Proposta progetto : Position Convolution Experts

Idea

Le CNN tradizionali applicano gli stessi filtri ovunque, ma oggetti diversi potrebbero beneficiare di operatori diversi in posizioni diverse, per poter sfruttare questo possibile beneficio si potrebbe usare un routing posizionale per decidere degli esperti convoluzionali mantenendo il contesto globale.

L'idea è quella di adottare un approccio simile a quello della Dynamic convolution, solo che invece di scegliere in maniera dinamica le dimensioni dei kernel scegliamo degli esperti convoluzionali che lavoreranno non sull'intera immagine ma su delle patch con la loro relativa posizione.
All'interno di ogni livello della rete c'è un *router MLP* (con pesi codivisi) e n *esperti convoluzionali* (Conv -> batch norm -> ReLU), il router prende in input posizione e contenuto della patch in modo da poter scegliere gli esperti in base sia alla posizione che al contenuto. La scelta dell'esperto avviene tramite softmax.

Una volta ottenuti gli output degli *esperti* si effettua una *somma pesata* dove i pesi sono le probabilità dei rispettivi esperti, per esempio se il router sceglie E_1, E_5, E_{11} con le rispettive probabilità $P_{E_1} = 0.3, P_{E_5} = 0.4, P_{E_{11}} = 0.3$ allora la somma avrà come pesi 0.3, 0.4, 0.3 per le rispettive feature map degli esperti, verrà poi passata in un canale convoluzionale 1x1.

Per mantenere contesto usiamo la *Deformable attention* che prenderà come input la *somma pesata* delle feature map degli *esperti* (dopo la conv 1x1); questa nuova feature map arricchita con l'attenzione sarà ulteriormente divisa in patch, il singolo patch con la relativa posizione viene passato al router che deciderà gli esperti convoluzionali e così via.

Training

Per addestrare la rete vorrei usare MNIST, CIFAR-10, Tiny-ImageNet, Pascal VOC e Camelyon per testare la rete su diversi task.

La rete verrebbe addestrata seguendo questo schema di training:

prendiamo per esemoio il training su CIFAR-10, stabiliziamo i parametri degli esperti dandogli il dataset per una decina di epoche, senza il coinvolgimento di attention e routing, in questa fase la

patch verrebbe passata o a un sottoinsieme randomico degli esperti nel layer o a tutti, concatenando le feature map attraverso una somma senza pesi o tramite una media, una volta stabilizzati i parametri verebbero introdotti attention e routing.

Per rafforzare la rete per alcuni batch si può usare data augmentation, alzando uno dei canali RGB, patch nere sull'immagine per nascondere dettagli.

Essendo tutto differenziabile, la loss usata sarebbe CrossEntropy, in aggiunta il router verrebbe monitorato osservando l'entropia delle sue scelte che può essere anche utile per regolare le scelte, inserendola nella loss totale : $L_{tot} = CE + \lambda L_{entropy}$, si possono usare anche altre loss di "bilanciamento" in base ai "problemi" del router

Possibili risultati

Ci si aspetta un miglioramento delle performance generali, in particolare sulla Object Detection grazie al contesto e al positional routing, che scelgono quali esperti usare per la determinata patch, la rete potrebbe comprendere che un patch in cui è visibile un albero non c'è una persona, e il router potrebbe comprendere che un esperto che localizza i bordi è importante se devo riconoscere un oggetto qualsiasi.

Anche in compiti di segmentazione la specializzazione locale e la deformable attention dovrebbero aiutare la rete a trovare le parti di immagini più pertinenti.
Avendo esperti che si specializzano ci si aspetta anche una maggiore interpretabilità della rete con un analisi di una mappa semantica delle scelte del router.