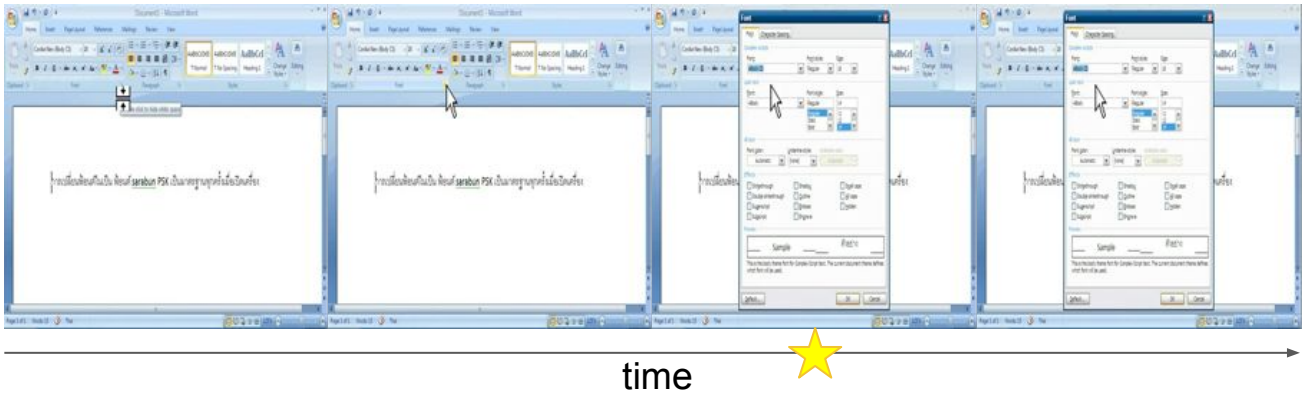
DeepTutorial

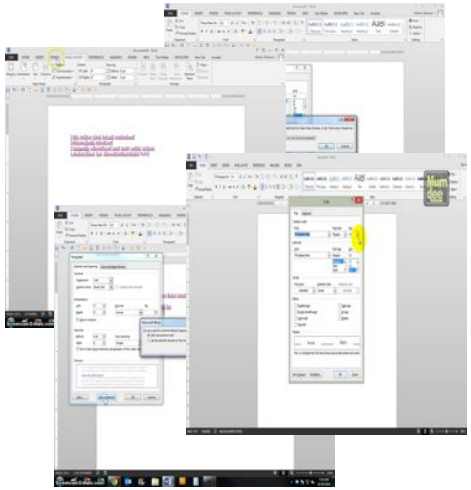
Detecting Key Instances in a Tutorial Video
and
Classifying Tutorial Type

Key Locations in Video

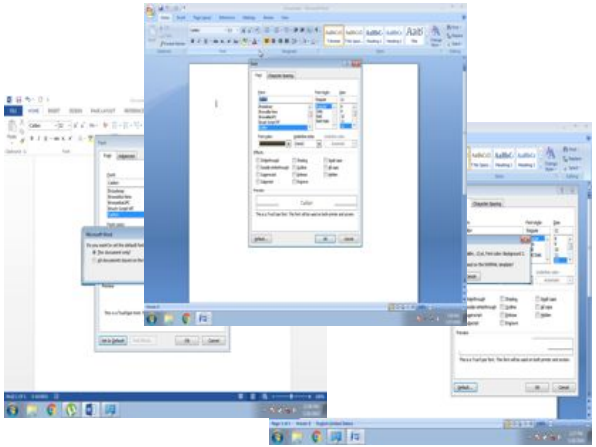
At which second did the author set the default font in Microsoft Word?



Default Font Dataset



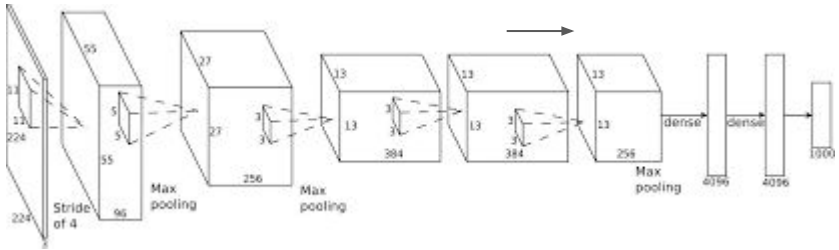
Youtube: 9 acceptable, 4 decent, 6 questionable,
2 unacceptable.



Domenic: 16 perfect (13 performing action, 3 not).

Alexnet

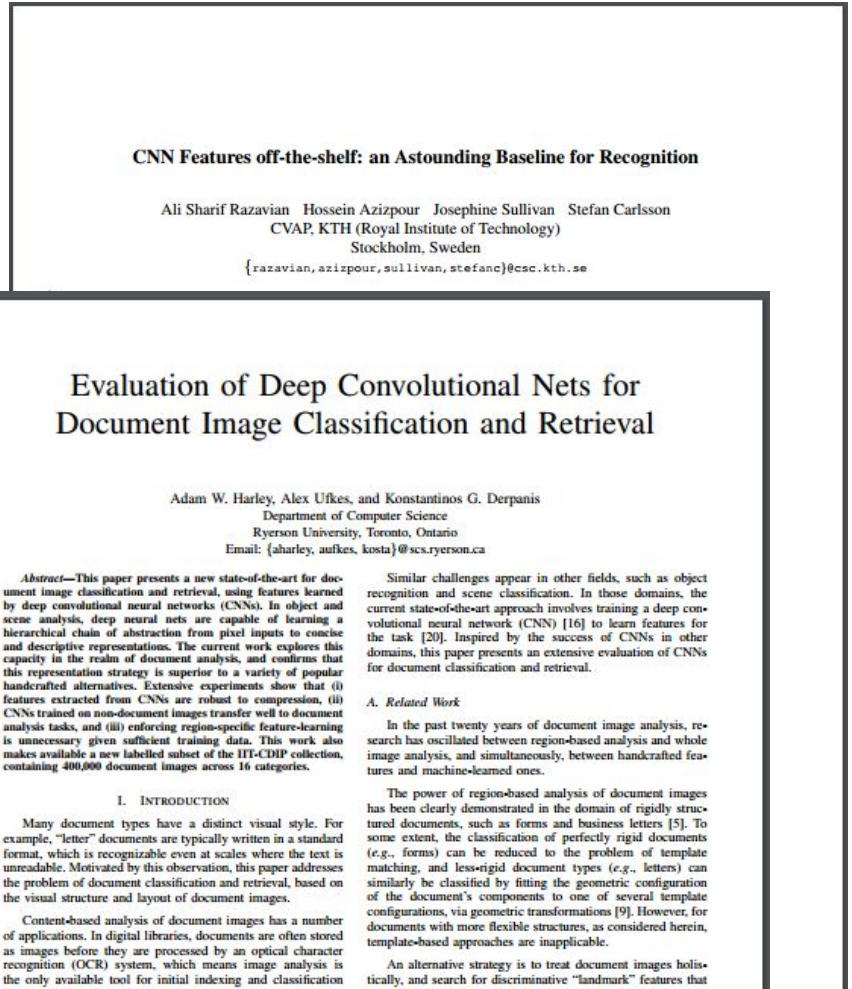
Alexnet: Deep Convolutional Network



But do we have enough data with enough variation?

Transfer Learning

27 Videoes is not enough Data,
so let's borrow!



CNN Features off-the-shelf: an Astounding Baseline for Recognition

Ali Sharif Razavian Hossein Azizpour Josephine Sullivan Stefan Carlsson
CVAP, KTH (Royal Institute of Technology)
Stockholm, Sweden
{razavian, azizpour, sullivan, stefanc}@csc.kth.se

Evaluation of Deep Convolutional Nets for Document Image Classification and Retrieval

Adam W. Harley, Alex Ufkes, and Konstantinos G. Derpanis
Department of Computer Science
Ryerson University, Toronto, Ontario
Email: {aharley, aufkes, kosta}@scs.ryerson.ca

***Abstract*—This paper presents a new state-of-the-art for document image classification and retrieval, using features learned by deep convolutional neural networks (CNNs). In object and scene analysis, deep neural nets are capable of learning a hierarchical chain of abstraction from pixel inputs to concise and descriptive representations. The current work explores this capacity in the realm of document analysis, and confirms that this representation strategy is superior to a variety of popular handcrafted alternatives. Extensive experiments show that (i) features extracted from CNNs are robust to compression, (ii) CNNs trained on non-document images transfer well to document analysis tasks, and (iii) enforcing region-specific feature-learning is unnecessary given sufficient training data. This work also makes available a new labelled subset of the IIT-CDIP collection, containing 400,000 document images across 16 categories.**

I. Introduction

Many document types have a distinct visual style. For example, "letter" documents are typically written in a standard format, which is recognizable even at scales where the text is unreadable. Motivated by this observation, this paper addresses the problem of document classification and retrieval, based on the visual structure and layout of document images.

Content-based analysis of document images has a number of applications. In digital libraries, documents are often stored as images before they are processed by an optical character recognition (OCR) system, which means image analysis is the only available tool for initial indexing and classification

Similar challenges appear in other fields, such as object recognition and scene classification. In those domains, the current state-of-the-art approach involves training a deep convolutional neural network (CNN) [16] to learn features for the task [20]. Inspired by the success of CNNs in other domains, this paper presents an extensive evaluation of CNNs for document classification and retrieval.

A. Related Work

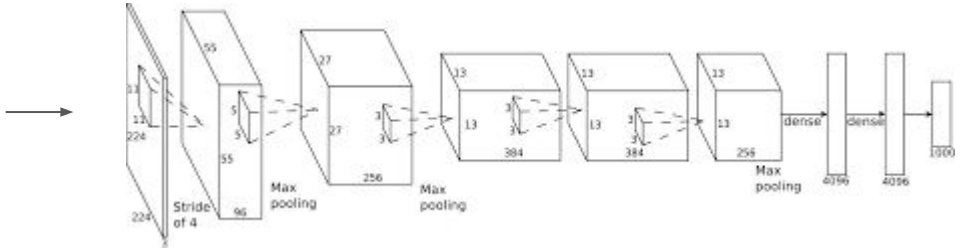
In the past twenty years of document image analysis, research has oscillated between region-based analysis and whole image analysis, and simultaneously, between handcrafted features and machine-learned ones.

The power of region-based analysis of document images has been clearly demonstrated in the domain of rigidly structured documents, such as forms and business letters [5]. To some extent, the classification of perfectly rigid documents (*e.g.*, forms) can be reduced to the problem of template matching, and less-rigid document types (*e.g.*, letters) can similarly be classified by fitting the geometric configuration of the document's components to one of several template configurations, via geometric transformations [9]. However, for documents with more flexible structures, as considered herein, template-based approaches are inapplicable.

An alternative strategy is to treat document images holistically, and search for discriminative "landmark" features that

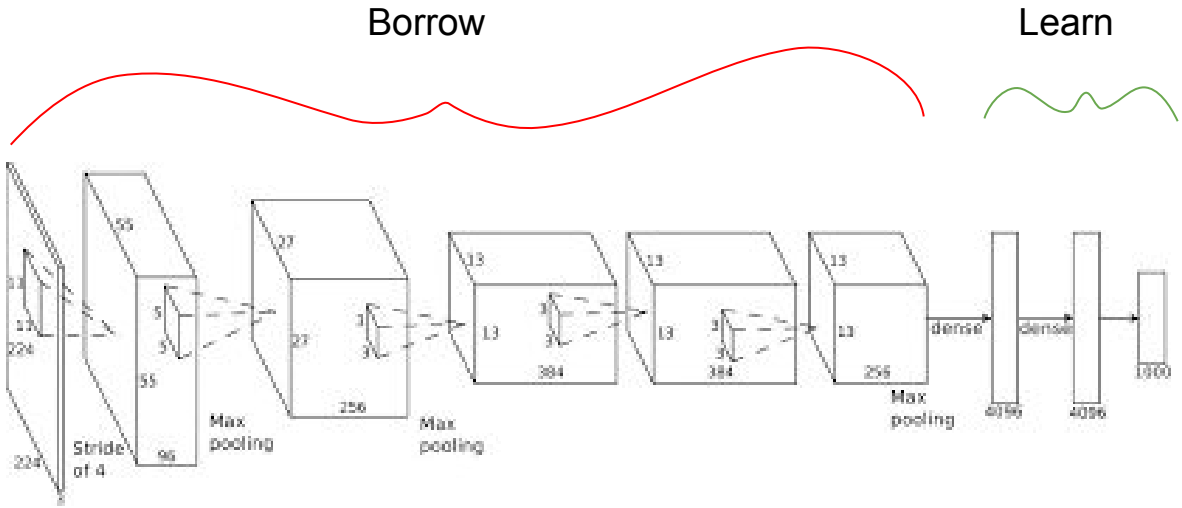
Alexnet Trained on Imagenet

Imagenet: 1 Million images with 1 Thousand class labels



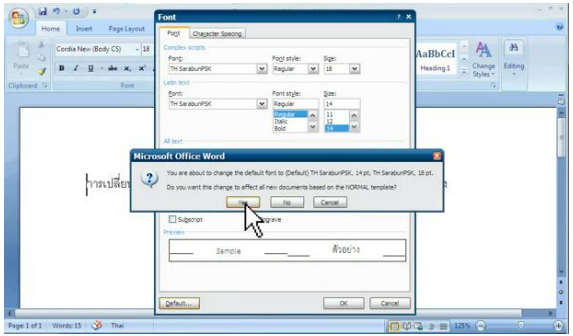
Why the Alexnet Architecture? Prior success, and convenience.

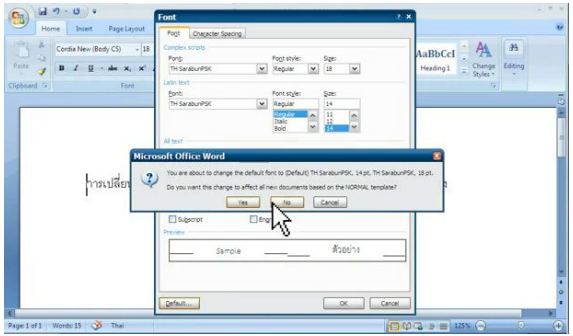
Keep the Features Learn Discrimination



Qualitative Results - Easy

Good results, but not perfect.

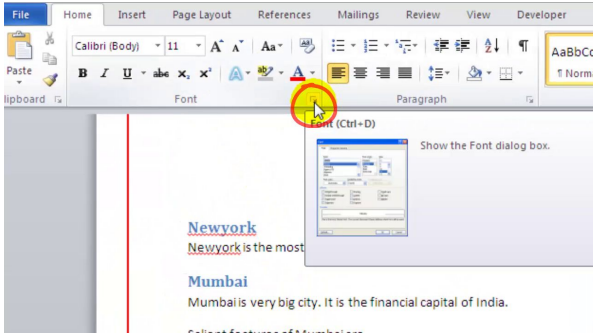




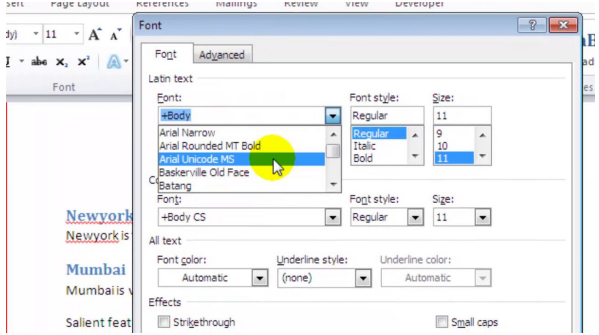
Qualitative Results - Hard

I had considered this video unusable due to zoom.

Incorrect



Correct



Data Split

Training: All of my video + 12 Youtube (9 acceptable, 3 decent).

Validation: 2 Youtube (1 acceptable, 1 questionable).

Test: 6 Youtube (questionable) - didn’t want to reduce the good sets further.

Precision, Accuracy, F1 Score (accuracy)

Iteration	Precision	Recall	F1 Score
60	84.5%	49.3%	**62.0%**
70	79.1%	32.0%	45.5%
80	88.1%	10.8%	19.2%
90	77.5%	28.6%	41.8%
100	89.8%	11.5%	20.4%
110	89.7%	17.8%	30.0%
140	93.0%	7.7%	14.3%
210	91.2%	9.1%	16.5%

Interpreting Results

When it says it’s a tutorial frame, it’s **84.5%** right about it.

It is able to recover **49.3%** of the tutorial frames.

Not enough invariance in dataset to find all of the frames?

Didn’t train on “questionable”-like set, but tested on “questionable” set.

DeepTutorial

Detecting Key Instances in a Tutorial Video
and
Classifying Tutorial Type

Continuing Forward...

Current proof of concept for binary (Default Font) classification.

Need better recall - with more data.

More kinds of Word Tutorials.