OCR4all

Eine semi-automatische Open-Source-Software für die OCR historischer Drucke

Installationsanleitung via VirtualBox

Version 1.0, Mai 2020

Um mit Blick auf zukünftige Versionen der Software und sonstige technische Neuerungen immer auf dem aktuellen Stand zu bleiben, empfehlen wir Ihnen, unsere Mailingliste OCR4all zu abonnieren.

Vorbereitungen

Die vorliegende Installationsanleitung ermöglicht es Ihnen, OCR4all schnell und einfach auf Ihrem Gerät einzurichten. Um ein reibungsloses Arbeiten mit der Software zu garantieren, wird OCR4all deshalb in eine sog. Virtuellen Maschine integriert. Diese funktioniert wie ein eigenes, kleines Betriebssystem auf Ihrem Gerät, welches immer dann aktiviert wird, wenn Sie OCR4all starten.
Die folgenden Vorbereitungsschritte beziehen sich auf den Download der für die Einrichtung einer Virtuellen Maschine notwendigen Software (VirtualBox) und den Download des OCR4all-Pakets:

1. Download der neuesten Version von VirtualBox für Ihr Betriebssystem auf www.virtualbox.org/wiki/Downloads. Klicken Sie dazu im Abschnitt „VirtualBox ... platform packages“ einfach auf „Windows hosts“ oder „OS X hosts“. Nach dem Download der Installationsdatei führen Sie diese aus und folgen den Installationsschritten der VirtualBox-Software.

2. Download des OCR4all-Pakets über www.kallimachos.de/ocr4all/vm-download.php. Die entsprechende Datei heißt „OCR4all_0.3.zip“. Nachdem Sie diese Datei heruntergeladen haben, entpacken Sie sie. Der Ihnen nun vorliegende Ordner „OCR4all_0.3“ enthält alle weiteren relevanten Daten.

3. Legen Sie den Ordner „OCR4all_0.3“ nun an einer beliebigen Stelle innerhalb Ihres Betriebssystems ab, z. B. in C:\.

In „OCR4all_0.3“ finden Sie folgende Struktur vor:

- Der Ordner „guides“ enthält ausführliche Anleitungen zur Benutzung der Software.

- Unter „image“ sind die Daten zur Einrichtung der OCR4all-Software abgelegt.

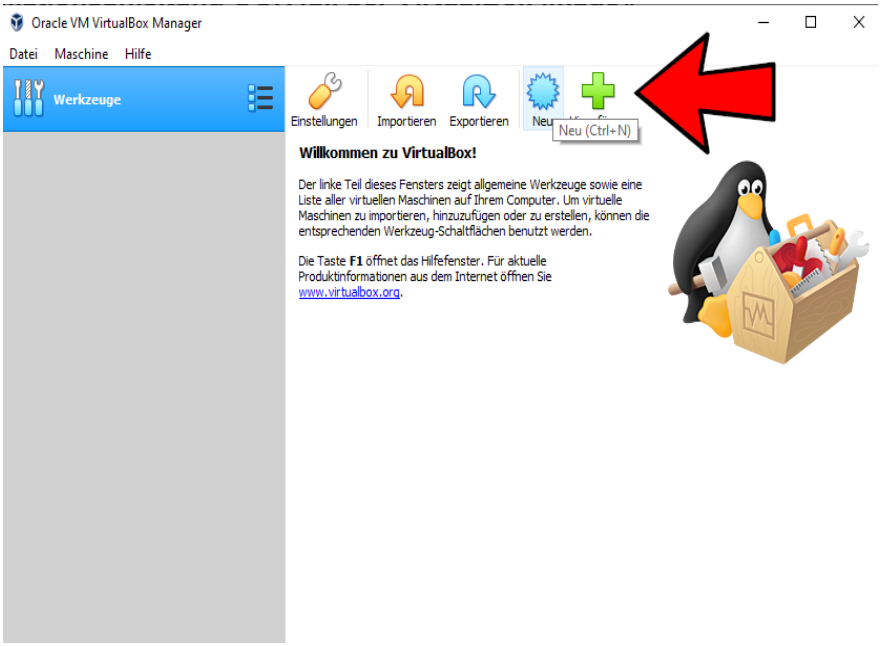
- In „ocr4all“ wiederum finden Sie die Ordner „data“ und „models“:
  - „data“ beinhaltet erste Testdaten für die Benutzung von OCR4all, d. h. die Scandateien unterschiedlicher Drucke. Diese sind genauso vorstrukturiert, wie Sie später auch Ihre eigenen Daten und Projekte anlegen sollten: ocr4all/data/*Werktitel*/input/*Scandateien*.
  - „models“ dagegen enthält OCR-Modelle, die zur Texterkennung genutzt werden. Auch hier liegen Ihnen bereits Daten vor, damit Sie OCR4all direkt verwenden können. OCR-Modelle, die Sie während Ihrer späteren Arbeit mit OCR4all selbst erstellen, werden automatisch in diesem Ordner gespeichert.

Aufsetzen der Virtuellen Maschine

Um für OCR4all eine neue Virtuelle Maschine anzulegen, starten Sie bitte die oben installierte VirtualBox-Software.

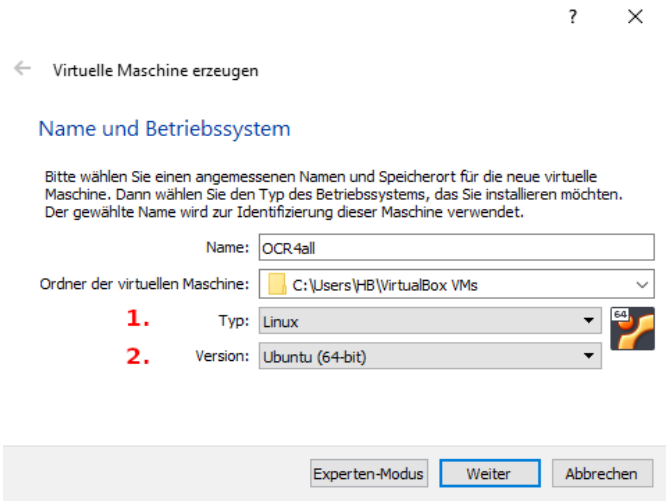
1. Anlegen einer neuen Maschine

Um eine neue Virtuelle Maschine anzulegen, klicken Sie auf „Neu“.



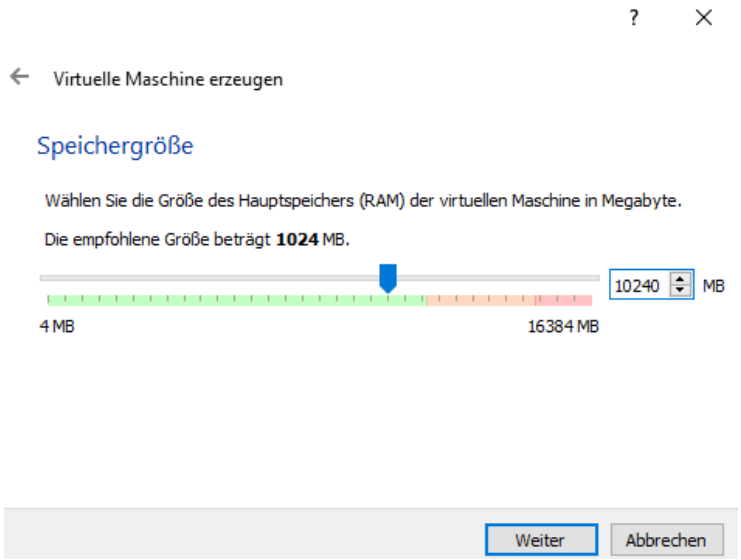
2. Parameter der Maschine setzen

Bitte beachten Sie hierbei, dass unabhängig von Ihrem eigenen Betriebssystem bei „Typ“ (1.) **Linux** und bei „Version“ (2.) **Ubuntu 64-Bit** ausgewählt werden muss. Unter „Name“ tragen Sie **OCR4all** ein, bei „Ordner der virtuellen Maschine“ muss keine Änderung durch Sie vorgenommen werden.



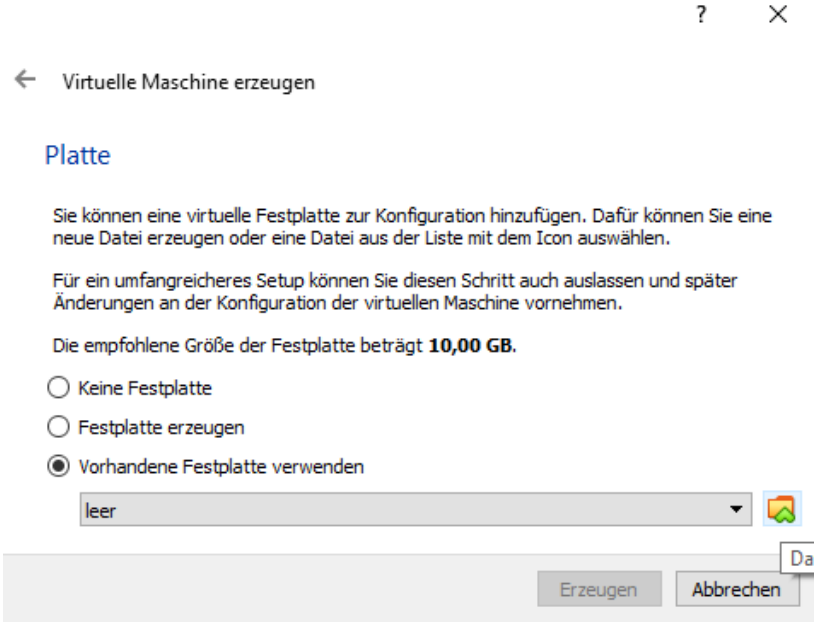
3. Ressourcenmanagement

In diesem Arbeitsschritt legen Sie fest, wie viel Arbeitsspeicher Ihres Betriebssystems der Virtuellen Maschine zur Verfügung gestellt werden soll. Wir empfehlen Ihnen, hier entsprechend Ihres eigenen Gerätes maximal die Größe des Hauptspeichers abzüglich etwa 2 GB anzugeben. Im vorliegenden Beispiel entspräche dies einer maximalen Angabe von 14384 MB.

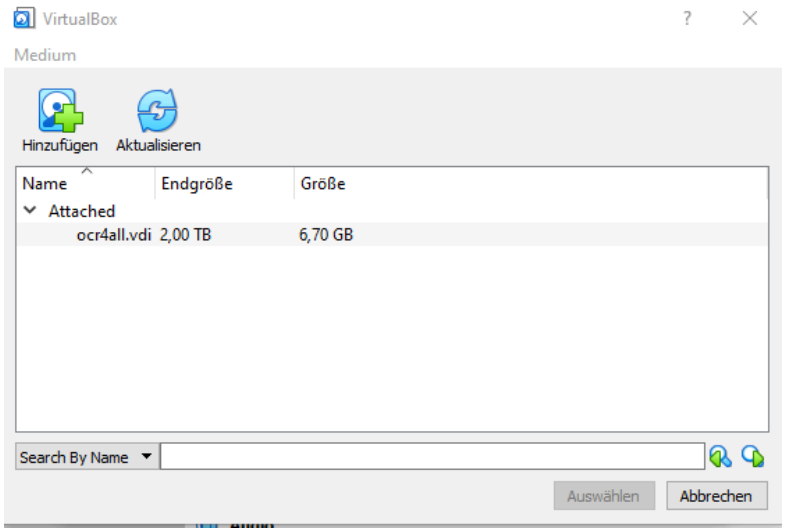


4. Festplatte einbinden

Um diesen Einrichtungsschritt durchzuführen, klicken Sie wie unten dargestellt zuerst auf „Vorhandene Festplatte verwenden“. Noch können Sie die Virtuelle Maschine nicht erzeugen.



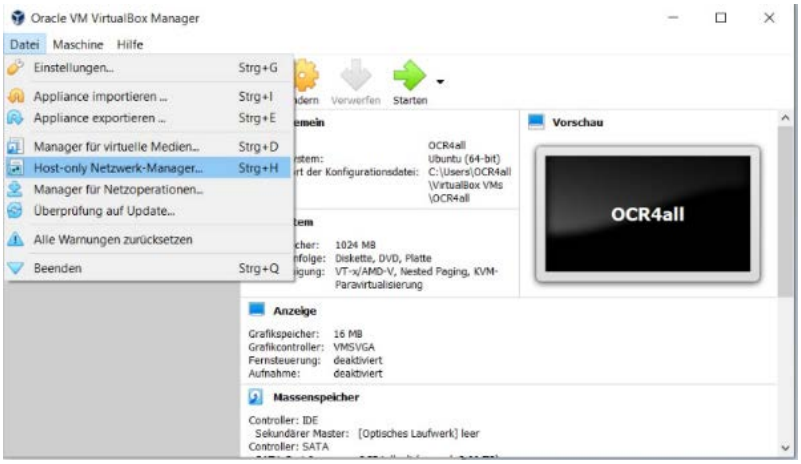
Wählen Sie deshalb anschließend über das kleine Ordner-Icon hinter „Leer“ das weiter oben heruntergeladene OCR4all-Image aus dem Ordner „OCR4all_0.3/image“ aus. Über den „Hinzufügen“-Button können Sie es nun der „vorhandenen Festplatte“ zuweisen.



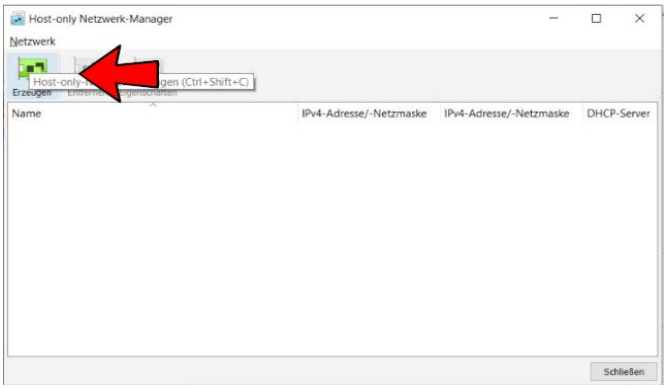
Klicken Sie danach auf „Erzeugen“, um die Einrichtung der Virtuelle Maschine namens OCR4all abzuschließen.

5. Einstellungen in der Virtuellen Maschine anpassen

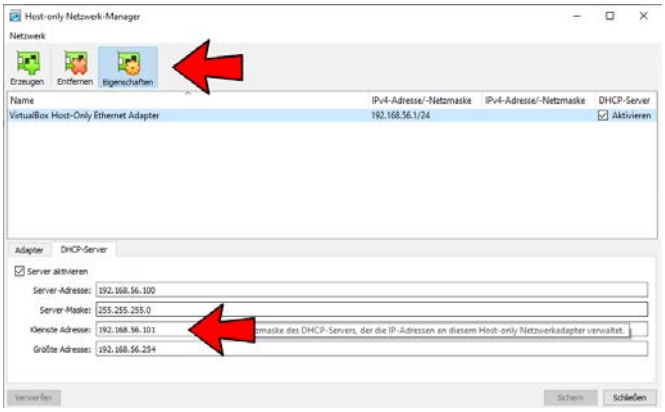
Um ein reibungsloses Funktionieren der Virtuellen Maschine zu garantieren, müssen nun noch einige weitere Einstellungen getätigt werden: Damit die Virtuelle Maschine und Ihr Gerät kommunizieren können, muss über den Menüpunkt **Datei** > **Host-only Netzwerk-Manager** ein sog. Netzwerkadapter erstellt werden.



Klicken Sie dazu wie hier dargestellt auf „Erzeugen“.

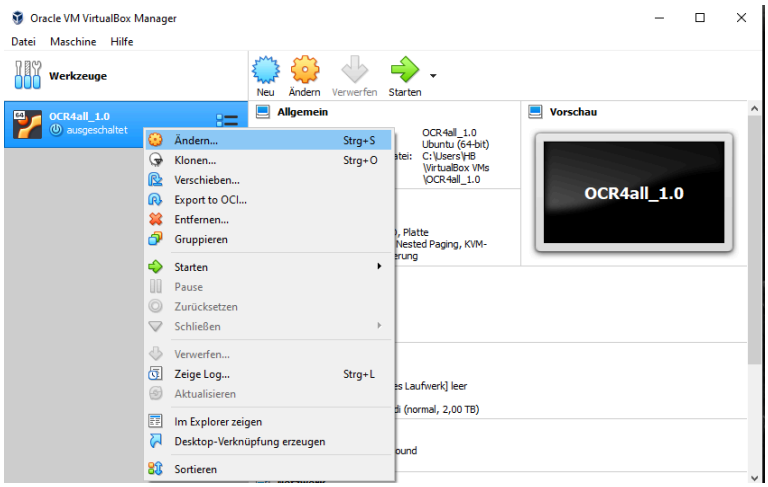


Nachdem Sie den Netzwerkadapter so erzeugt haben, klicken Sie wie nachfolgend dargestellt auf „Eigenschaften“. Nun notieren Sie sich für das spätere Starten von OCR4all die IP-Adresse der Virtuellen Maschine. Sie ist unter „Kleinste Adresse“ angegeben. In diesem beispielhaften Fall wäre diese IP also: 192.168.56.101

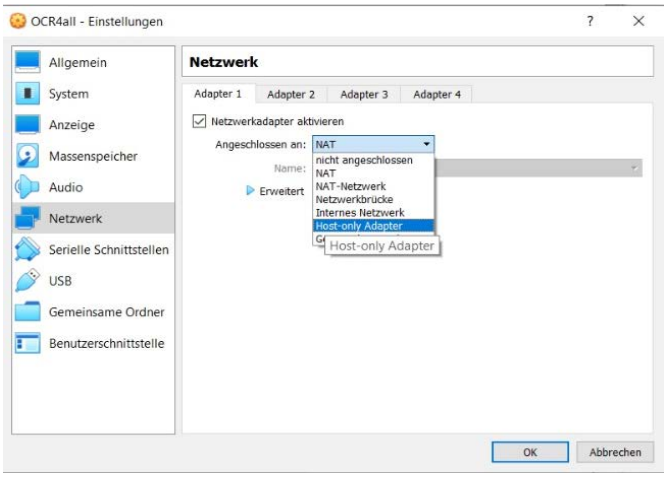


Schließen Sie nun den Host-only Netzwerk-Manager.

Sie schließen diesen Einrichtungsschritt ab, indem Sie nun in den Einstellungen der Virtuellen Maschine OCR4all in den Bereich **Netzwerk** navigieren. Klicken Sie dazu mit der rechten Maustaste auf die Virtuelle Maschine und dann auf „Ändern“.

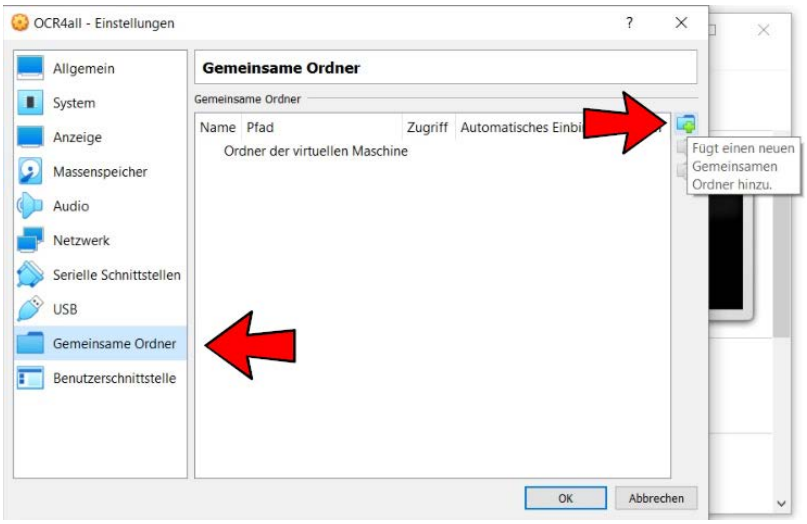


In der linken Seitenleiste der Einstellungen finden Sie den Bereich Netzwerk. Dort wählen Sie wie unten dargestellt noch den „Host-only Adapter“ aus.



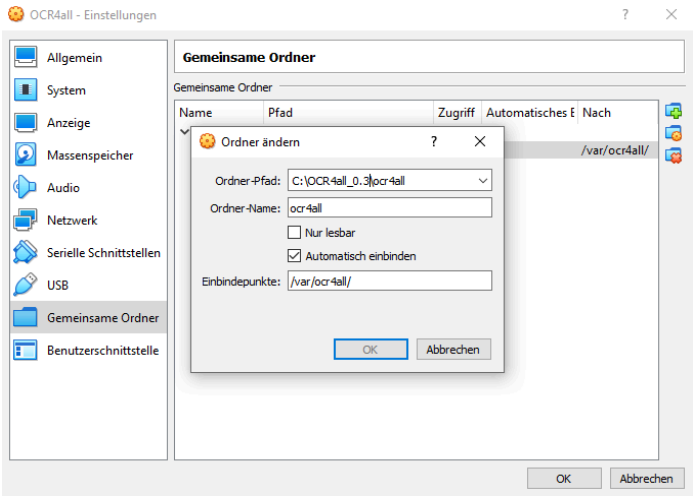
6. Einbindung des Gemeinsamen Ordners

Damit Daten zwischen OCR4all und Ihrem Gerät ausgetauscht werden können, muss nun noch ein neuer sog. „Gemeinsamer Ordner“ angelegt werden. Dies wird wieder über den Punkt **Einstellungen** im Unterpunkt **Gemeinsame Ordner** der OCR4all-Maschine erreicht. Über das kleine Ordner-Icon oben rechts wird ein neuer Ordner angelegt.



Daten-Ordner:

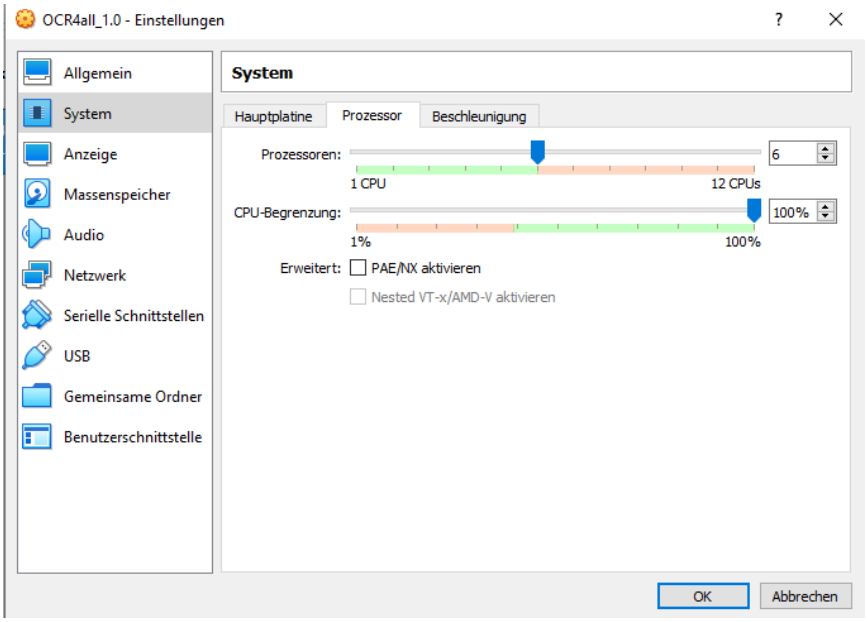
- Option „Automatisch einbinden“ auswählen
- „Einbindepunkte“ für Daten: /var/ocr4all/
- „Ordner-Pfad“ muss den Pfad enthalten, der innerhalb Ihres Betriebssystems zu dem von Ihnen weiter oben abgelegten Ordner „OCR4all_0.3/ocr4all“ führt.



Bestätigen Sie mit „OK“.

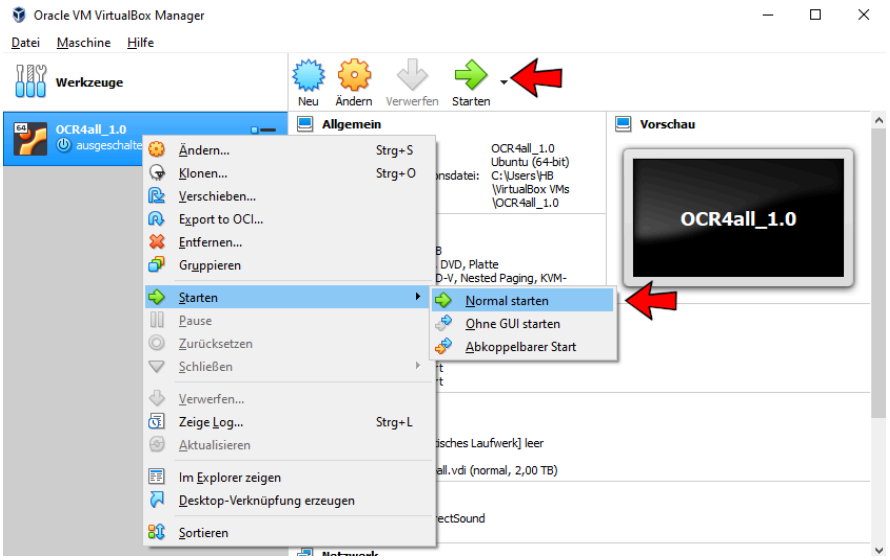
7. Zuweisung von CPU-Kernen

Um die Leistungsfähigkeit Ihrer Virtuellen Maschine und damit OCR4alls zu verbessern, sollten Sie abschließend noch weitere CPU-Kerne einbinden. Eine entsprechende Zuweisung weiterer Kerne nehmen Sie wie unten dargestellt über die Einstellungen der OCR4all-Maschine unter dem Reiter **System** und der Karteikarte **Prozessor** vor. Wir empfehlen, alle Ihnen zur Verfügung stehenden Kerne abzüglich eines Kerns zuzuweisen.



8. Starten der Virtuellen Maschine und Öffnen von OCR4all

Um die erzeugte und eingerichtete Virtuelle Maschine zu starten und mit der Arbeit in OCR4all zu beginnen, wählen Sie die Virtuelle Maschine einfach per Mausklick aus und klicken sie entsprechend der folgenden Darstellung auf den grünen Startpfeil.



Haben Sie den Start der Virtuellen Maschine initiiert, öffnet sich ein Fenster namens „OCR4all“. Warten Sie nun ab, bis die Maschine komplett hochgefahren ist. Dies ist der Fall, wenn im schwarzen Feld konstant die Angabe „ocr4all login ... Up ... seconds“ erscheint.

Um nun OCR4all selbst zu starten und mit der Arbeit an Ihren Projekten zu beginnen, öffnen Sie einen Browser (empfohlen wird Chrome) und geben folgende Adresse ein.

http://*IP-Adresse der Virtuellen Maschine*:8080

Die entsprechende IP-Adresse Ihrer Virtuellen Maschine haben Sie sich in Abschnitt 5 notiert. Für unser Beispiel bedeutete dies die Eingabe:

http://192.168.56.101:8080

Bestätigen Sie nach der Eingabe mit „Enter“. In Ihrem Browser sollte nun der sog. Project Overview, die Startseite OCR4alls zum Laden Ihrer Projekte, erscheinen.

Kontakt sowie weitere Anleitungen und Materialien zu OCR4all

Haben Sie Fragen zur Einrichtung oder Probleme bei der Installation der Software, Anmerkungen oder Anregungen, so zögern Sie bitte nicht, uns direkt und persönlich über ocr4all@uni-wuerzburg.de oder durch einen Beitrag auf GitHub zu kontaktieren.
Zusätzliche Anleitungen zur konkreten Anwendung der Software und anschauliche, detaillierte Beschreibungen zum Vorgehen bei einzelnen Arbeitsschritten, bspw. bei der Layoutsegmentierung oder dem Training werkspezifischer OCR-Modelle, finden Sie diese unter OCR4all/getting_started.