Assisting human experts in the interpretation of their visual process: A case study on assessing copper surface adhesive potency

Tristan Hascoet
Kobe University
tristan@people.kobe-u.ac.jp

Deng Xuejiao
Kobe University
dengxuejiao1005@yahoo.co.jp

Kiyoto Tai
MEC

Sachiko Nakamura
MEC

Tomoko Hayashi
MEC

Mari Sugiyama
MEC

Yasuo Ariki
Kobe University

Tetusya Takiguchi
Kobe University

Abstract

Deep Neural Networks are often though to lack interpretability due to the distributed nature of their internal representations. In contrast, humans can generally justify, in natural language, for their answer to a visual question with simple common sense reasoning. However, human introspection abilities have their own limits as one often struggles to justify for the recognition process behind our lowest level feature recognition ability: for instance, it is difficult to precisely explain why a given texture seems more characteristic of the surface of a finger nail rather than plastic bottle. In this paper, we showcase an application in which deep learning models can actually help human experts justify for their own low-level visual recognition process: We study the problem of assessing the adhesive potency of copper sheets from microscopic pictures of their surface . Although highly trained material experts are able to qualitatively assess the surface adhesive potency, they are often unable to precisely justify for their decision process. We present a model that, under careful design considerations, is able to provide visual clues for human experts to understand and justify for their own recognition process. Not only can our model assist human experts in their interpretation of the surface characteristics, we show how this model can be used to test different hypothesis of the copper surface response to different manufacturing processes.

1. Introduction

Humans are experts in communicating the reasoning process behind their answer to visual questions. For instance, on typical Visual Question Answering (VQA) samples, human annotators are often able to precisely justify in natural language the reason behind their answer to a certain visual question using simple common sense reasoning. In contrast, deep Learning models are often viewed as black box predictors lacking interpretability in the sense that existing tools often fail to explain the decision making process behind the models predictions. For instance, a deep learning model trained end-to-end on a VQA dataset may be able to provide the same answer as its human counterpart, but xxx.

While it is true that humans can justify for their answers on high level reasoning tasks, humans also often fail to explain the process behind their low-level feature recognition ability: for example, precisely defining the nature of a specific texture (what are the defining features of a plastic or wooden surface?) or specific low-level part attributes exhibiting large intra-class variations (what is the defining features of a ”leg” or a ”wing”?). Humans constantly perform such low-level visual recognition tasks while being unable to precisely justify for their own recognition process.

In this paper, we present one very practical instance of such situation in the micro-processor chip industry, in which expert material scientists are tasked with assessing the adhesive potency of copper sheets. We propose a model that, under careful design considerations, is able to provide visual clues for human experts to understand and justify for their decision process.

The proposed model is designed so that a subset of its internal representations carry semantically meaningful Information that can be visualized and easily interpreted by humans. Providing these visual clues, however, comes with the cost of imposing additional constraints on the architecture, which we found to degrade the model accuracy: Indeed, we found that networks with unrestricted architectures, (which do not provide interpretable features) perform better than network architectures restricted so as to provide semantically meaningful representations. This is because, as we restrict the architecture of the model, we formulate an assumption on the impact of the manufacturing process on the surface statistics which may not hold in reality. This result suggest an inherent trade-off between the expressivity of an architecture and the explainability of its process.

While the degradation of the model accuracy is problematic from a performance perspective, it offers an interesting opportunity from an explainability perspective: As the model accuracy degrades due to the inadequacy of the assumption implicitly formulated by the model architecture, we can use the model accuracy as a proxy metric for the adequacy of different assumptions. This allows us to quantitavely assess different assumptions regarding the impact of manufacturing processes on the copper surface. This may prove useful to quantify the impact of manufacturing process on cooper surface adhesive potency and eventually help optimize the manufacturing process.

In essence, the argument this paper is aiming for is as follows: although deep learning models lack the ”common sense reasoning abilities and the powerful formalism of natural language to communicate and justify for their decision making process, they can provide powerful to explain low-level recognition processes.

In practice, the contribution of this paper is as follows:

1. We formalize a segmentation procedure based on a probabilistic weak label segmentation framework.

2. We introduce a formalism to show how the model accuracy can be used as a proxy metric to quantify the validity of different assumptions on the dataset.

The remainder of this paper is organized as follows: In Section 2, we present some background information on the motivation for this project: We start by discussing the importance of copper surface adhesive potency, and detail the dataset used in our experiments. Section 3 details our contribution. Section 4 briefly relates our work to different research topics and Section 5 presents the results of our experiments. Finally, Section 6 further discusses the relevance of our results, insisting on the limitations of our assumptions to conclude this paper.

2. Background

Electric circuits are core to a wide variety of electronic devices including industrial and household appliances (e.g. fridges and microwaves), mobile communication devices and automobiles. To propagate electric signals between various components, electronic circuits rely on an electronic substrate made of copper wires through which electricity flows, isolated from each other by insulators (resin, resist, prepreg, etc.). Electronic substrates are organized in a multi-layered structure in which each layer contains different wire connections that need to be properly isolated from each other. Figure 1 illustrates the organization of such an electric circuit.

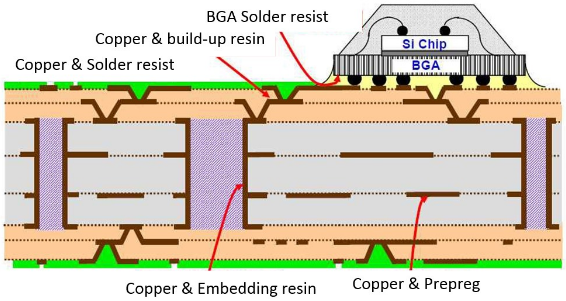

Figure 1. Illustration of a typical eletric circuit board.

The bulit-in copper microstructures can be easily peeled off from their isolating resin due to weak adhesion. For example, this can happen because of the impact of dropping a smartphone, or due to an excess of heat generated by running heavy computations. Such disadhesion of copper wires with their isolating resin may result in electrical short circuits in the electronic substrate and break the device. Hence a strong metal-to-resin bonding is necessary in order to enhance the reliability of electronic devices.

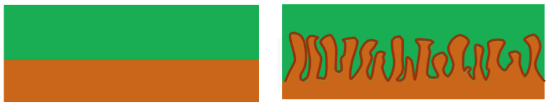

Figure 2. Illustration of xxx

Among other factors, the achievable strenght of the bond between the copper and the isolating resin is determined by the characteristics of the copper surface, as illustrated in Figure 2. In very broad terms, rough surfaces allow for stronger bonds as the asperities of the surface provide mechanical support against friction forces. In contrast, smooth surfaces provide little support against friction so that they have lower adhesive potential. In the remaining of this paper, we will refer to the potential bonding strength of a copper surface as its ”adhesive potency”. It is also important to note that the adhesive potency of a copper surface is related to its ”roughness”, which is observable at the microscopic scale.

Electronic substrate manufacturers have developped advanced manufacturing processes to shape the surface of copper sheets in order to increase their adhesive potency. This is typically achieved by applying a very small amount of a corrosive solution on the copper surface. Being able to accurately assess the adhesive potency of a copper surface would allow to further optimize manufacturing processes to increase the reliability of electronic devices. However, assessing the adhesive potency of a copper surface is a complex task, even for the most expert practinioners. Hence the motivations of this study is two fold: First we aim to automate the evaluation of a copper sheet adhesive potency from microscopic imaging of its surface. Second, we aim to understand the underlying principles by which the xxx Towards this goal, we built a dataset of microscopic images of copper surfaces, which we detail below.

We imaged copper surfaces using Scanning Electron Microscopy (SEM) at a resolution of xxx micro meters. To investigate the impact of different manufacturing processes on the copper surface, we applied 16 different corosive solutions, with decreasing corrosive power, to the copper surface. For each of these solutions, we captured 50 SEM images of $xxx \times xxx$ pixels so that the full dataset consists of 800 (16 $\times$ 50) images. Each image is annotated with a label $y_i \in Y$ corresponding to the corrosive solution used to shape the copper surface. Each solution y_i was obtained by submitting the original solution y_0 to an extreme stress test for a period of time $i \times T$. Hence, we know that for all i images of copper surfaces with label y_i show higher adhesive potency than the images labelled with y_{i+1}. However, we do not know the *exact* impact of the stress test on the surface adhesive potency.

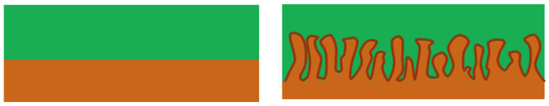

Figure 3. Illustration of xxx

3. Method

3.1. Dataset and Notations

In all our following experiments, we have split the full dataset $\mathcal{D}$ into a training, validation and test set $\mathcal{D} = \mathcal{D}_{tr} \cup \mathcal{D}_{val} \cup \mathcal{D}_{te}$ so that the number of images per label y_i in each set is 40, 5 and 5 respectively.

3.2. Baseline

We start by establishing a strong baseline for our study. The baseline architecture follows standard convolutional network designs for image classification. This architecture, illustrated in Figure 4, is made of several residual blocks sequentially interleaved with max pooling operations. Each residual block consists of n repetitions of a sequence of 3×3 Convolution, Batch Normalization and ReLU layers, followed by a residual skip connection. We set N residual blocks between every max pooling layer and we denote by d the number of pooling layers. Hence, the full depth D of the network (in number of convolution layers) is given by $D = d \times n \times N + 1$, where the term 1 corresponds to the initial 3×3 Convolution layer happening before the first pooling operation. Finally, the top layer of the network is made of a global average pooling layer followed by a linear softmax layer with output dimension 15 corresponding to our number of classes. For simplicity and contrary to standard practices, we keep the number of channels c constant in all layers of the network.

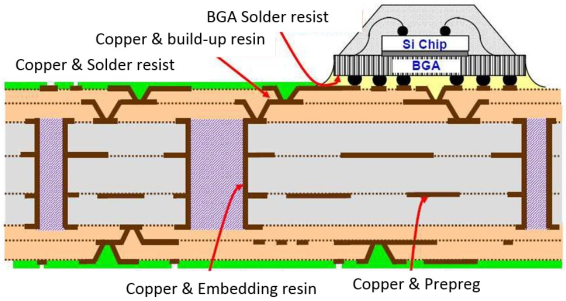

Figure 4. Illustration of our baseline architecture.

With this parameterization, our network is fully specified by the four hyper parameters c, d, n and N. We performed a grid search over these hyper parameters to select the best performing architecture. The details of this architecture search are given in the experiment section, and, as we shall see then, the best performing architecture performs significantly better than human experts. However, the decision process through which this model reaches such a high accuracy is entirely opaque as the distributed nature of the model's internal representations provides little interpretability. The remainder of this section details our attempt to design an architecture that can provide useful explanations of the process through which high accuracy recognition can be performed.

3.3. Assumptions

The basic idea behind our method is similar to the visual insights provided by the visualization method of [] and the attention mechanisms in visual models for image captioning and question answering tasks: We would like to evaluate the contribution of each input pixel to the final classification decision.

To do so, we modify our initial problem formulation into a segmentation task: Given an input $x \in \mathbb{R}^{h \times w}$, we want to design a model that outputs a segmentation mask $s \in \mathbb{R}^{h \times w}$ assessing the contribution of each inidivual pixel to the output adhesive potency score (i.e., the output class y). Training a typical segmentation model for this task would require ground truth segmentation masks s for each image x of the training set. However, manually annotating ground truth segmentation masks for this task is not feasible, as human experts are not able to provide such fine-grained annotations. Instead, we have to train the segmentation given a single groud truth label y_i per image.

In order to do so, we make the following assumption: **Assumption 1**: For a given image x with label y, the pixel-wise values of the true (unknown) segmentation mask s follow a spatially stationnary binomial distribution whose expected value, averaged over the spatial dimensions of x, is given by y following:

$$s_{hw}\tilde{a}\mathbb{B}(f(y)),$$

in which we introduced a target function f, which we will discuss below.

In other words, we suppose that there exists a true binary segmentation map s assigning to each individual pixel of x an adhesive potency score of 0 in regions of the surface providing low adhesive potency (i.e. smooth copper surface areas) and a value of 1 in regions of the surface providing high adhesive potency (i.e. rough copper surface areas). The adhesive potency score of an entire image is then given by the average of its pixel-wise score, and this average value is given by the image label y.

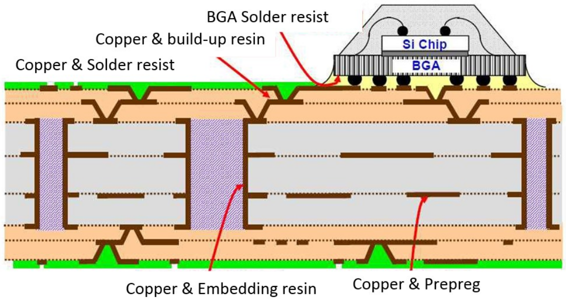

Figure 5. Illustration of the segmentation model architecture.

Figure xxx provides a visual illustration of this idea. For each image, we suppose the existence of such binary mask s quantifying the adhesive potency of local areas of the surface. The average value of the binary mask s amounts to the ratio of the copper surface covered in white (i.e. with value 1, corresponding to high adhesive potency). Our assumption means that this ratio stays constant for different images x sharing a similar label y, and takes a value given by the target function f.

3.4. Architecture

In this section, we present the architecture we use to compute binary segmentation masks from input images. Our architecture extends the baseline architecture presented in Section xxx, with an ascending path that progressively upsamples the output of the descending path, similar to the Unet architecture []. The expanding path is the exact symmetric of the descending path, with bilinear upsampling layers instead of max pooling. Different from the UNet architecture, and following previous works [], we merge the outputs of the descending path modules with the inputs of the ascending path by summation instead of concatenation and use valid convolutions to preserve the spatial resolution of the output. Finally, we add a sigmoid layer after the last convolutions to bound the output values between 0 and 1. Figure xxx illustrates this architecture.

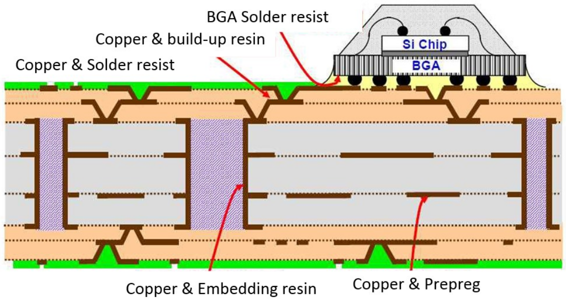

Figure 6. Illustration of the segmentation model architecture.

3.5. Loss function

Given a model M_θ, and a training dataset of labeled samples $\mathcal{D}_{\sqcup\nabla} = (x_i, y_i)$, learning is done by minimizing the following expectation over the model's parameters θ:

$$\theta^* = argmin_\theta \mathbb{E}_{x,y\in\mathcal{D}_{\sqcup\nabla}}||m_\theta(x) - f(y)||^2$$

AS we shall see in the next subsection, we use a hypothesis function g as an approximation of the true target function f, to which we do not have access, so that the actual training loss used to learn the model parameters θ is:

$$\theta^* = argmin_\theta \mathbb{E}_{x,y\in\mathcal{D}_{\sqcup\nabla}}||m_\theta(x) - g(y)||^2$$

3.6. Hypothesis function

In section xxx, we have made the assumption that labels y provide us with the expected value of the true binary segmentation masks through a target function f. In this section, we discuss the role of this target function.

The target function f defines the evolution of the copper surface adhesive potency with y as the surface ratio of high adhesive potency $\mathbb{E}_{h,w}s_{hw}$. f is an ideal function of which we have supposed the existence, but to which we can not have access, as we do not know the influence of the manufacturing process on the copper's surface.

We thus introduce a hypothesis function g approximating f, which we aim to estimate from the labeled data. Although we do not have access to f, we know several of its characteristics, which we can use to reduce the search space of hypothesis functions g.

In section xxx, we established that for all i, copper surfaces with label $i+1$ have lower adhesive $i-1$ have lower adhesive potency than copper sheets of label i. Hence, we know that f is a monotonically decreasing function. We can thus restrict our search of hypothesis function g to monotonically decreasing function. Second, as f describes the ideal, true evolution of the copper surface statistics with y, so that training an ideal model with the supervision signal provided by f should yield to zero test errors. Hence training a segmentation model with a hypothesis $g \approx f$ should lead to low test errors. This means that we can use the error of the model on the held out validation dataset as a proxy measure on the validity of the hypothesis function g.

In the experiment section, we evaluate the model errors with different hand-crafter hypothesis functions g. In future work, we aim to learn the hypothesis function g together with the model parameters.

4. Related Work

We identify three different lines of research that share similarities with our study: Explainabality of learned visual representations, weak supervision of segmentation models and machine learning applications for material science. Our can be seen as a weakly supervised approach to provide material experts with strong supports for interpretable decision, which falls at the intersection of these three research lines. We briefly present some of these works in the following subsections

4.1. Explaining CNN representations

Although deep neural networks have achieved a great success on a variety of challenging visualization tasks in recent years, our understanding of how these neural network models is far from enough to interpret. The pursuit of figuring out what is learned for each layer of models and how trained neural network models really "think" never stops.

Jason Yosinski et al. provide two useful tools for visualizing and interpreting neural nets. One is to visualize the activations produced on each layer of a trained convnet as it processes an image or video, and the other enables visualizing features at each layer of a DNN via regularized optimization in image space. Shan Carter et al. create an explorable activation atlas of features the network has learned, by using feature inversion to visualize millions of activations from an image classification network, which can reveal how the network typically represents some concepts. There is a new dataset called Visual Question Answering(VQA) containing open-ended questions about images.These questions require an understanding of vision, language and commonsense knowledge to answer. Trevor Darrel et al. build models (such as Pointing and Justification-based explanation model) to explain their decisions, generating convincing explanations. TensorFlow provides an attention-based model, which enables us to see what parts of the image the model focuses on as it generates a caption.

4.2. Weakly Supervised Segmentation

An other point we should pay attention to is different labelling strategies in conventional supervised learning. Typically speaking, it's often assumed that each instance is associated with one single label. However, there should usually be more than one labels for one instance in real-world tasks, if it is multi-label learning.

Eugene et al. propose a semi-supervised framework that employs image-to-image translation between weak labels. Yuxing Tang et al. build a similarity-based knowledge transfer mode trying to investigate whether knowledge about visual and semantic similarities of object categories can help improve the peformance of detectors trained in a weakly supervised setting.

The most related work with ours is Rihuan Ke's[*] in which they present a semi-supervised learning strategy for segmentation with lazy labels and develop a multi-task learning framework to integrate the instance detection, separation and segmentation within a deep neural network.

4.3. Machine Learning for Material Science

Deep learning has showed remarkable success on many important learning problems in chemistry, drug discovery, biology and materials science.

In the field of materials science, deep neural networks have also been receiving inreasing attention and have achieved great improvements, for example, in material property prediction and new materials discovery for batteries.

5. Experiments

5.1. Recognition results

We start by evaluating the baseline classification model described in Section xxx and illustrated in Figure xxx. The hyperparameters n, N, d, and c of the classifier architecture were obtained by a grid search within the limits of a 12GB Nvidia GPU memory size . The model was trained on the training dataset for 200 epochs using the Adam optimizer with default parameterization.

We compare the classification outputs of the baseline model to those of an expert material scientist on a blind test. The human expert was given the sample images of the training set to practice, and we evaluated the expert's answers on the samples of the test set.

Figure xxx shows the results achieved by the model and Figure xxx shows the results achieved by our baseine model. As can be seen in these figures, the model tends predict the manufacturing process more accurately than the human evaluator. However, these results should be taken with a grain of caution as the human expert was given little to practice on this specific dataset, while the model was selected as the best performing baseline from an extensive

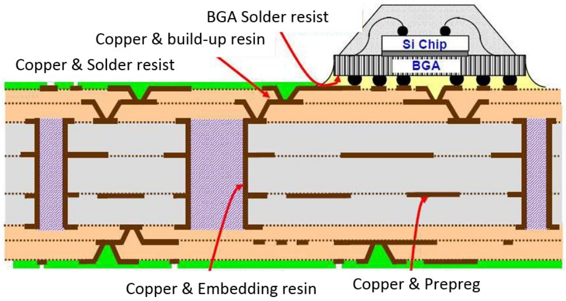

Figure 7. Results of the human evaluation.

parameter search. We plan on reconducting the human evaluation in the updated version of this paper.

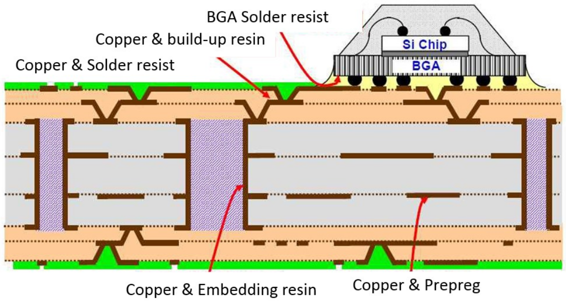

Figure 8. Results of the model evaluation.

Interestingly, however, the human expert seems to accurately recognize the surfaces of highest adhesive potency as seen in the upward trends of his results for low y values. For example, he easily identified the surface for $y = 0$. On the other hand, his guesses for low adhesive potency are much more random.

This is in stark contrast with the classifier accuracy, which perfectly identifies the low adhesive potency surfaces (high y values), but consistently misclassifies the images of $y = 1$ and $y = 2$.

5.2. Hypothesis testing

In this section, we

6. Discussion & Conclusion

xxx

References