Οικονομικό Πανεπιστήμιο Αθηνών, Τμήμα Πληροφορικής
Μάθημα: Αλληλεπίδραση Ανθρώπου – Υπολογιστή, 2024–25
Διδάσκων: Ι. Ανδρουτσόπουλος

Ασκήσεις μελέτης της ενότητας Β7 (υπολογιστική όραση με συνελικτικά νευρωνικά δίκτυα)

1. Θέλουμε να χρησιμοποιήσουμε μια τροποποιημένη μορφή του συνελικτικού νευρωνικού δικτύου της διαφάνειας 23 (LeNet), για να εντοπίζουμε τις συντεταγμένες (x, y) του κέντρου του κεφαλιού και των κέντρων των καρπών των δύο χεριών σε εικόνες (ή video frames) που περιλαμβάνουν έναν μόνο άνθρωπο μπροστά από μια κονσόλα ηλεκτρονικών παιχνιδιών εφοδιασμένη με έγχρωμη κάμερα και κάμερα βάθους. Η κάθε εικόνα έχει ανάλυση 256x256 και τέσσερα κανάλια (RGB και βάθος), δηλαδή είναι ένας τανυστής (tensor) τριών αξόνων, με σχήμα (shape) (256, 256, 4). Όπως στο σχήμα της διαφάνειας 23, υπάρχουν δύο συνελικτικά στρώματα (convolutional layers) που παράγουν 6 και 16 χάρτες χαρακτηριστικών (feature maps) αντίστοιχα αλλά οι συνελίξεις χρησιμοποιούν πυρήνες (kernels) με παράθυρο 3x3 και είναι ευρείες (wide, same), δηλαδή χρησιμοποιούν padding και διατηρούν την ανάλυση της αρχικής εικόνας σε κάθε κανάλι (βλ. και διαφάνεια 9). Τα δύο στρώματα υπο-δειγματοληψίας (pooling) χρησιμοποιούν max-pooling με παράθυρο 4x4 και βήμα (stride) 4 και στους δύο άξονες. Τα δύο πρώτα (τα κρυφά) πυκνά (dense) στρώματα του τελικού MLP εξακολουθούν να έχουν 120 και 84 νευρώνες αντίστοιχα.

α) Πόσους πυρήνες θα χρησιμοποιεί το πρώτο συνελικτικό στρώμα και τι σχήμα θα έχει ο καθένας;

Απάντηση: Το πρώτο συνελικτικό στρώμα θα χρησιμοποιεί 6 πυρήνες, ώστε να προκύπτουν 6 χάρτες χαρακτηριστικών, όπως φαίνεται στο σχήμα της διαφάνειας 23. Ο κάθε πυρήνας θα έχει 4 φέτες (slices), αφού η είσοδος έχει τώρα 4 κανάλια. Γνωρίζουμε από την εκφώνηση ότι κάθε πυρήνας εφαρμόζει σε κάθε κανάλι της εισόδου παράθυρο 3x3. Επομένως κάθε ένας από τους 6 πυρήνες θα είναι ένας τανυστής (tensor) τριών αξόνων, με σχήμα (shape) (3, 3, 4).

β) Τι ανάλυση θα έχουν οι χάρτες χαρακτηριστικών που θα προκύπτουν από το πρώτο στρώμα max-pooling;

Απάντηση: Αφού τα συνελικτικά στρώματα που χρησιμοποιούμε διατηρούν την ανάλυση σε κάθε κανάλι, κάθε ένας από τους 6 χάρτες χαρακτηριστικών (κανάλια) που προκύπτουν από το πρώτο συνελικτικό στρώμα, δηλαδή κάθε κανάλι στην είσοδο του πρώτου στρώματος max-pooling θα εξακολουθεί να έχει ανάλυση 256x256. Αφού κάθε στρώμα max-pooling χρησιμοποιεί παράθυρο 4x4 με βήμα (stride) 4 και στους δύο άξονες, ο κάθε ένας από τους 6 χάρτες που εξέρχονται από το πρώτο στρώμα max-pooling θα έχει ανάλυση (256/4)x(256/4), δηλαδή 64x64.

γ) Πόσους πυρήνες θα χρησιμοποιεί το δεύτερο συνελικτικό στρώμα και τι σχήμα θα έχει ο καθένας;

Απάντηση: Το δεύτερο συνελικτικό στρώμα θα χρησιμοποιεί 16 πυρήνες, ώστε να προκύπτουν 16 χάρτες χαρακτηριστικών, όπως φαίνεται στο σχήμα της διαφάνειας 23. Ο κάθε πυρήνας θα έχει 6 φέτες (slices), αφού η είσοδος του συνελικτικού στρώματος (η έξοδος του πρώτου στρώματος max-pooling) έχει 6 κανάλια (χάρτες). Γνωρίζουμε από την εκφώνηση ότι κάθε πυρήνας εφαρμόζει σε κάθε κανάλι της εισόδου του παράθυρο 3x3. Επομένως κάθε ένας από τους 16 πυρήνες θα είναι ένας τανυστής (tensor) τριών αξόνων, με σχήμα (shape) (3, 3, 6).

δ) Τι ανάλυση θα έχουν οι χάρτες χαρακτηριστικών που θα προκύπτουν από το δεύτερο στρώμα max-pooling;

Απάντηση: Αφού τα συνελικτικά στρώματα που χρησιμοποιούμε διατηρούν την ανάλυση σε κάθε κανάλι, κάθε ένας από τους 16 χάρτες χαρακτηριστικών (κανάλια) που προκύπτουν από το δεύτερο συνελικτικό στρώμα, δηλαδή κάθε κανάλι στην είσοδο του δεύτερου στρώματος max-pooling θα εξακολουθεί να έχει ανάλυση 64x64 (όπως στην έξοδο του πρώτου στρώματος max-pooling). Αφού κάθε στρώμα max-pooling χρησιμοποιεί παράθυρο 4x4 με βήμα (stride) 4 και στους δύο άξονες, ο κάθε ένας από τους 16 χάρτες που εξέρχονται από το δεύτερο στρώμα max-pooling θα έχει ανάλυση (64/4)x(64/4), δηλαδή 16x16.

ε) Πόσους νευρώνες θα έχει η είσοδος του τελικού MLP;

Απάντηση: Οι 16 χάρτες ανάλυσης 16x16 που εξέρχονται από το δεύτερο στρώμα max-pooling θα συνενώνονται σε ένα διάνυσμα 16x16x16 = 4096 χαρακτηριστικών, που θα δίνεται ως είσοδος στο τελικό MLP (τρία πυκνά στρώματα, dense layers) του σχήματος της διαφάνειας 23.

στ) Πόσους νευρώνες θα έχει το τελικό στρώμα εξόδου του MLP; Τι συνάρτηση ενεργοποίησης θα έχουν;

Απάντηση: Το στρώμα εξόδου του MLP θα έχει 6 νευρώνες, δύο για τις συντεταγμένες (x, y) του κέντρου του κεφαλιού και τέσσερις για τις συντεταγμένες των δύο κέντρων των καρπών. Οι νευρώνες αυτοί δεν θα έχουν συνάρτηση ενεργοποίησης, ώστε να μπορούν να παράγουν οποιονδήποτε πραγματικό αριθμό ο καθένας.

2. Μια εταιρεία κατασκευής οικιακών συσκευών ετοιμάζει έναν νέο τύπο (μοντέλο) φούρνου μικροκυμάτων που θα διαθέτει κάμερα. Η εταιρεία θέλει ο φούρνος να έχει τη δυνατότητα να αναγνωρίζει μέσω της κάμερας τον χρήστη που στέκεται μπροστά του, ώστε να προσαρμόζονται οι ρυθμίσεις του φούρνου στις προτιμήσεις του συγκεκριμένου χρήστη. Η εταιρεία σχεδιάζει να χρησιμοποιήσει ένα συνελικτικό νευρωνικό δίκτυο (CNN), το οποίο θα τροφοδοτείται με μια φωτογραφία του χρήστη που στέκεται μπροστά στη συσκευή. Το CNN θα έχει 10 νευρώνες εξόδου, γιατί η εταιρεία θεωρεί ότι κάθε συσκευή του συγκεκριμένου τύπου θα χρησιμοποιείται σε ένα σπίτι ή γραφείο όπου οι χρήστες θα είναι το πολύ δέκα. Η εταιρεία διαθέτει 1.000 φωτογραφίες 50 ενδεικτικών χρηστών (20 από κάθε ενδεικτικό χρήστη) που έχουν τραβηχτεί με την κάμερα του νέου φούρνου. Κάθε μία από τις 1.000 φωτογραφίες είναι επισημειωμένη με τον κωδικό (id, 1–50) του αντίστοιχου ενδεικτικού χρήστη. Αλλά η εταιρεία δεν διαθέτει εκ των προτέρων φωτογραφίες όλων των χρηστών (σε κάθε σπίτι, γραφείο) που θα χρησιμοποιήσουν την κάθε μία συσκευή του συγκεκριμένου νέου τύπου. Όταν μία συσκευή του συγκεκριμένου τύπου εγκαθίσταται σε ένα σπίτι ή γραφείο, θα ζητείται από κάθε έναν από τους (το πολύ 10) χρήστες της να τραβήξει 5-10 φωτογραφίες του με την κάμερα της συσκευής, χρησιμοποιώντας ειδική επιλογή της διεπαφής χρήστη. Εξηγήστε πώς θα μπορούσε η εταιρεία να χρησιμοποιήσει τις 1.000 φωτογραφίες ενδεικτικών χρηστών που διαθέτει, καθώς και μια γενική συλλογή εκατομμυρίων επισημειωμένων εικόνων (π.χ. εικόνες ζώων, τοπίων κ.λπ., όπως στο ImageNet), ώστε να προ-εκπαιδεύσει (από το εργοστάσιο) το CNN του νέου τύπου φούρνου και να καταφέρει η κάθε συσκευή του νέου τύπου να αναγνωρίζει (με ελάχιστη πρόσθετη εκπαίδευση) τους συγκεκριμένους χρήστες της (σε συγκεκριμένο σπίτι ή γραφείο) έχοντας στη διάθεσή της μόνο 5–10 φωτογραφίες του καθενός.

Απάντηση: Η εταιρεία θα μπορούσε να χρησιμοποιήσει έναν κωδικοποιητή CNN προ-εκπαιδευμένο στη συλλογή των εκατομμυρίων επισημειωμένων εικόνων (π.χ. προ-εκπαιδευμένο στο ImageNet). Από τον προ-εκπαιδευμένο κωδικοποιητή, θα κρατούσε μόνο τα συνελικτικά επίπεδα (μαζί με τα επίπεδα υπο-δειγματοληψίας max-pooling), όπως στις διαφάνειες 25–26. Πάνω από αυτά θα πρόσθετε ένα MLP με 50 νευρώνες εξόδου (έναν νευρώνα εξόδου για κάθε χρήστη του συνόλου των 1.000 φωτογραφιών ενδεικτικών χρηστών, με softmax συνάρτηση ενεργοποίησης στο επίπεδο εξόδου). Θα εκπαίδευε (fine-

tuning) το συνολικό σύστημα στις 1.000 φωτογραφίες ενδεικτικών χρηστών, εφαρμόζοντας και επαύξηση δεδομένων (data augmentation, διαφάνεια 27), ξεπαγώνοντας σταδιακά τα τελευταία συνελικτικά επίπεδα (όπως στη διαφάνεια 26), ώστε να προσαρμοστούν στο πρόβλημα της αναγνώρισης προσώπων από εικόνες της κάμερας του φούρνου. Κατόπιν θα αντικαθιστούσε το MLP με ένα νέο MLP με 10 μόνο νευρώνες εξόδου (έναν για κάθε πιθανό χρήστη ενός συγκεκριμένου σπιτιού ή γραφείου, πάλι με softmax στο επίπεδο εξόδου), χωρίς να εκπαιδεύσει το νέο MLP. Κατά την εγκατάσταση του φούρνου σε ένα νέο σπίτι ή γραφείο, το σύστημα με το νέο MLP (συνελικτικά επίπεδα, επίπεδα δειγματοληψίας, νέο MLP) θα εκπαιδευόταν (πρόσθετο fine-tuning) με τις φωτογραφίες των χρηστών του συγκεκριμένου σπιτιού ή γραφείου (5–10 φωτογραφίες για τον καθένα), εφαρμόζοντας πάλι και επαύξηση δεδομένων. Στο τελευταίο αυτό στάδιο εκπαίδευσης, ενδέχεται να ήταν προτιμότερο να κρατηθούν παγωμένα (αμετάβλητα) τα συνελικτικά επίπεδα, λόγω των σχετικά λίγων δεδομένων (φωτογραφιών) εκπαίδευσης που θα είχαμε ανά σπίτι ή γραφείο. Θα μπορούσε, όμως, η εταιρεία να διερευνήσει και την περίπτωση να ξεπαγώνει πάλι τα τελευταία συνελικτικά επίπεδα. Καλό θα ήταν, επίσης, να χρησιμοποιηθούν σε όλα τα στάδια εκπαίδευσης (και στην προ-εκπαίδευση) residual connections, dropout, batch normalization (βλ. προηγούμενες ενότητες).

3. Πώς θα μπορούσε να χρησιμοποιηθεί το νευρωνικό δίκτυο Faster R-CNN (διαφάνειες 29–31) στο πρόβλημα της άσκησης 1 (εντοπισμός κέντρου κεφαλιού και κέντρων καρπών);

Απάντηση: Στην περίπτωση αυτή θα είχαμε τρεις τύπους αντικειμένων (object types): κεφάλι, αριστερός καρπός, δεξιός καρπός. Το Region Proposal Network (RPN, διαφάνειες 29–30) θα παρήγαγε «προτάσεις» (regions of interest), δηλαδή κουτιά (bounding boxes) της εικόνας εισόδου, που θα οριοθετούσαν το καθένα με μεγάλη βεβαιότητα ένα αντικείμενο οποιουδήποτε από τους τρεις τύπους (κεφάλι, αριστερός, καρπός, δεξιός καρπός). Ο ταξινομητής (CNN-based classifier, διαφάνεια 31), θα ταξινομούσε το κάθε προτεινόμενο κουτί σε μία από τις τρεις κατηγορίες (τύπους αντικειμένων).

4. (**Προαιρετική**) Εξερευνήστε πώς θα μπορούσατε να χρησιμοποιήσετε συνελικτικά νευρωνικά δίκτυα στην εργασία του μαθήματος (κυρίως στη διεπαφή κινητού), υλοποιώντας τα σε PyTorch (καλύπτεται στα εργαστήρια/φροντιστήρια), ώστε η διεπαφή χρήστη που αναπτύσσετε να αντιλαμβάνεται π.χ. αν βρίσκεται μπροστά της ένας χρήστης (ή όχι) ή/και να καταλαβαίνει ποιος συγκεκριμένος χρήστης βρίσκεται μπροστά της. Προσπαθήστε να χρησιμοποιήσετε έναν ήδη προ-εκπαιδευμένο CNN κωδικοποιητή εικόνων (διαφάνειες 25–26, το PyTorch κάνει εύκολη τη χρήση προ-εκπαιδευμένων κωδικοποιητών), περαιτέρω εκπαίδευση με φωτογραφίες συγκεκριμένων χρηστών (όπως στην άσκηση 2, καλύτερα και φωτογραφίες χωρίς χρήστη μπροστά στη συσκευή), καθώς και επαύξηση δεδομένων (data augmentation, διαφάνεια 27). Αν είστε ιδιαίτερα φιλόδοξοι, σκεφτείτε επίσης μήπως θα ήταν χρήσιμο στην εργασία σας ένα μοντέλο αναγνώρισης αντικειμένων (object detection, διαφάνειες 28–31) ή image captioning (διαφάνειες 32–33). Δείτε π.χ. την ενότητα «Computer Vision» του ιστοτόπου «Papers with Code» (https://paperswithcode.com/sota) και τις ενότητες (tasks) «Image Classification», «Object Detection» και «Image-to-Text» του Hugging Face (https://huggingface.co/).