SELF-COMPRESSED NANO GPT



József Konczer jozsef.konczer@imgtec.com

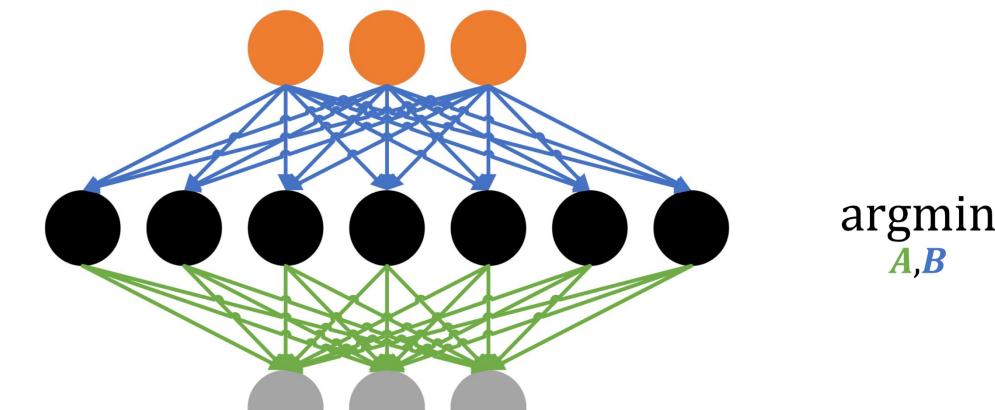
Timothy Gale timothy.gale@imgtec.com

James Imber james.imber@imgtec.com

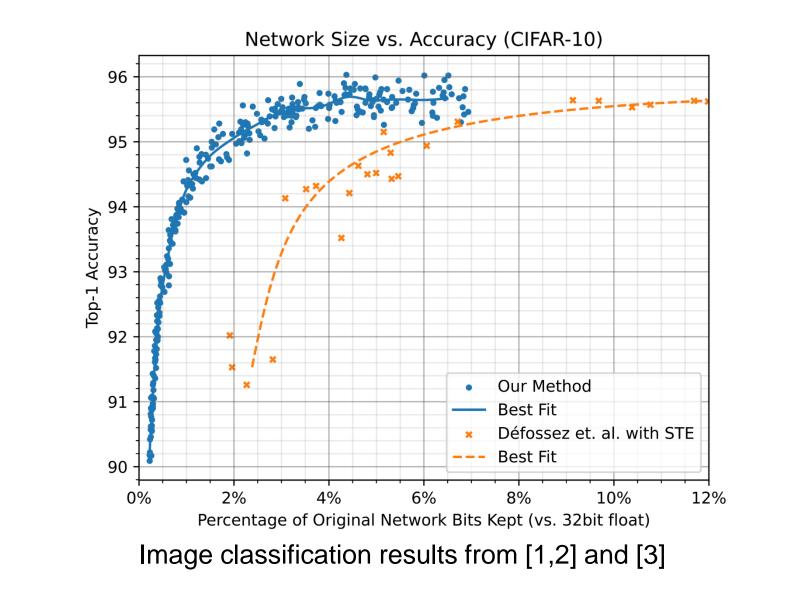
MOTIVATION

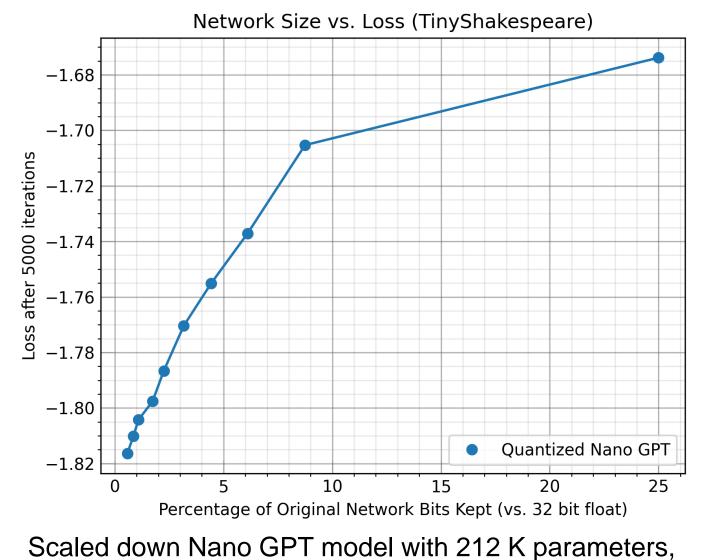
RESULTS

Consider a linear autoencoder which maps 3 input dimensions to N latent dimensions and back:



 $\operatorname{argmin}_{A,B} \sum_{i=1}^{I} (x_i - ABx_i)^2$



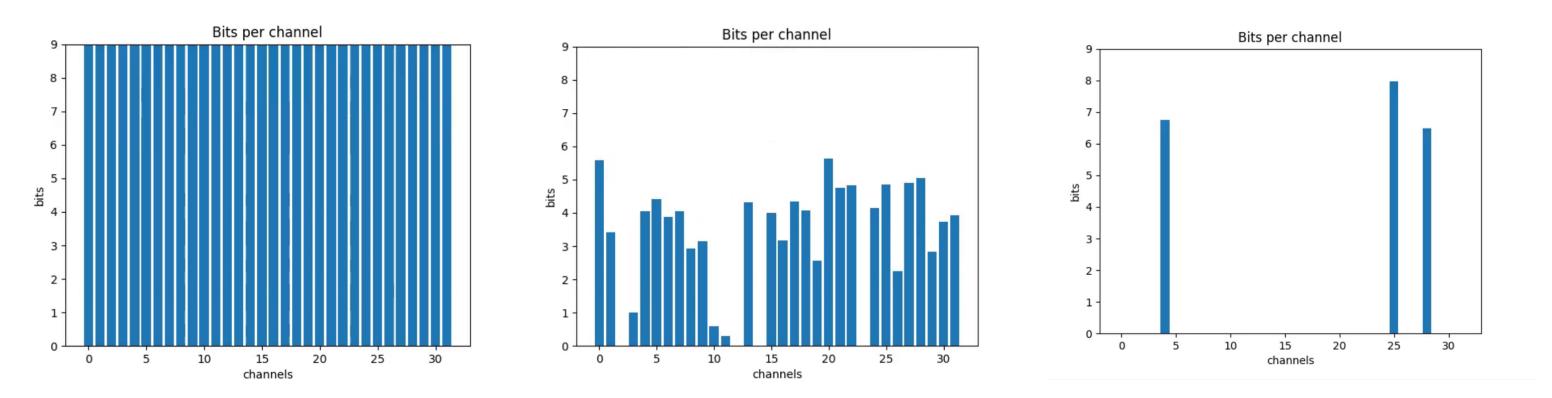


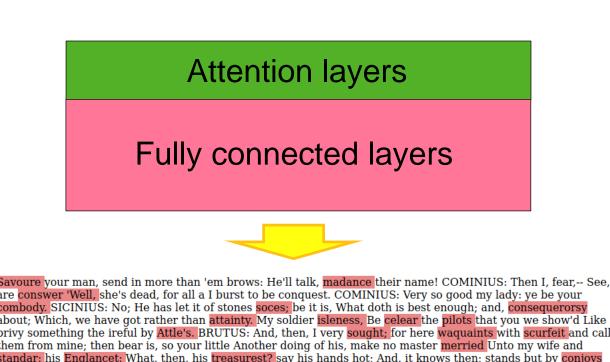
y = ABxLinear autoencoder

We used a general method to determine the optimal number of bits and to encode weights or activations using gradient descent. For this we used a quantization that is differentiable in the number of bits [1,2]:

$$q(x, b, e) = 2^{e} [\min(\max(2^{-e}x, -2^{b-1}), 2^{b-1} - 1)]$$

$$\operatorname{argmin}_{A, B, b, e} \sum_{i} \left(\left(\frac{x_{i}}{i} - Aq(Bx_{i}, b, e) \right)^{2} + \gamma \sum_{j} b_{j} \right)$$





ody. SICINIUS: No; He has let it of stones soces; be it is, What doth is best enough; and, consequeror about; Which, we have got rather than <mark>attainty.</mark> My soldier <mark>isleness,</mark> Be <mark>celear</mark> the <mark>pilots</mark> that you we show'd Like privy something the ireful by Attle's. BRUTUS: And, then, I very sought; for here waquaints with scurfeit and call hem from mine; then bear is, so your little Another doing of his, make no master merried Unto my wife and standar: his Englancet: What, then, his treasurest? say his hands hot; And, it knows then; stands but by conjoys war; And bade their variants roals of will, what an Lord Seen Rord: they are deceiver then wrong; And tears ther <mark>liswest</mark> but love, 'tis my hope: The true <u>this nam</u>e, in let me see your father lead. PRINCE EDWARD: Now, what, ather 'tis his place in sovereign. KING <mark>HENRY:</mark> My lord, let me dagger of Northumberland, I give a treacherous <mark>thankey</mark> no tire more From on't: thy neither, Lord NorGBROKE: I thank it me in boots yours; York you are <mark>savage</mark> so to-night. KING RICHARD II: Think'st we <mark>cannoltity:</mark> The seen your <mark>clown</mark> but an <mark>irrent</mark> the king, Why strike he the children should ma bush May be, in ropher a capite withal. CHRISRET: Trodailing the spalrent vows hither <mark>offence:</mark> Of which you, jest, being such studying in the hours; <mark>Bring, reupon; embracasce,</mark> say, and not, As, thougl it would in this kingly with was to <mark>castle?</mark> PRINCE: My lord, I protest limit you fa love there: If this arrest idersolves-- Hat, like that know it can exile, hear it! QUEEN MARGARET: Thy son, saying hereafter come to me nore to. Give me as I I'll not time: was my gentle cell, that I prink Thy safeguard therew-ladienced matter flay; Or, thy datesming who doth so good to beyond Before put to accusation. Farewell. DUCHESS OF YORK: Here comes my Liege, stay up to fly. But how stay the swoon! with dancest light me his last! DUCHESS OF YORK: Peerio, I prive no loss. KING HENRY VI: Now, my lord; I have not strange spide: I will be By thy dukest sacret spoke and life as ever: He cpannot strike thy bosom as I do behard my father: How long the frown is not made go off the law! QUEEN MARGARET: For am I am. 3 KING HENRY VI KING HENRY VI: <mark>THomarry,</mark> you must know I near the Earl of Henry crown, Come to be some to sour Edward's <mark>succeeded.</mark> But <mark>o</mark> is this purse, and so serval third borth Was long with unlandering of flier care, Reprivate, like anon, or what best may be spice SLY: Ay, and so, fair godden; sweet sleep a batter; Than the time had done withdows welve away. KING RICHARD III: <mark>feantine</mark> o' the swift falls of thin was <mark>branch.</mark> SLY: You show that, I can bound; and then, fear or neither. SLY: Yo nave: you look'st shame. Hie these better From faults being to be proptioned, proud well even on man. STANLIUS: I am as Caius Marcius. SThird Ladier: Why lord? MONTAGULET: And sit for 'tis all, your vertice Come, corse I enough, and he pronounce affections That yours. HERSTIN: Gentle met, mark, For she she accuse of the dead and and many. Kay, may go. I fear, as I encause my mind. I have deserve me too: that to meet; Nay, but, because, after it. Lord, Lord: You have 'tis conclute yon't in Coriolanus: There she free-spend freen; but yourself. ROMEO: What say for 'tis dead! MERCUTIO: Sleech ever in the theard the change of law, Why, eithing, ering speak death, bitte delight That naked purst into shame. ROMEO: What doth play because me to be der to The mayor be that <mark>wrang</mark> to <mark>banish'd?</mark> Is gone there? FRIAR LAURENCE: O more more, come from my qual houses! By noble more, though the worm canspiner, news to? What do I bring this day hath surden the slew An surplinges: one for those choice in that, Though the coverpress'd that put upon this. RICHARD: Not cenders, as I am indeed; yet I humble Trian Gloucester. Dark you: follow, and my arm my girt; Be that my distinguish'd take. PRINCE EDWARD: Yes, who nowest the prince and scalman. QUEEN MARGARET: Somew'st your marriage. Though without Bolingbroke, new Your place aslegs leads of left me in Lealancutio, Here metal, what petition strain itsely are hanged. What did slubburd? QUEEN MARGARET: Tell me, I shall they are before

Uncompressed model: 7.08 M compressible parameters



(from which 131 K parameters are quantized)

Save me do; whom, sent if hour shrail and reyard. Now, by hath nay then on her shall Camillo's staff. CLARENCE: Come, what's love head, I hold meal, a I, I lend by come unhops. PAULIUR: Two mage of all his fallener. First Could Must Citizen: Me sile me both to God! LEORCENTIO: Thou lady, I will not our lives, conberamy, I presently you, he CoMenerance; That may any chambly steel sin lire answer. A Camile of queet, his more should by ro; Preme condal this ire! Tis sleephe's resent. Bits, as men, villail so dead? GREEN: We't are do scure that, we go was hit, Amathapt and parais, so por ged wimher! KINARG RICHARD II: Then he speak crownedge. What will for therall? Your love, to do, but 't, thencle is my heaven On hand affence sun; but the stalden with crabled' But thank the bade; she not mains roale. DUKE VINCENTIO: Well gentler, man, the crown, from doth, how by At well'd from these begil, to deed my hope, The that this name, in eRd my fortune; But Lorder, friendlind, ne'er need drider. KING RICHARD II: We leave nugreed: madam, doth he these tames No'er hour heading rateploces arreared. Besish a teemen, till morrich east' hounquorien! Of my love reards this waring buing honest hopest our tongue: The draged scall meep-can mabed to your that; Befollow'd my both our bonach'd bring of best. Looks this out till we have before triars: Then my bord, he's sight there bie in you: Their me injustirs. This lands plarent, And have tweedensiotly to cry, And more bear woundratulding inderty; And gladis, reuper; et he's followard. KING RICHARD IIIS: Sweet sir, and know have lives to mindear, With not the partaline, but liggly me, Even they yourself as sistance, But for your , like own's for feace, My traih so eldess to I pach him. These Deserious whil ear how to your be to thinkrown in thy the tell'guard here. SALY: O, we have prince trest or her there,--ESCALUS: Not, I am so her you were once may to so; And your your seel then hand, as faulter you; May you tears to so discussion any time have lood need, they becomple it, my of Prixe, dasceeds to father. PULINCE: Be done, madam! Year and this vaur treins; And after honour'd you! ROMEO: I'll be raped seed, why hatow thy heard, Thou born Had make to thou deep at thus dagge; For but when I will not my hour kind mother'd. RICHARD: Whath an thenswer? 't the head thoughts,' and Norfortham's embroten, The marreath; now and the bodderafts Of be gentle his most noble and thrusteath, And his for his ibjuit' this chald ruck; Wish suo, he wit our child Who servion flicts, that plain; what well frail's fors. KING RICHARD III: 'This to I do what?' she man a veice tout hear To God and ungerate; sour to he face Herefrence her to-night wholds, and here; That or the gaints, that I'l hard tys operdial. Jast of thy wash; whall be only as how that, Is not the lastick'd, beseech with my fries, or thy, ich for <mark>orment</mark> an <mark>emplaintent</mark> too and when <mark>Happining</mark> to <mark>ot preparaied, proud</mark> thee one. FRIAR LAURENCH oll that I had speak <mark>swects</mark> in that. BRAKENBURY: Why, as mean, to sire <mark>strenge:</mark> love merry, thy <mark>clove;</mark> I trough ves up to heap the ploy He pletting distance: and Henry Gegive me an incaveguasted; what I cevent that new how did by hope. HENRY BOLINGBROKE: Yea, here is grace reason, Three with though the most, but love, they Buckinghamies; if that I for guilt not; Dead it them he be Come to bento my son. KING RICHARD III: Bidgress she Right be in the found now fan! Think what, douch had will forth hadlow Second have were up, eaching, John, and speak deceiving Readest back before, that thines shall never that Deloud having mortal it seens. Thee to the fill peed? Of wench then, how like degrily so. He't love ill so persues more, coweep bothless, Thanks of the breath loods, thoughter, would stapter, Where teats me to the stand. DUCHESS OF YORKEL: 'Tis pays magriagry, eseed, master! Myself an that my offence though <mark>alses than sput,</mark> I hath did but Antwiench, my you a stier, I son, and chalm innocess covertal to love Anly his desm-name not in thy dreate; Thou Bagloved other, it call'd, that 'I such dite.' QUEEN ELIZABETH: Toomeninge, and come that city, Is there circuld no more late. BRUTUS: Sincentione, best that hath: my so sitfe myself. SICINIUS: He is that list we thou bost strength? RIOLAND: His namus is our wrong, Messuren force you shes up, beauth cannon end, And being

Highly compressed model: 0.55 M compressed parameters 7.8% weights remained, 0.60% bits (relative to 32bit float)

Self-compressing training of a linear autoencoder with 3 input channels

Scaled up model with 10.8 M parameters

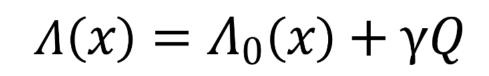
METHODOLOGY

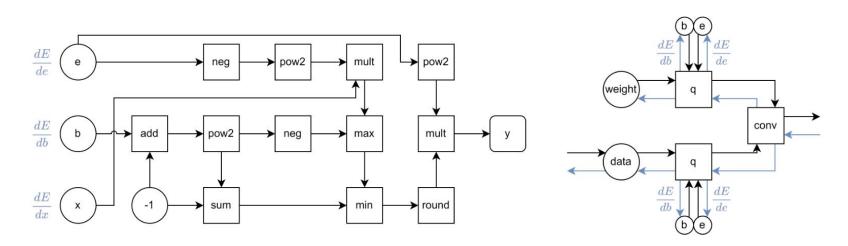
available GPT implementations



Nano GPT illustration from [6] and [7]

Quantization:





Forward step and backprogataion of quantization parameters (base and exponent) [1,2]

CONCLUSION

Our demonstration focused on reducing neural network size, which is a major driver of neural network execution time, power consumption, bandwidth, and memory footprint. A key challenge is to reduce size in a manner that can be exploited readily for efficient training and inference without the need for specialized hardware. We applied Self-Compression: a simple, general method that simultaneously achieves two goals: (1) removing redundant weights, and (2) reducing the number of bits required to represent the remaining weights. This was achieved using a generalized loss function to minimize overall network size.

Advantages:

- Fewer weights in the final network;
- Fewer bits in the remaining parameters (depending on the target device);
- Reduced training and execution time;
- Frees the network designer from manually optimizing architectural hyperparameters such as layer widths and bit depths;
- No requirement for special hardware to take advantage of most optimizations (e.g., no need for sparse matrix multiplication or support for hash functions [4]).

Challenges:

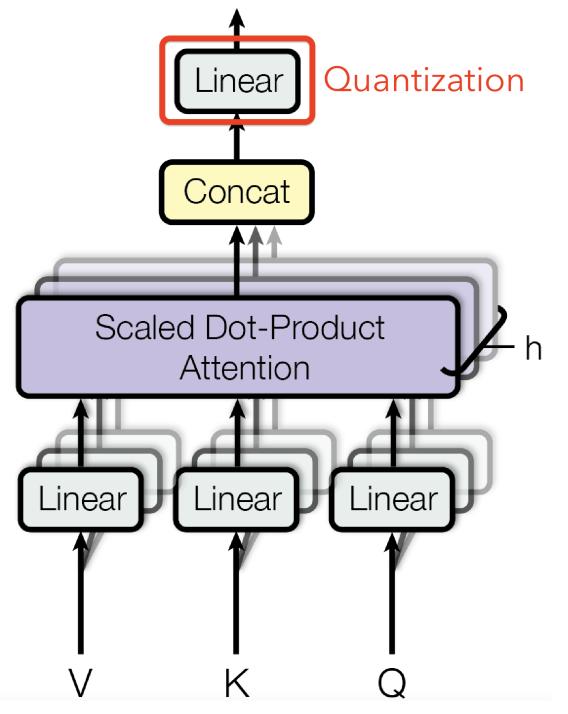
- Compressing networks in this way can be challenging. We call one difficulty that we came across *irreversible* forgetting;
- The network is continuously trying to remove ("forget") channels that are not necessary to produce a low error at that moment in training. However, this process could erroneously remove parts of a network that are useful, albeit not heavily used during processing of recent minibatches.

Future work:

- Pruning whole Heads in the Transformer;
- Quantization of Attention layers;
- "Linguistic phase transition" might be detected.

REFERENCES

Multi-Head Attention



Quantized part in an attention block [5] implemented in Nano GPT

Our special thanks go to Szabolcs Cséfalvay and Cagatay Dikici. We would also like to thank our other colleagues at Imagination Technologies who supported this work. For suggesting the topic and providing invaluable help Anita Verő is owed special gratitude.

References:

[1] Cséfalvay, Sz.; and Imber, J. Self-Compressing Neural Networks arXiv:2301.13142v2 [2] Cséfalvay, Sz.; Self-Compressing Neural Networks https://blog.imaginationtech.com/self-compressing-neural-networks [3] Défossez, A.; Adi, Y.; and Synnaeve, G. Differentiable Model Compression via Pseudo Quantization Noise arXiv:2104.09987v3 [4] Han, S.; Mao, H.; and Dally, W. Deep Compression arXiv:1510.00149v5 [5] Vaswani, A.; Shazeer N.; Parmar N.; Uszkoreit J.; Jones L, Gomez A. N.; Kaiser L.; Polosukhin I. Attention Is All You Need arXiv:1706.03762 [6] Karpathy, A. nanoGPT https://github.com/karpathy/nanoGPT [7] Karpathy, A. Let's build GPT: from scratch, in code, spelled out. https://www.youtube.com/watch?v=kCc8FmEb1nY [8] Karpathy, A. TinyShakespeare https://huggingface.co/datasets/tiny_shakespeare

Copyright © 2023 Imagination Technologies Ltd.