TECHNISCHE UNIVERSITÄT DRESDEN

ZENTRUM FÜR INFORMATIONSDIENSTE
UND HOCHLEISTUNGSRECHNEN
PROF. DR. WOLFGANG E. NAGEL

Master-Arbeit

zur Erlangung des akademischen Grades
Master of Science

Image Retrieval für Historische Bilder

Philipp Langen
(Geboren am 26. Dezember 1994 in Münsterlingen)

Hochschullehrer: Prof. Dr. Wolfgang E. Nagel
Betreuer: Dr. Christoph Lehmann & Dr. Taras Lazariv

Dresden, 30. Juni 2020

Hier Aufgabenstellung einfügen!

Selbstständigkeitserklärung

Hiermit erkläre ich, dass ich die von mir am heutigen Tag dem Prüfungsausschuss der Fakultät Informatik eingereichte Master-Arbeit zum Thema:

Image Retrieval für Historische Bilder

vollkommen selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt sowie Zitate kenntlich gemacht habe.

Dresden, den 30. Juni 2020

Philipp Langen

Kurzfassung

Abstract

Inhaltsverzeichnis

1 Motivation

Präzise und effiziente Suchwerkzeuge sind essenziell um große Datenmengen für einen Nutzer sinnvoll verwertbar zu machen. Dies gilt insbesondere auch im Bereich der Bildersuche. Die klassische Bildersuche basiert auf vom Nutzer formulierten Anfragen, mit deren Hilfe das Suchsystem eine Liste an passenden Bildkandidaten zusammenstellt und zurückgibt. Hierbei nutzt das System eine Reihe von Zusatzinformationen, sogenannten Metadaten, wie Tags, Titel, Aufnahmeort oder Datum. Eine alternativer Ansatz der Suche, der in dieser Arbeit behandelt wird, ist die inhaltsbasierte Bildersuche (engl. Content-Based Image Retrieval, kurz CBIR). Hierbei werden vom Nutzer keine Anfragen formuliert. Stattdessen dient ein Bild als Suchanfrage. Ziel ist es, Bilder mit gleichem oder ähnlichem Bildinhalt als Ergebnis zurückzugeben. Ein Vorteil dieser Herangehensweise ist, dass der Nutzer keine Informationen über den Inhalt des Suchbildes benötigt. Das System arbeitet ausschließlich mit den Pixelinformationen der Bilder. Ein weiterer Vorteil ist daher, dass weder im Suchbild noch in der Suchdatenbank Metadaten zu den Bildern vorhanden sein müssen, was die Einsatzmöglichkeiten von inhaltsbasierter Suche sehr flexibel gestaltet. Im Folgenden wird die inhaltsbasierte Suche auch als Image Retrieval bezeichnet.

Das Anwendungsgebiet dieser Arbeit ist die Suche auf historischen Bildern. Diese weit gefasste Domäne ist besonders herausfordernd, da sie sehr heterogene Daten enthält. Dabei gibt es nicht nur große Unterschiede in den abgebildeten Bildinhalten, wie Gebäude, Naturaufnahmen oder Portraits, sondern auch in den verwendeten Aufnahmeverfahren. Durch die Fortschritte der Aufnahmetechnik können historische Bilder sowohl in Form von Zeichnungen oder Malerei, aber auch als Druck oder in anfänglichen Formen der Photographie vorliegen. Da Metadaten zu historischen Bildern erst bei der Digitalisierung hinzugefügt werden können, sind diese oft gar nicht oder nur lückenhaft vorhanden. Dies macht die inhaltsbasierte Suche für diese Domäne im Vergleich zur klassischen Bildersuche zu einer besonders geeigneten Methode. Mit der Umsetzung einer unterstützenden Suche für das UrbanHistory4D Projekt [1] ergibt sich ein konkreter Anwendungsfall für diese Arbeit. Das UrbanHistory4D Projekt befasst sich mit der Erstellung interaktiver Stadtkarten. Wo vorhanden kann sich der Nutzer historische Aufnahmen anzeigen lassen, die den Ort zeigen, an dem er sich innerhalb der Karte befindet. Das Akkumulieren und Zuordnen von historischen Bildern zu diesen Plätzen ist ein wesentlicher Arbeitsanteil bei der Erstellung der Karten. Image Retrieval Systeme können helfen den Suchaufwand für die Ersteller der Karten signifikant zu reduzieren. Dabei handelt es sich um einen aktiven Forschungsbereich, in dem momentan unterschiedliche Suchsysteme analysiert werden.

Das Image Retrieval Verfahren DELF (attentive DEep Local Features) [2] entwickelt von Noh, Araujo et al., welches in dieser Arbeit untersucht wird ist ein Deep Learning Ansatz. Durch den raschen Fortschritt im Bereich tiefer neuronaler Netzwerkarchitekturen der letzten Jahre erfreuen sich gelernte Ansätze immer größerer Beliebtheit. DELF erzielt auf bekannten Benchmarkdatensätzen wie Oxford5k [3] und Paris6k [4] sehr gute Ergebnisse. Besonders gut schneidet DELF im Vergleich auf dem eigens erstellten

Google Landmarks Datensatz [5] ab. Dieser enthält mit über 1 Mio. Bilder und 13k unterschiedlichen Motiven eine deutlich heterogenere Mischung an Objekten als andere Benchmarks. Die gute Performanz auf diesem Datensatz lässt also hoffen, dass sich das DELF-Verfahren auch für die historische Domäne eignet.

2 Verwandte Arbeiten

Bei Information Retrieval handelt es sich um ein Problem aus dem Bereich der Computer Vision, welches bereits seit langem intensiv erforscht wird. In frühen Ansätzen versuchte man vor allem globale Beschreibungen von Bildern zu erstellen, um diese untereinander vergleichen zu können. Diese basierten zum Beispiel auf Farbhistogrammen oder Texturbeschreibungen [6]. Allerdings waren diese Ansätze oft sehr anfällig für Unterschiede in Beleuchtung, Skalierung und anderen Transformationen, wie sie bei unterschiedlichen Aufnahmen des selben Motivs auftreten können.
Ein wesentlicher Durchbruch gelang David G. Lowe 2004 mit der Entwicklung des SIFT-Verfahrens (Scale Invariant Feature Transform) [7]. Hierbei werden mehrere Konzepte vereint um Bildbeschreibungen zu erzeugen, die robuster gegenüber unterschiedlichen Transformationen sind. So arbeitet der SIFT Algorithmus beispielsweise nicht direkt auf den Bildern, sondern im sogenannten Scale Space. Dieser besteht auf unterschiedlich skalierten Versionen des Ursprungsbildes, auf welche wiederum unterschiedlich starke Gauß-Filter angewendete werden. Betrachtet werden schließlich Differenzbilder zwischen benachbarten Stärken der Gauß-Fitler Ergebnisse. Die Verwendung von unterschiedlich skalierten Bildversionen macht die berechneten SIFT-Merkmale deutlich robuster gegen Skalierungsunterschiede. Das SIFT-Verfahren besteht aus zwei Phasen. In der ersten Phase werden über die Suche nach lokalen Extrema bedeutsame Bildpunkte ausgewählt. Für diese werden in der zweiten Phase einzelne Deskriptoren berechnet. Das Bild wird also nicht global beschrieben, sondern über viele lokale Deskriptoren dargestellt. Die lokalen Deskriptoren ergeben sich aus Histogrammen der Gradientrichtungen umliegender Bildpunkte. Diese werden relativ zu der dominanten Gradientrichtung in der Umgebung berechnet, was die Deskriptoren invariant gegenüber Rotationen macht. Lowes Entwicklung bildet den Ursprung für viele abgeleitete Verfahren wie SURF[8], PCA-SIFT[9] und RIFT[10]. Auch in aktueller Forschung werden Image Retrieval Verfahren untersucht, die mit SIFT-Merkmalen arbeiten [11].
Der Trend bei der Entwicklung neuer Image Retrieval Systeme geht aktuell jedoch hauptsächlich in Richtung von gelernten Verfahren. Die Basis dieser Verfahren bilden tiefe CNN-Architekturen (Convolutional Neural Networks). Ein neuronales Netzwerk lässt sich als eine schichtweise Aneinanderreihung nichtlinearer Funktionen auffassen. Convolutional Neural Networks sind ein Sonderfall neuronaler Netze, welche sogenannte Convolutional Layer, zu deutsch faltende Schichten, enthalten. In diesen Schichten werden Faltungs/ bzw. Filteroperationen auf die Eingabedaten angewendet, um für das Netzwerk hilfreiche Merkmale in den Daten hervorzuheben. Dies ist durchaus vergleichbar mit den Filteroperationen, die im SIFT-Verfahren verwendet werden. Der Unterschied besteht jedoch darin, dass die Parameter der verwendeten Filtermasken sowie aller anderen Netzparameter nicht per Hand gewählt, sondern in einem Trainingsverfahren für den aktuellen Anwendungsfall optimiert werden. Der Entwickler bestimmt lediglich die grobe Architektur des Netzwerks, also die Anzahl, Größe und Reihenfolge der verwendeten Schichten sowie die Art der Operationen, die in ihnen durchgeführt werden. In Image Retrieval Systemen werden CNNs eingesetzt, um Bilddeskriptoren zu erstellen. Hierfür werden Zwischenergebnisse des Netzwerks, also die Ausgaben einer bestimmten Schicht genutzt. An welcher Stelle im Netzwerk die De-

skriptoren entnommen werden ist dabei von entscheidender Bedeutung. Zeiler und Fergus haben in ihrer Studie zur Visualisierung von CNNs gezeigt [12], dass die früheren Schichten von CNNs typischerweise einfache Konzepte wie Kanten oder Ecken hervorheben. Mit wachsender Tiefe der betrachteten Netzwerkschicht steigt auch die Komplexität der Konzepte, die von den Ausgaben der Schicht beschrieben werden können.

In dem in [13] beschriebenen Image Retrieval System von Babenko, Slesarev et al. wird als Modell ein CNN bestehend aus fünf faltenden gefolgt von drei vollvernetzten Schichten (im Englischen fully-connected layer) genutzt. Als Deskriptoren werden die Ausgaben der ersten bzw. zweiten vollvernetzten Schicht verwendet. In einer vollvernetzten Schicht hat jeder Wert der Eingabe Einfluss auf jeden Wert in der Ausgabe. Die Ausgaben solcher Schichten werden also von der gesamten Bildeingabe beeinflusst und können daher als globale Deskriptoren verstanden werden. Diese intuitive Herangehensweise erzielt leichte Verbesserung gegenüber den zur Zeit der Veröffentlichung gängigen algorithmischen Verfahren.
Razavian, Sullivan et al. stellen in [14] ein System auf Basis der in [15] beschriebenen Netzwerkarchitektur vor. Das Modell besteht ebenfalls aus fünf faltenden und drei vollvernetzten Schichten. Die Deskriptoren stammen aus den Ausgaben der letzten faltenden Schicht. Anders als bei Babenko, Slesarev et al. werden in diesem System mehrere Deskriptoren pro Bild erstellt. Hierfür werden systematisch Teilbilder aus Bildbereichen unterschiedlicher Größe generiert. Anschließend werden die Teilbilder auf eine feste Größe skaliert und als Eingabe in das Netzwerk gegeben. So wird für jeden betrachteten Bildbereich ein eigener lokaler Deskriptor erstellt. In ihren Experimenten stellen die Autoren fest, dass die Verwendung von lokalen Deskriptoren gegenüber einer globalen Betrachtung zu einer signifikanten Verbesserung der Retrievalperformanz führt. Der überwiegende Teil aktueller Retrieval Systeme setzt auf die Erstellung von lokalen Deskriptoren.
Eine interessante Frage bei der Konzeption von Image Retrieval Systemen, die mit lokalen Deskriptoren arbeiten ist, wie man entscheidet, welche Bildregionen am sinnvollsten zu betrachten sind. Das ONE-Verfahren [16] von Xie, Hong et al. nutzt ein VGG-19 [17] Modell und extrahiert Deskriptoren aus der vorletzten vollvernetzten Schicht. Als Eingaben in das Netzwerk dienen sogenannte Object Proposals. Dabei handelt es sich um Bildausschnitte, welche Regionen umschließen, in denen Objekte vermutet werden. Die Autoren testen sowohl manuell annotierte sowie automatisch extrahierte Object Proposals und erzielen mit beiden Ansätzen ähnlich gute Ergebnisse. Für die automatische Bestimmung von Object Proposals nutzen sie das Selective Search Verfahren [18].
Das Delf-Verfahren [2], welches in dieser Arbeit untersucht wird, basiert ebenfalls auf lokalen Deskriptoren. Die Deskriptoren werden aus einer faltenden Schicht aus dem hinteren Teil eines ResNet-50 [19] Modells extrahiert. Als Eingabe in das Netzwerk werden Bilder in ihrer Gesamtheit betrachtet. Da bis zur Extraktionsschicht keine vollvernetzten Schichten genutzt werden, kann für jeden extrahierten Wert zurückgerechnet werden, von welchen Bereichen des Ursprungsbildes er beeinflusst wurde. Dies erlaubt es die Ausgaben der Extraktionsschicht in einzelne lokale Deskriptoren zu unterteilen. Um auszuwählen welche der lokalen Deskriptoren zur Darstellung eines Bildes genutzt werden sollen, werden die lokalen Deskriptoren in ein weiteres neuronales Netz gegeben. Dieses Netz hat die Aufgabe zu bewerten, wie geeignet die einzelnen Deskriptoren zur Beschreibung des Gesamtbildes sind. Auf Grund dieser Bewertung werden die wichtigsten Deskriptoren zu jedem Bild ausgewählt, wogegen schlecht bewertete Deskriptoren verworfen werden. Der konzeptionelle Unterschied bei der Auswahl der Deskriptoren im Vergleich

zum ONE-Verfahren ist, dass die Auswahl auf Grund der bereits berechneten Deskriptoren geschieht anstatt auf Grund des Ursprungsbildes. Die Funktionsweise des Delf-Verfahrens wird in Kapitel 3 ab Seite 8 im Detail erklärt.
Bevor Neuronale Netze für die Erstellung von Deskriptoren genutzt werden können, müssen ihre Parameter in einem Trainingsverfahren optimiert werden. Während dem Training muss das Netzwerk eine Aufgabe lösen. Wie erfolgreich das Netzwerk dabei ist, wird mit Hilfe einer Fehlerfunktion dargestellt. Das Netz versucht seine Parameter so anzupassen, dass die Fehlerfunktion minimiert wird. Im Fall der bereits vorgestellten Verfahren wird dabei eine Dummy-Aufgabe, typischerweise die Klassifikation von Bildern, gelöst. In der letzten Zeit wurden jedoch einige Ansätze veröffentlicht, die versuchen neuronale Netze direkt an Image Retrieval Aufgaben zu trainieren. Radenović, Tolias und Chum stellen in [20] eine solchen Ansatz vor. Während dem Training arbeiten sie dabei mit Bildpaaren (i, j). Diese Paare werden als korrektes Match bezeichnet, falls sich ihre Bildinhalte überschneiden. Andernfalls handelt es sich um ein inkorrektes Match. Beide Bilder durchlaufen ein identisches Netz und erzeugen dabei jeweils eine Ausgabe $(\mathbf{d}(i), \mathbf{d}(j))$. Für die Optimierung wird eine spezielle Fehlerfunktion L definiert. Falls es sich bei den Bildern um ein korrektes Match handelt, sollten sich die Netzwerkausgaben der Bilder möglichst ähneln.

$$L_{korr}(i, j) = \frac{1}{2} \parallel \mathbf{d}(i) - \mathbf{d}(j) \parallel^2 \tag{2.1}$$

Handelt es sich jedoch um ein inkorrektes Match, sollten die Ausgaben einen Mindestabstand τ zueinander einhalten.

$$L_{inkorr}(i, j) = \frac{1}{2} \max \left[0, \tau - \parallel \mathbf{d}(i) - \mathbf{d}(j) \parallel^2\right] \tag{2.2}$$

Die Autoren testen ihr Verfahren auf unterschiedlichen CNN Architekturen wie VGG [17] und AlexNet [21] und erzielen damit sehr gute Ergebnisse auf gängigen Retrievalbenchmarks. Das direkte Training auf Retrievalaufgaben ist eine vielversprechende neue Forschungsrichtung im Retrievalbereich, an der momentan intensiv gearbeitet wird.

Da Image Retrieval Systeme meist auf großen Bilddatenbanken eingesetzt werden und somit für eine Suchanfrage viele Bilder miteinander verglichen werden müssen, ist es sinnvoll Bildrepräsentationen so kompakt wie möglich zu gestalten, um die Laufzeit der Suche zu verbessern. Insbesondere bei Verfahren, die lokale Deskriptoren erstellen und häufig hunderte oder tausende Merkmale pro Bild extrahieren, kann mit einer guten Kodierung viel Rechenzeit gespart werden. Ein beliebter Ansatz zur Erstellung kompakter Darstellungen aus lokalen Deskriptoren ist das BOVW-Modell (Bag-of-Visual-Words) [22], erstmals vorgestellt im Kontext von Textklassifikation von McCallum und Nigam. Hierbei werden zunächst alle aus einem Datensatz extrahierte Deskriptoren mittels Clusteranalyse (bspw. K-Means-Clustering [23]) in Gruppen eingeteilt. Deskriptoren, die dem gleichen Cluster zugeordnet werden, werden dabei auf das selbe "visuelle Wort" abgebildet. Als Beschreibung des Gesamtbilds dient ein Histogramm über die im Bild enthaltenen visuellen Wörter. Bei diesem Verfahren geht durch Quantisierung ein Teil der Information verloren. Das ebenfalls auf Clustering basierte VLAD-Verfahren [24] von Jégou, Douze et al. versucht diese Information nutzbar zu machen, indem es statt der Vorkommen die Quantisierungsfehler akkumuliert, die beim abbilden auf die nächsten visuellen Worte entstehen.
Um eine Suchanfrage mit einer Rangliste der ähnlichsten Bilder zum Suchbild beantworten zu können, werden die Deskriptoren der Bilder in der Datenbank mit denen des Suchbildes verglichen. Als Metrik

dient hierbei meist die euklidische Distanz. Auf kleinen Datensätzen ist es laufzeittechnisch sinnvoll alle Kombinationen von Such- und Datenbankbildern zu vergleichen. Häufig werden bei größeren Datensätzen jedoch Methoden der approximierten nächsten Nachbarsuche (ANN) verwendet. Diese garantieren zwar kein optimales Ergebnis, erlauben jedoch deutlich schnellere Verarbeitung von Suchanfragen. So gibt es zum Beispiel Ansätze Deskriptoren mit Hilfe spezieller Hashfunktionen zu vergleichen. Diese werden so konstruiert, dass ähnliche Deskriptoren auf die gleichen bzw. möglichst ähnliche Hashcodes abgebildet werden, während gleichzeitig die Kollisionswahrscheinlichkeit für sehr unterschiedliche Deskriptoren minimal gehalten wird. Wang et al. beschreiben in ihrer Studie [25] unterschiedliche Konzepte für die Erstellung solcher Hashfunktionen.

3 DELF

Das Delf-Verfahren [2] von Noh, Araujo et al. bildet die Basis für die Experimente, die in dieser Arbeit durchgeführt werden. Im folgenden Abschnitt wird das Verfahren schrittweise im Detail erklärt. Beschrieben wird hierbei die Neuimplementierung in ihrer Basiskonfiguration, wie sie für den Experimentalteil dieser Arbeit verwendet wird. Unterschiede zu der von den Autoren zu Verfügung gestellten Implementierung[1], sowie zu der Beschreibung des Verfahrens im Originalpapier [2] werden im hinteren Teil des Abschnitts erläutert.

Das Delf-Verfahren lässt sich in vier Phasen einteilen. Zu Beginn steht das sogenannte Fine-Tuning. Hierbei wird ein vortrainiertes Modell, in unserem Fall ein ResNet-50 Netzwerk, auf einem neuen Datensatz weiter trainiert. Die Domäne der Bilder dieses Datensatzes sollte dabei möglichst nahe der späteren Retrievalaufgabe sein, damit das Modell lernen kann aussagekräftige Deskriptoren für diese Art von Bilden zu berechnen. In der nächsten Phase wird auf dem Modell aufbauend ein Attention-Netzwerk trainiert welches die Güte berechneter Deskriptoren bewertet. In der dritten Phase werden für die Bilder der Datenbank, in der gesucht werden soll Deskriptoren extrahiert. Mit Hilfe des Attention-Netzwerks wird eine Vorauswahl besonders geeigneter Deskriptoren getroffen. Anschließend durchlaufen die Deskriptoren weitere Vorverarbeitungsschritte, mit denen sie in eine kompaktere Form überführt werden. In der finalen Phase kann Delf aktiv genutzt werden. Es können nun Bilder als Suchanfragen gestellt werden. Delf vergleicht eine Anfrage mit allen Bildern des Datensatzes anhand der vorverarbeiteten Deskriptoren. Potentielle Matches zwischen Deskriptoren werden in einem letzten Schritt geometrisch verifiziert. Das Ergebnis einer Anfrage ist eine Rangliste der ähnlichsten Bilder, sortierte nach der Anzahl verifizierte Deskriptoren-Matches mit dem Anfragebild.

3.1 ResNet

Das Delf-Verfahren nutzt zur Erstellung von Deskriptoren ein Residuales Netzwerk (kurz ResNet). Bei der im Jahre 2015 vorgestellten ResNet Architektur [19] von He, Zhang et al. handelt es sich um eine der meist genutzten tiefen CNN-Architekturen der aktuellen Forschung. ResNets finden Anwendung in unterschiedlichen Machine Learning Aufgaben, wie Klassifikation, Objektdetektion oder Image Retrieval.

Zeiler und Fergus haben gezeigt [12], dass CNNs mit wachsender Netzwerktiefe in der Lage sind komplexere Merkmale zu detektieren. Es scheint daher intuitiv zur Lösung immer komplexerer Aufgaben zunehmend tiefere Netzwerke zu konstruieren. Allerdings stellt sich heraus, dass ab einem gewissen Punkt keine Verbesserungen mehr mit dem bloßen aneinanderreihen von immer mehr Schichten erzielt werden können. Werden zu viele Schichten hinzugefügt kann die Trainingsperformanz sogar abnehmen. Mit dem rasanten Anstieg der Anzahl an Netzwerkparameter wird es immer schwieriger das Netzwerk

[1] `https://github.com/tensorflow/models/tree/master/research/delf`, zuletzt besucht 16.06.20

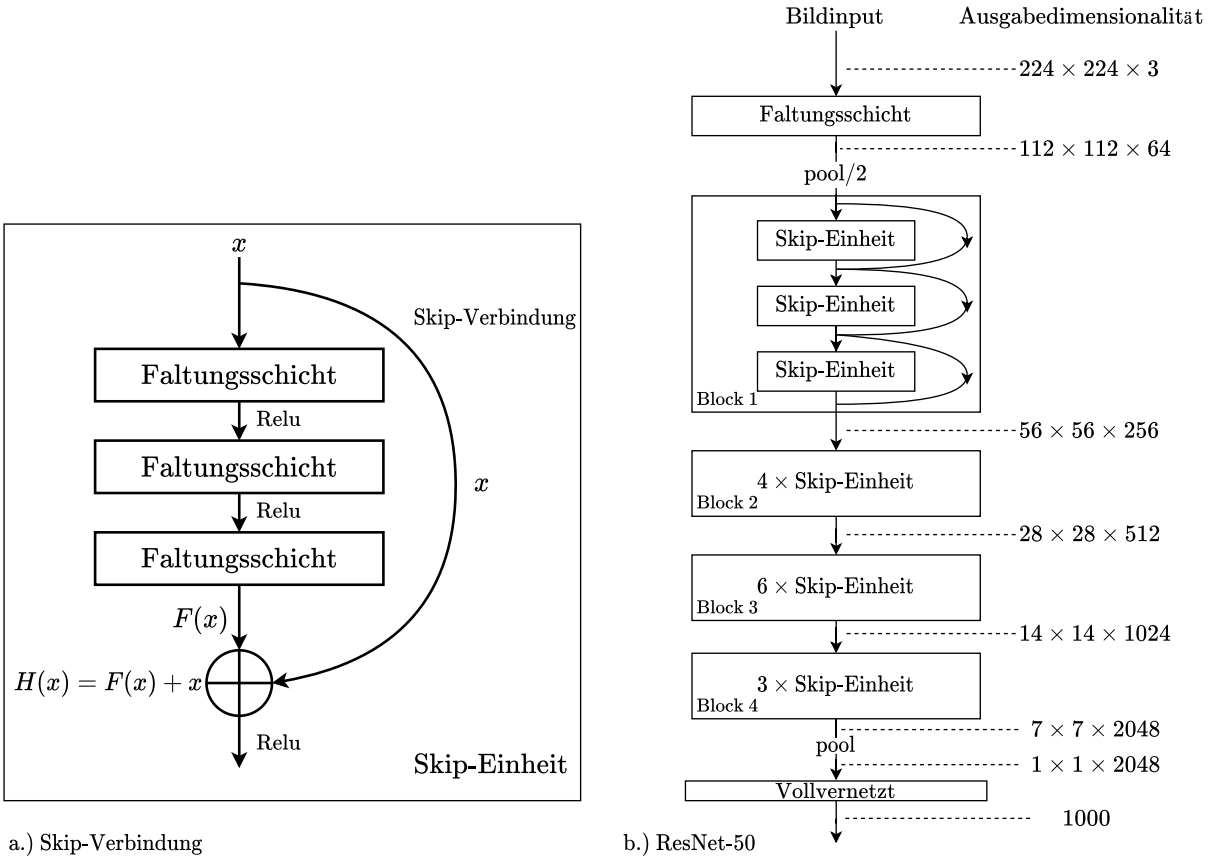

Abbildung 3.1: Aufbau der ResNet-Architektur (vgl. Fig.2, Fig.3 aus [19])

zu optimieren. Parameter konvergieren deutlich langsamer zu einem Optimum und es gibt mehr lokale Minima in denen ein Netzwerk im Optimierungsprozess stecken bleiben kann. ResNets wirken diesem Problem mit der Einführung sogenannter Skip-Verbindungen entgegen. Hierbei werden zusätzliche Direktverbindungen im Netzwerk geschaffen, bei denen einige Schichten übersprungen werden. Fließt eine Eingabe an den Beginn einer Skip-Verbindung, so wird auf dieser die Identität der Eingabe mitgeführt. Parallel durchläuft die Eingabe die übersprungenen Schichten. Am Ausgangspunkt der Verbindung wird schließlich die Ausgabe der übersprungenen Schichten mit der Identität aufsummiert (siehe Abb. 3.1a). Durch die Bereitstellung der Identität hat das Netzwerk eine bessere Grundlage zur Optimierung und einzelne schlecht optimierte Schichten weniger negative Auswirkung auf die Netzwerkausgabe. Die Autoren stellen fest, dass CNNs bei Verwendung von Skip-Verbindung schneller zu einem Optimum konvergieren und dabei bessere Minima gefunden werden.
ResNets können in unterschiedlichen Konfigurationen erstellt werden. Das für Delf verwendete ResNet-50 besteht aus 49 faltenden gefolgt von einer vollvernetzen Schicht. Skip-Verbindungen überspringen jeweils drei Schichten. Das Netzwerk kann in vier Blöcke unterteilt werden. Die Größe der einzelnen Featuremaps, der Ausgabe verringert sich nach jedem Block um den Faktor vier, wohingegen die Merkmalstiefe bzw. Anzahl der Featuremaps in der Ausgabe steigt (Siehe Abb.3.1b). In der Implementierung dieser Arbeit wird die von Torchvision zur Verfügung gestellte ResNet-50 Architektur genutzt[1].

[1] `https://github.com/pytorch/vision/blob/c2e8a00885e68ae1200eb6440f540e181d9125de/torchvision/models/resnet.py`, zuletzt besucht 16.06.20

3.2 Trainingsdaten

Um die Modelle für das Delf-Verfahren zu trainieren wird ein gelabelter Datensatz benötigt. Zum jetzigen Zeitpunkt steht kein solcher Datensatz von historischen Stadtaufnahmen mit ausreichender Größe zur Verfügung. Für diese Arbeit wird daher alternativ auf die Bilder der Google Landmark Challenge V2 [26] zurückgegriffen. Die Bilder entstammen einer Websuche auf der Wikimedia Datenbank[1] und zeigen Sehenswürdigkeiten aus der ganzen Welt. Der überwiegende Teil (72%) zeigt dabei menschengemachte Sehenswürdigkeiten, wie Kirchen, Museen oder Häuser. Auch wenn historische Aufnahmen keinen wesentlichen Teil der Bilder ausmachen enthält der Datensatz viele ähnliche Inhalte zu den historischen Datensätzen, die für das Retrieval genutzt werden. Der Trainingssatz der Landmark Challenge ist mit über 4 Millionen Bildern aus über 200k unterschiedlichen Kategorien sehr groß und heterogen. Da bei der Zusammenstellung keine Verifizierung der Bildinhalts durchgeführt wird kommt es häufig vor, dass Bilder in der falschen Kategorie einsortiert sind. Für das Delf-Training wird daher ein bereinigtes Subset des Datensatzes verwendet, welches von Yokoo, Ozaki et al. in Rahmen ihrer Arbeit [27] erstellt wurde. Aus dem bereinigten Datensatz werden die 40 häufigsten Kategorien gewählt und so ein Trainingsdatensatz über 39 790 gelabelten Bildern erstellt.

3.3 Fine-Tuning

Das Ziel der ersten Trainingsphase ist es das ResNet Modell so zu optimieren, dass das Modell bei der Verarbeitung eines Bildes Zwischenergebnisse erzeugt, die den Bildinhalt aussagekräftig beschreiben. Dies ist die Voraussetzung, um später leistungsstarke Deskriptoren erstellen zu können. Während der Optimierung versucht das Netzwerk die Klassifikationsaufgabe des Trainingsdatensatzes zu lösen. Zu Beginn werden die Netzwerkparameter dabei nicht zufällig initialisiert. Stattdessen wird ein Vortrainiertes Modell als Ausgangspunkt genutzt. Diese Art des Trainings wird als Fine-Tuning bezeichnet und ist eine gängige Methode, um den Trainingsprozess zu erleichtern. Auch in anderen Image Retrival Systemen [14] [20] wird diese Art des Trainings genutzt. Zeiler und Fergus zeigen in ihren Experimenten (vgl. [12] Kapitel 5.2), dass Netzwerke beim Training lernen allgemein nützliche Merkmale zu extrahieren, die sich auf unterschiedliche Datensätze anwenden lassen. Um ein vortrainiertes Netzwerk auf einen neuen Datensatz anzupassen sind daher nur kleine Veränderungen der Netzwerkparameter notwendig.
Als Ausgangspunkt für das Delf-Training wird ein auf ImageNet trainiertes ResNet-50 genutzt. Bei ImageNet handelt es sich um einen sehr große Klassifikationsdatensatz mit 1.4M Bildern aus 1000 sehr unterschiedlichen Kategorien. Durch die Vielfalt an Kategorien eigenen sich auf ImageNet trainierte Netzwerke als Ausgangspunkt für viele Klassifikationsaufgaben. Daher stellen die meisten Machine Learning Frameworks auf ImageNet trainierte Netzwerke zur Verfügung[2].
Während dem Training erwartet das Netzwerk quadratische Bilder mit einer Seitenlänge von 224 Pixeln und 3 Farbkanälen als Eingabe. Hierfür werden die Trainingsdaten zunächst quadratisch zugeschnitten und auf 250×250 Pixel skaliert. Anschließend wird ein zufälliger 224×244 Pixelbereich ausgewählt. Um das vortrainierte Netzwerk auf dem Trainingssatz weiter zu trainieren muss das Netzwerk so angepasst werden, dass die Ausgaben der letzten Schicht die korrekte Form für die zur Optimierung verwendete

[1] `https://commons.wikimedia.org`, zuletzt besucht 18.06.20
[2] `https://pytorch.org/docs/stable/torchvision/models.html`, zuletzt besucht 23.06.20

Fehlerfunktion hat. Als Fehlerfunktion wird hier der Cross-Entropy Loss berechnet:

$$\text{CrossEntropyLoss}(Y, \hat{Y}) = -\sum_{\forall c \in C} Y(c) * \log \hat{Y}(c) \tag{3.1}$$

$\hat{Y}$ ist hierbei die Verteilung der Klassenwahrscheinlichkeiten, die das Modell für eine Eingabe vorhergesagt hat. $\hat{Y}(c)$ ist die vom Netzwerk bestimmte Wahrscheinlichkeit, mit der die Eingabe der Klasse c zuzuordnen ist. Y beschreibt die tatsächliche Kategorie der Eingabe. Y ist also ein Verteilung, bei der die Wahrscheinlichkeit für jede, bis auf die korrekte Kategorie 0 und für die tatsächliche Klasse 1 ist. Als Ausgabe des Netzwerks wird also ein Vektor der Wahrscheinlichkeitsverteilung erwartet, dessen Dimensionalität der Anzahl der unterschiedlichen Klassen $|C|$ im Datensatz entspricht und dessen Einträge sich auf 1 summieren.
Damit die Ausgaben des Netzwerks die richtige Dimensionalität haben wird zunächst die letzte vollvernetzte Schicht entfernt. An ihre Stelle tritt eine faltenden Schicht mit 1×1 Filtermasken, die eine Merkmalstiefe von 2048 erwarten, was der Dimensionalität vorangehenden Schicht entspricht (vgl. Abb. 3.1b). In der faltenden Schicht werden $|C|$ dieser Filtermasken auf die Eingabe angewendet, wodurch die Ausgabe die gewünschte Dimensionalität erhält. Um die Netzwerkausgaben in den richtigen Wertebereich zu überführen, werden diese von einer Softmaxfunktion aktiviert, bevor die Fehlerfunktion berechnet wird:

$$\text{Softmax}(\hat{Y}')_c = \frac{e^{\hat{Y}'_c}}{\sum_{\forall x \in C} e^{\hat{Y}'_x}} \forall c \in C \tag{3.2}$$

Hierbei ist $\hat{Y}'$ ein Ausgabevektor des Netzwerks und $\hat{Y}'_c$ der Eintrag des Vektors welcher zur Klasse c zugeordnet ist. Der resultierende Vektor kann als Wahrscheinlichkeitsverteilung über die unterschiedlichen Klassen im Bezug zur Eingabe interpretiert werden. Mit den vorgenommenen Modifikationen kann das ResNet Modell auf dem Trainingsdatensatz optimiert werden. Die für das Fine-Tuning und Attention-Training verwendeten Hyperparamter werden im Experimentalteil in Sektion **??** erläutert.

3.4 Attention Training

Delf erzeugt eine große Anzahl an lokalen Deskriptoren, um Bilder zu beschreiben. Da jeder Deskriptor nur einen Ausschnitt des Originalbildes beschreibt, werden auch Deskriptoren für wenig aussagekräftige Bereiche, wie z.B. Teile des Himmels erstellt. Diese Deskriptoren beanspruchen nicht nur zusätzliche Rechenzeit während des Matchingprozesses sondern können auch zu falsch positiven Matches führen. Ziel der zweiten Trainingsphase ist es daher auf Basis dieser Deskriptoren ein Netzwerk zu trainieren, welches in der Lage ist die Qualität der Deskriptoren zu bewerten und so ungeeignete Kandidaten herauszufiltern. Zur Erstellung der Deskriptoren dient das ResNet-50, welches in der ersten Phase trainiert wurde. Als Deskriptoren werden dabei die Ausgaben aus dem dritten ResNet-Blocks genutzt (vgl. Abb. 3.1b). Die Ausgaben haben eine Dimensionalität von $w \times h \times 1024$, wobei w und h abhängig von der Breite und Höhe des Eingabebildes sind. Die einzelnen Koordinaten der 1024 Featuremaps lassen sich jeweils auf einen Bildbereich in der Eingabe zurückführen. So kann die Ausgabe in $w \times h$ Deskriptoren der Größe 1024 eingeteilt werden. Wie auch beim Fine-Tuning werden die Bilder für das Training zunächst quadratisch zugeschnitten. Anschließend werden sie auf eine zufällige Seitenlänge zwischen 255 bis 720 Pixel skaliert. Die Skalierung hat Einfluss auf die Beschaffenheit der entstehenden Deskriptoren.

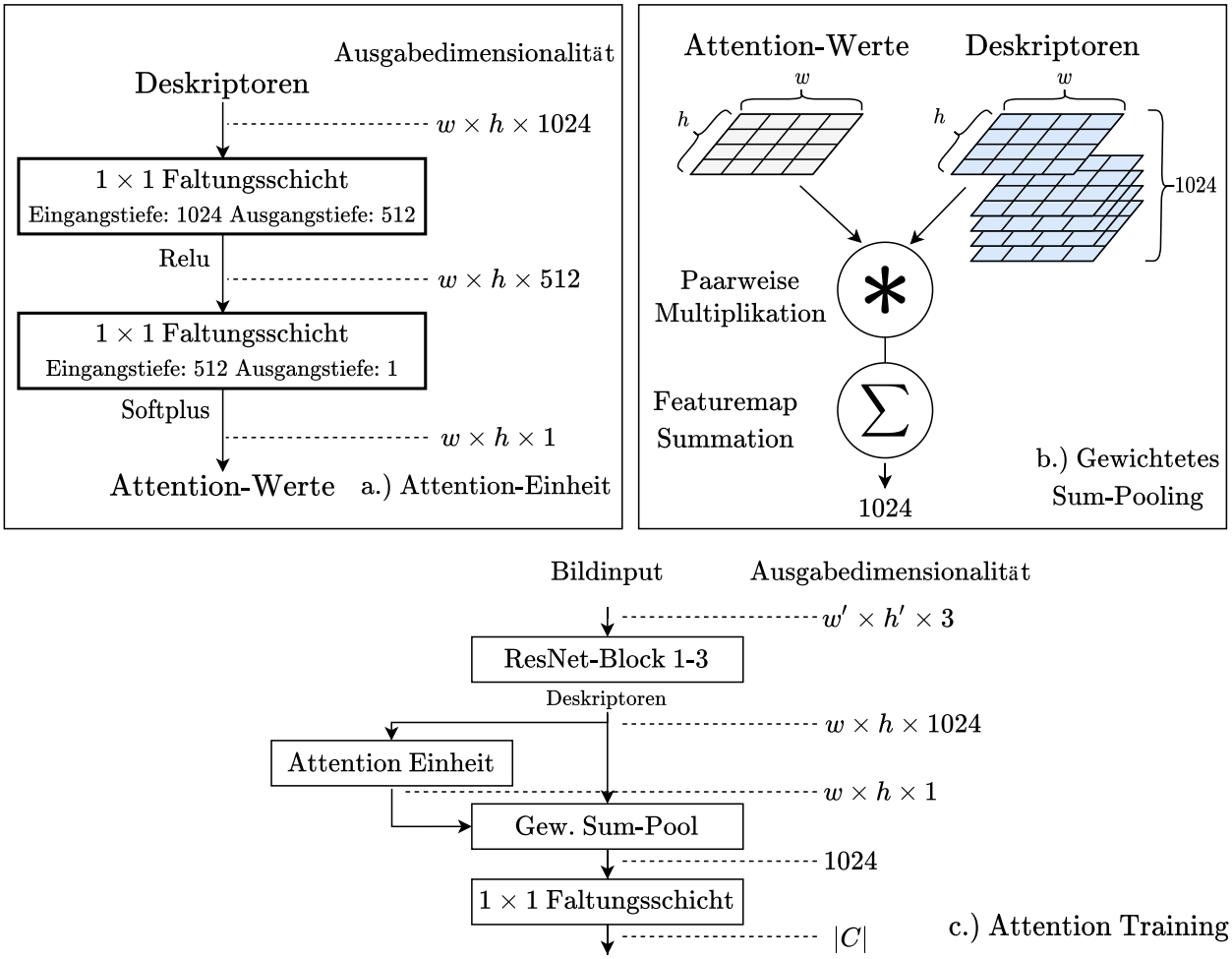

Abbildung 3.2: Architektur des Attention Trainings

Durch das zufällige Skalieren der Trainingsbilder lernt das Attention-Netzwerk mit Deskriptoren aus verschiedenen Skalen umzugehen. Während dem Attention-Training werden die Parameter des ResNets nicht mehr verändert. Die Schichten nach dem Extraktionspunkt in Block 3 werden nicht mehr benötigt und können verworfen werden.

Aufgabe des Attention-Netzwerks ist es für eine Eingabe an Deskriptoren der Form $w \times h \times 1024$ eine einzelne Featuremap der Größe $w \times h$ zu generieren, dessen Werte jeweils einen Deskriptor bewerten. Wichtig ist dabei, dass die Berechnung der einzelnen Attention-Werte nur von den Werte der dazugehörigen Deskriptoren abhängen dürfen. Dies kann durch faltenden Schichten mit 1×1-Filtermasken realisiert werden. Die Attention-Einheit besteht aus zwei solcher Schichten, welche die Merkmalstiefe der Deskriptoren sukzessive auf 1 reduzieren (vgl. Abb. 3.2a). Um die Parameter der Attention-Einheit zu optimieren müssen ihre Ausgaben zur Lösung der Trainingsaufgabe beitragen. Da die Attention-Werte später genutzt werden um zu entscheiden, welche Deskriptoren Einfluss auf die Lösung der Retrievalaufgabe haben ist es sinnvoll sie auch beim Training in einer Form zu nutzen, die den Einfluss der Deskriptoren zur Lösung der Klassifikationsaufgabe reguliert. Die Attention-Werte werden zur Gewichtung der Deskriptoren genutzt und dafür elementweise mit den Featuremaps der Deskriptoren Multipliziert. Anschließend werden die gewichteten Featuremaps zu jeweils einem Wert aufsummiert. Als Ausgabe ergibt sich ein 1024 dimensionaler Vektor (vgl. Abb. 3.2b). Abschließend muss die Ausgabe in eine pas-

sende Form für die Cross-Entropy Loss Fehlerfunktion gebracht werden. Dies geschieht analog wie im Fine-Tuning durch Verwendung einer 1×1-Faltungsschicht und anschließender Softmaxaktivierung der Ausgabe (vgl. 3.2c).

3.5 Extraktion und Verarbeitung

Nachdem das Training der Modelle abgeschlossen ist, kann mit der Extraktion aller benötigten Informationen über den Retrievaldatensatz begonnen werden. Netzwerkparameter werden ab jetzt nicht mehr modifiziert. Schichten und Operationen nach der Attention-Einheit erfüllen daher keine Zweck mehr und können entfernt werden.

3.5.1 Multi-Skalen-Extraktion

Für jedes Bild des Retrievaldatensatzes werden die lokalen Deskriptoren am Extraktionspunkt nach dem dritten ResNet-Block und die dazugehörigen Attention-Werte nach der Attention-Einheit extrahiert. Da die Skalierung des Bildinhalts Einfluss auf die resultierenden Deskriptoren hat, wird für jedes Bild eine Reihen von unterschiedlich skalierten Versionen betrachtet. Dies ist vergleichbar mit der Verwendung des Scale Spaces im SIFT-Verfahren (vgl. Kap.2 Abs.2) und soll zur Invarianz gegenüber Skalierungsoperationen beitragen. Für jedes Bild werden sechs unterschiedlichen Skalen mit Skalierungsfaktoren zwischen 2 und $\frac{1}{4}$ erstellt, wobei sich benachbarte Skalen um den Faktor $\sqrt{2}$ unterscheiden.

3.5.2 Deskriptorlokalisierung

Für den weiteren Verlauf des Verfahrens muss jeder lokale Deskriptor einem Bereich des Eingangsbildes zuortbar sein. Bei allen verwendeten Schichten des Modells bis zum Extraktionspunkt handelt es sich um Faltungs- oder Poolingschichten. Von welchen Bereichen der Eingabe die Ausgaben dieser Schichten abhängen lässt sich an drei Parametern festmachen. Die Größe der Filtermasken k bestimmt die Größe des Einflussbereiches einzelner Ausgaben. Die Verschiebung der Filtermasken bzw. die Schrittgröße s bestimmt die Verschiebung zwischen den Einflussbereichen der Ausgaben. Das Padding p bestimmt die Größe des Pufferbereichs, der der Eingabe hinzugefügt wird, und sorgt so für eine initiale Verschiebung der Einflussbereiche. Das Padding ist im folgenden immer symmetrisch und wird daher an jeder Seite der Eingabe hinzugefügt. Betrachtet man mehrere aufeinanderfolgende Schichten, so lassen sich die Einflüsse dieser Parameter wie folgt rekursiv berechnen:

$$\hat{k}_n = \hat{k}_{n-1} + ((k_n - 1) * \hat{s}_{n-1}) \tag{3.3}$$

$$\hat{s}_n = \hat{s}_{n-1} * s_n \tag{3.4}$$

$$\hat{p}_n = \hat{p}_{n-1} + (p_n * \hat{s}_{n-1}) \tag{3.5}$$

$$\hat{k}_0 = k_0 \tag{3.6}$$

$$\hat{s}_0 = s_0 \tag{3.7}$$

$$\hat{p}_0 = p_0 \tag{3.8}$$

Wobei $\hat{k}_n$, $\hat{s}_n$ und $\hat{p}_n$ die Größe der Einflussbereiche, Schrittgröße und Paddinggröße im Bezug zur ursprüngliche Eingabe nach n Schichten repräsentieren. k_n, s_n und p_n zeigen die selben Größen für Schicht

n im Bezug zur Ausgabe der vorangehenden Schicht. In Abbildung **??** werden diese Berechnungen exemplarisch erklärt. Berechnet man diese Werte für den Extraktionspunkt nach dem dritten ResNet-Block ergibt sich für jeden lokalen Deskriptor ein quadratischer Einflussbereich $\mathbf{k}$ mit einer Seitenlänge von 267 Pixeln im Ursprungsbild. Die Verschiebung zwischen Einflussbereichen benachbarter Deskriptoren $\mathbf{s}$ beträgt dabei 16 Pixel. Das Ursprungbild erhält bis zu dieser Schicht ein effektives Padding $\mathbf{p}$ von 133 Pixeln in jede Richtung. Aufgrund dieser Werte lassen sich die Einflussbereiche für alle $w \times h$ Deskriptoren, die am Extraktionspunkt anfallen wie folgt berechnen:

$$x_{min}(i,j) = i * \mathbf{s} - \mathbf{p} \tag{3.9}$$

$$x_{max}(i,j) = x_{min}(i,j) + \mathbf{k} \tag{3.10}$$

$$y_{min}(i,j) = j * \mathbf{s} - \mathbf{p} \tag{3.11}$$

$$y_{max}(i,j) = y_{min}(i,j) + \mathbf{k} \qquad \text{wobei } 0 \leq i < w \text{ und } 0 \leq j < h \tag{3.12}$$

Literaturverzeichnis

[1] Ferdinand Maiwald, Jonas Bruschke, Christoph Lehmann, and Florian Niebling. A 4D Information System for the Exploration of Multitemporal Images and Maps using Photogrammetry, Web Technologies and VR/AR. *Virtual Archaeology Review*, 10:1, 07 2019.

[2] Hyeonwoo Noh, Andre Araujo, Jack Sim, Tobias Weyand, and Bohyung Han. Large-Scale Image Retrieval with Attentive Deep Local Features. pages 3476–3485, 10 2017.

[3] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[4] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Lost in Quantization: Improving Particular Object Retrieval in Large Scale Image Databases. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[5] Yan-Tao Zheng, Ming Zhao, Yang Song, Hartwig Adam, Ulrich Buddemeier, Alessandro Bissacco, Fernando Brucher, Tat-Seng Chua, Hartmut Neven, and Jay Yagnik. Tour the World: A Technical Demonstration of a Web-Scale Landmark Recognition Engine. In *Proceedings of the 17th ACM International Conference on Multimedia*, MM '09, page 961–962, New York, NY, USA, 2009. Association for Computing Machinery.

[6] Arnold W. M. Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain. Content-Based Image Retrieval at the End of the Early Years. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(12):1349–1380, December 2000.

[7] David Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60:91–, 11 2004.

[8] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. SURF: Speeded up Robust Features. volume 3951, pages 404–417, 07 2006.

[9] Yan Ke and Rahul Sukthankar. PCA-SIFT: A more Distinctive Representation for Local Image Descriptors. volume 2, pages II–506, 05 2004.

[10] Cordelia Schmid and J. Ponce. Semi-Local Affine Parts for Object Recognition. *BMVC04*, 08 2004.

[11] Miaojing Shi, Yannis Avrithis, and Hervé Jégou. Early Burst Detection for Memory-Efficient Image Retrieval. 06 2015.

[12] Matthew Zeiler and Rob Fergus. Visualizing and Understanding Convolutional Neural Networks. volume 8689, 11 2013.

[13] Artem Babenko, Anton Slesarev, Alexandr Chigorin, and Victor Lempitsky. Neural Codes for Image Retrieval. volume 8689, 04 2014.

[14] Ali S. Razavian, Josephine Sullivan, Stefan Carlsson, and Atsuto Maki. Visual Instance Retrieval with Deep Convolutional Networks. *ITE Transactions on Media Technology and Applications*, 4(3):251–258, 2016.

[15] Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton. ImageNet Classification with Deep Convolutional Neural Networks. *Neural Information Processing Systems*, 25, 01 2012.

[16] Lingxi Xie, Richang Hong, Bo Zhang, and Qi Tian. Image classification and retrieval are one. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, ICMR '15, page 3–10, New York, NY, USA, 2015. Association for Computing Machinery.

[17] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv 1409.1556*, 09 2014.

[18] J. R. Uijlings, K. E. Sande, T. Gevers, and A. W. Smeulders. Selective Search for Object Recognition. *Int. J. Comput. Vision*, 104(2):154–171, September 2013.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. 7, 12 2015.

[20] Filip Radenović, Giorgos Tolias, and Ondřej Chum. CNN Image Retrieval Learns from BoW: Unsupervised Fine-Tuning with Hard Examples. volume 9905, pages 3–20, 10 2016.

[21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton. ImageNet Classification with Deep Convolutional Neural Networks. *Neural Information Processing Systems*, 25, 01 2012.

[22] Andrew McCallum and Kamal Nigam. A comparison of event models for naive bayes text classification. In *AAAI 1998*, 1998.

[23] James B. MacQueen. Some Methods for Classification and Analysis of Multivariate Observations. 1967.

[24] Hervé Jégou, Matthijs Douze, Cordelia Schmid, and Patrick Perez. Aggregating Local Descriptors into a Compact Image Representation. pages 3304 – 3311, 07 2010.

[25] Jingdong Wang, Heng Shen, Jingkuan Song, and Jianqiu Ji. Hashing for Similarity Search: A Survey. 08 2014.

[26] Tobias Weyand, Andre Araujo, Bingyi Cao, and Jack Sim. Google Landmarks Dataset v2 – A Large-Scale Benchmark for Instance-Level Recognition and Retrieval, 2020.

[27] Shuhei Yokoo, Kohei Ozaki, Edgar Simo-Serra, and Satoshi Iizuka. Two-stage Discriminative Reranking for Large-scale Landmark Retrieval, 2020.