Maskinlæringssystemer for gastrointestinale endoskopier

Av Michael Alexander Riegler, forsker, SimulaMet, Oslo

Pia Smedsrud, lege, ph.d.-student, SimulaMet og Augere Medical AS, Oslo

Thomas de Lange, lege, Oslo universitetssykehus, Oslo; førsteamanuensis, Institutt for klinisk medisin, Universitetet i Oslo, Oslo

Pål Halvorsen, forsker, SimulaMet, Oslo; professor, Oslo Metropolitan University, Oslo

Assistert diagnostikk med hjelp av kunstig intelligens (KI) har vært etterspurt lenge og kan bli et viktig hjelpemiddel innen medisin, godt hjulpet av den raske utviklingen innen maskinvare. Denne har gjort innføringen av slike hjelpemidler mulig på relativt kort sikt.

Sikrere påvisning og klassifisering av funn og lesjoner innen radiologi og endoskopi er i ferd med å bli et viktig forskningsområde innen KI, og det fokuseres spesielt på maskinlæring. Imidlertid krever vellykket utvikling et komplett system som kan brukes i sanntid i daglig praksis, og som begrenser seg til utvikling av algoritmer.

Det kreves også store randomiserte studier for å fastslå om kvaliteten og påliteligheten til systemene er god. Vi deler i denne artikkelen våre erfaringer fra utviklingen av et system for gastrointestinale endoskopier[1-5] og belyser viktige utfordringer for å skape en effektiv digital assistent.

Assistert diagnostikk ved gastrointestinale endoskopiundersøkelser

Vi har utviklet maskinlæringssystemer for flere typer applikasjoner, og innen medisin har vi blant annet fokusert på gastrointestinal (GI) endoskopi sammen med partnere på Bærum sykehus, Oslo universitetssykehus, Karolinska Institutet og Kreftregisteret. Til tross for betydelige tekniske forbedringer av endoskopiutstyret de siste 10-15 årene representerer en uønsket variasjon mellom endoskopører fremdeles en utfordring.

Dette medfører en vesentlig variasjon i påvisning og vurdering av forandringer i slimhinnen og anatomiske landemerker.[7,8] Studier har vist at rundt 20 % av polyppene i kolon overses.[9] Variasjonen skyldes blant annet forskjeller i ferdighet, perseptuelle faktorer, personlighet, erfaring og kunnskap.[6] Utvikling av et automatisert datastyrt støttesystem for påvisning og karakterisering av slimhinneforandringer kan bli et viktig bidrag til å redusere variasjonen mellom endoskopører.

«**Data is king»**

«Data is king» er et mantra innen maskinlæring og mye ressurser brukes på å samle gode data. For å lære opp modellene til å gjenkjenne forandringer må legenes kunnskap om hva som er normalt og unormalt overføres til datamaskinene. Denne kunnskapen, eller «fasiten», kalles gjerne «ground truth». Ved å samle og merke data viser man maskinen hva man ønsker at den skal lære og oppdage.

Utviklingen og treningen av modeller og algoritmer avhenger i stor grad av tilgjengeligheten og kvaliteten på dataene. Dette gjelder alle typer maskinlæring, men særlig innen feltet kalt «dyp læring», som er et mye brukt verktøy innen bildeanalyse.[10]

Maskinlæringsmodellen lærer ved å analysere tilgjengelig data, og både kvaliteten og mengden er derfor essensiell for å lykkes. Maskinen gjør som den har lært, og hvis

læringen er basert på data av dårlig kvalitet blir resultatene tilsvarende. For å sikre et korrekt datagrunnlag bør fasiten/ground truth kvalitetssjekkes av flere eksperter.[11]

Videre diskuteres det ofte hvor mye data som kreves, uten at det finnes noe universelt svar. Selv om datasettet ideelt sett bør være så stort som mulig,[12] mener andre at også mindre datasett kan brukes.[5] En tommelfingerregel kan være at man trenger 1 000 bilder per klasse med funn som skal detekteres. Det er utgangspunktet for størrelsen på klassene man for eksempel finner i Kvasir-datasettet[13] vi har laget.

Et potensielt problem i maskinlæring er det man kaller overtilpasning («overfitting»). Det høres kanskje rart ut, men maskinen kan lære noe *for* godt hvis datamaterialet som brukes for trening og testing er for likt. Dersom maskinen trenes utelukkende på åpenbare tilfeller med lite variasjon, vil den bli god på å gjenfinne nettopp slike tilfeller godt, men ikke klare å generalisere.

Dette vil føre til at andre funn blir oversett. Overtilpasning kan sammenlignes med pugging – man lærer ikke å generalisere kunnskapen. For mange like eksempler i dataene fører derfor til en type overtrening. Dette unngås ved å ha et variert datasett. Et generelt råd er derfor å samle et så mangfoldig datasett som mulig. Dette er viktigere enn størrelsen på datasettet.

For eksempel vil bilder av mange ulike polypper gi en mer generell modell enn mange forskjellige bilder av samme polypp. Variasjonen vil gjøre at algoritmen lærer seg en mer generell modell for polypper.

Maskinlæringsalgoritmer

Det finnes mange typer modeller og algoritmer som brukes innen dyp læring. Vanligvis skiller man mellom «supervised» og «unsupervised» læring. For medisin brukes ofte «supervised» læring, hvor dataene har korrekte annotasjoner som viser hva som er de korrekte funnene («ground truth»). Videre finnes det for eksempel søkebaserte algoritmer,[4] men de aller fleste bruker en type dype nevrale nettverk.[14]

Disse algoritmene bygges opp etter ulike lagdelte arkitekturer, med nettverk av nevroner i flere lag. Jo flere lag, jo dypere er det nevrale nettverket. Her eksisterer det mange ulike varianter, men for bilde- og videoanalyse brukes ofte «convolutional» nevrale nettverk (CNN) eller generative adversarielle nettverk (GAN). Slike nettverk brukes både til lokalisering av objekter i bilder (segmentering)[2] og til å bestemme hva slags objekt som er funnet (klassifisering).[3,15,16]

Det er ikke bare typen nettverk som skal velges og evalueres, men også konfigurasjonen av nettverkene. Det finnes nettverk som er forhåndstrent på mer generelle databaser med mange millioner annoterte bilder av hverdagslige gjenstander. Deler av denne læringen er generell og gjenbrukes, overført til ønsket felt, som for eksempel, medisin. Alternativt kan man trene nettverket fra bunnen av. Det tar mye lengre tid med behov for mer data enn forhåndstrente nettverk. I tillegg velger man hvor mange lag man vil ha og hva hvert enkelt lag skal gjøre. Alt dette er eksempler på det man kaller hyperparametere, og for å velge riktig må man prøve mange kombinasjoner og evaluere hva som gir best resultat.

Evaluering

Det er en utfordring å sammenligne metoder og modeller fordi det som oftes brukes forskjellige datagrunnlag, ulikt utstyr og forskjellige statistiske måleenheter, slik at en direkte sammenligning ikke er mulig. Utfordringen kan reduseres ved å tilgjengeliggjøre alle data (open access) slik at flere kan bruke samme grunnlaget for reproduserbar

sammenligning. Dessuten må man også beskrive hvordan dataene er splittet i trenings- og testsett for å gjøre analysen reproduserbare.

Vanligvis deles datasettet inn i tre mindre sett for trening, evaluering og testing, alternativt to sett bare for trening og avsluttende testing, mens kryssvalidering er den mest robuste metoden. Her deles settet for eksempel inn i n deler, hvor ideen er at en del av dataene er utelatt fra treningen og bare brukt til testing, mens de resterende n-1 delene brukes til treningen.

Slik utfører man da n eksperimenter hvor man rullerer hva som brukes til trening og hva som brukes til testing. Etter trening og testing må modellen valideres. Dette bør først gjøres på et annet datasett, og etter hvert i en klinisk setting. Først da vet man om modellen kan brukes som klinisk støtteverktøy.

Videre er det viktig at man bruker samme statistiske metoder og måleenheter når resultatene presenteres. Medisinere bruker ofte sensitivitet, spesifisitet og positiv prediktiv verdi (PPV), mens informatikere ofte bruker presisjon, «recall» (samme som sensitivitet), Matthews korrelasjonskoeffisient og F1 score.[13]

Alle disse statistiske målene bør vise mer enn 90 % nøyaktighet for å være på nivå med de beste ekspertene. Disse statistiske målingene beregnes ut i fra hvor mange sanne og falske positive og negative funn systemet returnerer, så alle resultatene kan med fordel rapporteres.

Komplette systemer

Det er viktig å innse at maskinlæringen bare er en komponent i en lang rekke av byggeklosser som må passe sammen for å skape et system for automatisk deteksjon av lesjoner. Hvis et slikt system skal være praktisk brukbart i klinikken må hele kjeden av systemer utvikles til å fungere i sanntid. Bilde og video må tas direkte fra det medisinske utstyret og mates inn i den ferdigtrente modellen der de analyseres for å påvise forandringer.

Anatomiske landemerker er også viktige å analysere for å dokumentere at undersøkelsene er komplette. Det endelige resultatet fra systemet må vises til legen på en lettfattelig og enkel måte for en siste kvalitetssjekk og vurdering om hvilke beslutninger resultatet gir for videre utredning og behandling. For at dette skal fungere må systemet tilpasses arbeidsflyten i klinikken.

Endoskopier vurderes fortløpende mens undersøkelsen gjøres. Det krever rask respons fra systemet med sanntidstilbakemelding. I tillegg til å evaluere nøyaktigheten av et system ved å undersøke andelen av forandringer som klassifiseres, må også hastigheten til systemet beregnes. Systemet må prosessere bildene minst like fort som endoskopet leverer, som vanligvis er mellom 30 og 50 bilder i sekundet. Med andre ord – hvis ikke hvert bilde kan analyseres på under 25 millisekunder, kan ikke systemet brukes i klinikken.

Konklusjon

Grunnet den raske maskinvareutviklingen er tiden inne for å introdusere maskinlæringsverktøy innen assistert medisinsk diagnostikk. Her vektlegges «assistert» – vi har liten tro på at systemet håndterer hele undersøkelsen alene. Det er imidlertid mange parametere å ta hensyn til, og det er viktig å metodisk lete seg frem til den beste løsningen.

For eksempel er høykvalitetsdata en viktig del i treningen av en algoritme, og tid og ressurser må settes av til å innhente og annotere mye og korrekte data. For å muliggjøre fullstendige sammenligninger mellom metoder, bør de samme datasettene benyttes, og så mange som mulig av de vanlige statistiske måleenhetene bør brukes.[13]

Aktuelle studier innen påvisning av forandringer, særlig polypper, viser lovende sensitivitet og spesifisitet, mens påfølgende klinisk testing har vært mindre overbevisende. Mange har sannsynligvis et «overfitting»-problem, som sannsynligvis kan løses ved større variasjon i treningsdatasettene. Det er foreløpig ikke vist at systemene reduserer variasjonen mellom endoskopørene eller andelen oversette forandringer.

Klinisk validering er helt nødvendig og det er fremdeles en vei å gå før slike systemer kan tas i bruk. Retningslinjene gitt over kan være til hjelp for å skape sammenlignbare resultater.

Interessekonflikter

Noe av arbeidet er støttet av Norges Forskningsråd (AutoCap, #282315). Forfatterne er i tillegg involvert i et medisinsk oppstartselskap (Augere Medical).

Referanser

1. de Lange T, Halvorsen P, Riegler M. Methodology to develop machine learning algorithms to improve performance in gastrointestinal endoscopy. World Journal of Gastroenterology, 2018; 24(45):5057-5062
2. Pogorelov K, Riegler M, Eskeland SL, de Lange T, Johansen D, Griwodz C, Schmidt PT, Halvorsen P. Efficient disease detection in gastrointestinal videos - global features versus neural networks. Multimed Tools Appl 2017; 76: 22493-22525
3. Pogorelov K, Ostroukhova O, Jeppsson M, Espeland H, Griwodz C, de Lange T, Johansen D, Riegler M, Halvorsen P. Deep learning and hand-crafted feature based approaches for polyp detection in medical videos. Proc. of IEEE CBMS, 2018
4. Riegler M, Pogorelov K, Halvorsen P, de Lange T, Griwodz C, Johansen D, Schmidt PT, Eskeland SL. EIR - efficient computer aided diagnosis framework for gastrointestinal endoscopies. Proc. of CBMI, 2016
5. Riegler M, Lux M, Griwodz C, Spampinato C, de Lange T, Eskeland SL, Pogorelov K, Tavanapong W, Schmidt PT, Gurin C, Johansen D, Johansen H, Halvorsen P. Multimedia and medicine: Teammates for better disease detection and survival. Proc. of ACM MM, 2016
6. Hewett DG, Kahi CJ, Rex DK. Efficacy and effectiveness of colonoscopy: how do we bridge the gap? Gastrointest Endosc Clin N Am 2010; 20: 673-684
7. Lee SH, Jang BI, Kim KO, Jeon SW, Kwon JG, Kim EY, Jung JT, Park KS, Cho KB, Kim ES, Park CG, Yang CH; DeaguGyeongbook Gastrointestinal Study Group. Endoscopic experience improves interobserver agreement in the grading of esophagitis by Los Angeles classification: conventional endoscopy and optimal band image system. Gut Liver 2014; 8: 154-159
8. van Doorn SC, Hazewinkel Y, East JE, van Leerdam ME, Rastogi A, Pellisé M, Sanduleanu-Dascalescu S, Bastiaansen BA, Fockens P, Dekker E. Polyp morphology: an interobserver evaluation for the Paris classification among international experts. Am J Gastroenterol 2015; 110: 180-187
9. Lanspa SJ, Lynch HT. Quality indicators for colonoscopy and the risk of interval cancer. N Engl J Med 2010; 363
10. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, van der Laak JAWM, van Ginneken B, Sánchez CI. A survey on deep learning in medical image analysis. Med Image Anal 2017; 42: 60-88

11. Gottlieb K, Hussain F. Voting for image scoring and assessment (VISA)--theory and application of a 2 + 1 reader algorithm to improve accuracy of imaging endpoints in clinical trials. BMC Med Imaging 2015; 15: 6
12. Chen XW, Lin X. Big data deep learning: challenges and perspectives. IEEE Access 2014; 2: 514-525
13. Pogorelov K, Randel KR, Griwodz C,Eskeland SL, de Lange T, Johansen D, Spampinato C, Dang-Nguyen DT, Lux M, Schmidt PT, Riegler M, and Halvorsen P. Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection. Proc. of ACM MMSys, 2017
14. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015; 521: 436-444
15. Shin Y, Balasingham I. Automatic polyp frame screening using patch based combined feature and dictionary learning. Comput Med Imaging Graph 2018; 69: 33-42
16. Alammari A, Islam AR, Oh J, Tavanapong W, Wong J, De Groen PC. Classification of ulcerative colitis severity in colonoscopy videos using CNN. Proc. of ACM ICIME, 2017