Sparse Coding

Hyunjoong Kim

soy.lovit@gmail.com

github.com/lovit

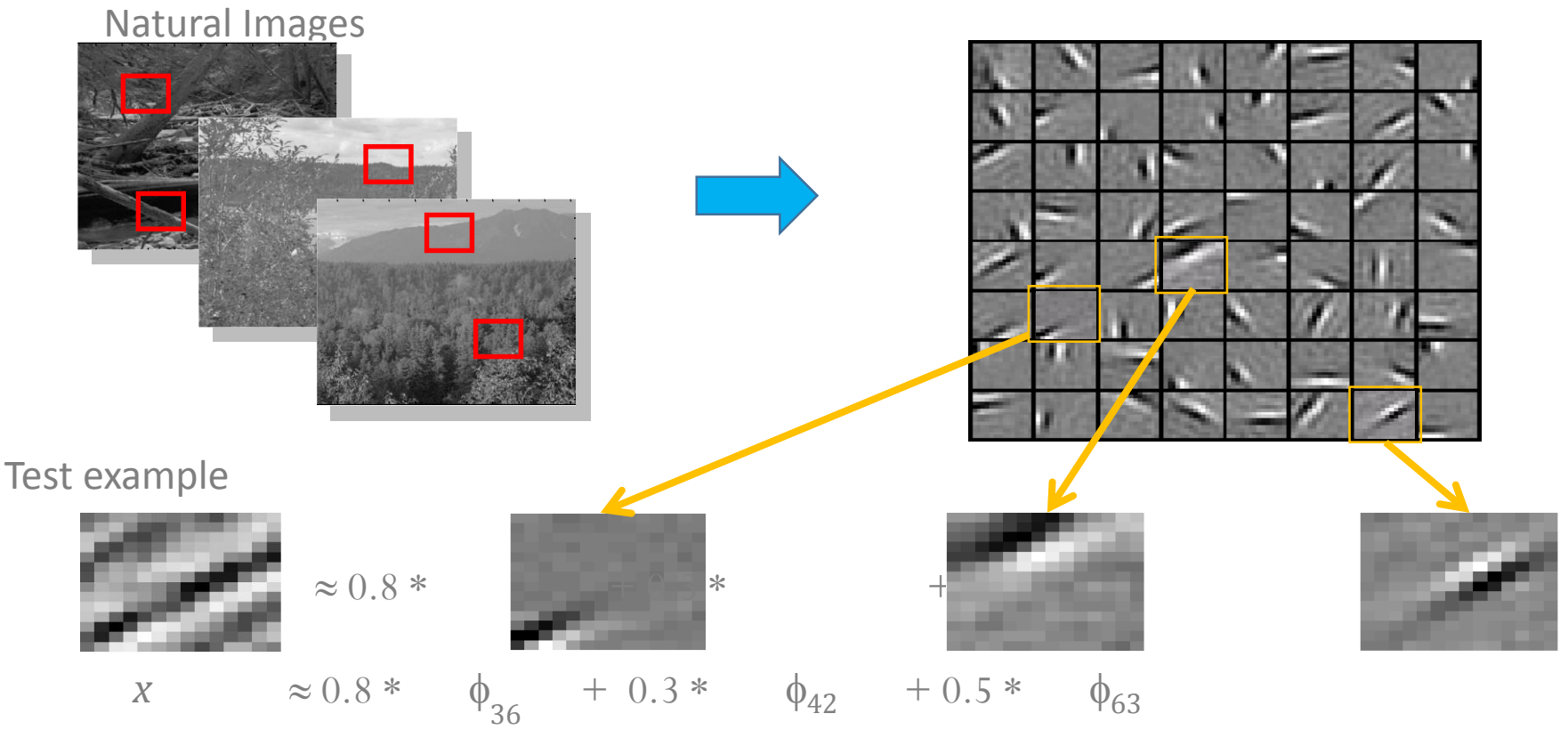
Sparse Coding

- Dictionary learning 이라고도 불리며, representation learning 을 위해 자주 이용되던 방법입니다.
  - 학습데이터 X 로부터 dictionary D 와 sparse representation β 를 학습합니다.
  - β 는 sparse vector 이기 때문에 해석이 용이합니다.

$$argmin_{D,\beta} |X - \beta^T D|^2 + \lambda|\beta|_1$$

Sparse Coding

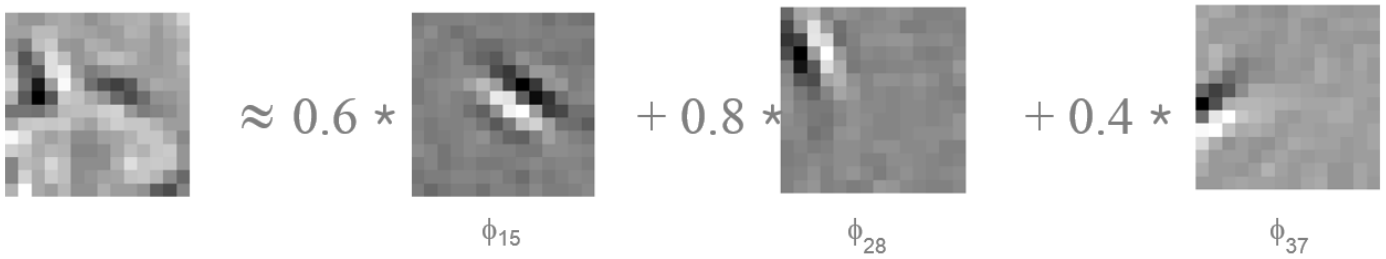
- Images 의 sparse representation 에 자주 이용되었습니다.



- http://ufldl.stanford.edu/eccv10-tutorial/

Sparse Coding

- Images 의 sparse representation 에 자주 이용되었습니다.



Represent as: [0, 0, ..., 0, **0.6**, 0, ..., 0, **0.8**, 0, ..., 0, **0.4**, ...]

Represent as: [0, 0, ..., 0, **1.3**, 0, ..., 0, **0.9**, 0, ..., 0, **0.3**, ...]

- http://ufldl.stanford.edu/eccv10-tutorial/

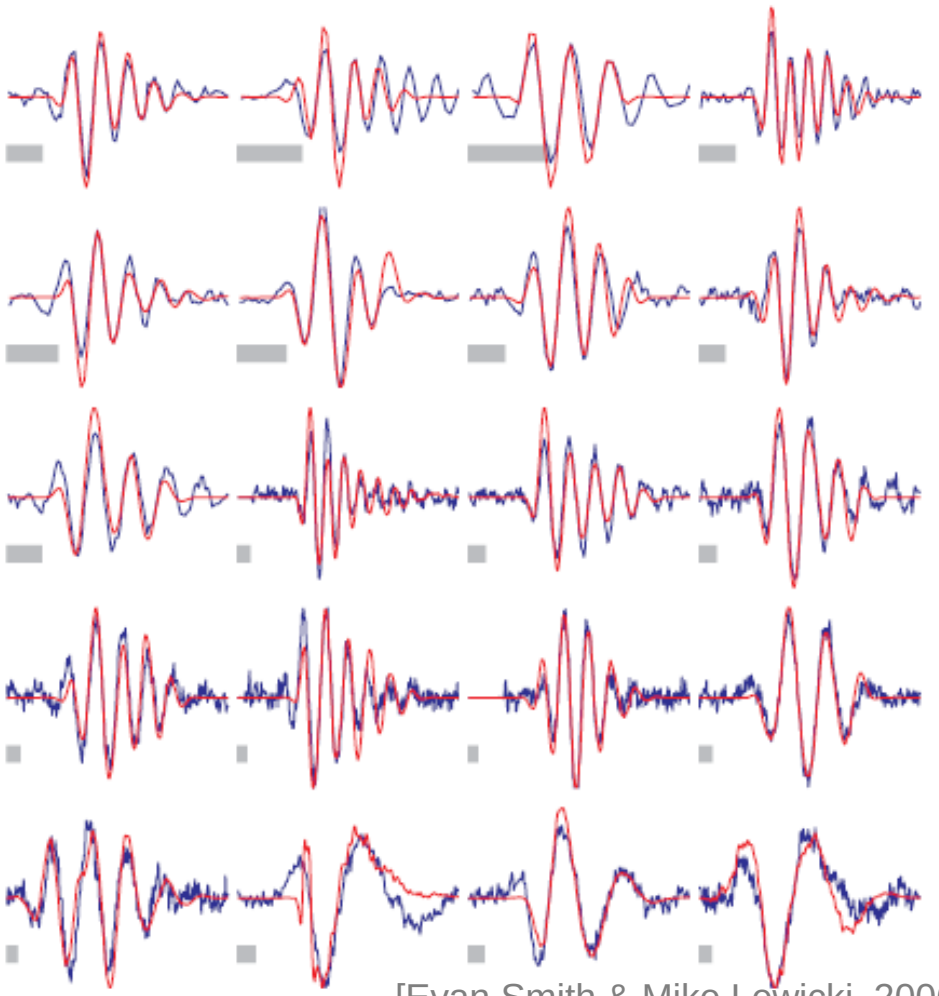
Sparse Coding

- $argmin_{D,\beta} |X - \beta^T D|^2 + \lambda|\beta|_1$
  - X 가 (n, m) 크기의 행렬일 때,
  - D 는 (p, m) 행렬이며, $p > m$ 입니다. 이를 over-completed 라 합니다.
  - β^T 는 (n, p) 의 행렬이며, X 의 새로운 sparse representation 입니다.
    - β_i 는 sparse vector 이며, β_{ij} 는 D_j 의 coefficients 입니다.

Sparse Coding

- Sparse coding 은 데이터에 주로 나타나는 패턴들을 학습하기 때문에 noise 를 제거하는 능력이 있습니다.
  - $|X - \beta^T D|^2 + \lambda|\beta|_1$ 의 $\beta^T D$ 를 이용하면 denoising 이 가능합니다.

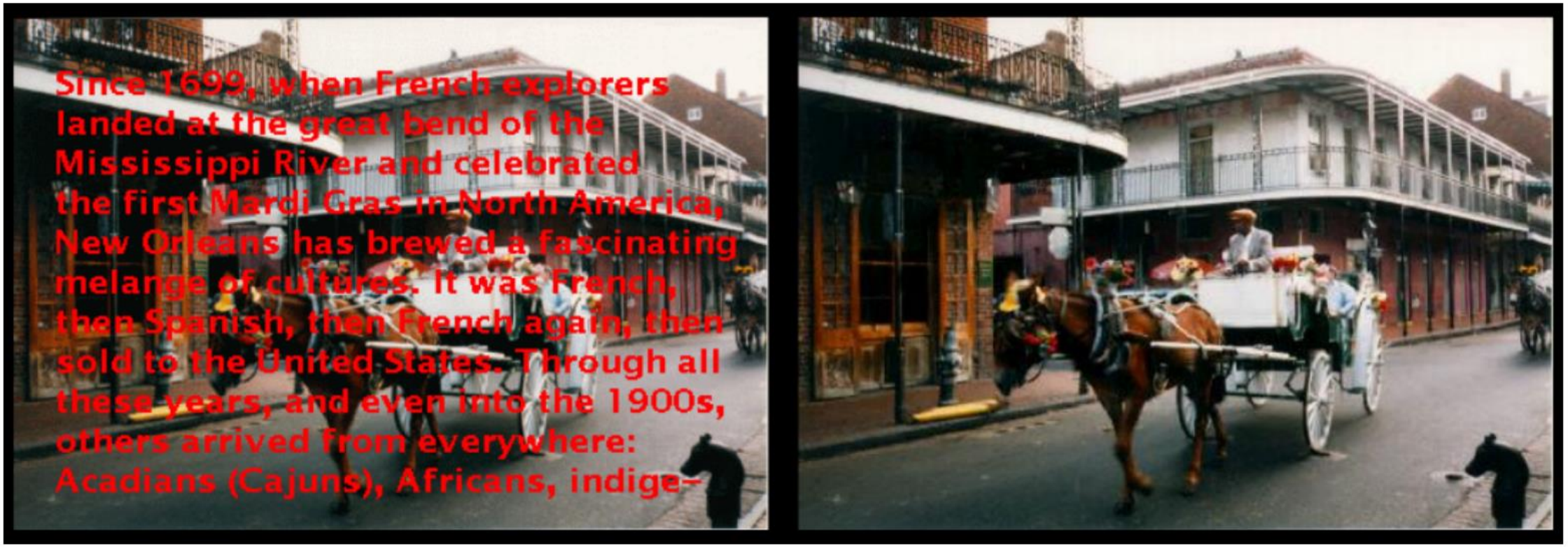
Sparse Coding : denoising



[Evan Smith & Mike Lewicki, 2006]

- http://ufldl.stanford.edu/eccv10-tutorial/

Sparse Coding : denoising



- https://www.cs.ubc.ca/~schmidtm/MLRG/sparseCoding.pdf

Sparse Coding : denoising

- https://www.cs.ubc.ca/~schmidtm/MLRG/sparseCoding.pdf
- http://scikit-learn.org/stable/auto_examples/decomposition/plot_image_denoising.html#sphx-glr-auto-examples-decomposition-plot-image-denoising-py

Sparse Coding

- 학습을 위해서는 최적화 방법들이 이용되며, 계산 시간이 긴 편입니다.
- 학습 데이터 X 를 한번에 이용하기 때문에 메모리 사용량도 많습니다.
  - Mini-batch dictionary learning 방법등이 대안으로 이용됩니다.

http://scikit-learn.org/stable/modules/decomposition.html#sparsecoder

Sparse Coding

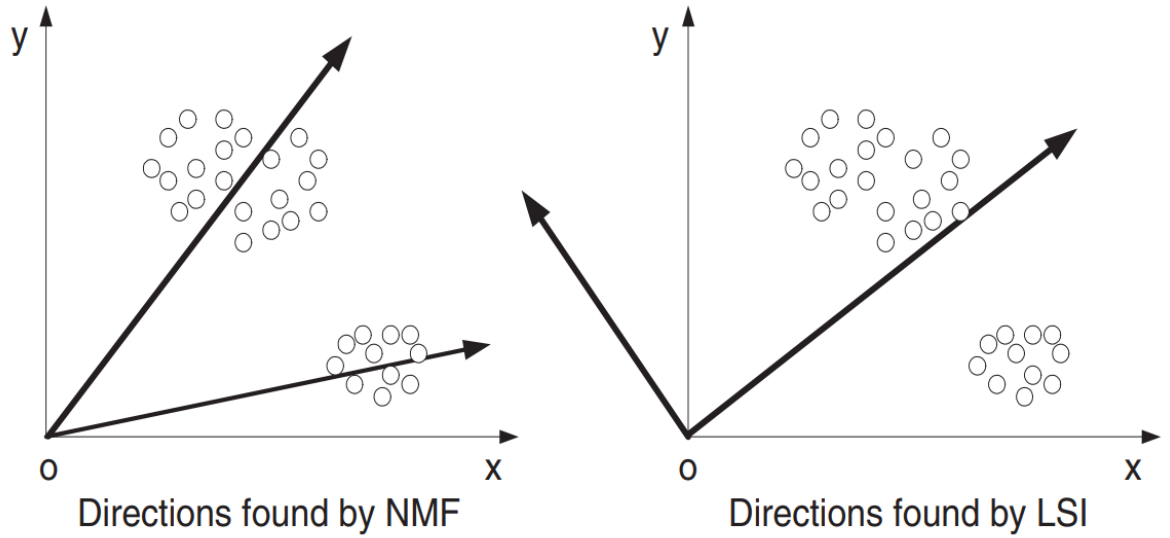
- Sparse coding 은 topic modeling 을 위해서도 이용될 수 있습니다.
- $|X - \beta^T D|^2 + \lambda|\beta|_1$
  - X : (n, m) 크기의 document – term vector
  - D 는 (p, m) 행렬이며, topic 의 word representation
  - β^T 는 (n, p) 의 행렬이며, 문서의 topical representation

Nonnegative Matrix Factorization

- Sparse coding, Latent Semantic Indexing 은 음의값을 포함하는 벡터로 표현되기 때문에 해석이 어렵습니다.
  - 그러나 term frequency representation 에서는 음의값이 없습니다.
  - 특히 LSI 는 각 차원이 서로 독립이라는 가정 때문에 음의값을 포함한 벡터가 학습됩니다.

Nonnegative Matrix Factorization

- NMF 는 각 축이 서로 독립이라는 가정을 하지 않음으로써, 자연스러운 topic representation 을 얻을 수 있도록 도와줍니다.



Xu, W., Liu, X., & Gong, Y. (2003, July). Document clustering based on non-negative matrix factorization. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval* (pp. 267-273). ACM.

Nonnegative Matrix Factorization

- NMF 는 dictionary 와 encoded vector 를 학습합니다.

$$minimize \quad 0.5 \cdot |X - DY|_F^2$$

$$where \quad D \geq 0, Y \geq 0$$

Frobenius norm

- Frobenius norm 은 벡터의 각 elements 의 제곱의 합입니다.

$$|A|_F = \sqrt{\sum_i \sum_j |a_{ij}|^2}$$

Nonnegative Matrix Factorization

- Scikit learn 의 NMF 는 L1 regularization 도 함께 구현되어 있습니다.

$$minimize \quad 0.5 \cdot |X - DY|_F^2 + \alpha \cdot (|D|_1 + |Y|_1)$$

$$where \quad D \geq 0, Y \geq 0$$

http://scikit-learn.org/stable/modules/generated/sklearn.decomposition.NMF.html

Nonnegative Matrix Factorization

- NMF 의 components 들은 nonnegative weights 로 이뤄졌기 때문에 topic modeling 이나 document clustering 용도로 이용할 수 있습니다.
  - 비슷한 문서에서 함께 등장한 단어들은 하나의 components 에 묶입니다.

Nonnegative Matrix Factorization

- NMF 를 이용한 2016-10-20 뉴스의 topic modeling 예시입니다.

Components	keywords
component#71	불독 데뷔 걸그룹 쇼케이스 키미 형은 무대 소라 롤링 세이 오전 마포구 매력 101 멤버들 싱글 프로듀스 강렬 20일 표현
component#11	공개 모습 화보 캔디 공유 캐릭터 도깨비 매력 선보 영상 마음 촬영 한편 메이크업 소리 조안 장근석 특히 변신 기대감
component#60	기록 1위 트와이스 스트리밍 방탄소년단 차트 누적 발표 뮤직비디오 올해 가온차트 2016년 데뷔 유튜브 최고 미니앨범 조회수 차지 1주년 부문
component#23	방송 출연 이날 프로그램 전현무 무대 예능 웃음 오후 시청률 지상파 노래 김지민 아프리카 광고 라고 샤이니 한편 20일 이에
component#79	신화 앨범 발매 13집 팬들 정규 컴백 활동 11월 그룹 이번 공개 콘서트 데뷔 예정 멤버들 기대 곡들 19년 아이돌

Nonnegative Matrix Factorization

- NMF 를 이용한 2016-10-20 뉴스의 topic modeling 예시입니다.

Components	keywords
component#1	트럼프 클린턴 토론 대선 후보 힐러리 주장 공화당 미국 3차 선거 민주당 도널드 대선후보 결과 라스베이거스 지지 이날 이라고 발언
component#39	후보 지금 힐러리 미국 말씀 트럼프 있습니다 이런 때문 제가 우리 사실 클린턴 생각합니다 경제 생각 발언 국가 대통령 도널드
component#57	여성 남성 여성들 남성들 혐오 자신 사회 임신 남자 표현 운동 말하는 이상 낙태 차별 사건 비율 생각 여자 군대
component#14	중국 필리핀 두테르테 양국 남중국해 베이징 주석 정상회담 분쟁 협력 영유권 시진핑 방문 경제 투자 성장률 갈등 수출 로드 확대
component#42	영화 감독 개봉 작품 배우 연기 걷기왕 이야기 관객 관객들 제작 심은경 흥행 만복 주연 출연 배우들 스크린 럭키 영화제