Interpretability at a Glance

C Henrik R Åslund
c.aslund19@imperial.ac.uk

Harleen Hanspal
h.hanspal21@imperial.ac.uk

Avinash Kori
a.kori21@imperial.ac.uk

Introduction

- Interpretability has been suggested as a method to make AI safe
    - Especially in circumstances where it is hard to write down an entire specification and rely on verification or optimisation
- Interpretable begs the question: Interpretable to whom? We need to consider assumptions on the cognitive limitations of the human meant to do the interpretation.
- Contribution:
    - Define a metric for when a layer in a neural network is interpretable. We also propose a metric for an entire network.
    - Evaluate using 3 networks
        - Visual inspection
        - Another common analysis, the fourier transform
- Organised as follows

Related Work

- PCA done to visualise weights before[1]

Defining Metric: ‘Normalised Eigen-Area’

- Desiderata
    - We should be able to talk apply it to layers
    - A metric that tells you whether something is interpretable is a high bar. At least, it should tell you whether a layer is uninterpretable
    - Should relate to information content
    - Should relate to limits to human cognitive capabilities
    - We want to be able to compare different networks as well

[1] https://distill.pub/2020/circuits/equivariance/

Formal Definition

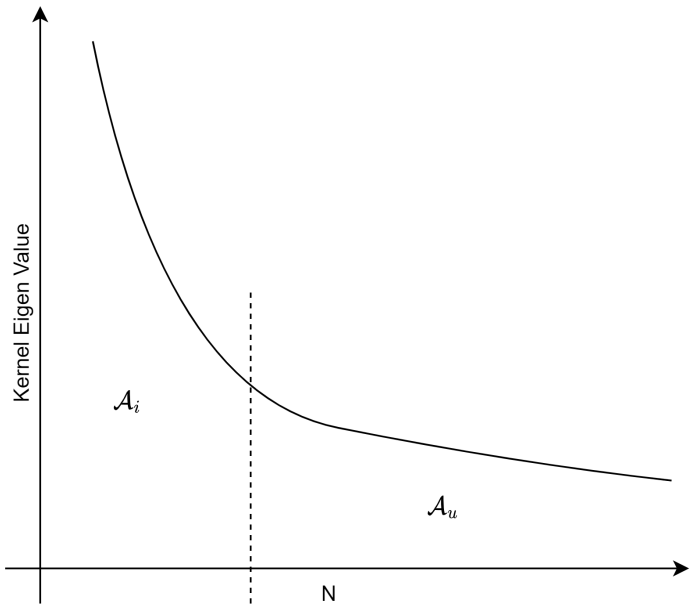

Figure ?: Spread of eigen values for convolutional weight kernels, according to our formulation A_i and A_u defines interpretable and uninterpretable eigen areas respectively.

- Furthermore consider a threshold k, which is an estimate of a human's cognitive limitations. More specifically, it is the number of principal components they can evaluate
- Represent the eigenvalues as the index ordered by size (with largest first) along the x-axis and the eigenvalues on the y-axis.
- Score is in [0,1]

Layer-wise scores:

- Consider the eigenvalue decomposition over the filters in one layer
- If score is close to 0 the layer is uninterpretable.
    - There could be safety-relevant things hidden away in a large blob of values which is not inspectable by any human.
- If score is close to 1, it is interpretable.
    - It is a necessary condition for safety, but not sufficient
- If there is one layer that is uninterpreatable in the network, then the entire network is uninterpretable

$$score^{\Phi}_{l} := \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{k} \lambda_i + \sum_{i=k+1}^{n} \lambda_i}$$

Network-wise scores:

- Let us consider two different networks Φ_1 and Φ_2 with similar performance, we claim the network Φ_1 is more interpretable than Φ_2 if:
    - All the layers in Φ_1 generate more eigen-area score w.r.t all the layers in network Φ_2
    - If a particular layer is network is uninterpretable we consider entire network to be uninterpretable
- Based on above two properties we formally define our network-wise scores as *Geometric mean* of all the layer-wise scores in a particular network

$$score^{\Phi} = (\Pi^{|\Phi|}_{l} score^{\Phi}_{l})^{1/|\Phi|}$$

Adding Laplacian Filters

- After network weights but before eigenvalue decomposition
- Example: increase the contrast for edges

Evaluation: Fourier Transform and Visual Inspection[2]

- 2 pretrained networks, trained on IMAGENET dataset
    - AlexNet
    - VGG
- We use two methods to evaluate
    - Visual inspection:
        - We consider simplicity of feature activations generated
    - Fourier transform:
        - We consider number of frequency components present in the feature activation map generated
        - Less number of frequency components implies more interpretable layer

[2] The repository for the experiments is avalailable at https://github.com/hh10/InterpretabilityHackathon2022.

Quantitative and qualitative comparisions between networks

Table 1: Our layer-wise score compared against 4 different networks. Layer *l* refers to the last layer in this case. The networks with randomised weights are used for sanity check.

Network	$score_l$ when *k*=5
AlexNet (pretrained) Conv2D_7	0.32
AlexNet (random) Conv2D_7	0.05
Vgg19Net (pretrained) Conv5_4	0.22
Vgg19Net (random) Conv5_4	0.02

For qualitative evaluation we manually find a filter activating for snakes and snails and use the feature activation maps generated to maximise that particular filter to perform subjective evaluation.

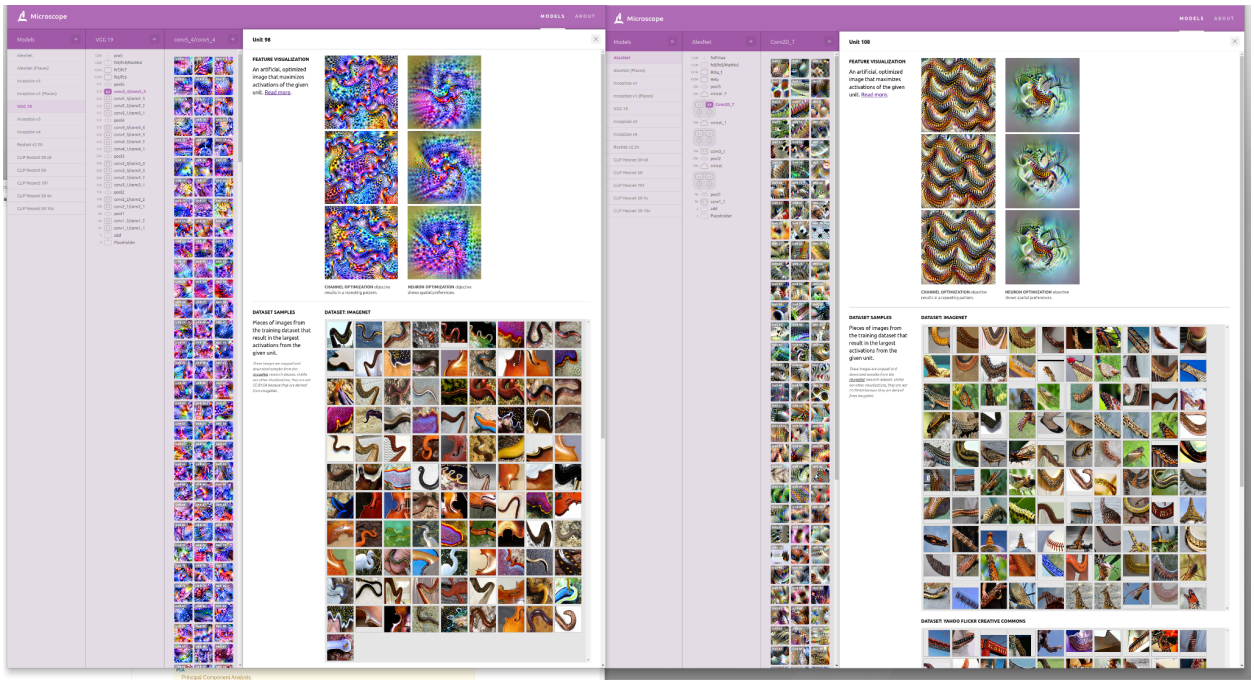

Figure 1: Visual inspection of intepretability for the networks from Table 1.

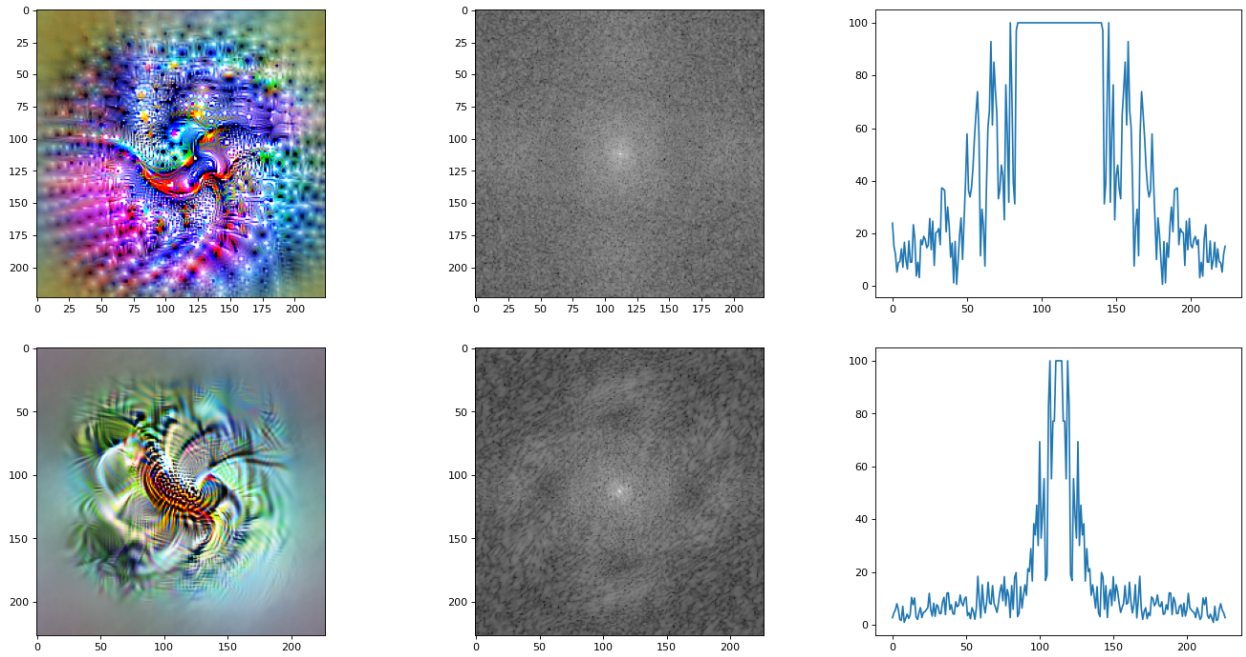

Figure 2: To validate the visual inspection, we compute fourier transform of the activation maps of the two models: The frequency analysis of the activations also indicate the same with higher frequencies in the vgg_net activations and lower frequences in the alexnet acts.

Discussion:

- Network-wise scores for AlexNet and VGG19 are **0.467** and 0.401 respectively, indicating alexnet to be more interpretable given both the networks have same performance
- We observe that the explanations generated using alexnet have fewer frequency components and are slightly visually simpler

AlexNet analysis

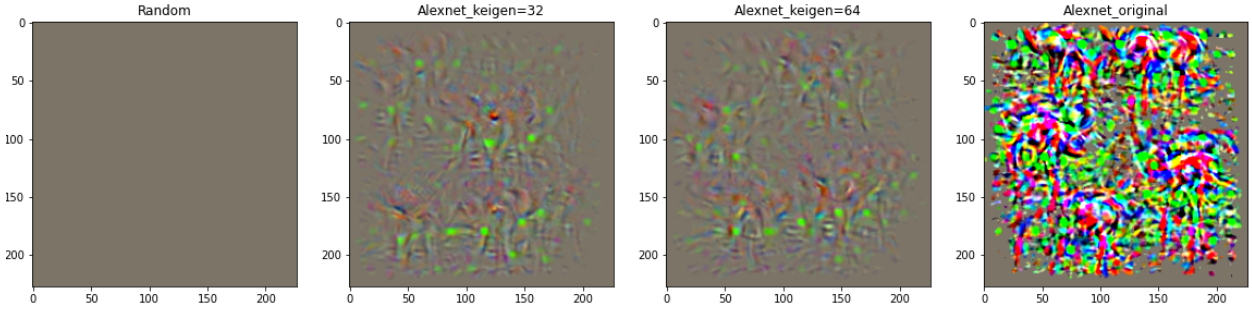

Figure 3: Inputs to networks that maximise the activation in the final layer. In the two middle panels, the weight vectors with the *k* largest eigenvalues have been kept and the remaining weight vectors have been replaced with copies of those. From left to right, the networks are (1) is randomly initialised AlexNet, (2) AlexNet where *k*=32, (2) AlexNet with *k*=64, (4) network with trained AlexNet weights.

Table 2: Our layer-wise scores for different layers of alexnet along with scores for the networks with randomised weights are used for sanity check.

Layer	Score_l (pretrained)	Score_l (random)
0: Conv2d	0.77	0.38
3: Conv2d	0.56	0.13
6: Conv2d	0.45	0.089
8: Conv2d	0.48	0.088
10: Conv2d	0.46	0.1
1: Linear	0.25	0.01
4: Linear	0.35	0.02
6: Linear	0.62	0.038

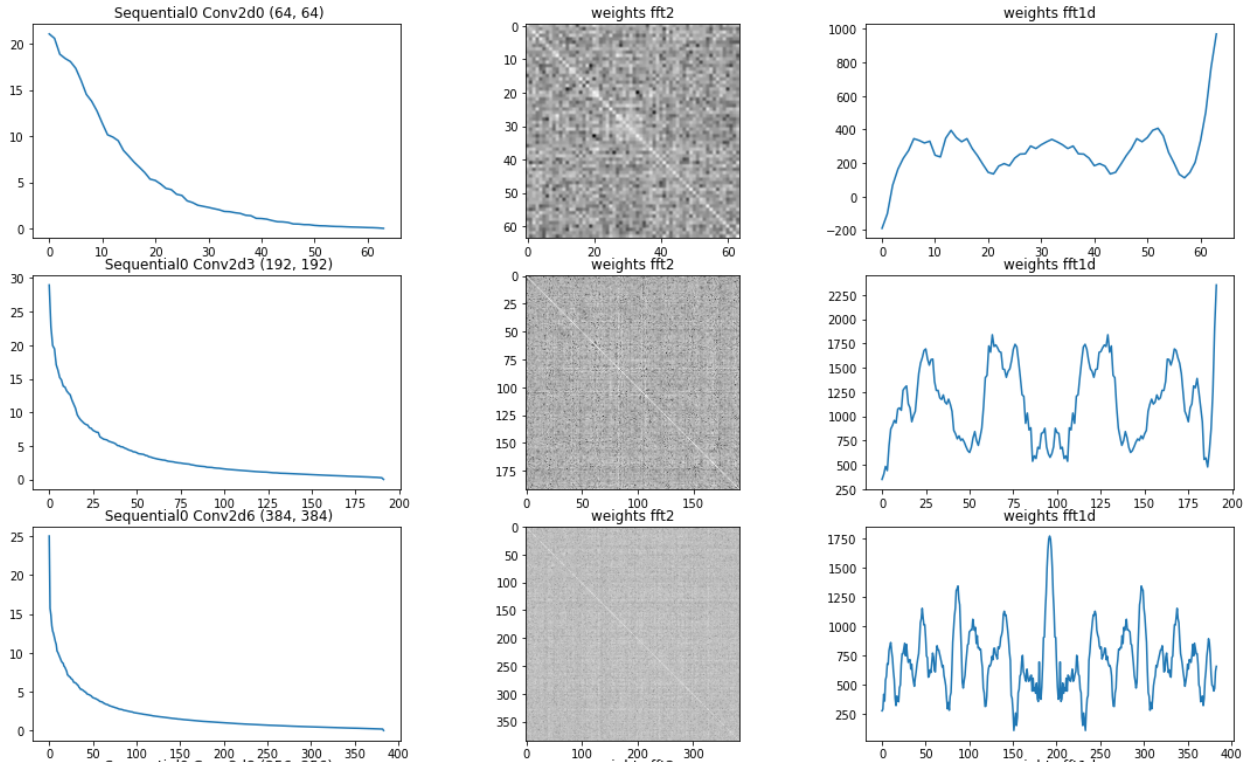

Figure 4: On similar lines, we look at the ffts of the individual layer features to find correlation between 'the variance explained by lesser number of eigenvectors' and 'lesser (high) frequencies in its spectrum'. Notice conv filters of 3 Sequential layers in AlexNet have 64, 128, and 363 eigenvalues respectively. Notice from the FFTs of the covariance matrix of the weights (3rd column) that there is only 1 major frequency in the the firstrow and at least 3 major frequencies in the second and multiple frequency peaks in third row. Indicating high explainability scores from row 1 to row 3 as expressed by our defined metric in Table 2.

Discussion:

- According to the layerwise score, early and late layers are more interpretable than middle layers. It would be interesting to see if domain experts would agree.
- Unintuitively, the first layer was also interpretable to some degree for the randomised network

VGG19 Analysis

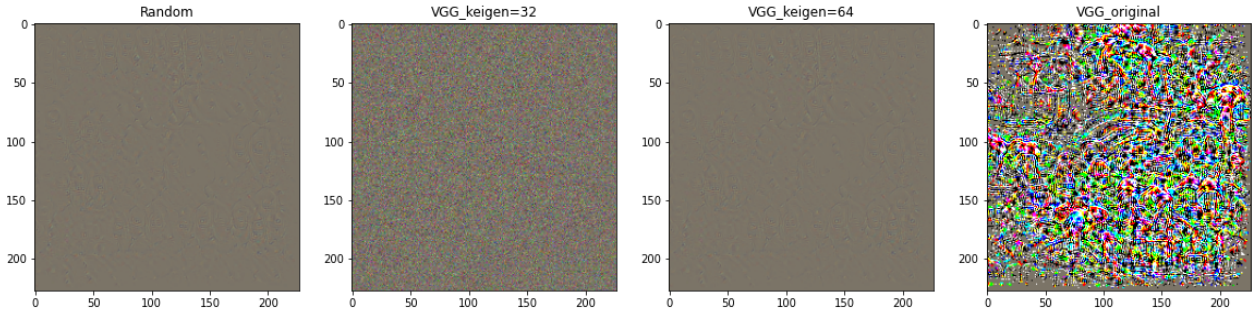

Figure 5: Inputs to networks that maximise the activation in the final layer. In the two middle panels, the weight vectors with the *k* largest eigenvalues have been kept and the remaining weight vectors have been replaced with copies of those. From left to right, the networks are (1) is randomly initialised VGG19, (2) VGG19 where *k*=32, (2) VGG19 with *k*=64, (4) network with trained VGG19 weights.

Table 3: The same as Table 2 except for VGG instead of AlexNet.

Layer	Score_l (pretrained)	Score_l (random)
0: Conv2d	0.99	0.8
2: Conv2d	0.87	0.34
5: Conv2d	0.67	0.23
7: Conv2d	0.51	0.18
10: Conv2d	0.44	0.13
12: Conv2d	0.40	0.09
14: Conv2d	0.43	0.01
16: Conv2d	0.36	0.1
19: Conv2d	0.29	0.07
21: Conv2d	0.29	0.05

23: Conv2d	**0.33**	**0.05**
25: Conv2d	**0.29**	**0.05**
28: Conv2d	**0.24**	**0.05**
30: Conv2d	**0.27**	**0.05**
32: Conv2d	**0.34**	**0.05**
34: Conv2d	**0.29**	**0.05**
0: Linear	**0.20**	**0.007**
3: Linear	**0.24**	**0.02**
6: Linear	**0.65**	**0.04**

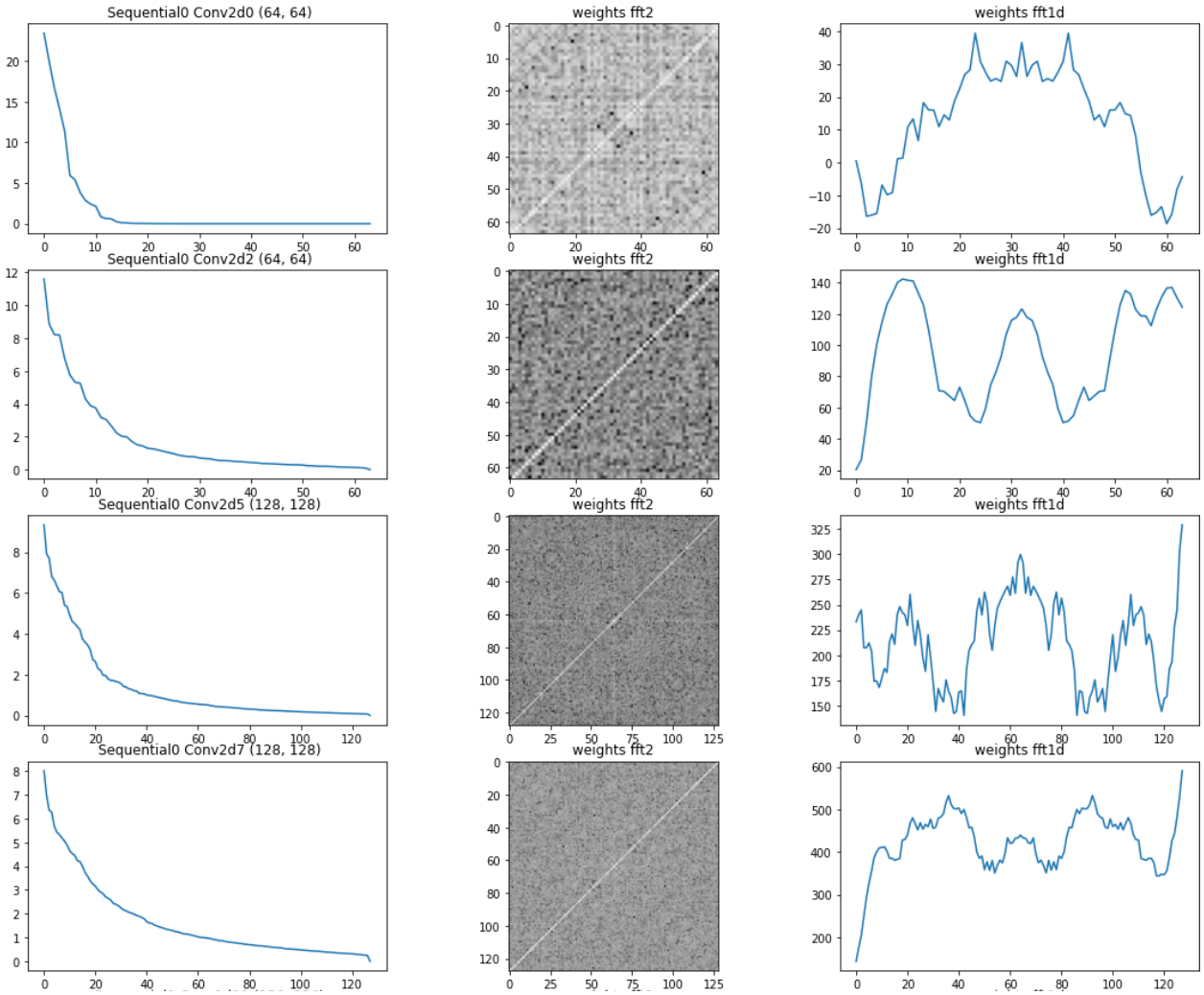

Figure 6: On similar lines, we look at the ffts of the individual layer features to find correlation between 'the variance explained by lesser number of eigenvectors' and 'lesser (high) frequencies in its spectrum'. Notice 2 conv filters of 2 Sequential layers in VggNet below: (a) the first 2 rows have 78 eigenvalues each, with the top 5 eigenvectors accounting for 94% and 66% of variance. Notice from the FFTs of the covariance matrix of the weights (3rd column) that there is only 1 major frequency in the first row and at least 2 major frequencies in the second. (b) for the third, fourth rows, the top 5 eigenvectors (out of 127) account for 43% and 32% of total variance in the layer filters. Notice from FFTs, how third row filter has 2 major frequencies while 4th being more flat, has more frequency components.

Conclusion

- Alexnet is more interpretable than VGG19
- Initial layers and final layers are more interpretable as compared to intermediate layers

Limitations

- Need to test score's behaviour across multiple networks
- Subjective evaluation can be made as a survey to capture qualitative consistency