Möglichkeiten und Herausforderungen bei der Verwendung von Bildern in Suchmaschinen

Timo Schulz¹, Björn Zessack², Bengt Rabe³ und Sebastian Kielmann⁴

Abstract: Das Datenvolumen, das Suchmaschinen indexieren, ist in den letzten Jahren rasant angestiegen und zusätzlich zu der klassischen Domäne Text, bekommen Multimedia-Dateien zunehmend eine wichtige Bedeutung. Dazu zählen in erster Linie Bilder, aber auch Sprache und Filme. Zusätzlich zum kontinuierlichen Anstieg des Volumens bedeutet die Erschließung dieser neuen Domänen aber auch, dass neue Verfahren benötigt werden, um Nicht-Text-Daten ausreichend zu beschreiben und sinnvoll durchsuchbar zu machen und das zu der Erschließung der Inhalte auch neue Verfahren für die Indexierung benötigt werden, da im Gegensatz zu Text, Multimedia-Dateien ein wesentlich größeres Datenvolumen beinhalten. Nachfolgend soll eine Übersicht gegeben werden, welches Potenzial eine Einbindung der neuen Domain in Suchmaschinen bringt, aber auch die Herausforderungen und Probleme, wie z.B. die Einbindung in bestehende User-Interfaces und die Mehrdeutigkeit bei der Bestimmung der Nutzerabsicht. Als konkreter Fall wird dabei der Bereich e-Commerce mit Schwerpunkt Shopping betrachtet.

Keywords: Bild, Deep Learning, GPU, e-Commerce

1 Einleitung

Mit der zunehmenden Verbreitung von Breitband-Internetzugängen, nimmt das Datenvolumen im Internet rasant zu, da es immer einfacher möglich ist, mehr und größere Daten schnell auszutauschen und zu präsentieren. Schon 2001 gab es Suchmaschinen, die über 250 Millionen [Sm10] Bilder indexiert hatten und die heutige Zahl dürfte bei über 10 Milliarden liegen. Im Gegensatz zu den Anfängen der Suchmaschinen, wo oft nur Text indexiert wurde, stellt die Indexierung beliebiger Multimedia-Dateien eine deutlich größere Herausforderung dar, da mehr Speicherplatz und mehr Rechenleistung für die Erschließung und Indexierung verwendet wird.

Im Gegensatz zur Textsuche, wo die Eingabe lediglich aus einer Menge Stichworten, bzw. kurzen Sätzen besteht, stellt die Suche in Multimedia-Dateien eine zusätzliche Herausforderung dar, da es mehrere Modalitäten bei der Eingabe geben kann, die mit der Absicht des Nutzers in Einklang gebracht werden müssen. Im Beispiel von Bildern gilt es festzulegen, wie/was genau gesucht werden sollten. Die folgende Liste enthält einige Beispiele:

¹ picalike GmbH, Stresemannstraße 23 D-22769, t.schulz@picalike.com.

² picalike GmbH, Stresemannstraße 23 D-22769, b.zessack@picalike.com.

³ picalike GmbH, Stresemannstraße 23 D-22769, b.rabe@picalike.com.

⁴ picalike GmbH, Stresemannstraße 23 D-22769, s.kielmann@picalike.com.

- Stichworte: visueller Inhalt ist mit Schlagworten verknüpft
- Freitext: Text wird aufbereitet und in einem semantischen Raum transformiert
- Skizze: Erfordert Hochladen und Skelett-Indexierung der Bilder
- Beispielbild: Semantische, globale Ähnlichkeitssuche
- Kombination mehrerer Möglichkeiten

Des Weiteren gibt es verschiedene Anwendungsfälle der Bildersuche:

- Exakte Suche: Finde alle Bilder, die z.B. eine bestimmte Person/Objekt zeigen
- Suche von (Fast-)Duplikaten: Finde alle Bilder, beliebig nah an der Referenz sind
- Konzept-Suche: Finde alle Bilder, die zu einem Konzept passen (Rotes Ballkleid mit Rüschen und Punkten)

Im weiteren beziehen wir uns auf die Konzept-Suche anhand eines Beispielbildes. Für diesen Fall ist unsere These, dass ein vollständiges, maschinelles Bildverständnis einer reinen Textsuche oft überlegen ist. Weiterhin ist eine Ähnlichkeitssuche auf Bildern, gerade im Bereich Mode, einer reinen Textsuche in den meisten Fällen überlegen, da Produktbeschreibungen begrenzt, bzw. fehleranfällig sind und damit das Suchvokabular begrenzt ist. Zeigt z.B. ein Bild eine Bluse mit blauen, auffälligen Knöpfen, aber keine Erwähnung in der textuellen Beschreibung, kann dieses Detail für eine Suche nicht verwendet werden.

2 Nachteile der klassischen Textsuche für die Bildersuche

Bilder spielen im Internet eine immer größere Rolle. Gerade im e-Commerce gibt es Bereiche wie Mode, in denen Entscheidungen oft rein visuell, bzw. ästhetischer Natur sind. Ohne eine genaue Erfassung der Bilddaten kann eine Suche aus Sicht von Nutzern nicht zufriedenstellend durchgeführt werden, da der Erfolg von einer vollständigen, manuellen Annotation des Bildmaterials abhängt. Das heißt, es muss eine Suche auf konzeptioneller Ebene möglich sein, die über eine reine Stichwortsuche hinausgeht. Ein Vorteil ist, dass zur Beschreibung der Konzepte das Bild selbst verwendet werden kann und es somit nicht notwendig ist, dass eine manuelle Annotation für ein Bild vorliegt.

In der Vergangenheit, wurde oft der umliegende oder Anker-Text verwendet, um Bildinhalte auf Seiten zu beschreiben[He09, Me09, FC09]. Der Ansatz ist allerdings nicht zuverlässig, da es keine Garantie gibt, dass der Text eine ausreichende Relevanz zu dem Bild hat. Auch werden Mehrdeutigkeiten wie Jaguar (Auto oder Raubkatze) auf dieser Ebene nicht aufgelöst und es ist möglich, dass nur ein kleiner Aspekt des Bildes beschrieben wird. Generell ist es bei manueller Annotation auch möglich, dass sich die

Relevanz nachträglich oder zeitlich ändert. Dass in frühen Zeiten vermieden wurde, Bildinhalte zu analysieren, ist auch der mangelnden damaligen Leistung der CPUs geschuldet, da selbst moderne Verfahren ohne GPU schnell an ihre Grenzen geraten, wenn es darum geht Millionen von Bildern mittels Deep-Learning-Verfahren zu verarbeiten.

Ein anderes Problem bei der manuellen Annotation von Bildern ist die Mehrdeutigkeit von Metadaten durch individuelle Schemata, aber auch Details wie die Nutzung verschiedener Sprachen, wodurch katalogübergreifende Lösungen oft nicht durchführbar sind, oder nur mit erheblichem Mehraufwand. Auch ist nicht auszuschließen, dass Daten fehlerhaft oder unvollständig vorliegen, was gravierende Einschränkungen bei der Suche durch Nutzer haben kann. Insgesamt ist eine Erschließung des Bildmaterials durch eine manuelle Annotation fehleranfällig und zeitaufwendig und nicht zeitnah möglich. Eine einheitliche Taxonomie wäre in jedem Fall von Vorteil, um alternative Fachbegriffe und unscharfe Formulierungen zu behandeln, aber auch um unterschiedliche Konzepte klar abzugrenzen.

Allerdings kann es auch bei einer direkten Verwendung des Bildes eine Taxonomie notwendig sein, um eine konsistente Darstellung zu ermöglichen. Für eine mögliche Stichwortsuche in Bildkonzepten ist es somit notwendig, Eingaben von Nutzer direkt in die Taxonomie zu überführen, z.B. durch eine automatische Vervollständigung der Eingabe, was allerdings voraussetzt, dass Mehrdeutigkeiten, wie Sprache, bereits auf anderer Ebene behandelt wurden. Hierbei wäre eine Möglichkeit, die Konzepte für das Bildmaterial sprachneutral zu codieren und dann für jede Sprache ein Mapping anzulegen, was allerdings nicht immer konsistent ist, da Konzepte in einigen Sprachen sehr vielfältig beschrieben werden können.

3 Vorteile durch die Verwendung von Bilddaten

Da ein Bild unabhängig von Sprache selbst-beschreibend ist, also alle relevanten Informationen enthält, um ein Objekt zu beschreiben, besteht die Herausforderung darin, eine geeignete Methode zu finden, die rohen Pixelwerte in eine konzeptionelle Sicht zu überführen. In der klassischen Bildverarbeitung, vor dem Siegeszug von Deep Learning, wurden Bilder oft durch lokale Features beschrieben, wie z.B. HoG[DT05], SIFT[Lo04], SURF[Ba08] oder globale Beschreibungen wie GIST[Do09], die dann in eine kompakte Repräsentation überführt wurden. Da diese Verfahren allerdings nicht für einen speziellen Zweck optimiert wurden, sind die Ergebnisse oft nicht zufriedenstellend gewesen. Seit der Beschleunigung durch GPUs werden deshalb immer öfter Modelle verwendet, die gezielt für einen speziellen Zweck trainiert werden können. Der Vorteil hierbei ist, dass Bildfeatures nicht mehr generisch sind, sondern passend zum Problem erzeugt werden. Allerdings sind auch bei Deep Learning die meisten Verfahren auf Annotationen angewiesen, so dass Veränderungen bei Taxonomien eine Anpassung und ein erneutes Training der Modelle zur Folge haben können. Weiterhin ist ein Abwägen notwendig, um ein Mittelmaß zwischen Genauigkeit und vorhandenen Ressourcen zu finden, da selbst mit Deep Learning die Extraktion von allen relevanten Bildinformationen in einer

angemessenen Zeitspanne ein Thema für zukünftige Forschung ist. Bei der Klassifikation haben die Modelle allerdings schon übermenschliche Genauigkeit[Ru14] erreicht und durch einen großen Label-Raum, zum Beispiel ImageNet[De09], sind die Features solche Modelle zum Teil vielseitig einsetzbar, wodurch die Einführung neuer Attribute nicht zu einem kompletten neuen Training führt, sondern stückweise Erweiterungen erlauben. Es soll nun ein konkreter Fall für eine Ähnlichkeitssuche betrachtet werden.

3.1 Use Case

Mode Im Fall von einer e-Commerce Shopping-Seite gehen wir von einer Seite aus, die eine einfache Suche nach Stichworten oder Artikelnummer erlaubt. Von Interesse ist allerdings nur der Fall, dass ein Nutzer nur eine sehr grobe Vorstellung hat, was er eigentlich sucht. Das könnte z.B. ein Paar neue Sneaker sein, wobei allerdings noch nicht klar ist, was für genaue Eigenschaften das Produkt haben soll. Ähnlich der Textsuche, ist ein Startpunkt notwendig, der dann verfeinert wird. Dies könnte durch die Auswahl der Ober-Kategorie geschehen, oder aber durch die Auswahl von vordefinierten Einsprungspunkten, wie z.B. Trends. Je nach Ziel ist eine gewisse Vielfalt der Ergebnisse notwendig, um dem Nutzer seine Möglichkeiten aufzuzeigen. So ist es möglich, dass der Nutzer ein (globales) Konzept gefunden hat, dass er weiter verfeinern möchte, ohne aber vordefinierte Filter, wie Marke, zu verwenden. Das ausgewählte Bild fungiert dann als Referenz, um ähnliche Produkte zu finden, die verwandte Konzepte enthalten. Wie genau die Bewertung der Bilder stattfindet ist variabel, es kann z.B. Form/Farbe/Textur unterschiedlich bewertet werden, was ein klarer Vorteil gegenüber textlicher Beschreibung ist, da sich z.B. die Wahrnehmung von Farben nur schwer mit festen Filtern abbilden lässt. Genauso ist eine vollständige Aufzählung sinnvoller Texturen für eine spezielle Suche oft überhaupt nicht möglich. Obwohl bei Modebildern oft nur ein Objekt abgebildet ist, kann die Absicht eines Nutzers nicht ohne explizites Feedback inferiert werden. Eine Verfeinerung der Suche ist nämlich nur dann möglich, wenn klar ist, welches Detail auf dem Bild für die weitere Suche relevant ist, z.B. hoher Absatz, aber keine Schnürsenkel. Die Entscheidungen bei jeder Iteration können ebenfalls als Feedback in die Suche einfließen, Stichwort Ranking. Am Ende der Suche wurden die Produkte dann hoffentlich so eingegrenzt, dass ein Nutzer leicht seinen Favoriten wählen kann.

4 Zusammenfassung

Die These, dass die inhaltliche Verwendung von Bilddaten einen Mehrwert für nutzergetriebene Suchen darstellt, wurde exemplarisch belegt. Allerdings ist der Mehrwert stark davon abhängig, wie vollständig das Bild konzeptionell durch Features beschrieben werden kann. Hierzu ist einerseits ein gutes Modell notwendig, andererseits eine ausreichende, manuelle, Annotationen der Trainingsdaten. Das heißt, die Verwendung der Bildmodalität setzt weiterhin Text als Modalität voraus, um Bilder zu beschreiben. Des weiteren stellt die Verwendung von Bilddaten auch neue Anforderungen an

Suchmaschinen selbst dar, da es notwendig ist, große Mengen an Bilddaten kompakt zu beschreiben, ohne große Einbußen bei der Qualität, diese zu speichern und schnell abrufbar zu halten. Im speziellen Fall von Katalogbildern, die durch den Fokus auf nur ein Produkt, sehr oft keine Mehrdeutigkeit erlauben, ist eine Bestimmung der Nutzerabsicht einfacher, allerdings kann dieser Fall nicht auf natürliche Bilder verallgemeinert werden, was dazu führt, dass nicht immer klar ist, ob ein spezielles Konzept vom Benutzer bevorzugt wird. Das wirft die Frage auf, wie Suchmaschinen bei Verwendung von Bildern als Eingabe eine Verfeinerung oder Korrektur der Suchabsicht erlauben. Auch ist eine weitere Herausforderung in Hinsicht auf die Ressourcen, der kontinuierliche Bildupload von Nutzer, wenn betrachtet wird, dass eine normale Suchanfrage im Schnitt nur wenige kurze Worte enthält, wobei selbst ein kleines Farb-Bild (3*256*256) eine Größe von 192 KB hat. Auch wenn die Ressource Bild in den vergangenen Jahren sehr stark an Bedeutung gewonnen hat und die Fortschritte, zum Teil durch Deep Learning, sehr eindrucksvoll sind, bleiben noch viele offene Fragen und Herausforderungen, um die Modalität Bild in Suchmaschinen einzubinden.

Literaturverzeichnis

- [Ba08] Bay, Herbert; Ess, Andreas; Tuytelaars, Tinne; Van Gool, Luc: Speeded-Up Robust Features (SURF). *Comput. Vis. Image Underst.*, 110(3):346–359, Juni 2008.
- [De09] Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; Fei-Fei, L.: ImageNet: A Large-Scale Hierarchical Image Database. In: *CVPR09*. 2009.
- [Do09] Douze, Matthijs; Jégou, Hervé; Sandhawalia, Harsimrat; Amsaleg, Laurent; Schmid, Cordelia: Evaluation of GIST Descriptors for Web-scale Image Search. In: *Proceedings of the ACM International Conference on Image and Video Retrieval. CIVR '09*, ACM, New York, NY, USA, S. 19:1–19:8, 2009.
- [DT05] Dalal, Navneet; Triggs, Bill: Histograms of Oriented Gradients for Human Detection. In: *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1 - Volume 01*. *CVPR '05*, IEEE Computer Society, Washington, DC, USA, S. 886–893, 2005.
- [FC09] Filipe, Joaquim; Cordeiro, José, Hrsg. *WEBIST 2009 - Proceedings of the Fifth International Conference on Web Information Systems and Technologies*, Lisbon, Portugal, March 23-26, 2009. INSTICC Press, 2009.
- [He09] Hearst, Marti A.: *Search User Interfaces*. Cambridge University Press, 1. Auflage, 2009.
- [Lo04] Lowe, David G.: Distinctive Image Features from Scale-Invariant Keypoints. *Int. J. Comput. Vision*, 60(2):91–110, November 2004.
- [Me09] Metaxas, Panagiotis Takis: On the Evolution of Search Engine Rankings. In: (Filipe; Cordeiro, J.) [FC09], S. 200–207.

- [Ru14] Russakovsky, Olga; Deng, Jia; Su, Hao; Krause, Jonathan; Satheesh, Sanjeev; Ma, Sean; Huang, Zhiheng; Karpathy, Andrej; Khosla, Aditya; Bernstein, Michael S.; Berg, Alexander C.; Li, Fei-Fei: ImageNet Large Scale Visual Recognition Challenge. CoRR, abs/1409.0575, 2014.

- [Sm10] Smith, Nate: , Ooh! Ahh! Google Images presents a nicer way to surf the visual web. <https://googleblog.blogspot.de/2010/07/ooh-ahh-google-images-presents-nicer.html>, July 2010.