Question 3: Theoretical Part

(a) $\frac{\partial f(x)}{\partial z_1^{[2]}} = w_1^{[3]} \frac{\partial a_1^{[2]}}{\partial z_1^{[2]}} = w_1^{[3]} \sigma\left(z_1^{[2]}\right)\left(1 - \sigma\left(z_1^{[2]}\right)\right) = w_1^{[3]} a_1^{[2]}(1 - a_1^{[2]})$

$$f = [w_1^{[3]} w_2^{[3]}] A^{[2]}, \quad A^{[2]} = \sigma(Z^{[2]})$$

(b) $\frac{\partial f}{\partial Z^{[2]}} = \frac{\partial f}{\partial A^{[2]}} \frac{\partial A^{[2]}}{\partial Z^{[2]}} = \left[w_1^{[3]}\ w_2^{[3]}\right]^T o\ A^{[2]} o (1 - A^{[2]})$

(c) $\frac{\partial f(x)}{\partial Z^{[1]}} = \frac{\partial f(x)}{\partial Z^{[2]}} \frac{\partial Z^{[2]}}{\partial Z^{[1]}} = \frac{\partial f(x)}{\partial Z^{[2]}} \frac{\partial Z^{[2]}}{\partial A^{[1]}} \frac{\partial A^{[1]}}{\partial Z^{[1]}} = \begin{bmatrix} w_{11}^{[2]} & w_{12}^{[2]} \\ w_{21}^{[2]} & w_{22}^{[2]} \end{bmatrix} \delta_2 o\ A^{[1]} o (1 - A^{[1]})$

(d) $\delta_4 = \frac{\partial f(x)}{\partial w_{11}} = \frac{\partial f(x)}{\partial Z^{[1]}} \frac{\partial Z^{[1]}}{\partial w_{11}} = \delta_3^T \begin{bmatrix} x_1 \\ 0 \end{bmatrix}$

Question 4: Bonus Question

What's wrong with convolutional networks?

Summary and Comments:

In the speech Professor Geoffrey Hinton explains the cons of Convolutional Neural Networks and talks about his project named as Capsule Network. He explained how bad the pooling is and the fact that it is working so well is indeed a big disaster.

The shortcomings he discussed in his talk were:

1. **Backpropagation:** He explains that backpropagation requires a lot of data and it's an inefficient way of deep learning.

2. **Translational invariance:** He explains that CNNs have poor translational invariance and CNNs behave poorly when objects are rotated as they have no information of orientation ("pose") and when lighting conditions are changed. He gives an example of a square rotated to a diamond shape which will be viewed differently by observers and the fact that it's transparent will make it effected by lighting as well. So, in order to keep the model simpler and not to go in much complexity he said to assume it to be translucent. So, the 3D orientation is dependent on viewers, colors and lighting. This is known as "Pose". Hence the same object which is slightly rotated will not fire the neuron that will recognize the object. So CNNs cannot extrapolate geometric relationship to new viewpoints.

3. **Pooling Layer:** In CNNs, each feature detector is local and hence are repeated across layers as CNNs have multiple layers of feature detectors. CNNs use pooling to do routing. In Pooling, the relationship between the part and the whole is ignored so hence if we talk about the face, we need to combine few features like eyes, nose, 2 ears and mouth only then we can say it is a face. So, the precise location of the features is not captured in pooling in order to reduce number of parameters. The inability to capture the precise location of features results in output of two images to be similar which is not correct. He says that pooling solves the wrong problem of invariance and psychology of shape perception. So, it fails to utilize the underlying linear structure and is a very poor way to dynamic routing.

Sub-sampling results in no precise relationship between higher parts like nose and mouth and if overlapping of subsampling is done then this can be reduced.

Owing to the above issue of sub-sampling being invariant to small changes in viewpoint, Hinton proposes to aim for Equivariance as viewpoint changes lead to corresponding neural activities changes. He explains that place-coded equivariance is when different neurons are activated when significant changes happen while rate coding equivariance is that same neurons have different activation when there are small changes.

In between the talk Professor Geoffrey also explains how the human vision imposes a rectangular coordinate frame on objects so brain can do linear translation, but the rotation is difficult. However as highlighted in point 2. CNNs do not work when there is rotation.

In order to explain the power of the coordinate frames he also talks about the tetrahedron puzzle which took many people a lot of time to solve and interestingly one MIT professor gave a mathematical proof how it was impossible to solve after trying for 10 minutes. Then he also poses the class, another inverse tetrahedron puzzle which he describes is easy if we think tetrahedron from a non-standard way.

Then Professor Geoffrey discusses more about "Capsules" that they encode information like scale, velocity, orientation, color, deformation etc. A capsule gives the probability of presence of an entity as output along with additional metadata which is different from CNN. Hence this special feature helps the brain to distinguish which feature is mouth and which is eye of underlying face (component). Capsules do coincidence filtering that is they receive multidimensional prediction vectors from underlying layers capsules and then tries to find a tight cluster for predications. Hence, they work the best to filter out noise as high dimensional coincidences are much better than neuron.

He also explains about what the right representation of images is where he discusses that Computer Vision is inverse of computer graphics. The vision systems should take appearance and get the matrix that gives pose according to the respective viewer. This will make it easy to calculate the relationship between a part and a retina from relationship between whole and retina.

At the end of the lecture, he depicts a system which is combination of above which is not very clear because of the slide at front in the video. What he says is that his system is same in terms of classification accuracy to CNN but it is slower than CNN as it doesn't have CNN's property of matrix multiplies. However, his model removes all the shortcomings of CNN.

Conclusion

The dangers of CNNs can be removed by data augmentation and by re-implementing Professor Geoffrey Hinton's MATLAB written code in a much-optimized manner by utilizing parallel hardware's and GPUs. The problem of Pearson Re-id can be solved by using more than one neuron/model for transformation as explained by Professor Geoffrey. We should be open to pay additional penalty to include more information to improve equivariance.