\text{Q}\): PCA whitening transformation
VLAD vs. Triangulation Embedding

$$\mathcal{X} = \{ x_i \in \mathbb{R}^D, i = 1, \ldots, T \}, \quad \mathcal{D} = \{ \mu_k \in \mathbb{R}^D, k = 1, \ldots, K \}$$

**VLAD**

$$\phi_{\text{VLAD},k}(x) = \text{NN}(x, \mu_k)(x - \mu_k)$$

**T-Emb**

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

$Q$: PCA whitening transformation
VLAD vs. Triangulation Embedding

\[ \mathcal{X} = \{ \mathbf{x}_i \in \mathbb{R}^D, i = 1, \ldots, T \}, \quad \mathcal{D} = \{ \mathbf{\mu}_k \in \mathbb{R}^D, k = 1, \ldots, K \} \]

**VLAD**

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

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

**T-Emb**

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

\(Q\): 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

Sum pooling

- 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
    - ...

Sum pooling
Visual Burstiness

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

Generalized max pooling [3]
Generalized Max Pooling

- Seek encoding $\xi$ which weights each embedding $\phi$

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

Generalized max pooling [3]
Generalized Max Pooling

- Seek encoding $\xi$ which weights each embedding $\phi$

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

- Max pooling: equally similar to frequent and rare patches

$\Rightarrow$ Enforce similarity between any patch encoding and aggregated representation to be constant

$$\Phi^T \xi_{gmp} = 1_n,$$
Generalized Max Pooling

• Seek encoding $\xi$ which weights each embedding $\phi$

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

• Max pooling: equally similar to frequent and rare patches

$\rightarrow$ Enforce similarity between any patch encoding and aggregated representation to be constant

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

$\rightarrow$ Optimization problem can be cast as a ridge regression problem

$$\xi_{\text{gmp}} = \text{argmin}_{\xi} ||\Phi^T \xi - 1_n||^2 + \lambda ||\xi||^2,$$

$\lambda \to 0$: max pooling
$\lambda \to \infty$: sum pooling
Sum Pooling vs. Generalized Max Pooling

- GMP gives only slight improvements

- **Identity**
  - mAP: 78.8

- **VLAD**
  - mAP: 89.9

- **VLAD++**
  - mAP: 90.4

- **ResNet-22**
  - mAP: 87.7

- **T-Emb16**
  - mAP: 89.1
Further Improvements
Exemplar SVMs + (local descriptor) PCA whitening

| Method       | Sum  | GMP  |
|--------------|------|------|
| Identity     | 78.8 | 78.8 |
| VLAD         | 89.9 | 89.9 |
| VLAD++       | 90.4 | 90.7 |
| T-Emb16      | 87.7 | 89.1 |

mAP

ResNet-22
Exemplar SVMs + (local descriptor) PCA whitening

| Method      | ResNet-22 | VLAD | VLAD++ | T-Emb16 |
|-------------|-----------|------|--------|---------|
| Identity    | 78.8      | 84.5 | 91.7   | 87.7    |
| Sum         | 78.8      | 89.9 | 90.7   | 89.1    |
| GMP         | 78.8      | 89.9 | 91.7   | 91.6    |
| GMP + E-SVM | 84.5      | 91.7 | 91.6   | 91.6    |
### Exemplar SVMs + (local descriptor) PCA whitening

|         | Identity | VLAD   | VLAD++ | T-Emb16 |
|---------|----------|--------|--------|----------|
| ResNet-22 | 78.8     | 89.9   | 91.7   | 90.4     |
|          | 78.8     | 89.9   | 91.6   | 87.7     |
|          | 84.5     |        |        |          |
|          |          |        |        | 91.6     |

### mAP

|         | ResNet-22 | ResNet-22 + PCA wh. |
|---------|-----------|---------------------|
| mAP     | 86.8      | 89.3                |
|         | 88.1      | 90.0                |
|         | 89.3      | 90.2                |
|         | 93.2      | 90.4                |
|         | 93.2      | 91.5                |
|         | 91.5      | 92.7                |

|         | Sum     | GMP      | GMP + E-SVM |
|---------|---------|----------|-------------|
| ResNet-22 | 86.8    | 88.1     | 89.3        |
|          | 90.0    | 90.2     | 91.5        |
|          | 90.4    | 91.5     | 92.7        |
| T-Emb16  | 87.7    | 89.1     | 91.6        |
|          | 91.6    |          |             |
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

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

Top-1

If we desire to avoid insult we must be able to repel it. If we desire to see the instruments of our rising 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
[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.