WIKIPEDIA

Bag-of-words model in computer vision

In **computer vision**, the **bag-of-words model** (BoW model) can be applied to image classification, by treating image features as words. In document classification, a bag of words is a sparse vector of occurrence counts of words; that is, a sparse histogram over the vocabulary. In computer vision, a *bag of visual words* is a vector of occurrence counts of a vocabulary of local image features.

Contents

Image representation based on the BoW model
- Feature representation
- Codebook generation

Learning and recognition based on the BoW model
- Generative models
  - Naïve Bayes
  - Hierarchical Bayesian models
- Discriminative models
  - Pyramid match kernel

Limitations and recent developments

See also

References

External links

Image representation based on the BoW model

To represent an image using the BoW model, an image can be treated as a document. Similarly, "words" in images need to be defined too. To achieve this, it usually includes following three steps: feature detection, feature description, and codebook generation.[1] A definition of the BoW model can be the "histogram representation based on independent features".[2] Content based image indexing and retrieval (CBIR) appears to be the early adopter of this image representation technique.[3]

Feature representation

After feature detection, each image is abstracted by several local patches. Feature representation methods deal with how to represent the patches as numerical vectors. These vectors are called feature descriptors. A good descriptor should have the ability to handle intensity, rotation, scale and affine variations to some extent. One of the most famous descriptors is Scale-invariant feature transform (SIFT).[4] SIFT converts each patch to 128-dimensional vector. After this step, each image is a collection of vectors of the same dimension (128 for SIFT), where the order of different vectors is of no importance.

Codebook generation

The final step for the BoW model is to convert vector-represented patches to "codewords" (analogous to words in text documents), which also produces a "codebook" (analogy to a word dictionary). A codeword can be considered as a representative of several similar patches. One simple method is performing k-means clustering over all the vectors.[5] Codewords are then defined as the centers of the learned clusters. The number of the clusters is the codebook size (analogous to the size of the word dictionary).

Thus, each patch in an image is mapped to a certain codeword through the clustering process and the image can be represented by the histogram of the codewords.

Learning and recognition based on the BoW model

Computer vision researchers have developed several learning methods to leverage the BoW model for image related tasks, such as object categorization. These methods can roughly be divided into two categories, generative and discriminative models. For multiple label categorization problem, the confusion matrix can be used as an evaluation metric.

Generative models

Here are some notations for this section. Suppose the size of codebook is V.

- w: each patch w is a V-dimensional vector that has a single component that equals to one and all other components equal to zero (For k-means clustering setting, the single component equal one indicates the cluster that w belongs to). The vth codeword in the codebook can be represented as $w^v = 1$ and $w^u = 0$ for $u \neq v$.
- $\mathbf{w}$: each image is represented by $\mathbf{w} = [w_1, w_2, \cdots, w_N]$, all the patches in an image
- d_j: the jth image in an image collection
- c: category of the image
- z: theme or topic of the patch
- π: mixture proportion

Since the BoW model is an analogy to the BoW model in NLP, generative models developed in text domains can also be adapted in computer vision. Simple Naïve Bayes model and hierarchical Bayesian models are discussed.

Naïve Bayes

The simplest one is Naïve Bayes classifier.[6] Using the language of graphical models, the Naïve Bayes classifier is described by the equation below. The basic idea (or assumption) of this model is that each category has its own distribution over the codebooks, and that the distributions of each category are observably different. Take a face category and a car category for an example. The face category may emphasize the codewords which represent "nose", "eye" and "mouth", while the car category may emphasize the codewords which represent "wheel" and "window". Given a collection of training examples, the classifier learns different distributions for different categories. The categorization decision is made by

$$c^* = \arg\max_c p(c|\mathbf{w}) = \arg\max_c p(c)p(\mathbf{w}|c) = \arg\max_c p(c)\prod_{n=1}^{N} p(w_n|c)$$

Since the Naïve Bayes classifier is simple yet effective, it is usually used as a baseline method for comparison.

Hierarchical Bayesian models

The basic assumption of Naïve Bayes model does not hold sometimes. For example, a natural scene image may contain several different themes. Probabilistic latent semantic analysis (pLSA)[7][8] and latent Dirichlet allocation (LDA)[9] are two popular topic models from text domains to tackle the similar multiple "theme" problem. Take LDA for an example. To model natural scene images using LDA, an analogy is made with document analysis:

- the image category is mapped to the document category;
- the mixture proportion of themes maps the mixture proportion of topics;
- the theme index is mapped to topic index;
- the codeword is mapped to the word.

This method shows very promising results in natural scene categorization on 13 Natural Scene Categories (http://vision.stanford.edu/resources_links.html).[1]

Discriminative models

Since images are represented based on the BoW model, any discriminative model suitable for text document categorization can be tried, such as support vector machine (SVM)[6] and AdaBoost.[10] Kernel trick is also applicable when kernel based classifier is used, such as SVM. Pyramid match kernel is newly developed one based on the BoW model. The local feature approach of using BoW model representation learnt by machine learning classifiers with different kernels (e.g., EMD-kernel and X^2 kernel) has been vastly tested in the area of texture and object recognition.[11] Very promising results on a number of datasets have been reported. This approach[11] has achieved very impressive results in the PASCAL Visual Object Classes Challenge (http://www.pascal-network.org/challenges/VOC/).

Pyramid match kernel

Pyramid match kernel[12] is a fast algorithm (linear complexity instead of classic one in quadratic complexity) kernel function (satisfying Mercer's condition) which maps the BoW features, or set of features in high dimension, to multi-dimensional multi-resolution histograms. An advantage of these multi-resolution histograms is their ability to capture co-occurring features. The pyramid match kernel builds multi-resolution histograms by binning data points into discrete regions of increasing size. Thus, points that do not match at high resolutions have the chance to match at low resolutions. The pyramid match kernel performs an approximate similarity match, without explicit search or computation of distance. Instead, it intersects the histograms to approximate the optimal match. Accordingly, the computation time is only linear in the number of features. Compared with other kernel approaches, the pyramid match kernel is much faster, yet provides equivalent accuracy. The pyramid match kernel was applied to ETH-80 database (https://web.archive.org/web/20080124115650/http://www.mis.informatik.tu-darmstadt.de/Research/Projects/categorization/eth80-db.html) and Caltech 101 database (https://web.archive.org/web/20080121104826/http://vision.cs.princeton.edu/resources_links.html) with promising results.[12][13]

Limitations and recent developments

One of the notorious disadvantages of BoW is that it ignores the spatial relationships among the patches, which are very important in image representation. Researchers have proposed several methods to incorporate the spatial information. For feature level improvements, correlogram features can capture spatial co-occurrences of features.[14] For generative models, relative positions[15][16] of codewords are also taken into account. The hierarchical shape and appearance model for human action[17] introduces a new part layer (Constellation model) between the mixture proportion and the BoW features, which captures the spatial relationships among parts in the layer. For discriminative models, spatial pyramid match[18] performs pyramid matching by partitioning the image into increasingly fine sub-regions and compute histograms of local features inside each sub-region. Recently, an augmentation of local image descriptors (i.e. SIFT) by their spatial coordinates normalised by the image width and height have proved to be a robust and simple Spatial Coordinate Coding[19][20] approach which introduces spatial information to the BoW model.

The BoW model has not been extensively tested yet for view point invariance and scale invariance, and the performance is unclear. Also the BoW model for object segmentation and localization is not well understood.[2]

A systematic comparison of classification pipelines found that the encoding of first and second order statistics (Vector of Locally Aggregated Descriptors (VLAD)[21] and Fisher Vector (FV)) considerably increased classification accuracy compared to BoW, while also decreasing the codebook size, thus lowering the computational effort for codebook generation.[22] Moreover, a recent detailed comparison of coding and pooling methods[20] for BoW has showed that second order statistics combined with Sparse Coding and an appropriate pooling such as Power Normalisation can further outperform Fisher Vectors and even approach results of simple models of Convolutional Neural Network on some object recognition datasets such as Oxford Flower Dataset 102.

See also

- Part-based models
- Fisher Vector encoding
- Segmentation-based object categorization
- Vector space model
- Bag-of-words model
- Feature extraction

References

1. Fei-Fei Li; Perona, P. (2005). *A Bayesian Hierarchical Model for Learning Natural Scene Categories*. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. **2**. p. 524. doi:10.1109/CVPR.2005.16 (https://doi.org/10.1109%2FCVPR.2005.16). ISBN 978-0-7695-2372-9.
2. L. Fei-Fei; R. Fergus & A. Torralba. "Recognizing and Learning Object Categories, CVPR 2007 short course" (http://people.csail.mit.edu/torralba/shortCourseRLOC/index.html).
3. Qiu, G. (2002). "Indexing chromatic and achromatic patterns for content-based colour image retrieval" (http://www.cs.nott.ac.uk/~qiu/webpages/Papers/ColorPatternRecognition.pdf) (PDF). *Pattern Recognition*. **35** (8): 1675–1686. doi:10.1016/S0031-3203(01)00162-5 (https://doi.org/10.1016%2FS0031-3203%2801%2900162-5).
4. Vidal-Naquet; Ullman (1999). "Object recognition with informative features and linear classification" (http://www.cs.ubc.ca/~lowe/papers/iccv99.pdf) (PDF). *Proceedings Ninth IEEE International Conference on Computer Vision*. pp. 1150–1157. CiteSeerX 10.1.1.131.1283 (https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.131.1283). doi:10.1109/ICCV.2003.1238356 (https://doi.org/10.1109%2FICCV.2003.1238356). ISBN 978-0-7695-1950-0.
5. T. Leung; J. Malik (2001). "Representing and recognizing the visual appearance of materials using three-dimensional textons" (http://www.cs.berkeley.edu/~malik/papers/LM-3dtexton.pdf) (PDF). *International Journal of Computer Vision*. **43** (1): 29–44. doi:10.1023/A:1011126920638 (https://doi.org/10.1023%2FA%3A1011126920638).
6. G. Csurka; C. Dance; L.X. Fan; J. Willamowski & C. Bray (2004). "Visual categorization with bags of keypoints" (http://www.xrce.xerox.com/Research-Development/Publications/2004-010). *Proc. of ECCV International Workshop on Statistical Learning in Computer Vision*.
7. T. Hoffman (1999). "Probabilistic Latent Semantic Analysis" (https://web.archive.org/web/20070710083034/http://www.cs.brown.edu/~th/papers/Hofmann-UAI99.pdf) (PDF). *Proc. of the Fifteenth Conference on Uncertainty in Artificial Intelligence*. Archived from the original (http://www.cs.brown.edu/~th/papers/Hofmann-UAI99.pdf) (PDF) on 2007-07-10. Retrieved 2007-12-10.
8. Sivic, J.; Russell, B.C.; Efros, A.A.; Zisserman, A.; Freeman, W.T. (2005). "Discovering objects and their location in images" (http://www.robots.ox.ac.uk/~vgg/publications/papers/sivic05b.pdf) (PDF). *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*. p. 370. CiteSeerX 10.1.1.184.1253 (https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.184.1253). doi:10.1109/ICCV.2005.77 (https://doi.org/10.1109%2FICCV.2005.77). ISBN 978-0-7695-2334-7.
9. D. Blei; A. Ng & M. Jordan (2003). Lafferty, John, ed. "Latent Dirichlet allocation" (https://web.archive.org/web/20080822212053/http://www.cs.princeton.edu/~blei/papers/BleiNgJordan2003.pdf) (PDF). *Journal of Machine Learning Research*. **3** (4–5): 993–1022. doi:10.1162/jmlr.2003.3.4-5.993 (https://doi.org/10.1162%2Fjmlr.2003.3.4-5.993). Archived from the original (http://www.cs.princeton.edu/~blei/papers/BleiNgJordan2003.pdf) (PDF) on 2008-08-22. Retrieved 2007-12-10.

10. Serre, T.; Wolf, L.; Poggio, T. (2005). "Object Recognition with Features Inspired by Visual Cortex" (http://cbcl.mit.edu/projects/cbcl/publications/ps/serre-PID73457-05.pdf) (PDF). *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. **2**. p. 994. CiteSeerX 10.1.1.71.5276 (https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.71.5276). doi:10.1109/CVPR.2005.254 (https://doi.org/10.1109%2FCVPR.2005.254). ISBN 978-0-7695-2372-9.
11. Jianguo Zhang; Marcin Marszałek; Svetlana Lazebnik; Cordelia Schmid (2007). "Local Features and Kernels for Classification of Texture and Object Categories: a Comprehensive Study" (http://lear.inrialpes.fr/pubs/2007/ZMLS07/ZhangMarszalekLazebnikSchmid-IJCV07-ClassificationStudy.pdf) (PDF). *International Journal of Computer Vision*. **73** (2): 213–238. doi:10.1007/s11263-006-9794-4 (https://doi.org/10.1007%2Fs11263-006-9794-4).
12. Grauman, K.; Darrell, T. (2005). "The pyramid match kernel: discriminative classification with sets of image features" (http://www.cs.utexas.edu/~grauman/papers/grauman_darrell_iccv2005.pdf) (PDF). *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*. p. 1458. CiteSeerX 10.1.1.644.6159 (https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.644.6159). doi:10.1109/ICCV.2005.239 (https://doi.org/10.1109%2FICCV.2005.239). ISBN 978-0-7695-2334-7.
13. Jianchao Yang; Kai Yu; Yihong Gong; Huang, T. (2009). "Linear spatial pyramid matching using sparse coding for image classification" (http://www.ifp.illinois.edu/~jyang29/ScSPM.htm). *2009 IEEE Conference on Computer Vision and Pattern Recognition*. p. 1794. doi:10.1109/CVPR.2009.5206757 (https://doi.org/10.1109%2FCVPR.2009.5206757). ISBN 978-1-4244-3992-8.
14. Savarese, S.; Winn, J.; Criminisi, A. (2006). "Discriminative Object Class Models of Appearance and Shape by Correlatons" (http://johnwinn.org/Publications/papers/Savarese_Winn_Criminisi_Correlatons_CVPR2006.pdf) (PDF). *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2 (CVPR'06)*. **2**. p. 2033. CiteSeerX 10.1.1.587.8853 (https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.587.8853). doi:10.1109/CVPR.2006.102 (https://doi.org/10.1109%2FCVPR.2006.102). ISBN 978-0-7695-2597-6.
15. Sudderth, E.B.; Torralba, A.; Freeman, W.T.; Willsky, A.S. (2005). "Learning hierarchical models of scenes, objects, and parts" (http://ssg.mit.edu/~esuddert/papers/iccv05.pdf) (PDF). *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*. p. 1331. CiteSeerX 10.1.1.128.7259 (https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.128.7259). doi:10.1109/ICCV.2005.137 (https://doi.org/10.1109%2FICCV.2005.137). ISBN 978-0-7695-2334-7.
16. E. Sudderth; A. Torralba; W. Freeman & A. Willsky (2005). "Describing Visual Scenes using Transformed Dirichlet Processes" (http://ssg.mit.edu/~esuddert/papers/nips05.pdf) (PDF). *Proc. of Neural Information Processing Systems*.
17. Niebles, Juan Carlos; Li Fei-Fei (2007). "A Hierarchical Model of Shape and Appearance for Human Action Classification" (http://vision.stanford.edu/posters/NieblesFeiFei_CVPR07_poster.pdf) (PDF). *2007 IEEE Conference on Computer Vision and Pattern Recognition*. p. 1. CiteSeerX 10.1.1.173.2667 (https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.173.2667). doi:10.1109/CVPR.2007.383132 (https://doi.org/10.1109%2FCVPR.2007.383132). ISBN 978-1-4244-1179-5.
18. Lazebnik, S.; Schmid, C.; Ponce, J. (2006). "Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories" (http://www-cvr.ai.uiuc.edu/ponce_grp/publication/paper/cvpr06b.pdf) (PDF). *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2 (CVPR'06)*. **2**. p. 2169. CiteSeerX 10.1.1.651.9183 (https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.651.9183). doi:10.1109/CVPR.2006.68 (https://doi.org/10.1109%2FCVPR.2006.68). ISBN 978-0-7695-2597-6.
19. Koniusz, Piotr; Yan, Fei; Mikolajczyk, Krystian (2013-05-01). "Comparison of mid-level feature coding approaches and pooling strategies in visual concept detection". *Computer Vision and Image Understanding*. **117** (5): 479–492. doi:10.1016/j.cviu.2012.10.010 (https://doi.org/10.1016%2Fj.cviu.2012.10.010). ISSN 1077-3142 (https://www.worldcat.org/issn/1077-3142).
20. Koniusz, Piotr; Yan, Fei; Gosselin, Philippe Henri; Mikolajczyk, Krystian (2017-02-24). "Higher-order occurrence pooling for bags-of-words: Visual concept detection". *IEEE Transactions on Pattern Analysis and Machine Intelligence*. **39** (2): 313–326. doi:10.1109/TPAMI.2016.2545667 (https://doi.org/10.1109%2FTPAMI.2016.2545667). ISSN 0162-8828 (https://www.worldcat.org/issn/0162-8828). PMID 27019477 (https://www.ncbi.nlm.nih.gov/pubmed/27019477).
21. Jégou, H.; Douze, M.; Schmid, C.; Pérez, P. (2010-06-01). *Aggregating local descriptors into a compact image representation*. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. pp. 3304–3311. doi:10.1109/CVPR.2010.5540039 (https://doi.org/10.1109%2FCVPR.2010.5540039). ISBN 978-1-4244-6984-0.
22. Seeland, Marco; Rzanny, Michael; Alaqraa, Nedal; Wäldchen, Jana; Mäder, Patrick (2017-02-24). "Plant species classification using flower images—A comparative study of local feature representations" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5325198). *PLOS ONE*. **12** (2): e0170629. doi:10.1371/journal.pone.0170629 (https://doi.org/10.1371%2Fjournal.pone.0170629). ISSN 1932-6203 (https://www.worldcat.org/issn/1932-6203). PMC 5325198 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5325198). PMID 28234999 (https://www.ncbi.nlm.nih.gov/pubmed/28234999).

External links

- A demo for two bag-of-words classifiers (http://people.csail.mit.edu/fergus/iccv2005/bagwords.html) by L. Fei-Fei, R. Fergus, and A. Torralba.
- Caltech Large Scale Image Search Toolbox (https://web.archive.org/web/20101203074412/http://www.vision.caltech.edu/malaa/software/research/image-search/): a Matlab/C++ toolbox implementing Inverted File search for Bag of Words model. It also contains implementations for fast approximate nearest neighbor search using randomized k-d tree, locality-sensitive hashing, and hierarchical k-means.
- DBoW2 library (https://github.com/dorian3d/DBoW2): a library that implements a fast bag of words in C++ with support for OpenCV.

Retrieved from "https://en.wikipedia.org/w/index.php?title=Bag-of-words_model_in_computer_vision&oldid=883515461"

This page was last edited on 15 February 2019, at 21:30 (UTC).