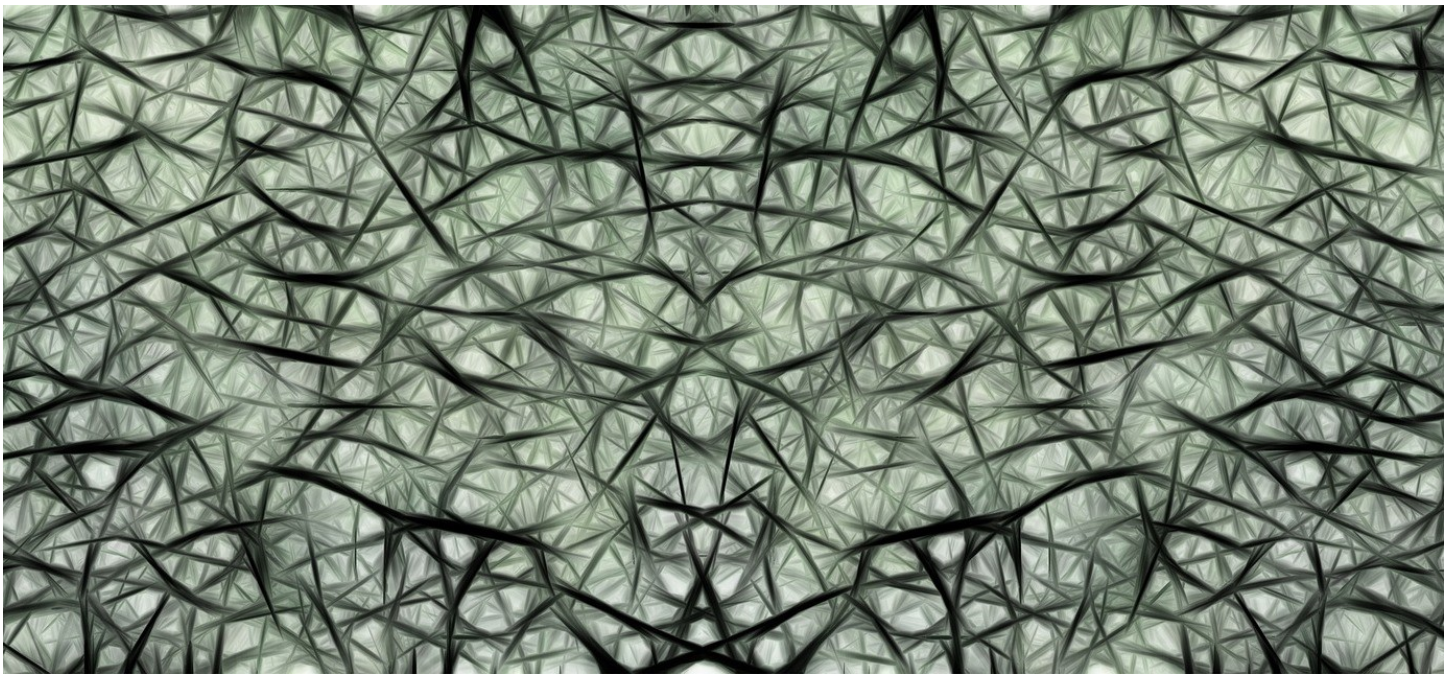
Nov 11 · 10 min read



Hinton++

Geoffrey Hinton is onto something. His model of machine intelligence, which relies upon neuron-clumps that he calls 'capsules', is the best explanation for how our own brains make sense of the world, and thus, how machines can make sense of it, too. Yet, there is always room for improvement. Capsules fail to account for our comprehension of *distortions*—we can tell 'what the toddler *meant*' when faced with an ungrammatical sentence, and we can see 'how to *re-arrange* Mr. Potato Head' when his features are out of alignment. **Capsules only recognize faces, for example, when their parts' poses are *proper*, with no suggestion for how to fix deformations**. I'll try to make some adjustments to his work, here.

Hold on, what are capsules?

It's okay if you've never heard of capsules. They are not a new topic—they've just been hiding inside Geoffrey Hinton's head since the seventies! At their core, capsules are an attempt to *overcome geometric changes*. When we observe a square on a sheet of paper, we recognize it as a square, even as the page *turns* or *tilts away from us*. The 'pixels' of

that square change dramatically, while our ***concept*** of 'square' remains static.

Artificial neural networks do not think in terms of static 'squares' like we do. Hinton created capsules so that CNNs can maintain *static concepts* of squares **even when those squares tilt and rotate**. Same goes for faces, cars, anything. Capsules *compensate for geometric distortion* by recording what Hinton calls 'pose'.

An object's **pose** describes *where* it is in the field of view, how *tilted* it is, its *relative size*, and how *skewed* it is. Together, these qualities describe the distortions that appear when 3D objects are reduced to a 2D image, and those objects move through the field of view. Hinton's core insight is this: if I see a face, but that face is *tilted and rotated*, then that *same* tilt and rotation is applied to *every part of the face*. So, if I see a 2D image with a tilted mouth, I expect to see *equally tilted eyes and nose*!

Hinton's capsules each look for their *part* of the face—one capsule seeks out noses, another searches for eyes, and a third capsule hopes to find a mouth. When each of those capsules finds an instance of their object, they light up, saying "I found an eye/nose/mouth!" They record their respective object's **pose**, and pass that information along to a *higher-layer capsule*, the face capsule. When that face capsule receives a signal from mouth, nose, and eye capsules, **it compares their poses**. A real face should have the same pose data for eyes, nose, and mouth. **If all three poses agree**, then the face capsule lights up, saying "I've found a face". That's all there is to it.

The face capsule in our example has its own pose, and it signals to a higher-layer capsule that detects the upper body. If capsules for shoulders, neck, and hair all light up, and all agree on a pose, then the higher-layer capsule lights up, too. This cascade of agreement continues until whole people are identified, and similar capsules operate for each other kind of object. The main idea is that larger objects are composed of smaller objects, and those smaller objects *expect each other to agree on their relative arrangement*. So far, Hinton's capsules work quite well, separating and identifying overlapping handwritten digits with accuracy far beyond existing neural networks. However…

"I remember that it didn't look right…"

Hinton's capsules stumble when their *expected agreement* doesn't appear. If the eyes are completely sideways, and the mouth is upside-down, then the 'face capsule' just won't register. The face capsule is as

silent as if there were **no face-parts *at all,*** even though those parts were present and mis-aligned. In contrast, *we* would look at the mis-aligned features and say "it's a *face*, but the eyes are sideways, and the mouth is upside-down." Our own 'face' neuron is lighting up, along with neurons for eyes and a mouth, *even though their poses do not agree*! Oops.

So, let's think this through. Some part of our brains see the eyes, nose, and mouth, and *they want to compress that information*. The **simplest compression** is to call that jumble 'a face', even though the features are mis-aligned. Our brains go right ahead and light up the 'face' neuron. Yet, our brains don't stop there. They *go back and* ***check*** that compression. "I am calling this madness a face, but *is it really a face*?" Our brains seem to project the 'face expectation' *back down to the lower layers*, asking, "if this were a **normal face**, what would the **normal poses** be for mouth, nose, and eyes?"

When that projection of pose occurs, the brain uses the pose of the 'face' that it settled upon during compression. "If this *really* is a face, in the orientation that I think it is, then the mouth should be here, oriented this way, and the eyes should be here and here, oriented this way…" The face's compression generates an *expectation of pose for its parts*!

The *actual* poses of the eyes and mouth are **compared** to the *expected* poses of those parts. That's when our brain registers an error! "I saw mouth, nose, and eyes, which led me to believe there was a face, ***but the face I assumed was there would have a different poses for these parts***." Our brains take note of those errors, and *remember* them. Additional neurons must fire, to record the distortions relative to expectation—the mouth is upside-down compared to the face's *expectation* of 'mouth', and the eyes are sideways compared to the face's *expectation* of 'eyes'. Those distortions are additional 'relative pose' data. And, our brains **keep track of those distortions**, alongside the neuron that detected the face, so that we are able to remember them later!

So, comparing Hinton's capsule to the our own brains, above:

The capsule sees eyes, nose, and mouth, records their poses, and sends those poses to the face neuron. The face neuron checks to see if those poses agree; because the poses *disagree*, **Hinton's face neuron *does not* fire**. It decides that *there is no face at all*!

Our brain model, however, records the presence of eyes, nose, and mouth, and sends those signals to the face neuron. **The face neuron *does* fire**, and it settles on a face pose that *minimizes dissonance*. That is, our brains take the pose of each part, and ask, crudely, "if all these parts make a face, then what face-pose is most likely?" Then, they **project that pose** back down to the eyes, nose, and mouth. Because the pose of the mouth and eyes are different from the face pose, *additional 'distortion' neurons are triggered*, recording the **change** to the face's pose that produce the eye and mouth poses. Our brains say "*there is a face*, but it's parts are distorted." We, unlike Hinton's capsules, remember what was wrong with the picture.

Chittering Monkey Brains

In addition to minimizing dissonance between poses, our brains also minimize dissonance between competing interpretations. This concept of dissonance has a biological equivalent. Our brains send many signals forward, each saying "look what I've found!" Yet, those signals regard the *same spot in our attention*, and they cannot all be right. So, our brains find the strongest among those signals, **by iterative suppression according to dissonance**. Imagine: you glance at a drawing which can be interpreted as an image of an old woman with a shawl or as a young woman with a hat…

One of those interpretations wins, at any given moment. Yet, our brains can hop between both viewpoints. Somehow, the signal that says "old woman's eye" competes with the signal that says "young woman's ear". Suppose that the picture was colored-in—in one coloring, the 'ear' is as

tan as the young woman's cheek, and the old woman illusion disappears; in another coloring, the 'eye' is white with a speck of green, dismissing the young woman interpretation. Each viewpoint is reinforced in turn, while the other is suppressed because it is *dissonant*. ("That might be an eye—wait, no, it's skin-colored." or "That might be an ear—wait, no, it has a white sclera and green iris.")

Be More Specific!

Looking at the optical illusion, our brains trigger neurons which correspond to both the young and old woman, up to a point. At the highest level of awareness, only one interpretation can exist. So, when both interpretations arrive, our brains begin to ***suppress*** *some of the signals that lead to each interpretation, until one interpretation is suppressed more than the other*. The brain takes some of the neurons that lit-up, and silences them. For the interpretation of 'young woman', we see her nose, her ear, jawline, and necklace. For 'old woman', those *same spots* register as an eyelid, eye, nose, and mouth. Each part has competing interpretations, and our brains begin ignoring one in favor of the other—calling the young woman's jawline a nose, instead, to see if that *reduces dissonance*, or turning the old woman's eye into an ear, in the hope of settling the disagreement.

This concept of 'knocking out' or 'reversing' inputs until disagreement disappears, applied to Hinton's capsules and poses, might look something like this: "I see a bowl of fruit"/ "I see a smiling face"… "They can't both be right… is this an eye, or an apple? I'll call it an eye, and see if one interpretation is more valid that the other, using my assumption"… "Yeah, if that apple is an eye, then this is likely a face, not a fruit bowl." Fundamentally, this process requires that higher-level neurons **send a signal back to the lower levels**, which opt for one signal or another, and those lower-level neurons **re-transmit their signals**, until the lower level neurons have minimal disagreement with their higher-level interpretation.

Disagreement may not disappear completely, though, as seen with the Mr. Potato Head example. Each lower-level neuron transmits it's guess of what it sees—"eye", "nose", "mouth", along with each part's *pose*. Those parts and poses pass to the higher layer, where they light-up the "face" neuron… yet, the parts' poses disagree! The mouth is upside-down, and the eyes are completely sideways! Our brains don't throw away the "face" neuron's signal —they just try to find a way to *minimize that disagreement*. "Perhaps the whole face is upside-down, explaining the inverted mouth? No, that can't be right, because the

nose is right-side up, and the eyes are in the correct position for an upright face." "Or, is the face sideways, explaining the unusual eyes? No, because that would put the mouth and nose in the completely wrong position." So, our brains say "**the face *really is* upright, because that orientation causes *the least disagreement*, but *parts* of the face are oriented wrong.**" The poses disagree, and that disagreement didn't go away—our brains just minimize that disagreement as much as possible.

What's Next?

Once our brains have minimized disagreement, and they've settled on a higher-level interpretation, they perform an extra step: they **project** that higher-level interpretation's *expectations* back down to the lower layers. This is a critical distinction from Hinton's capsules, and it is vitally important for moving toward generalized machine intelligence.

The Mr. Potato Head face, though distorted, has an 'average pose' for the entire face. On a normal face with that same average pose, there are *associated poses* for each part—"An upright face should have upright eyes, nose, and mouth". Our brains take the average pose from Mr. Potato Head, the pose which minimized disagreement, and project that pose back down as the *expected* pose of the parts. **Wherever the expected pose disagrees with the observed pose, our brains take note**: "I would expect, for this upright face, to see an upright mouth, but this mouth is upside-down!" (This method ensures that data receive **near-optimal compression**; our brains keep track of the 'main idea', along with any 'special instances'.) Later, when we think back to what we've seen, we don't remember ***only*** seeing a face—we ***also*** remember that its mouth was inverted. Capsules currently don't recall such distortions.

So, we have a *reverberation* between higher-level abstractions and lower-level features in two ways: 1) any 'tie' between competing abstractions ('young woman'/ 'old woman') is resolved by **suppressing some of their inputs until the tie is broken** (i.e. minimizing dissonance); 2) once an abstraction is agreed upon ('upright face'), its **idealized expectation** is projected back down to the lower layers ('upright eyes, nose, mouth'), to find where any disagreements remain ('upside-down mouth'). Our brains momentarily forget the *losing* abstraction, yet they remember the places where our higher-level expectations *disagreed* with lower-level observations. We recall seeing only the young woman OR the old woman, while we remember that Mr. Potato Head's mouth was upside-down.

In Practice

I suggest a simple fix for Hinton's capsules. Back-propagation by stochastic gradient descent still applies. Pose vectors are still compared as Hinton described. We only require a modification to the *feed-forward activation* of the network.

Normally, artificial neural networks march forward like a phalanx, stepping from the lowest layer of the network to the highest, never moving backward. However, the model I've described has many instances of backward and forward motion—when competing abstractions suppress inputs until settling on a single interpretation, and when the 'average pose' at a higher layer projects backwards to identify where its input's poses differ from expectation. This sort of neural network ***echoes***.

So, before arriving at a final answer, the network is excited *from low to high*, even activating multiple neurons in the *output layer*. Moving backward from the output layer, the network *suppress* mid-layer neurons until a single output neuron dominates. ("Yeah, it's an old woman.") Then, this output sends its pose expectation *backward* through the network, to record *where the idealized expectation differs from observation.* ("It's a face, but the mouth is upside-down.") The combination of output-layer activations AND low-level distortions is what the network 'really' sees, and that is what our neural networks should 'remember'… though, storing those memories efficiently is its own thorny problem. I'll leave that one for later.