

Inteligência artificial na análise de mamografia

Hae Yong Kim
Professor associado, Dept. Eng. Sist. Eletrônicos, EP-USP.
Seminário 13/08/2020, FM-USP.

1. Introdução

Em primeiro lugar, gostaria de agradecer à Profa. Maria Aparecida o convite para dar este seminário.

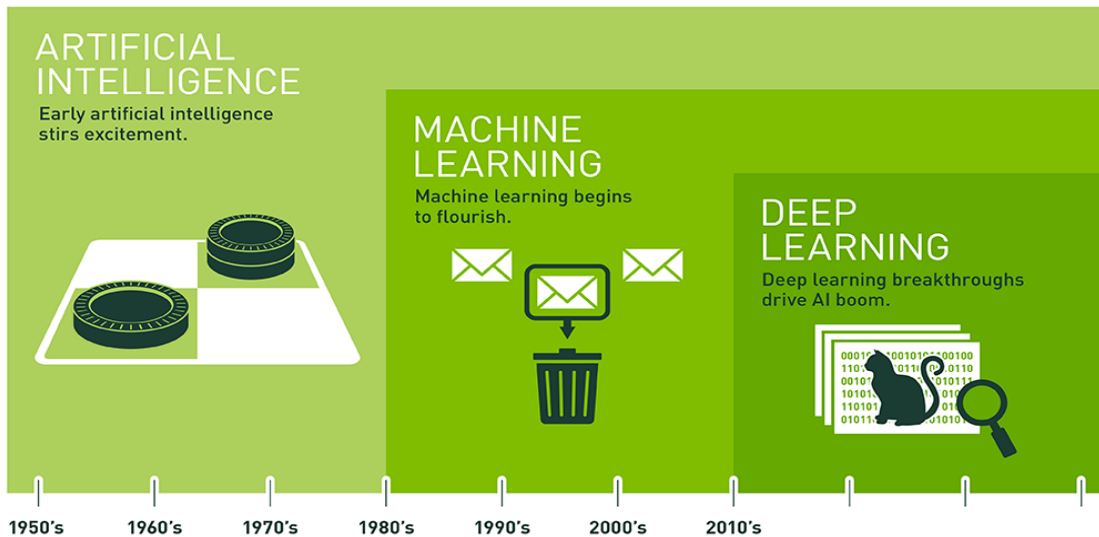
Faz aproximadamente um ano, sob coordenação da profa. Maria Aparecida e Dr. Shimizu, constituímos o “Grupo Radiômica” que está estudando o uso da inteligência artificial na análise de mamografia.

Houve recentemente uma revolução na inteligência artificial com a introdução da rede neural convolucional profunda [LeCun1989, Krizhevsky2012, LeCun2015]. Aprendizagem profunda está sendo usada com sucesso para analisar diferentes tipos de imagens médicas.

O objetivo deste seminário é explicar, intuitivamente usando o mínimo de matemática e computação, como funciona a rede neural convolucional profunda. Explicarei:

- a) O que é aprendizagem de máquina (inteligência artificial).
- b) Como funciona uma rede neural artificial.
- c) O que é convolução e como consegue extrair as características das imagens.
- d) Como funciona rede neural convolucional profunda.
- e) Análise de mamografia por computador.
- f) Análise de mamografia antes da aprendizagem profunda.
- g) Análise de mamografia usando aprendizagem profunda.

2. Aprendizagem de máquina supervisionada (ou inteligência artificial)



Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

“Aprendizagem de máquina” é uma sub-área de “inteligência artificial”. “Aprendizagem profunda” é uma sub-área de “aprendizagem de máquina”. Porém, muitas vezes os três termos são usados como sinônimos.

Aprendizagem de máquina supervisionada é, em essência, algo muito simples. Considere o problema de classificar um indivíduo em “Adulto”, “Bebê” ou “Criança” (ABC), dado o seu peso em kg. Para isso, o usuário fornece ao computador a tabela 1, com exemplos de entrada-saída. Por exemplo, a primeira linha (4, B) indica que a classificação de um indivíduo de 4 kg é “Bebê”.

Depois, o usuário pede ao computador que classifique o indivíduo com característica “16” (tabela 2). O computador pode adotar diferentes técnicas para classificar “16”, mas uma ideia razoável é procurar, entre os exemplos de treino, aquela instância mais parecida a “16”. Fazendo isto, computador encontra característica “15” cuja classificação é “C”. Então, o computador acredita que o indivíduo “16” deve ser da mesma categoria que “15” e atribui rótulo “C” a “16”. Este método chama-se “vizinho mais próximo”.

Tabela 1: Amostra de treino

<i>entradas de treino</i> (características)	<i>saídas de treino</i> (rótulos, classificações)
4	B
15	C
65	A

Tabela 2: Instâncias de teste a classificar.

<i>entrada de teste</i>	<i>saída do computador</i>
16	C

Evidentemente, este problema é simples demais. Um exemplo um pouco mais complexo está ilustrado na figura 1. Dadas duas características de um inseto (por exemplo, seu comprimento e peso), classificá-lo em “grilo”, “gafanhoto” ou “formiga”. Aqui também “vizinho mais próximo” pode ser usado.

Existem muitos outros algoritmos de aprendizagem além do “vizinho mais próximo”: árvore de decisão, boosting, classificador de Bayes, support vector machine, rede neural artificial, etc. Todos eles resolvem bem os problemas simples como os acima.

O problema é que uma mamografia não possui 1, 2 ou 3 características de entrada. Possui tipicamente $3000 \times 4000 = 12.000.000$ pixels e nenhum algoritmo de aprendizagem convencional funciona bem com esta quantidade de entradas.

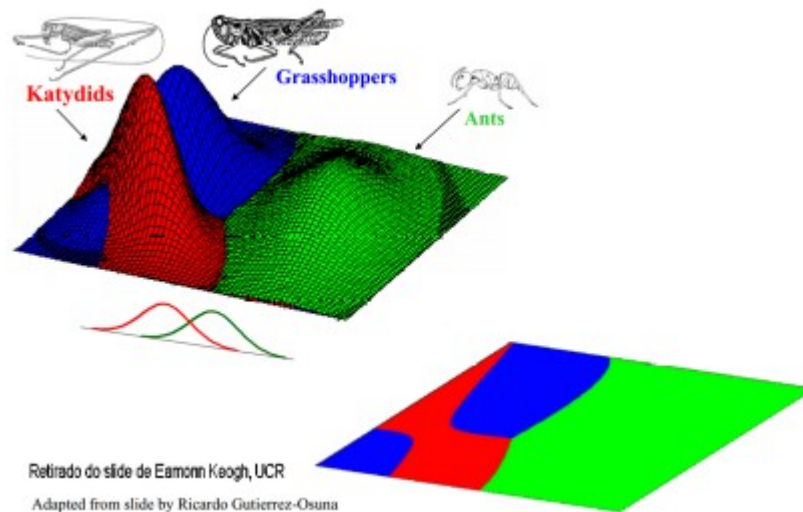


Figura 1: Classificação do inseto em grilo, gafanhoto ou formiga usando duas características.

3. Rede neural artificial

Antes da aprendizagem profunda, rede neural era somente uma entre muitas técnicas de aprendizagem. Porém, agora ela é a técnica de aprendizagem mais importante, pois é a base para a aprendizagem profunda. Rede neural também consegue resolver o problema ABC, mas o seu mecanismo de funcionamento é mais complexo que o “vizinho mais próximo”.

Um neurônio da rede (figura 2) recebe n entradas i_1, \dots, i_n (isto é, n números), multiplica-as pelos respectivos pesos w_i , soma um viés b , aplica uma função de ativação ao resultado para gerar a saída a (um número).

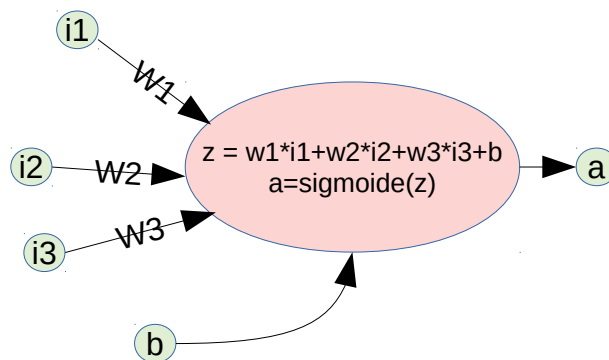


Figura 2: Um neurônio da rede neural artificial.

Uma rede neural consiste de vários neurônios interconectados. A rede neural que resolve o problema ABC está na figura 3. Cada “flecha” da rede possui um peso associado e cada neurônio da rede possui um viés associado. Para simplificar, vou chamar de “pesos” o conjunto formado pelos os pesos mais o vieses. O problema é encontrar os pesos que classifiquem corretamente os indivíduos em A, B ou C.

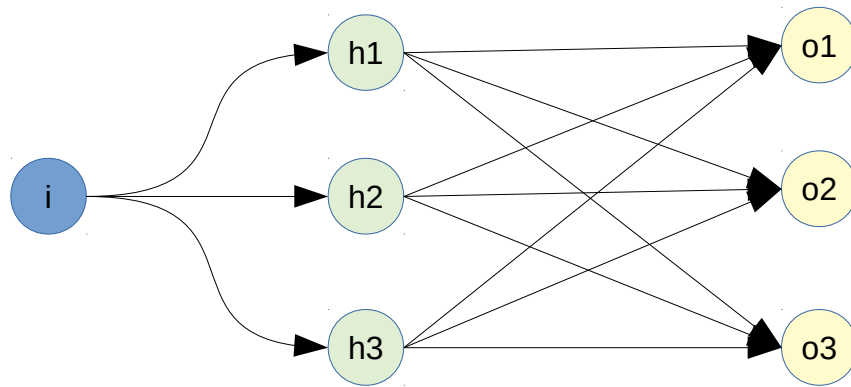


Figura 3: Estrutura da rede neural para resolver problema “Adulto”, “Bebê” e “Criança”.

Para usar rede neural, vamos converter os rótulos A, B e C em vetores (1, 0, 0), (0, 1, 0) e (0, 0, 1), pois a rede da figura 3 possui três saídas (tabelas 3 e 4). Precisamos encontrar os pesos que fazem a classificação desejada. Para isso, o programa inicializa aleatoriamente os pesos. Para treinar a rede, o programa apresenta à rede um exemplo de treinamento (por exemplo “4”) e pega a saída fornecida pela rede. A saída desejada é o vetor (0, 1, 0) significando “Bebê”. Então, o programa modifica um pouquinho cada peso para que a saída obtida fique um pouco mais próxima da desejada.

Tabela 3: Amostras de treinamento e vetor de categorias.

<i>entradas de treino</i>	<i>saídas de treino</i>	<i>categorias</i>
4	B	0 1 0
15	C	0 0 1
65	A	1 0 0

Tabela 4: Instâncias de teste e vetor de categorias.

<i>entrada de teste</i>	<i>saída do computador</i>
16	0 0 1

É como mexer um pouco a bolinha da figura 4 de lugar para que ela vá para um ponto com erro um pouco mais baixo. Este processo, chamado de retro-propagação, é repetido muitas vezes para todos os exemplos de treino. Após o treino, a rede aprende a classificar corretamente os indivíduos.

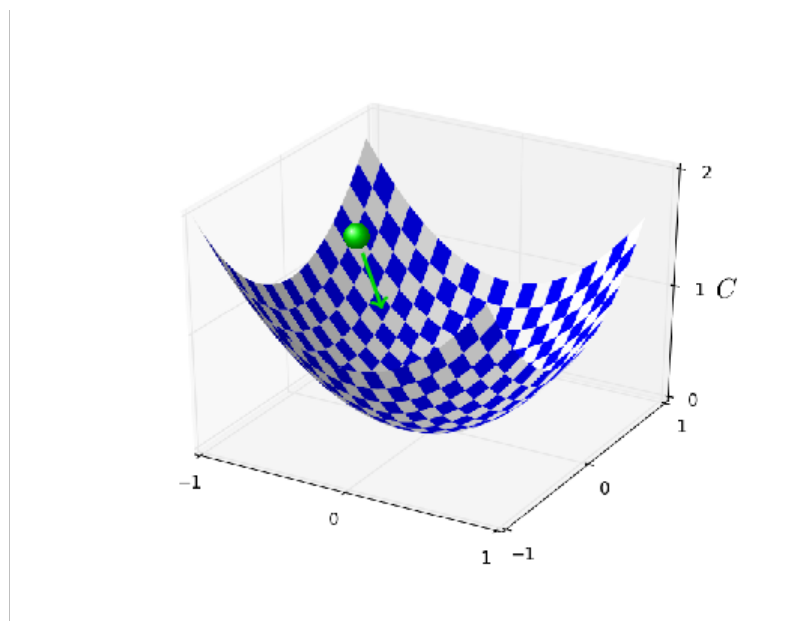


Figura 4: Retro-propagação muda os valores dos pesos para chegar ao ponto de baixo erro.

4. Extração de características e convolução

Qualquer algoritmo de aprendizagem consegue resolver problemas simples, como “ABC” ou “os 3 tipos de insetos”. Porém, nos problemas mais complexos, é necessário primeiro extrair características. Para classificar mamografia com 12 milhões de pixels, devemos resumir de alguma forma as informações dos pixels e alimentar o algoritmo de aprendizagem apenas com as informações resumidas.

É possível resumir informações de uma imagem fazendo convoluções (também conhecida como “filtro linear” ou “filtro” ou “casamento de modelo”, figura 5). Uma convolução percorre a imagem de entrada com uma matriz de pesos, calculando média aritmética dos valores dos pixels ponderados pelos pesos. Note que, por uma “feliz coincidência”, tanto rede neural quanto convolução calculam médias aritméticas com pesos.

A figura 6 mostra duas (das muitas convoluções) usadas para detectar faces humanas. Se uma face humana aparece numa certa região da imagem, essas duas convoluções devem dar respostas altas nessa região. A primeira convolução verifica o quanto a região dos olhos é mais escura do que a região das bochechas. A segunda convolução verifica se a área entre os olhos é mais clara do que as áreas nos olhos. A figura 5 ilustra o cálculo da segunda convolução. Há um pico (255) na saída correspondente à região clara entre duas regiões escuras na entrada.

Concluindo, é possível extrair característica das imagens executando várias convoluções. Porém, é muito difícil descobrir quais são as melhores convoluções para resolver um determinado problema.

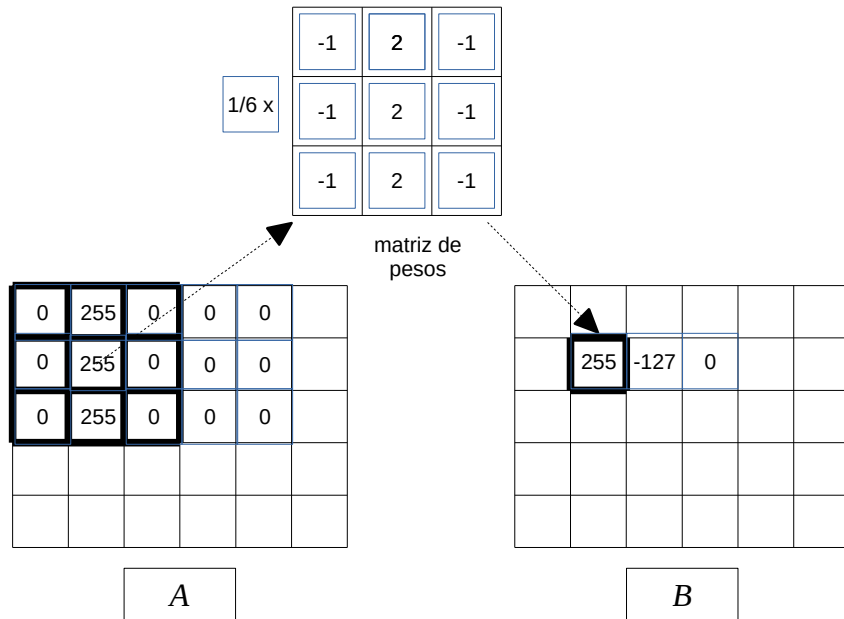


Figura 5: Convolução que detecta “região clara entre duas regiões escuras”.

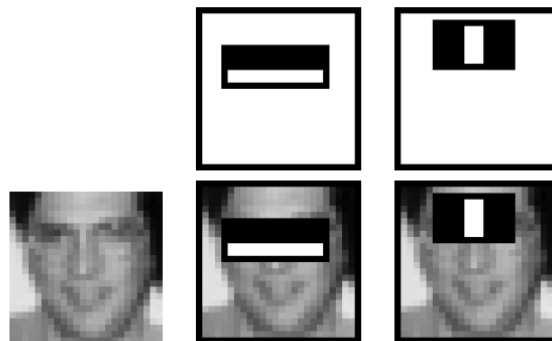


Figura 6: Duas (entre muitas) convoluções necessárias para detectar face.

5. Rede neural convolucional profunda

5.1 Classificação de dígitos manuscritos

Apresento uma visão intuitiva de rede neural convolucional profunda, usando o exemplo de classificar os dígitos manuscritos MNIST (figura 7). MNIST possui 60.000 imagens para treino e 10.000 imagens para teste. O “vizinho mais próximo” gera taxa de erro de 3% e “rede neural artificial” gera erro de 2%.

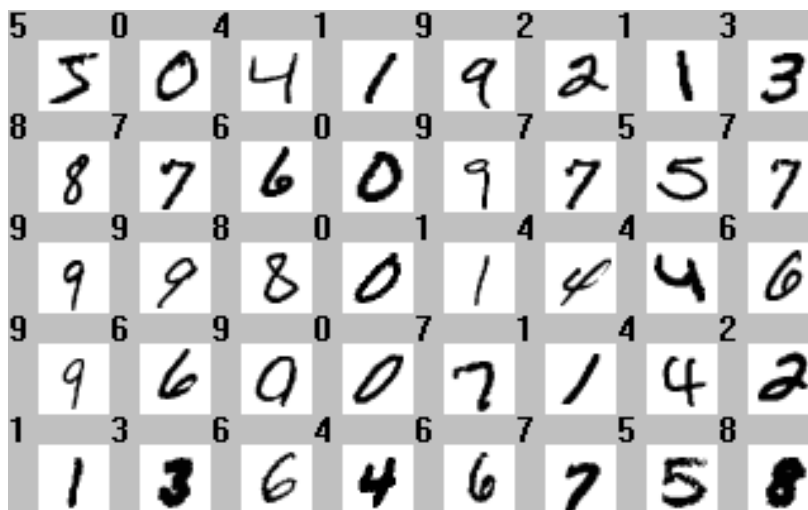


Figura 7: Alguns dígitos da base MNIST com os rótulos.

Como diminuir mais o erro? Talvez usar convoluções para detectar pontas de retas, linhas verticais e horizontais (figura 8)? Duas pontas de retas, uma linha vertical e duas linhas horizontais caracterizariam a forma “3” (contém) que poderia ser detectada fazendo uma outra convolução. Duas formas “3” caracterizariam o dígito “3” que seria detectada por uma ainda outra convolução. Rede convolucional profunda utiliza esta ideia de fazer convoluções em cascata. É convolucional porque usa convoluções. É profunda porque existem muitas camadas de convoluções.

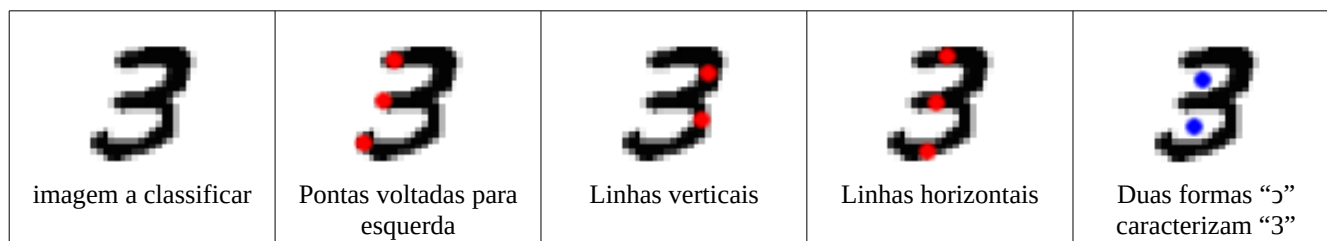


Figura 8: Seria possível usar convoluções para melhorar a classificação de MNIST?

Aqui, os pesos da rede neural são os próprios pesos das convoluções. A rede convolucional descobre quais são as convoluções adequadas usando retro-propagação. Ela primeiro inicializa aleatoriamente os pesos. Depois, efetua repetidamente retro-propagação, modificando aos poucos cada peso para diminuir o erro. Retro-propagação acaba escolhendo automaticamente as convoluções adequadas. Porém, quem projeta a estrutura da rede continua sendo um ser humano.

Figura 9 mostra a estrutura de rede profunda “LeNet” usada para classificar os dígitos. Ela pode ser dividida em duas partes: a primeira (em vermelho) extrai as características usando convoluções; a segunda (em azul) classifica as características usando uma rede neural clássica.

Figura 10 mostra as 20 convoluções da primeira camada projetadas automaticamente. Algumas delas possuem interpretações intuitivas. Por exemplo, o filtro em vermelho detecta as retas horizontais o filtro em azul detecta as retas verticais. As imagens filtradas por esses filtros na figura 11 confirmam que, de fato, eles detectam as retas horizontais e verticais.

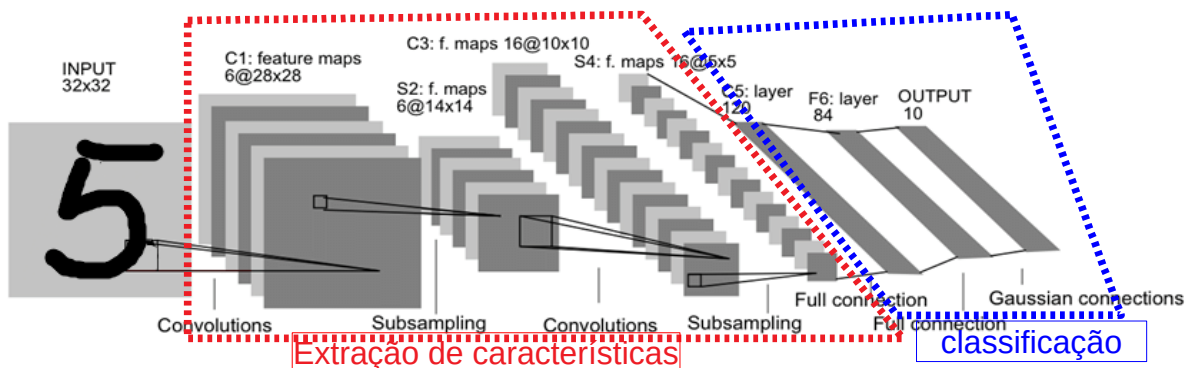


Figura 9: Rede neural convolucional profunda LeNet para classificar dígitos manuscritos.

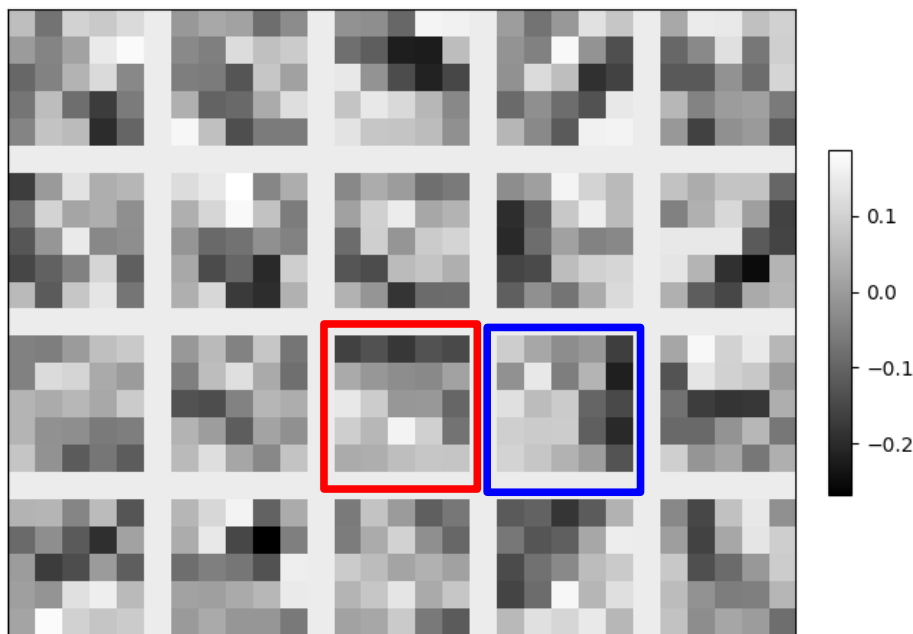


Figura 10: Os pesos das 20 convoluções 5x5 da primeira camada. Os filtros marcados em vermelho e azul detectam respectivamente retas horizontais e verticais.

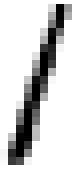
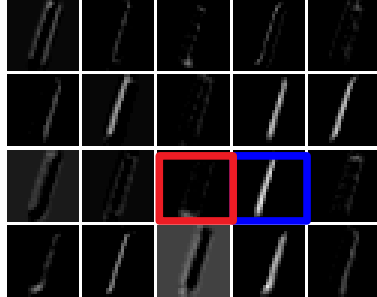
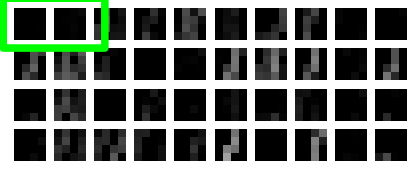
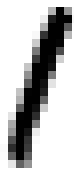
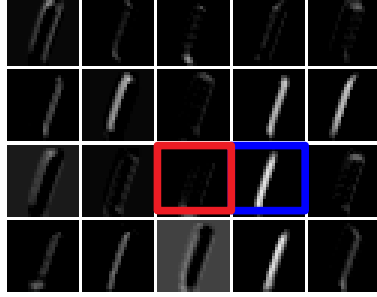
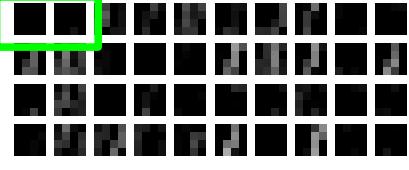
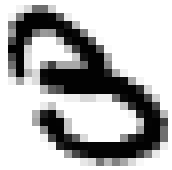
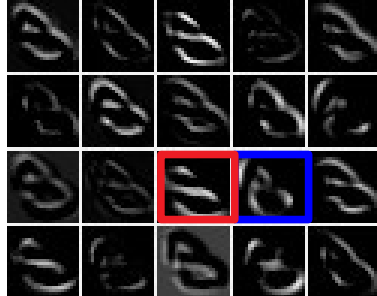
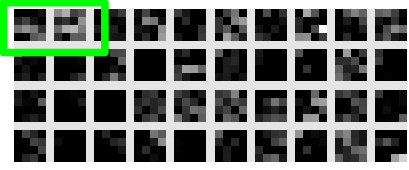
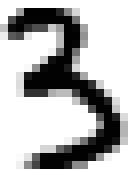
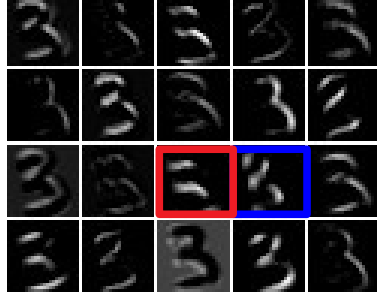
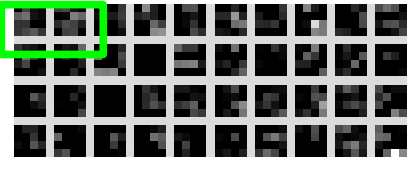
(a) imagem	(b) 20 saídas dos filtro da primeira camada	(c) características escolhidas
		
		
		
		

Figura 11: (a) Dígito a classificar. (b) Ativações da primeira camada. (c) As características extraídas automaticamente pela rede convolucional.

A coluna c da figura 11 mostra as características extraídas automaticamente. São as “informações resumidas” que permitem classificar os dígitos. É possível distinguir dígitos “1” e “3” olhando essas características? Sim, é só olhar (por exemplo) os dois primeiros blocos (marcados em verde). Eles são completamente pretos em “1”, mas estão cheios de ativações em “3”.

A taxa de erro de rede convolucional é 0,37%, muito menos do que 2% usando rede clássica. Figura 12 mostra os 37 dígitos classificados incorretamente. A taxa de erro de um ser humano é algo entre 2% e 2,5%. Portanto, rede convolucional erra substancialmente menos do que um ser humano.

2	3	9	6	1	4	4	9	7	4
27	83	89	65	71	46	94	95	71	94
3	9	6	7	4	0	3	6	3	7
53	49	68	72	94	20	53	61	35	79
6	6	7	7	9	6	3	0	4	9
60	68	79	73	94	61	35	60	94	49
2	3	6	7	8	2	6	[Redacted]		
72	53	65	71	82	72	56			

Figura 12: Os 37 dígitos classificados incorretamente. Número à esquerda é a classificação correta. Número à direita é a classificação dada pelo algoritmo.

Para problemas simples como MNIST, a diferença entre usar ou não aprendizagem profunda é obter 37/10000 ou 200/10000 erros. Porém, para problemas somente um pouco mais complexos, não usar aprendizagem profunda significa ter resultado equivalente à “chute”.

A figura 13 mostra o banco de dados Cifar-10 com pequenas imagens classificadas em 10 categorias: 0=airplane, 1=automobile, 2=bird, 3=cat, 4=deer, 5=dog, 6=frog, 7=horse, 8=ship, 9=truck. O “vizinho mais próximo” tem taxa de acerto de 15%, praticamente igual ao “chute” que obteria 10% de acerto. Enquanto isso, rede profunda acerta 93%.

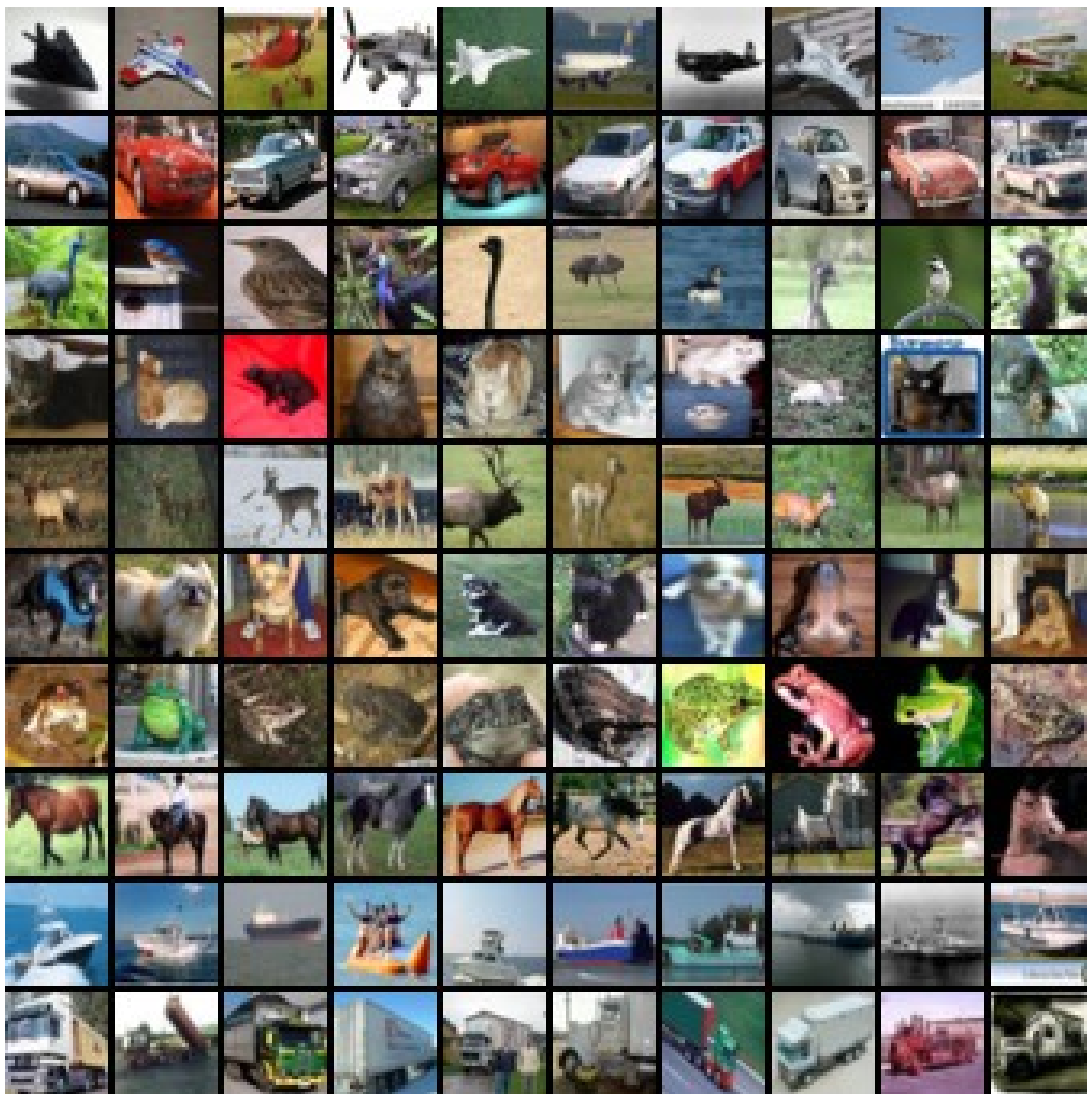


Figura 13: As 10 classes de Cifar-10. As classes são 0=airplane, 1=automobile, 2=bird, 3=cat, 4=deer, 5=dog, 6=frog, 7=horse, 8=ship, 9=truck.

Hoje, existem várias arquiteturas de redes mais sofisticadas que LeNet, por exemplo, AlexNet, VGG, ResNet, Inception, etc. Além da classificação, as redes profundas podem ser usadas para muitos outros problemas da Medicina como segmentação, detecção, etc.

Concluindo, a grande vantagem da rede convolucional profunda (em relação aos algoritmos de aprendizagem clássicos) é que ela consegue extrair automaticamente as características adequadas para classificar imagens. Nenhum outro método de aprendizagem fazia isso: sempre era necessário programar manualmente como extrair as características. A grande revolução de inteligência artificial atual é basicamente a possibilidade do computador escolher automaticamente quais são as características que mais ajudam a resolver um problema.

Por outro lado, a inteligência artificial ainda está muito longe de atingir a inteligência humana. Inteligência artificial necessita de uma quantidade enorme de exemplos de treinamento. Ser humano consegue aprender a partir de uns poucos exemplos. Portanto, a grande questão aberta é como fazer o computador aprender a partir de poucos exemplos.

5.2 Alguns problemas de imagens considerados insolúveis há apenas 8 anos atrás

Reconhece objeto ou animal na imagem.



89.79% chimpanzee



87.75% orangutan

Um programa simples comete 1% de erro ao classificar rostos em masculino/feminino.



Colorir automaticamente fotos preto e branco.



Colorado National Park, 1941



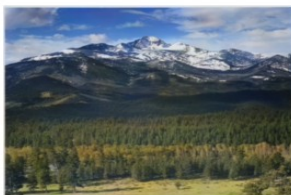
Textile Mill, June 1937



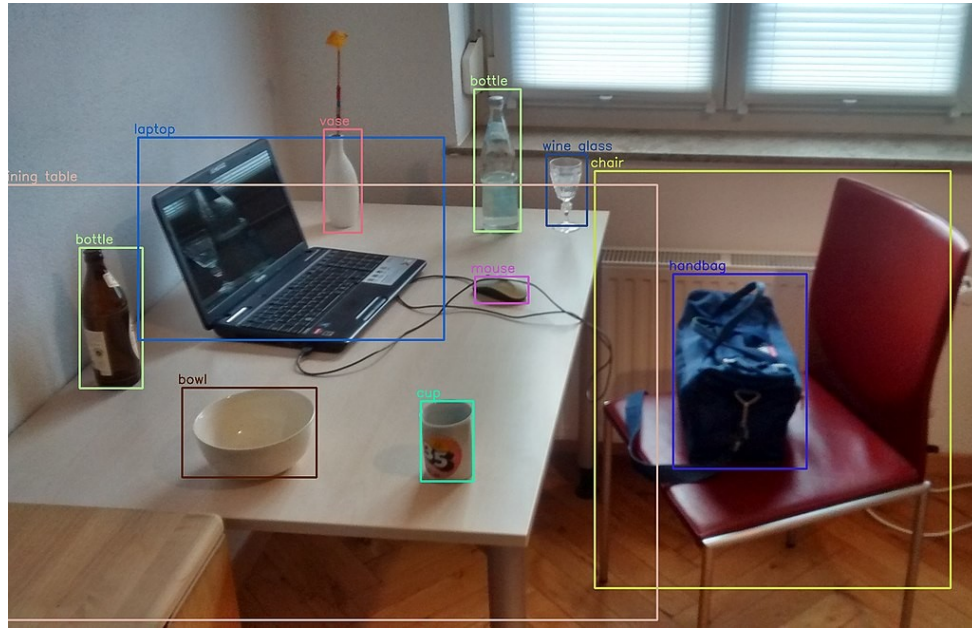
Berry Field, June 1909



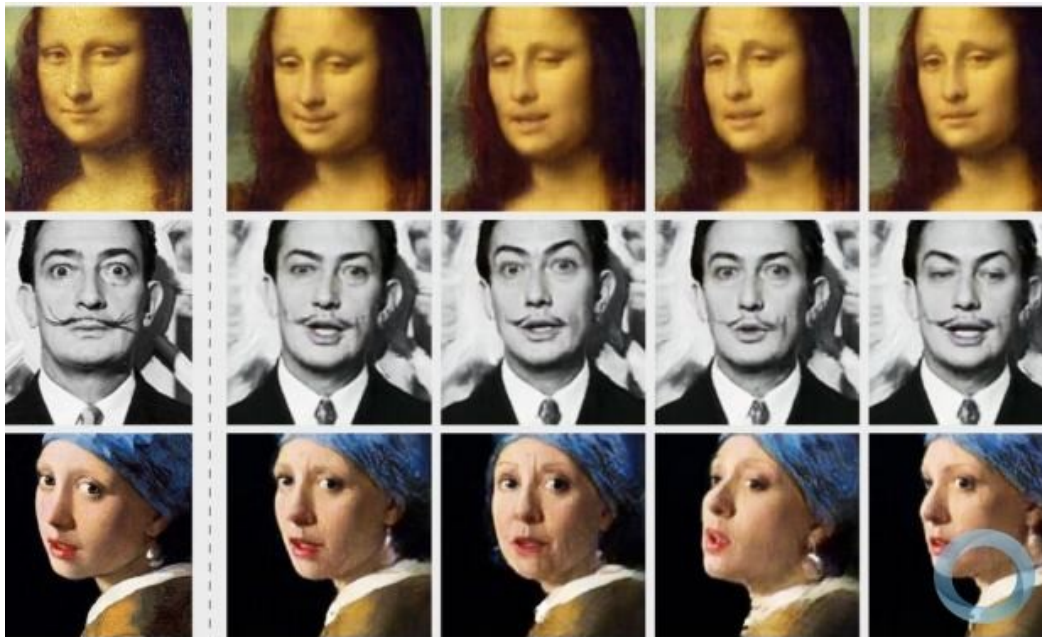
Hamilton, 1936



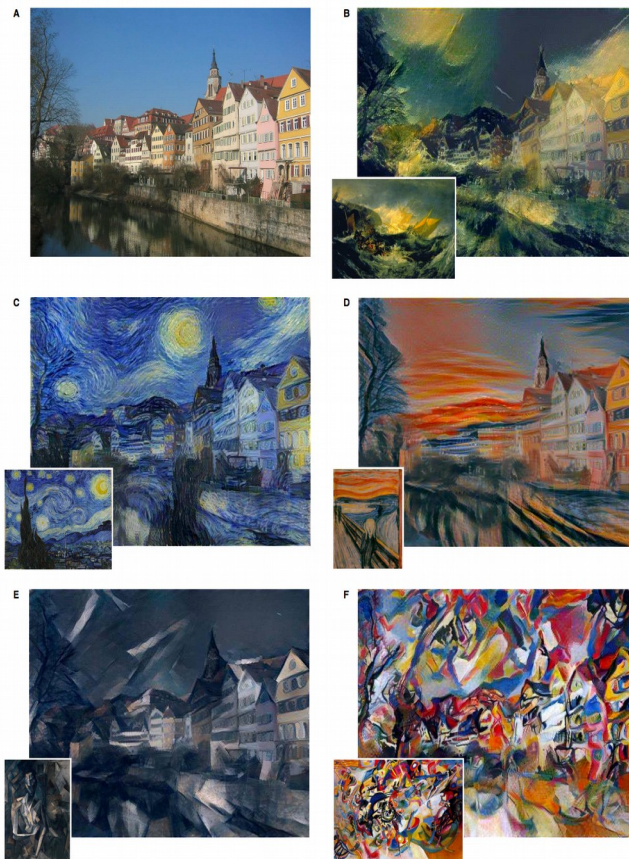
Detector objetos:



Gerar faces falsas:



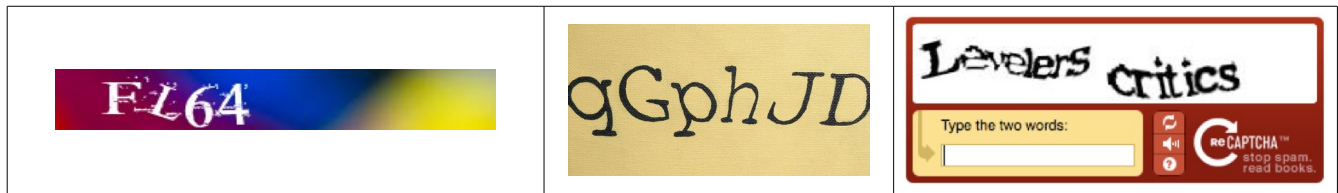
Transferir estilo de pintura



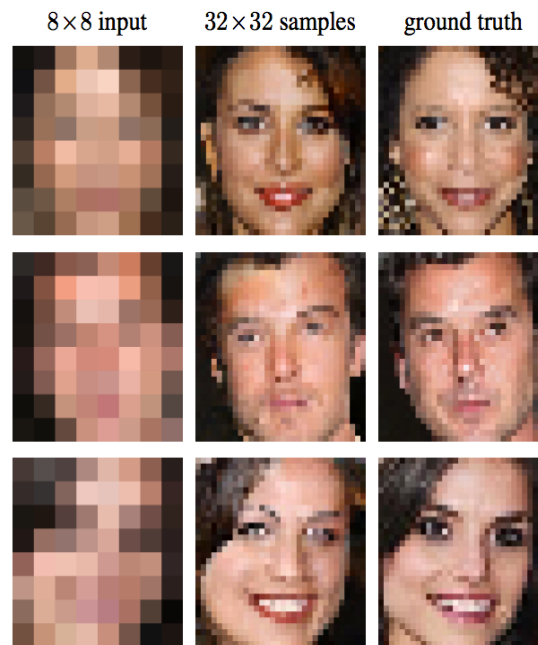
Aprender a jogar um jogo:



Computador soluciona melhor algumas “captchas” do que ser humano.



Super-resolução:



Aplicações que não são de imagens:

- Reconhecimento de fala (Siri, Cortana, etc)
- Tradução de texto (Google translate, Bing translator, etc.)
- Recomendação de vídeo no Youtube.
- Propaganda com sugestão de produtos que o usuário pode se interessar.
- Carro autônomo.
- Composição automática de música.
- Leitura labial (93% de acerto, melhor que ser humano 52%).

6. Análise de mamografia por computador

O problema que queremos resolver é: “Dada uma mamografia, dizer se a paciente tem câncer ou não”.

Este é um problema típico de classificação de imagens. Só que mamografia possui ≈ 12 milhões de pixels com 4096 níveis de cinza! Devemos diminuir a quantidade de informações extraindo algumas características.

Há duas soluções:

1. Métodos clássicos dividem o problema em dois sub-problemas (figura 14): CADe (detecção das lesões) e CADx (classificação das lesões). Os dois problemas são resolvidos tipicamente utilizando técnicas desenvolvidas manualmente.
2. A rede neural convolucional profunda resolve todos os sub-problemas de uma só vez usando exemplos de treinamento.

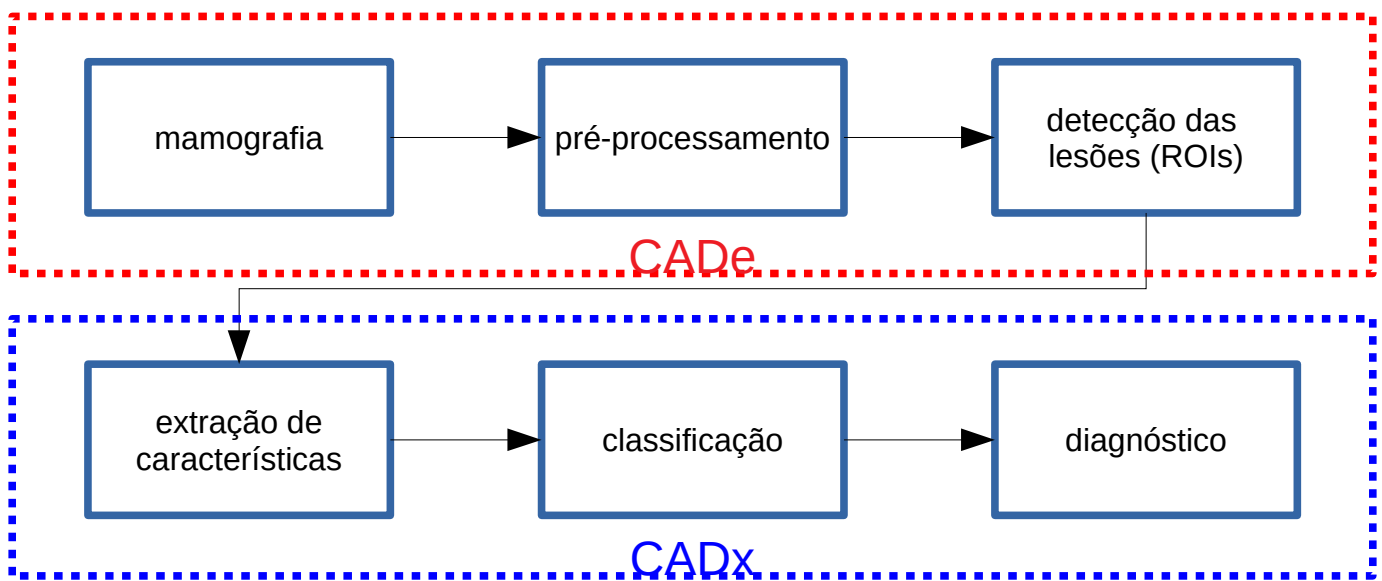


Figura 14: Etapas para classificação de mamografia clássica.

7. Análise de mamografia clássica

7.1 CADe e CADx

Como é muito difícil escrever manualmente um programa que classifique uma mamografia em câncer/não-câncer com alta taxa de acerto, antes da aprendizagem profunda os programas tentavam apenas ajudar o radiologista, chamando a atenção para as regiões suspeitas.

Dividia-se o problema em dois (figura 14): CADe (computer-aided detection) e CADx (computer-aided diagnosis). CADe localizava regiões suspeitas na mamografia (que podem ser câncer ou falso positivo). CADx classificava as regiões suspeitas em maligno ou benigno [Ayer2010].

Na figura 15, as flechas indicam verdadeiras lesões tipo “massa”. O programa “ImageChecker” marcou com “*” as possíveis lesões “massa” e com “Δ” as possíveis lesões “microcalcificação”. Versão 3 do programa errou todas as marcações, enquanto que versão 8 acertou uma única lesão “massa” mas gerou vários falsos-positivos [Kim2010, Dromain2013].

A especificidade desses CADe era muito baixa, gerando por volta de um falso-positivo por vista. Assim, esses sistemas melhoravam a sensibilidade do radiologista (de 4% a 15% [Domain2013]) mas pioravam na especificidade (de 5% a 35%).

7.2 CADe e CADx

Os artigos [Oliver2010 e Elter2009] apresentam os principais métodos clássicos para a detecção e segmentação de lesões.

Uma vez que as lesões são localizadas pro CADe, CADx as classifica em benigno ou maligno. O trabalho [Elter2009] apresenta algumas características usadas para essa tarefa. Depois, os métodos clássicos de aprendizagem eram usados para classificá-las.

O trabalho [Ayer2010] mostram tabelas com AUCs de vários CADx, cujo resumo está na tabela 5. AUCs da tabela 5 não podem ser comparadas com AUCs das técnicas modernas, pois os números da tabela 5 se referem à classificação de pequenas imagens ao redor das lesões, enquanto que AUCs de técnicas modernas normalmente se referem a classificar mamografias inteiras.

Tabela 5: O menor e o maior AUC reportado por [Ayer2010] em CADx para classificar lesões.

Modalidade	Menor AUC	Maior AUC
Mamografia	0,83	0,965

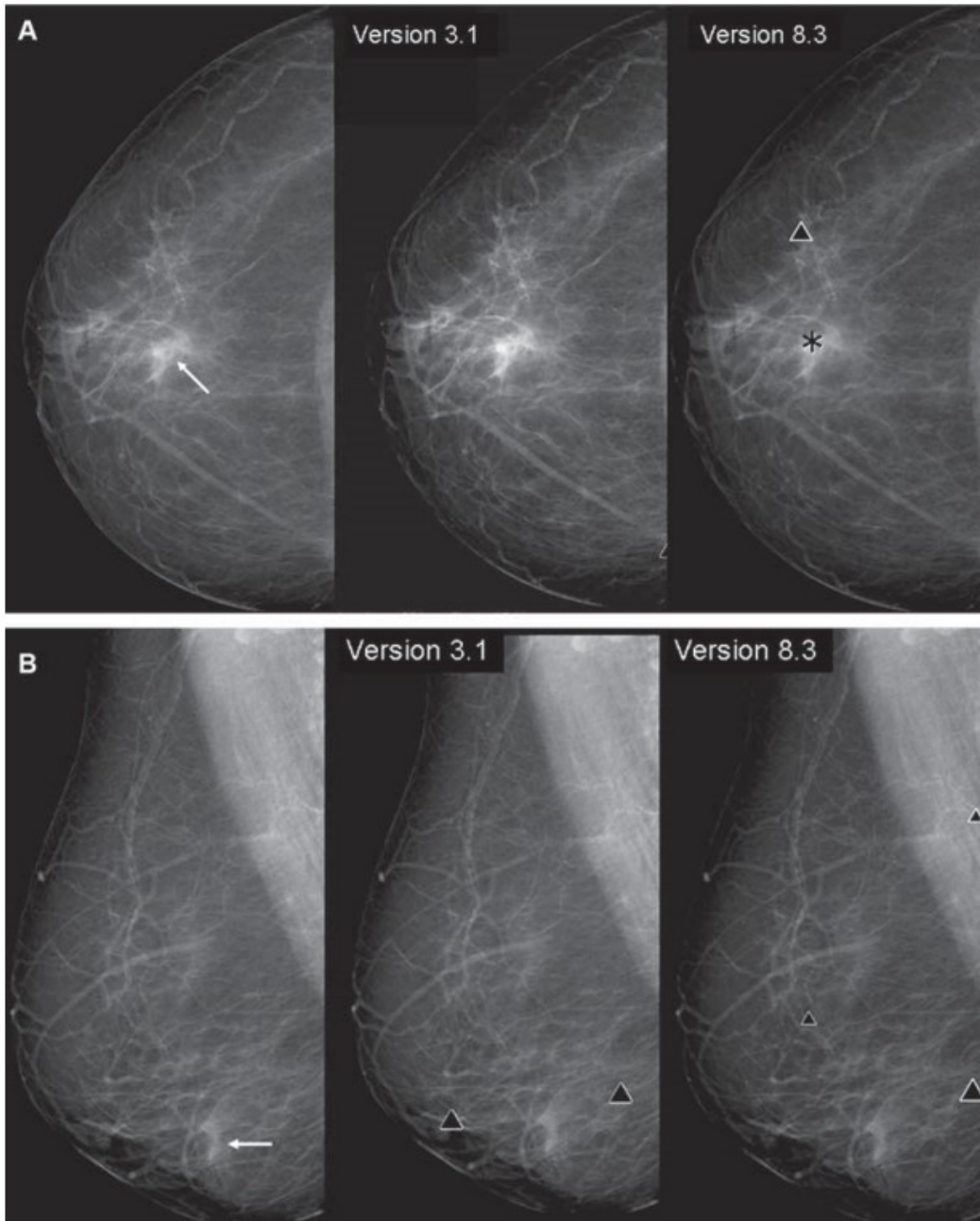


Figura 15: Flecha indica lesão tipo massa verdadeira. CADe indica com * lesão tipo massa e com Δ lesão tipo microcalcificação. Figura retirada de [Kim2010].

8. Análise de mamografia usando aprendizagem profunda

Descrevo abaixo alguns artigos já estudados pelo nosso grupo.

8.1 [Wu2020]

O trabalho [Wu2020] utiliza um BD com 230.000 exames dos quais 5.800 com biópsia e 985 com câncer. Relata ter obtido AUC 0.895. Este trabalho disponibilizou o programa obtido mas não o BD utilizado. O programa gera, como passo intermediário, “heatmaps” que indicam a probabilidade de lesão por região. A figura 16 ilustra um “heatmap”.

O nosso grupo construiu um BD com 135 pacientes jovens de até 40 anos. Mulheres jovens possuem mama densa, o que pode dificultar a análise por computador. Das 270 mamas, 170 tinham câncer e 100 não tinham. Testando o software, obtivemos AUC de 0.876.

Fizemos “transfer learning” para adequar a rede neural ao nosso BD. Fazendo “5-fold cross validation”, obtivemos AUC 0.9018.

Com isso concluímos que o programa consegue diagnosticar câncer em mulheres jovens sem grande perda de acuracidade. Também concluímos que é possível adaptar o programa original para um banco de dados específico para melhorar a acuracidade.

8.2 [Shen2019]

O artigo [Shen2019] descreve uma técnica nova de aprendizagem profunda denominada “end-to-end approach”. Basicamente, esta técnica treina primeiro um classificador de lesões em benigno/maligno. Depois, este classificador é generalizado para classificar mamografia inteira. Os autores relatam que obtiveram AUC de 0.98. O nosso grupo já conseguiu replicar este trabalho.

8.3 [McKinney2020]

O trabalho [McKinney2020] testou um sistema de inteligência artificial moderno para uso clínico, nos EUA e em UK. O sistema consiste de 3 sub-sistemas e as respostas dos 3 sub-sistemas são ponderadas para dar veredito final. Reporta uma redução de 5.7% e 1.2% (respectivamente EUA e UK) em falsos positivos e 9.4% e 2.7% em falsos negativos.

8.4 Conclusão

Diferentemente dos antigos sistemas, os novos parecem ser capazes de realmente ajudar os radiologistas a diminuir tanto falsos negativos quanto falsos positivos.

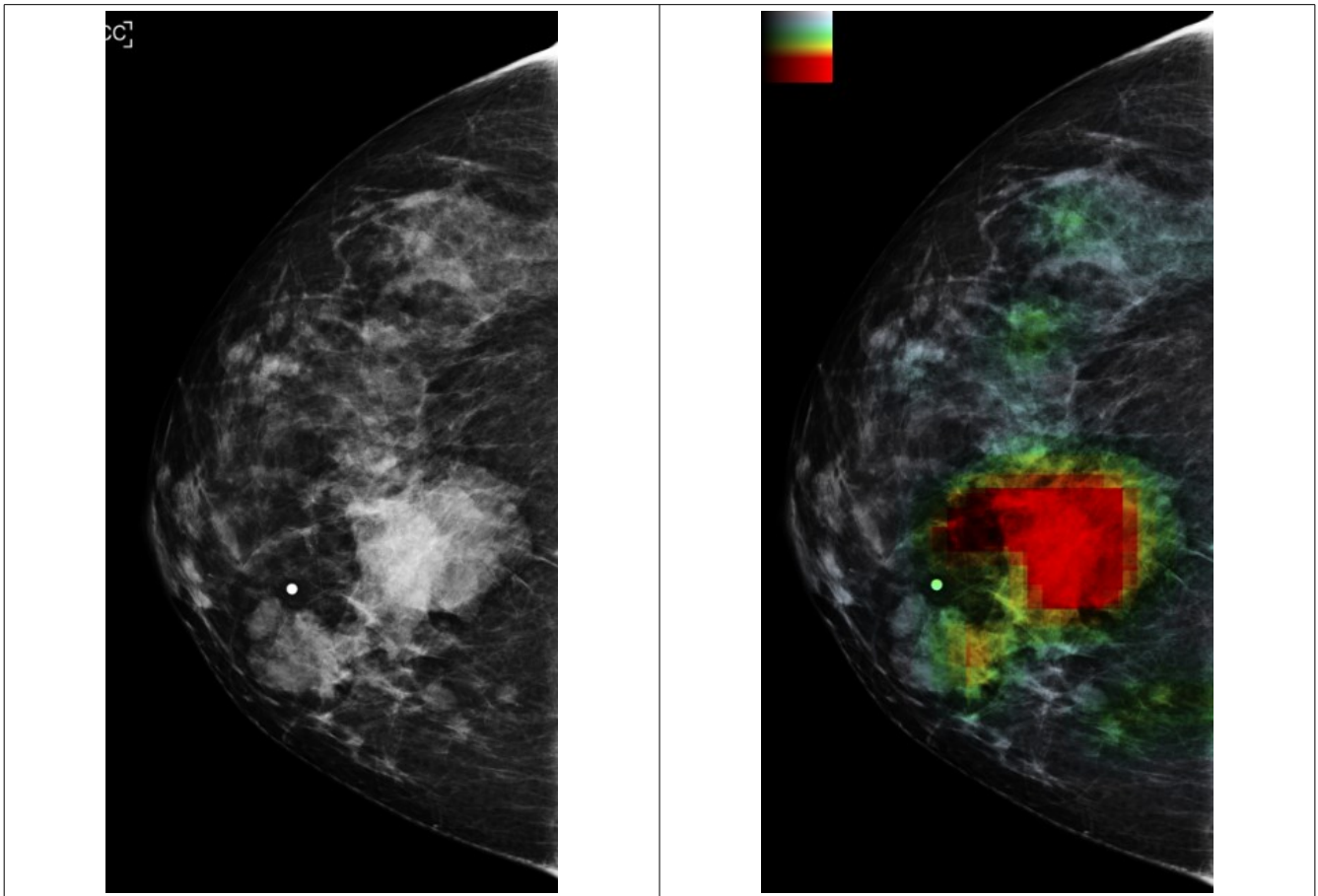


Figura 16: “Heatmap” gerado pelo programa [Wu2020] indicando regiões com maior possibilidade de lesão.

9. Problemas

A grande questão da inteligência artificial atual é como diminuir o número de exemplos de treinamento necessário. Um ser humano consegue aprender a partir de poucos exemplos, enquanto que o computador necessita de milhares ou milhões de exemplos.

O que falta nesta área são bancos de dados públicos de mamografias digitais com rótulos dados por biópsias. A sua ausência torna impossível comparar diferentes técnicas e impede que qualquer um possa testar suas ideias. Se o nosso grupo construir um bom BD público, seria uma grande contribuição para a Ciência.

Há outros problemas relacionados que merecem ser estudados, por exemplo:

- a) Classificar a densidade da mamografia e recomendar a necessidade de realizar outros exames.
- b) Associar defeito genético com mamografia.

Referências

[LeCun1989] LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4), 541-551.

[Krizhevsky2012] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097-1105).

[LeCun2015] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, 521(7553), 436-444.

[Ayer2010] Ayer, T., Ayvaci, M. U., Liu, Z. X., Alagoz, O., & Burnside, E. S. (2010). Computer-aided diagnostic models in breast cancer screening. *Imaging in medicine*, 2(3), 313.

[Kim2010] Seung Ja Kim, Woo Kyung Moon, Soo-Yeon Kim, Jung Min Chang, Sun Mi Kim & Nariya Cho (2010) Comparison of two software versions of a commercially available computer-aided detection (CAD) system for detecting breast cancer, *Acta Radiologica*, 51:5, 482-490.

[Dromain2013] Dromain, C., Boyer, B., Ferre, R., Canale, S., Delalogue, S., & Balleyguier, C. (2013). Computed-aided diagnosis (CAD) in the detection of breast cancer. *European journal of radiology*, 82(3), 417-423.

[Elter2009] Elter, M., & Horsch, A. (2009). CADx of mammographic masses and clustered microcalcifications: a review. *Medical physics*, 36(6Part1), 2052-2068.

[Shen2019] Shen, L., Margolies, L. R., Rothstein, J. H., Fluder, E., McBride, R., & Sieh, W. (2019). Deep learning to improve breast cancer detection on screening mammography. *Scientific reports*, 9(1), 1-12.

[Wu2020] Wu, N., Phang, J., Park, J., Shen, Y., Huang, Z., Zorin, M., ... & Wolfson, S. (2019). Deep neural networks improve radiologists' performance in breast cancer screening. *IEEE transactions on medical imaging*.

[McKinney2020] McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., ... & Etemadi, M. (2020). International evaluation of an AI system for breast cancer screening. *Nature*, 577(7788), 89-94.