Deep Learning: Interpretability?

QUANT 2019, Venice, Feb 22, 2019

Damiano Brigo
Dept. of Mathematics, Imperial College, London

Agenda I

1. How do we define interpretability
   - BLACK BOX problems with deep neural networks
   - Accuracy vs interpretability
2. Feature attribution methods
   - Local surrogates
   - Shapley Values and Cooperative game theory
3. Shapley values in machine learning
4. Example
   - Advantages
   - Disadvantages
5. Conclusions

Disclaimer and sources

I am not an expert on machine learning.

Here I present the perspective of an academic with experience in financial modeling and risk management approaching deep learning

Most of the material is taken from:

Ideas on machine learning interpretability, Navdeep Gill, H2O.ai
`https://www.slideshare.net/0xdata/ideas-on-machine-learning-interpretability`

But what *is* a neural network? 3Blue1Brown
`https://www.youtube.com/watch?v=aircAruvnKk`
in turn based on the book by Michael Nielsen,

`http://neuralnetworksanddeeplearning.com/`

Ideas on interpreting machine learning, Hall, Phan and Ambati
`https://www.oreilly.com/ideas/ideas-on-interpreting-machine-learning`

Disclaimer and sources

Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. Christoph Molnar. 2018/12/22 [Book, just came out]
https://christophm.github.io/interpretable-ml-book/

Interpretability of deep learning models, Eduardo Perez Denadai
https://towardsdatascience.com/interpretability-of-deep-learning-models-9f52e54d72ab

Interpretability of deep learning models: A survey of results, Chakraborty et al.
https://ieeexplore.ieee.org/document/8397411

One Feature Attribution Method to (Supposedly) Rule Them All: Shapley Values, Cody Marie Wild
https://towardsdatascience.com/one-feature-attribution-method-to-supposedly-rule-them-all-shapley-values-f3e04534983d

Interpretability?

What is interpetability? From Molnar:

'There is no mathematical definition of interpretability. A (non-mathematical) definition I like is Miller (2017).[1] "Interpretability is the degree to which a human can understand the cause of a decision." '

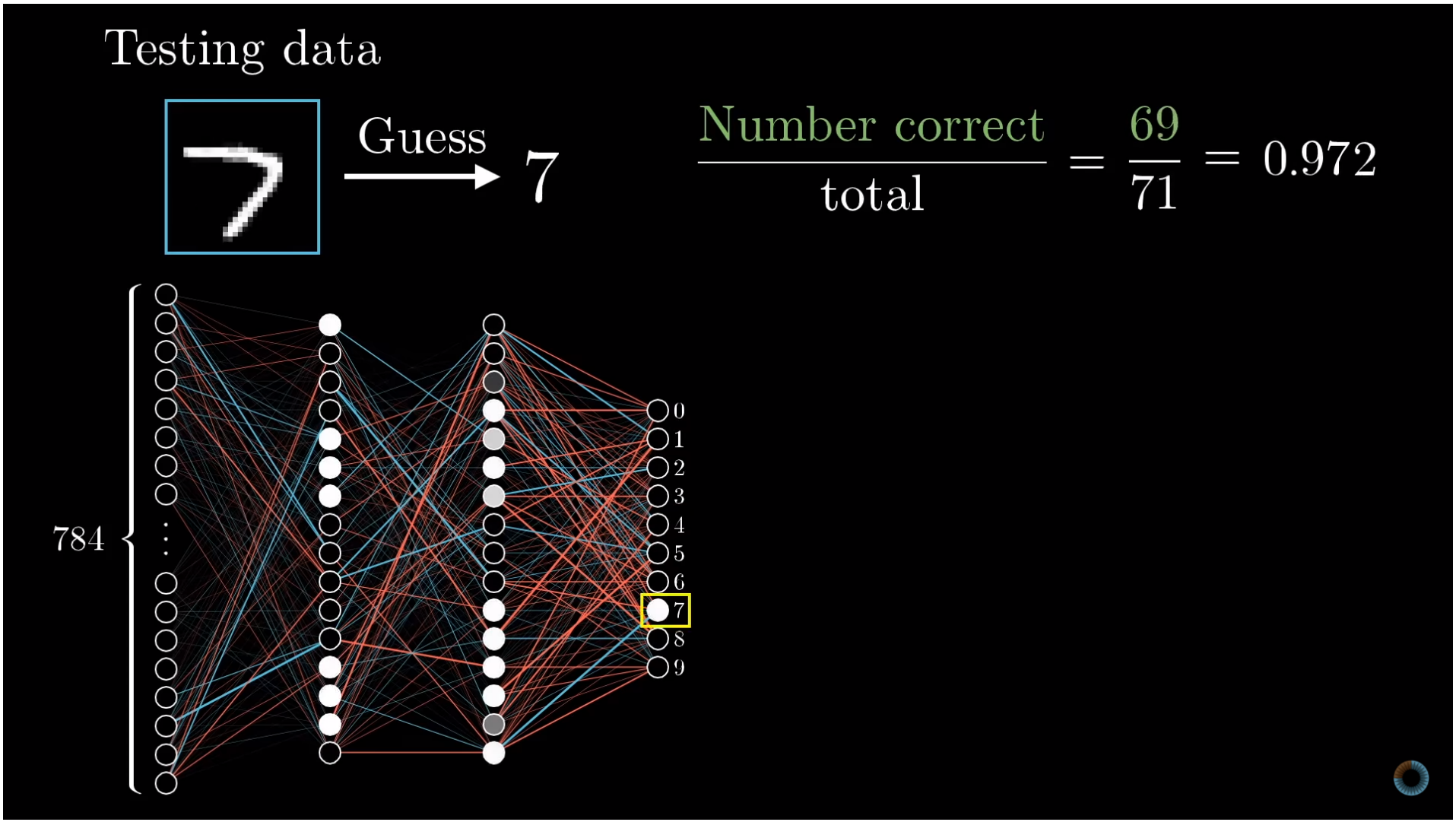
'Or Been et al (2017)[2] "Interpretability is the degree to which a human can consistently predict the models result" '

E.g: classic DNN recognizing scanned digit (28x28=784 pixels).

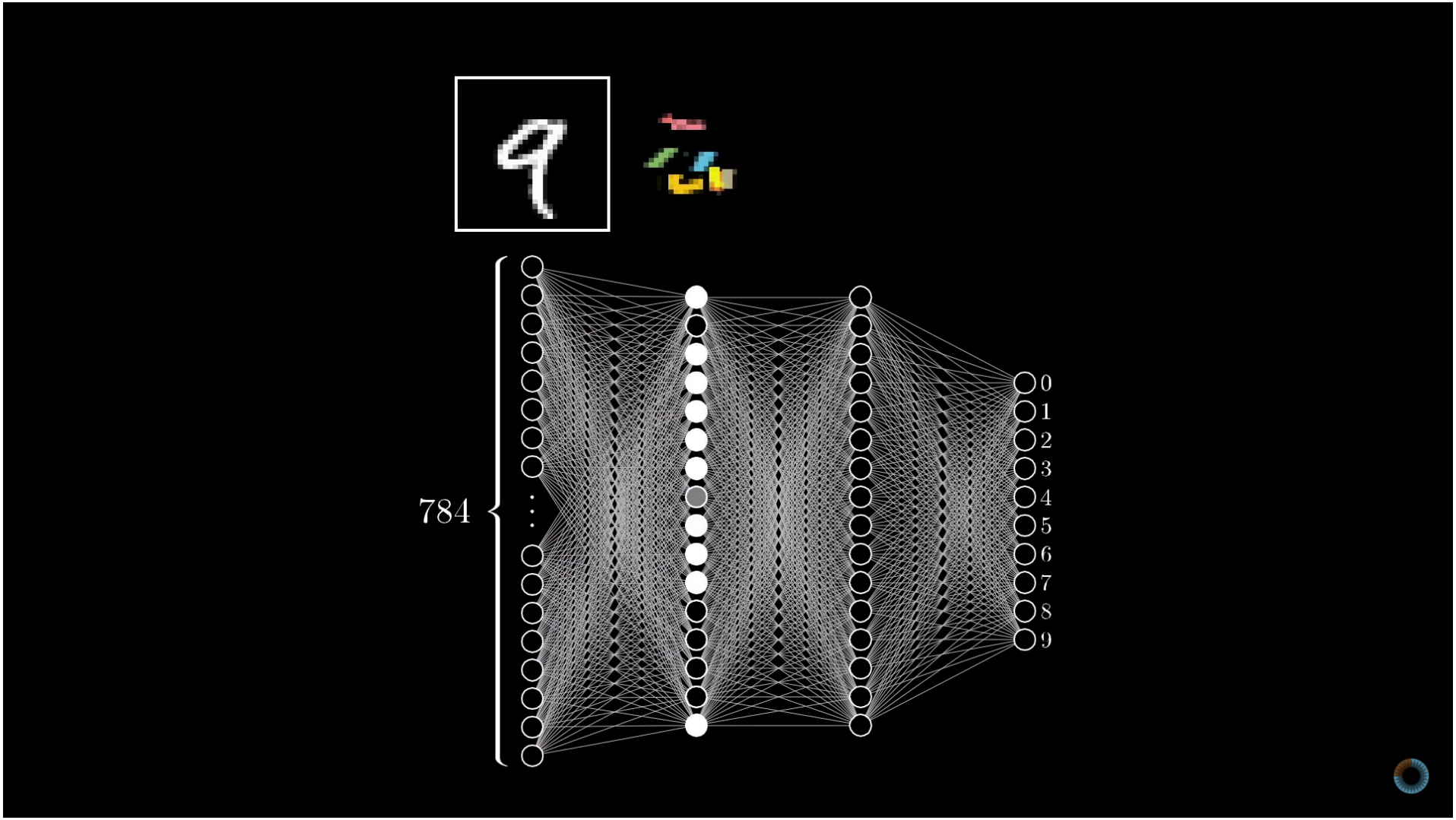
[1] Miller, Tim. 2017. "Explanation in Artificial Intelligence: Insights from the Social Sciences." arXiv Preprint arXiv:1706.07269

[2] Kim, Been, Rajiv Khanna, and Oluwasanmi O. Koyejo. "Examples are not enough, learn to criticize! criticism for interpretability." Advances in Neural Information Processing Systems. 2016

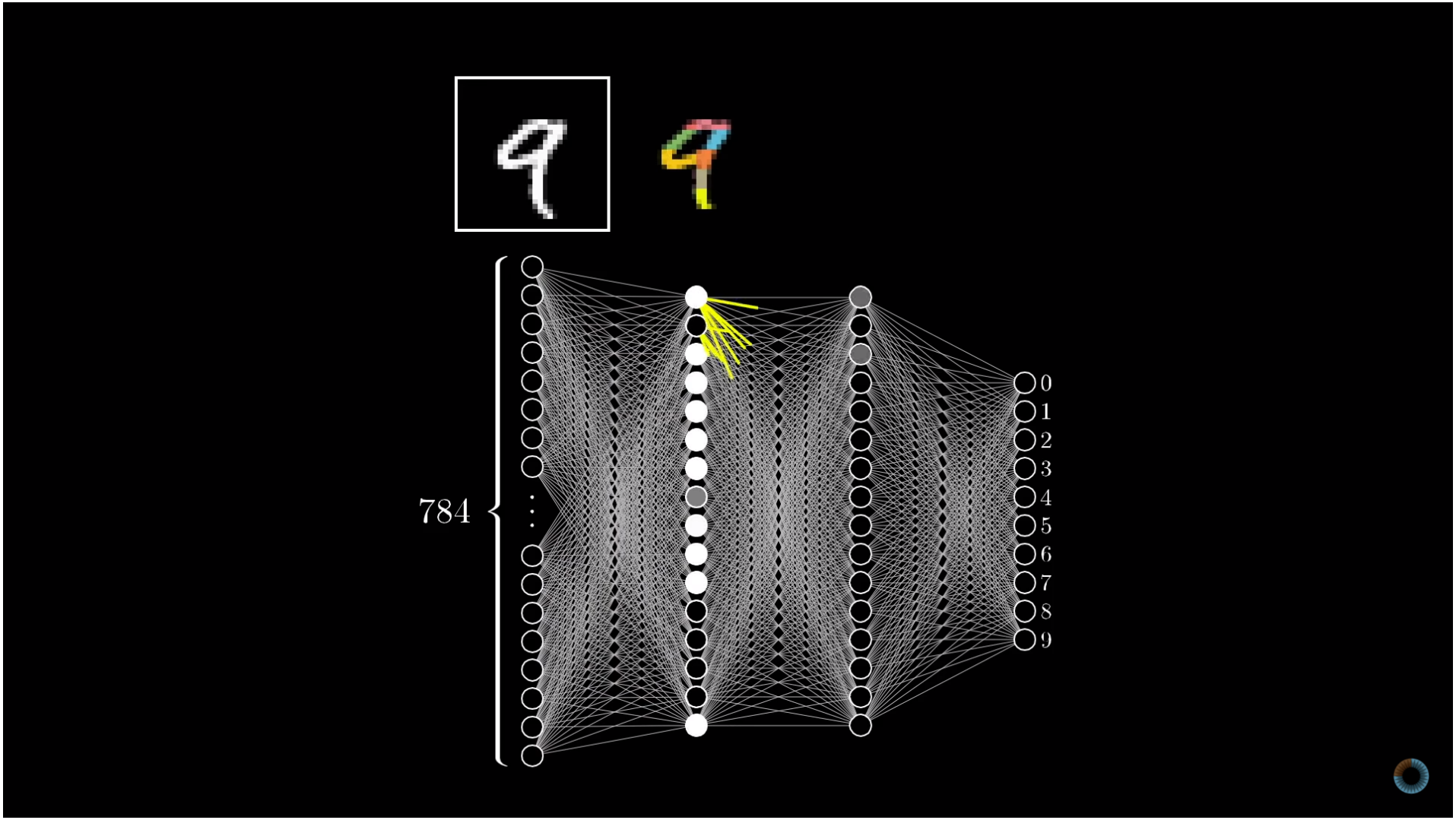
Example: deep learning example from 3Blue1Brown.



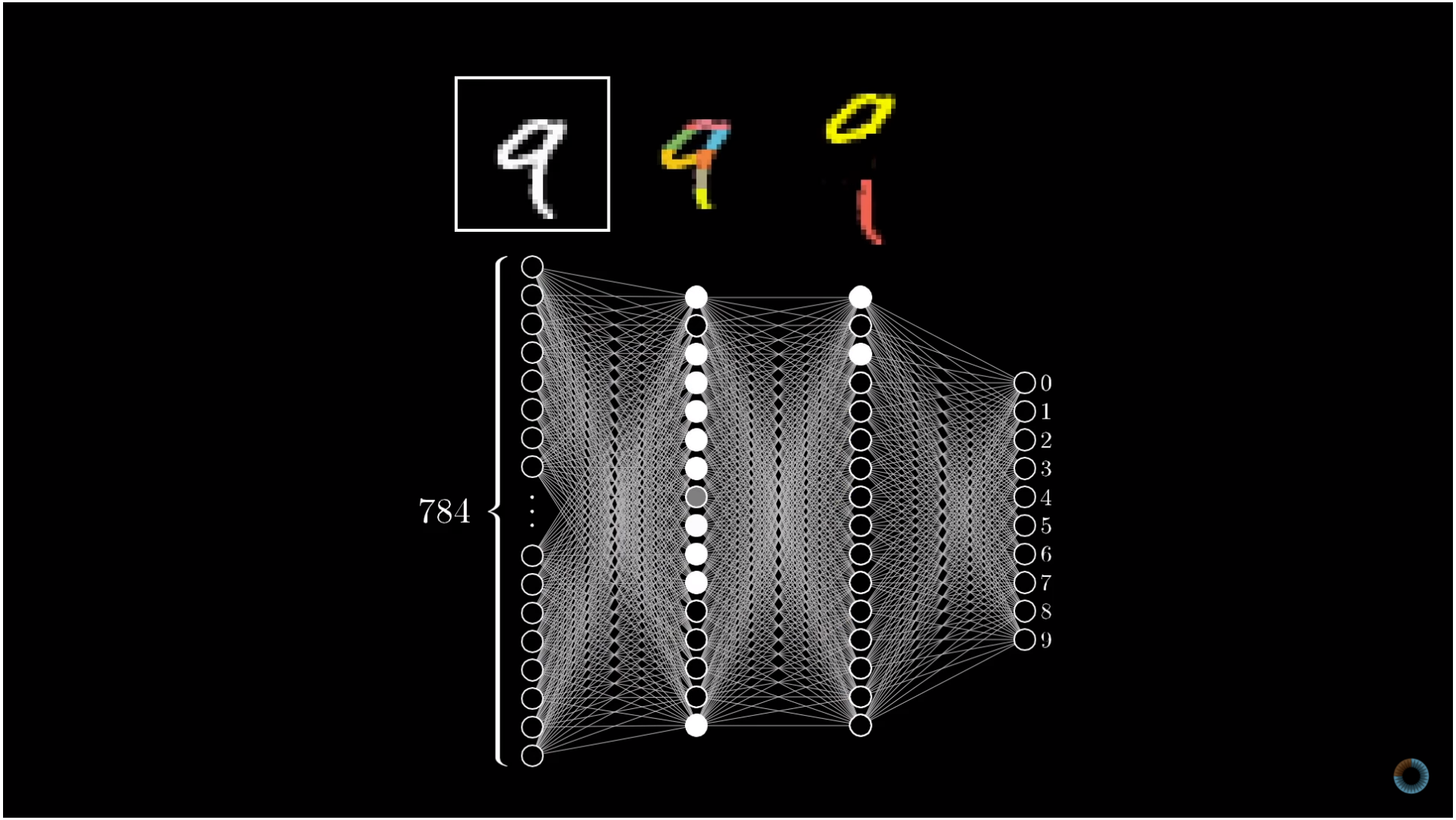
Example: deep learning example from 3Blue1Brown.



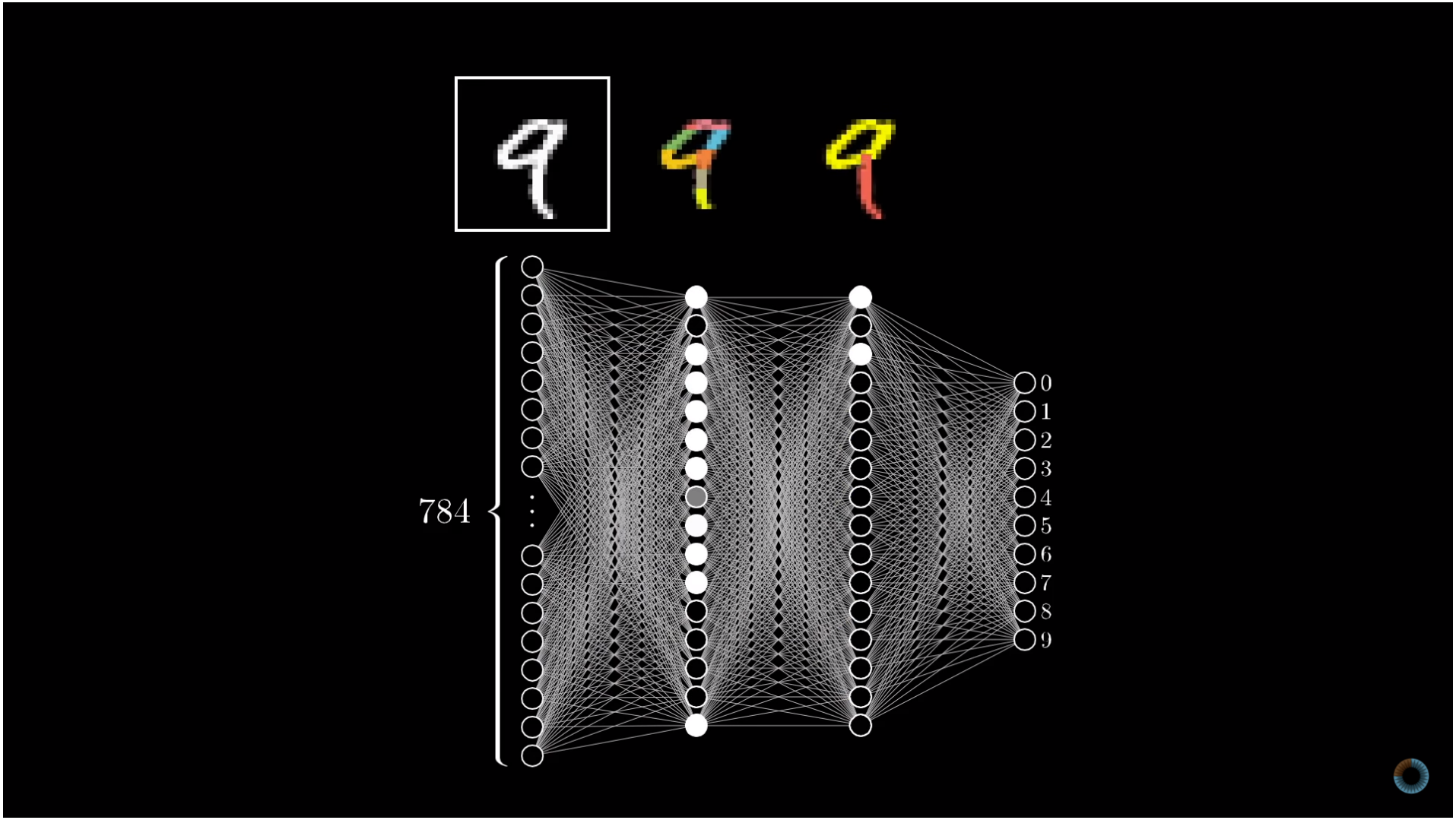
Example: deep learning example from 3Blue1Brown.



Example: deep learning example from 3Blue1Brown.



Example: deep learning example from 3Blue1Brown.



Example: deep learning example from 3Blue1Brown.

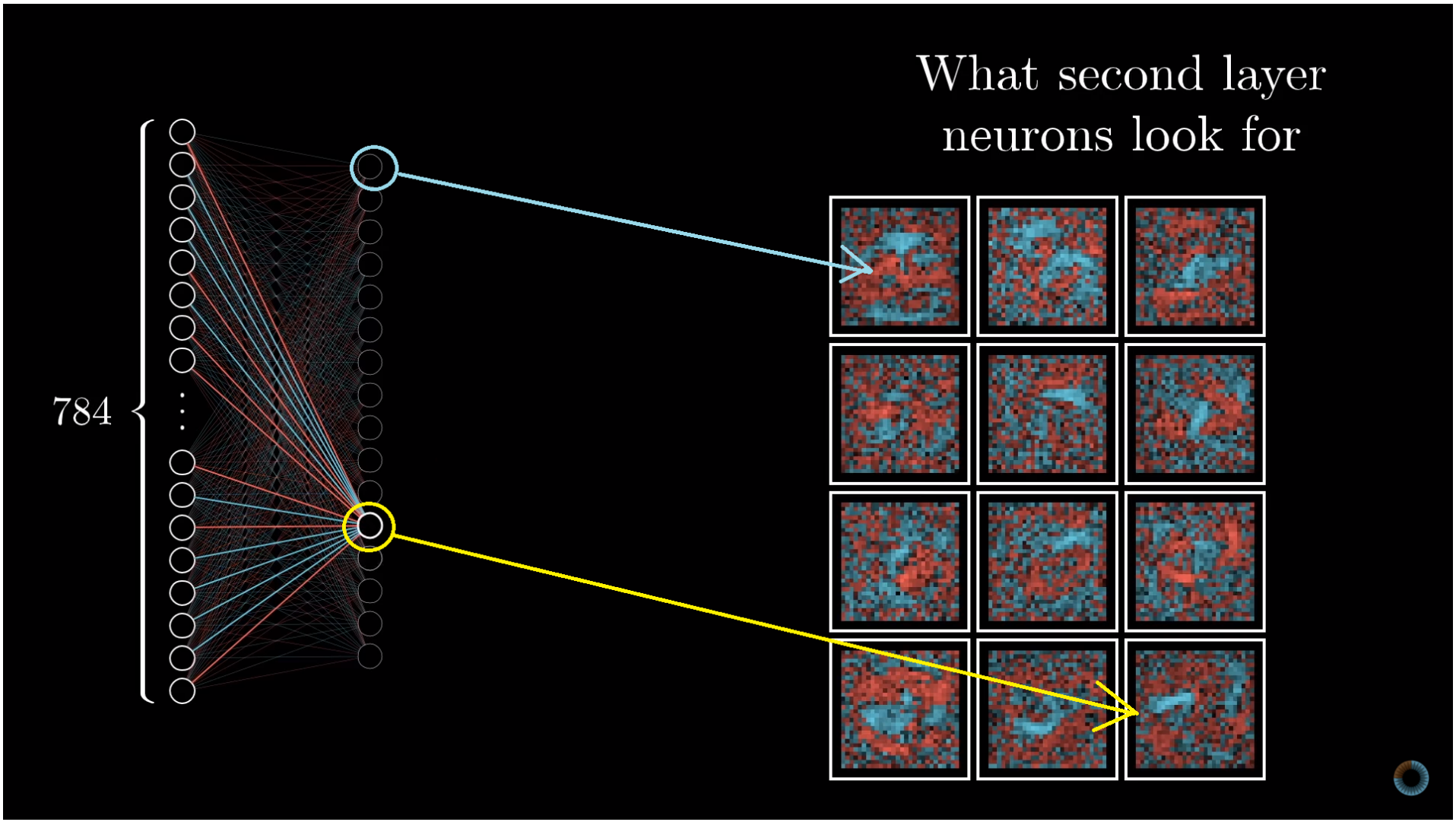
This would be very nice for interpretability and levels of abstractions.

Is this what happens?

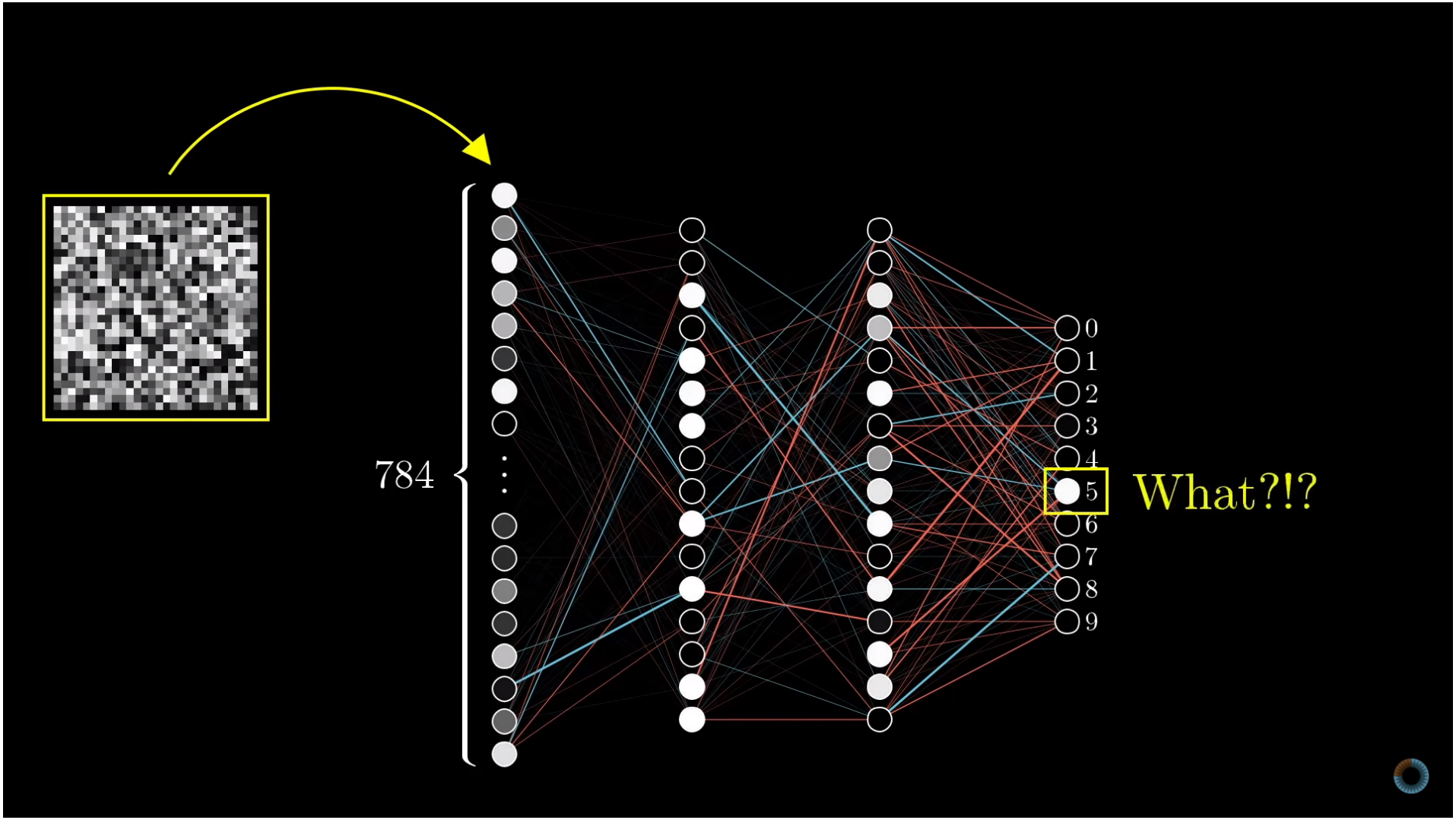
Unfortunately, **not at all**

We look at a grid plot of the weights from the first layer, contributing to each neuron in the second layer. Colors say whether weights are positive or negative, with brightest colors meaning higher quantities.

Example: deep learning example from 3Blue1Brown.



Example: deep learning example from 3Blue1Brown.



Example: deep learning example from 3Blue1Brown.

This was very old technology, from the 80's-90's, but it makes the "black box" problem quite clear.
(Data and code for the example are in the book by Michael Nielsen, `http://neuralnetworksanddeeplearning.com/`)

This was a feed-forward deep neural network. With structures like recurrent neural networks and LSTM networks, for example, things get more complicated, but a plain feed-forward network is enough to make the point.

Accuracy vs interpretability

- Deep neural network can be trained through cost function minimization on the last layer and then backpropagation;
- It can attain very accurate fit to input data for example in supervised learning;
- However understanding what's going on can be very challenging, even in the simple example above;
- One might try to interpret the weights in the first layer of the inputs as a one to many relationship;
- However, this is lost after layer one and one gets a messy many to many relationship;
- Non-local effects take over via backpropagation. It is hard to understand what's going on in connecting inputs and outputs.

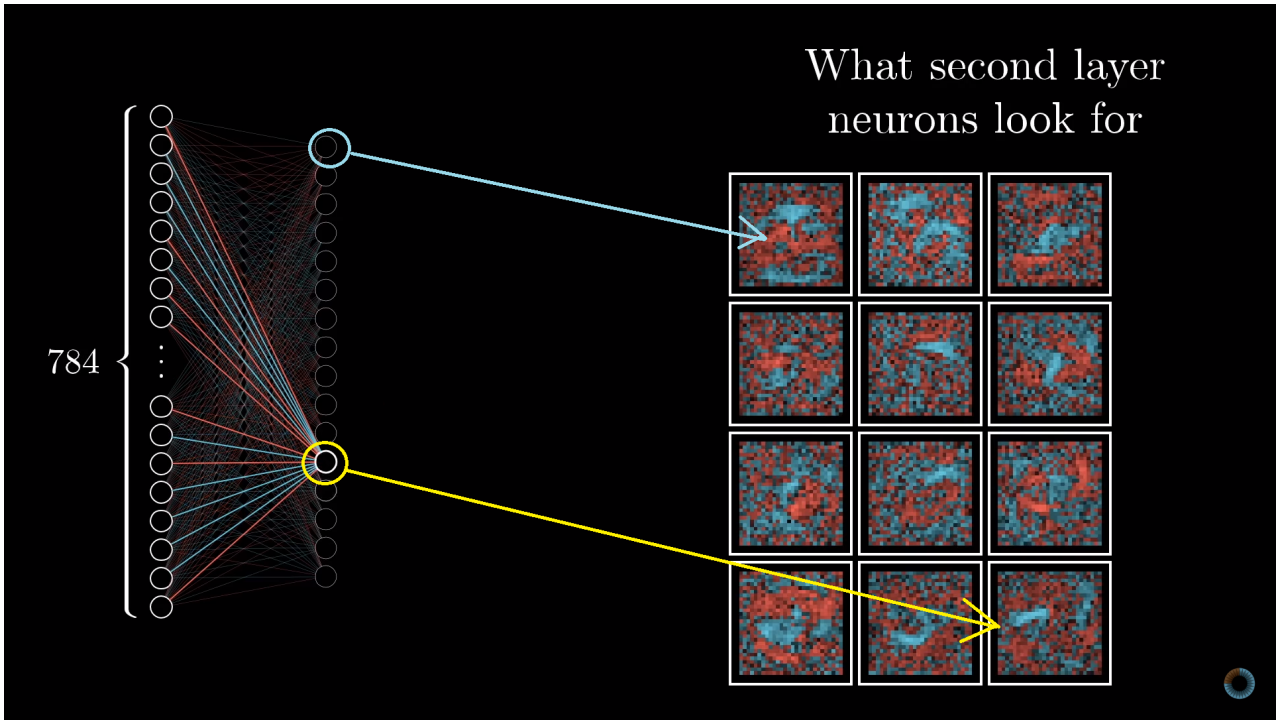
Accuracy vs interpretability

In the literature above I read, I came across two frameworks for interpreting the deep neural network.

- Saliency methods.
- Feature attribution

Interpretability: Saliency methods

- Good for visualizing what happens in the network.
- "Which weights are activated given some inputs?" Toy example:



- More realistic:"What region of an image is detected by a convolutional layer?" However:
- Does not say which feature "best" describes our final predictions

Interpretability: Feature Attribution

Feature attribution

- Find out explanatory power each variable has on output variable.
- Local surrogate models like
  - Layer-wise Relevance Propagator (LRP),
  - Local Interpretable Model-agnostic Explanations (LIME), and
  - Deep Learning Important Features (DeepLift)

  are part of a framework of "additive feature attribution" (AFA)(*). They try to fit an interpretable (surrogate) model LOCALLY to inputs and outputs of the deep neural network.
- Is this really an "explanation" or an "interpretation" of what the DNN does? More in the conclusions.
- Alternatively, use **Shapley values** from cooperative game theory to estimate feature contribution. SVs are model agnostic.

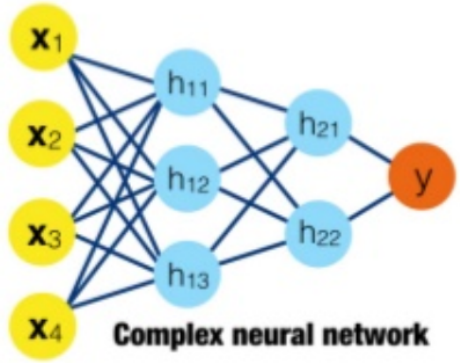
(*) S. Lundberg, S. Lee, "A Unified Approach to Interpreting Model Predictions", Neural Information Processing Systems 2017.

Local surrogates

Surrogate models

BAD	CUSTOMER_DTI	LOAN_PURPOSE	CHANNEL
0	0.18	MORT	7
1	0.42	HELOC	10
0	0.11	MORT	10
0	0.21	MORT	1

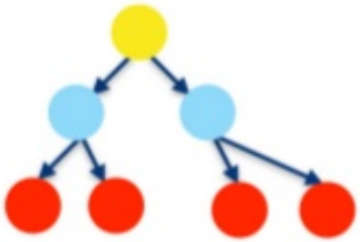
1. Train a complex machine learning model



Complex neural network

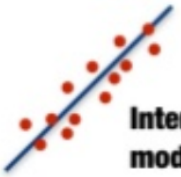
BAD	PREDICTED_BAD	CUSTOMER_DTI	LOAN_PURPOSE	CHANNEL
0	0.47	0.18	MORT	7
1	0.82	0.42	HELOC	10
0	0.18	0.11	MORT	10
0	0.12	0.21	MORT	1

2. Train an interpretable model on the original inputs and the predicted target values of the complex model



Interpretable decision tree

Or



Interpretable linear model

Home Equity Line of Credit and Mortgages. Source: `https://www.oreilly.com/ideas/ideas-on-interpreting-machine-learning`

Additive Feature Attribution: Local surrogates

Let f be model to be explained and $\hat{f}$ the DNN estimation.

We wish to explain the single prediction $y = \hat{f}(x)$ based on input x.

With x fixed, the simplified input x' has each component as binary, with 1 meaning that that input component is present, 0 absent.

"Absence" means a feature value is equal to the dataset mean (removing a feature changes DNN/model, would need to retrain)

We have a function h_x such that $x = h_x(x')$. Input x will have some components k equal to the mean of that component across the dataset (absent, $x'_k = 0$) and other components j different (present, $x'_j = 1$). For example, denoting the input averages with m and other values with z,

$$x = [z_1, z_2, m_3, m_4] \Rightarrow x' = [1, 1, 0, 0].$$

$$\text{Then} \quad h_{[z_1,z_2,m_3,m_4]}([1, 1, 0, 0]) = [z_1, z_2, m_3, m_4].$$

Additive Feature Attribution: Local surrogates

Given input x, an explanation model is a local model g on simplified inputs such that, ideally, $g(z') \approx \hat{f}(h_x(z'))$ for $z' \approx x'$.

AFA: Explanation model is an additive g

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i$$

This model attributes an effect ϕ_i to each feature i. ϕ_0 is the response when all z_i are set to their dataset means, so that all z' are zero.

LRP, LIME and DeepLift use the equation above locally in x to fit the ϕ and get possible local explanations.

Additive Feature Attribution: Shapley Values

From cooperative game theory, **Shapley values**.

Game with N players. For coalition of players S, $f(S)$ is expected sum of payoffs (gain) members of S can obtain by cooperation. $f(\Phi) = 0$.

Shapley values distribute total gains to collaborating players. "i" gets

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [f(S \cup \{i\}) - f(S)]$$

$$\phi_i = \frac{1}{N} \sum_{\text{coalitions without } i} \frac{\text{marginal contribution of } i \text{ to coalition}}{\text{number of coalitions sans } i \text{ of this size}}$$

Additive Feature Attribution: Shapley Values

This is the only definition of the ϕ_i satisfying the four properties of Shapley values:

- Efficiency: Total gain is recovered $\sum_i \phi_i = f(N)$,
- (ii) Dummy: if a player never adds marginal value, their ϕ should be 0;
- (iii) Symmetry: If 2 players add the same marginal value to any subset to which they are added, their payoff portion ϕ is the same;
- (iv) Additivity. If a game is composed of two subgames, you should be able to add the payoffs calculated on the subgames, and that should match the payoffs calculated for the full game

Additive Feature Attribution: Shapley Values

What does this have to do with deep learning?

- *Game* is prediction task for single instance of dataset.
- *Gain* is the actual prediction for this instance minus the average prediction for all instances.
- *Players* are the feature values of the instance that collaborate to receive the gain (= predict a certain value).

$$AFA: \quad g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i$$

$$\text{Shapley:} \quad \phi_i(\hat{f}, x) = \sum_{z' \subseteq x' \setminus i} \frac{|z'|!(M - |z'| - 1)!}{M!} [\hat{f}(h_x(z' \cup i)) - \hat{f}(h_x(z'))]$$

Intuition: feature values $(x_j)_j$ enter in random order. All values contribute to prediction. Shapley value of x_i is average (on orders) of change in the prediction that current coalition receives when x_i joins.

The four properties revisited

- Efficiency: The feature contributions add up to the difference of prediction for x and the average.

$$\sum_i \phi_i(\hat{f}, x) = \hat{f}(x) - E[\hat{f}(X)].$$

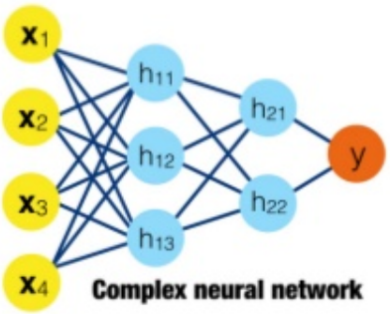
- Dummy: A feature j that does not change the predicted value - regardless of in which coalition of feature values it is replaced with its mean - has Shapley value 0.
- Symmetry: The contributions of two features values j and k is the same if they contribute equally to all possible coalitions.
- Additivity: For a combined prediction of f and F, the respective Shapley values are $\phi_j(f) + \phi_j(F)$. Suppose you trained a random forest, so that prediction is average of many decision trees. Additivity guarantees that for a feature value, you can calculate the Shapley value for each tree individually, average, and get the Shapley value for the feature value for the random forest.

Example I

Suppose we fit a DNN to predict good and bad loans

BAD	CUSTOMER_DTI	LOAN_PURPOSE	CHANNEL
0	0.18	MORT	7
1	0.42	HELOC	10
0	0.11	MORT	10
0	0.21	MORT	1

1. Train a complex machine learning model



Complex neural network

We have four features: x_1: Customer Debt To Income (DTI), x_2: Loan Purpose (Mortg/Heloc), x_3: Channel (1–10), x_4: annual income;

The output is a probability that the loan is bad.
Input client $DTI = 0.18$, $LP =$ "$Mortg$", $Ch = 7$, $AI = 70K$.
Suppose that the probability prediction for that is 0.7
Suppose that the average prediction across the dataset is 0.5.
Hence $f(x) - E[f(X)] = 0.7 - 0.5 = 0.2$.

Example II

Suppose we wish to find out how DTI contributed to this 0.2.
We need to look at all ways to add DTI to coalitions. We will thus consider "add DTI to:"

$$\Phi, \{LP\}, \{Ch\}, \{AI\}, \{LP, Ch\}, \{LP, AI\}, \{LP, Ch\}, \{LP, AI, Ch\}$$

As we explained earlier, we don't want to actually remove inputs and thus re-train the network.

Hence when we say for example $\{LP, Ch\}$, meaning that DTI and AI are missing, we mean to say that DTI and AI have been set to their average value across the dataset, denoted $\overline{DTI}, \overline{AI}$. Hence $\{LP, Ch\} = \{\overline{DTI}, LP, Ch, \overline{AI}\}$

Example III

Hence, to get DTI's contribution, ϕ_{DTI}, average

$$\hat{f}\{DTI, \overline{LP}, \overline{AI}, \overline{Ch}\} - \hat{f}\{\overline{DTI}, \overline{LP}, \overline{AI}, \overline{Ch}\}$$

$$\hat{f}\{DTI, LP, \overline{AI}, \overline{Ch}\} - \hat{f}\{\overline{DTI}, LP, \overline{AI}, \overline{Ch}\}$$

$$\hat{f}\{DTI, \overline{LP}, AI, \overline{Ch}\} - \hat{f}\{\overline{DTI}, \overline{LP}, AI, \overline{Ch}\}$$

$$\hat{f}\{DTI, \overline{LP}, \overline{AI}, Ch\} - \hat{f}\{\overline{DTI}, \overline{LP}, \overline{AI}, Ch\}$$

$$\hat{f}\{DTI, LP, AI, \overline{Ch}\} - \hat{f}\{\overline{DTI}, LP, AI, \overline{Ch}\}$$

$$\hat{f}\{DTI, LP, \overline{AI}, Ch\} - \hat{f}\{\overline{DTI}, LP, \overline{AI}, Ch\}$$

$$\hat{f}\{DTI, \overline{LP}, AI, Ch\} - \hat{f}\{\overline{DTI}, \overline{LP}, AI, Ch\}$$

$$\hat{f}\{DTI, LP, AI, Ch\} - \hat{f}\{\overline{DTI}, LP, AI, Ch\}$$

Example IV

Assume you get $\phi_{DTI} = 0.04$. This means that *DTI* contributed 0.04 to the difference $\hat{f}\{DTI, LP, AI, Ch\} - E(\hat{f}) = 0.2$.

If we repeat the procedure for LP, AI and Ch we may find that they contribute respectively $0.06, 0.13, -0.01$ so that

$$0.2 = 0.02_{DTI} + 0.06_{LP} + 0.13_{AI} - 0.01_{Ch}$$

This can be obtained implicitly with a simulation to avoid the exponential number of coalitions in cases with higher dimensions.

Instead of computing all coalitions exhaustively, we sample.

Shapley values: advantages

Advantages of Shapley values. Christoph Molnar
(Interpr. ML, `https://christophm.github.io/interpretable-ml-book/`)

Efficiency. Prediction "gain" is fairly distributed among the features values of the instance (LIME does not guarantee this - local).

Contrastive explanations. Instead of comparing a prediction to the average prediction of entire dataset, one could compare it to a subset or even to a single data point. Impossible with local models (LIME etc). Shapley value can be defined for arbitrary subsets, not just for *i*'s.

Solid theory (see 4 properties above). LIME assume linear behavior of the machine learning model locally, but there is **no theory** behind

Legal requirements. *Shapley value might be the only method to deliver a full explanation. In situations where the law requires explainability - like EU's "right to explanations" - the Shapley value might be the only legally compliant method (solid theory).*

Shapley values: disadvantages

Computing time. Exponential n. of coalitions is dealt with by sampling coalitions and limiting n. of scenarios. Decreasing MC scenarios reduces time, but increases the variance. n. scenarios?

Prone to misinterpreting. For a given input, Shapley value of feature *j* is *not* the difference of predicted values with and without *j* (removing *j* entirely would be a new DNN that would have to be retrained).

Rather, it is the *average marginal contribution of j across all possible coalitions*. Marginal does not mean removing *j* but replacing it with its average across data (either explicitly or via simulation).

The Shapley value is the wrong method for **sparse explanations** (always uses all features). SHAP (SHapley Additive exPlanations, Lundberg & Lee 2016) can provide explanations with few features. DeepSHAP in particular combines DeepLift and Shapley values, so as to be used for DNN (alternatively, Kernel SHAP can be used for DNN).

Shapley values: disadvantages

No prediction models, only contributions. The Shapley value returns no prediction model like LIME. It cannot be used to make statements about changes in prediction for changes in the input, e.g. If I earn 500USD more a year, my credit score increases by 5 points.

Need access to training data if you want to calculate the Shapley value for a new data instance. Not enough to access the prediction function because you need the data to replace parts of the instance of interest with values from randomly drawn instances of the data.

Inclusion of unrealistic data instances for correlated features. To simulate that a feature value is missing from a coalition, we marginalize the feature by sampling values from the feature's marginal distribution (equivalent to replacing with mean). OK if features are independent but if not we might sample feature values that do not make sense for this instance as their joint probability could be extremely low or even zero.

Do we really have an interpretation of a DNN?

- Are the above really interpretations/explanations?
- A bundle of tangent models is not “interpretation” / “explanation”.
- The whole point of the DNN is that it is nonlinear and that it has nonlocal effects. That’s why it is powerful. Local models ignore those.
- Shapley values are a combinatorial approach that has some good properties, and are a step forward, but does it really allow a human to predict or understand how the model will respond?
- I doubt regulators in finance, for example, will accept these methods to justify the use of DNN in risk management
- The black box aspect remains
- My gut feeling as a novice is that the reason why DNN are so powerful is the same why they are very difficult to interpret.
- Given progress in the area we can expect a breakthrough. Without that, DNN will be hardly accepted in heavily regulated areas (e.g. EU right to explanation)

QUESTIONS?

Thank you for your attention

Questions?