Βελτιστοποίηση πολλαπλασιασμού αραιού πίνακα με διάνυσμα σε επεξεργαστές γραφικών με τη χρήση Συνελικτικών Νευρωνικών Δικτύων

Διπλωματική Εργασία

Αναστασιάδης Πέτρος

Υπεύθυνος καθηγητής : Γεώργιος Γκούμας
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Εθνικό Μετσόβιο Πολυτεχνείο

Ιούλιος, 2018

Δομή παρουσίασης

Αραιοί Πίνακες

Ορισμός αραιού πίνακα

- Κάθε πινάκας που έχει πολύ μεγάλο αριθμό μηδενικών στοιχείων συγκριτικά με τις διαστάσεις του
- Αρκεί το ποσοστό αυτό να μπορεί να αξιοποιηθεί

Αραιοί Πίνακες

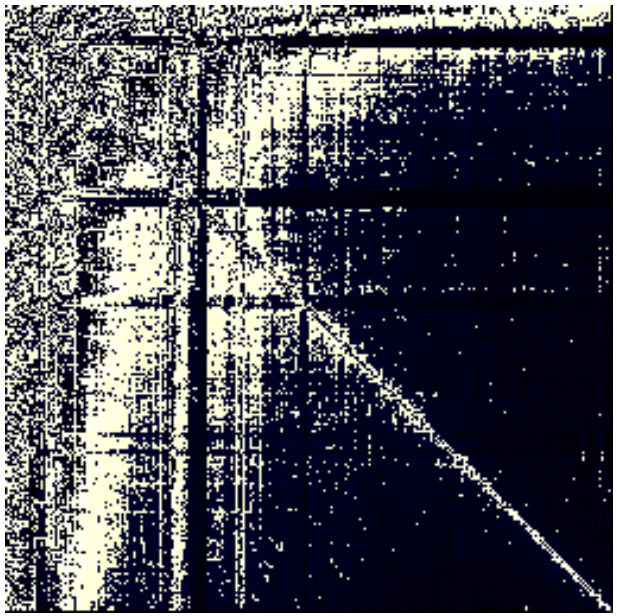
Ορισμός αραιού πίνακα

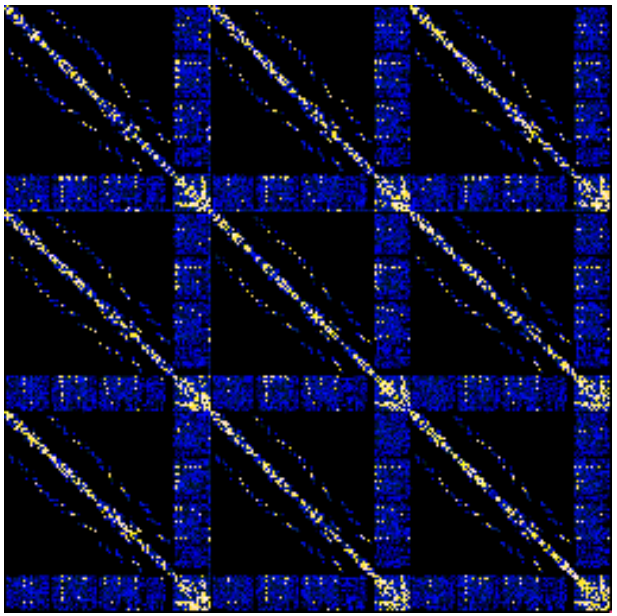
- Κάθε πινάκας που έχει πολύ μεγάλο αριθμό μηδενικών στοιχείων συγκριτικά με τις διαστάσεις του
- Αρκεί το ποσοστό αυτό να μπορεί να αξιοποιηθεί

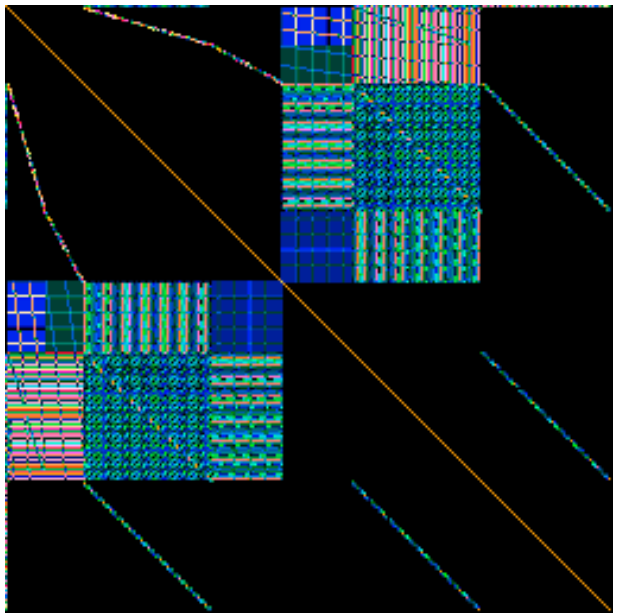
Μερικές Εφαρμογές

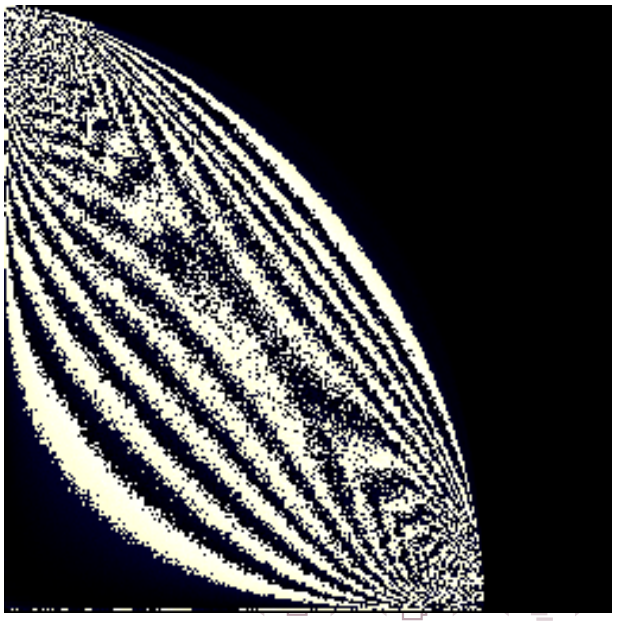
- Εφαρμογές Γράφων
- Προσομοίωση ηλεκτρικών κυκλωμάτων
- Προσομοίωση χημικών αντιδράσεων
- 2D η 3D γεωμετρικές απεικονίσεις επιστημονικών προβλημάτων

Μερικά παραδείγματα









Πολλαπλασιασμός αραιού πίνακα με διάνυσμα (SpMV)

- Η πράξη $y = A \cdot x$ όπου:
  - x είναι το διάνυσμα εισόδου
  - y είναι το διάνυσμα εξόδου/αποτέλεσμα
  - Α είναι ο πίνακας εισόδου, ο οποίος είναι **αραιός**

Πολλαπλασιασμός αραιού πίνακα με διάνυσμα (SpMV)

- Η πράξη $y = A \cdot x$ όπου:
  - x είναι το διάνυσμα εισόδου
  - y είναι το διάνυσμα εξόδου/αποτέλεσμα
  - A είναι ο πίνακας εισόδου, ο οποίος είναι **αραιός**
- Συχνά κομμάτι πολύ υπολογιστικά απαιτητικών εφαρμογών
- Μεγάλο ποσοστό του συνολικού χρόνου εκτέλεσης
- Αποθήκευση ολόκληρου πίνακα A τεράστια σπατάλη μνήμης

Δομές αποθήκευσης αραιών πινάκων

Ιδέα

- Αποθήκευση **μόνο** των μη-μηδενικών στοιχείων στη μνήμη
  - Επιπλέον αναγκαία πληροφορία οι θέσεις τους

Δομές αποθήκευσης αραιών πινάκων

Ιδέα

- Αποθήκευση **μόνο** των μη-μηδενικών στοιχείων στη μνήμη
  - Επιπλέον αναγκαία πληροφορία οι θέσεις τους
- Απαιτούμενη μνήμη συναρτήσει μη-μηδενικών στοιχείων

Δομές αποθήκευσης αραιών πινάκων

Ιδέα

- Αποθήκευση **μόνο** των μη-μηδενικών στοιχείων στη μνήμη
  - Επιπλέον αναγκαία πληροφορία οι θέσεις τους
- Απαιτούμενη μνήμη συναρτήσει μη-μηδενικών στοιχείων

Δομές αποθήκευσης αραιών πινάκων

Ιδέα

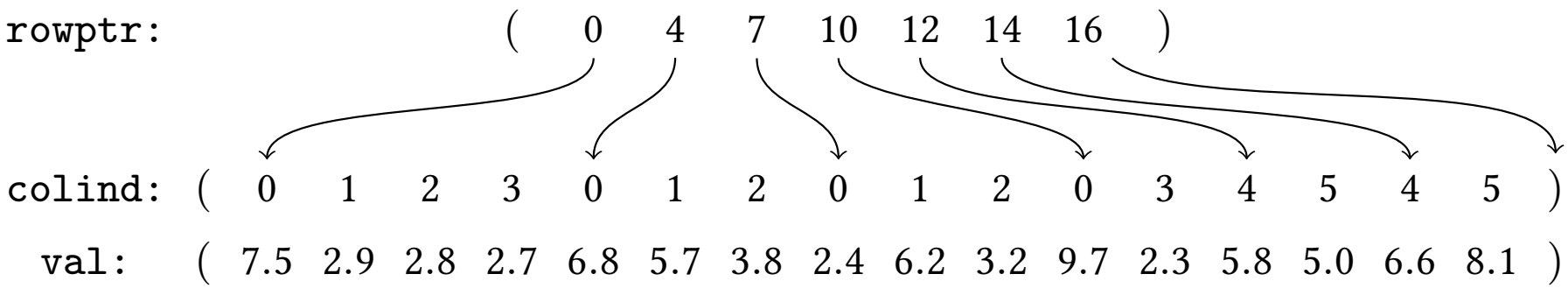
- Αποθήκευση **μόνο** των μη-μηδενικών στοιχείων στη μνήμη
  - Επιπλέον αναγκαία πληροφορία οι θέσεις τους
- Απαιτούμενη μνήμη συναρτήσει μη-μηδενικών στοιχείων

Βασικές Δομές

- Δομή Συντεταγμένων (COO)
- Δομή Συμπιεσμένης Γραμμής (CSR)
- Δομή Συμπίεσης ELLPACK (ELL)
- Τετραγωνική Δομή Συμπίεσης (BSR)
- Διαγώνια Δομή (DIA)
- Υβριδική Δομή (HYB)

Παράδειγμα: Δομή CSR

$$A = \begin{pmatrix} 7.5 & 2.9 & 2.8 & 2.7 & 0 & 0 \\ 6.8 & 5.7 & 3.8 & 0 & 0 & 0 \\ 2.4 & 6.2 & 3.2 & 0 & 0 & 0 \\ 9.7 & 0 & 0 & 2.3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 5.8 & 5.0 \\ 0 & 0 & 0 & 0 & 6.6 & 8.1 \end{pmatrix}$$



Διαφορές αραιών δομών

Γιατί όμως υπάρχουν τόσες διαφορετικές δομές;

Διαφορές αραιών δομών

Γιατί όμως υπάρχουν τόσες διαφορετικές δομές;

- Κάθε δομή έχει διαφορετικό στόχο.

Διαφορές αραιών δομών

Γιατί όμως υπάρχουν τόσες διαφορετικές δομές;

- Κάθε δομή έχει διαφορετικό στόχο.
  - Ελαχιστοποίηση μνήμης
  - Βελτιστοποίηση σε συγκεκριμένους πίνακες
  - Ικανοποιητική επίδοση ανεξαρτήτως εισόδου

Διαφορές αραιών δομών

Γιατί όμως υπάρχουν τόσες διαφορετικές δομές;

- Κάθε δομή έχει διαφορετικό στόχο.
  - Ελαχιστοποίηση μνήμης
  - Βελτιστοποίηση σε συγκεκριμένους πίνακες
  - Ικανοποιητική επίδοση ανεξαρτήτως εισόδου

Ποιά είναι όμως η **καλύτερη** για μια δεδομένη είσοδο;

Διαφορές αραιών δομών

Γιατί όμως υπάρχουν τόσες διαφορετικές δομές;

- Κάθε δομή έχει διαφορετικό στόχο.
  - Ελαχιστοποίηση μνήμης
  - Βελτιστοποίηση σε συγκεκριμένους πίνακες
  - Ικανοποιητική επίδοση ανεξαρτήτως εισόδου

Ποιά είναι όμως η καλύτερη για μια δεδομένη είσοδο;

- Η απάντηση δεν είναι καθόλου απλή, καθώς εξαρτάται από:
  - Την αρχιτεκτονική εκτέλεσης
  - Τα χαρακτηριστικά της εισόδου
  - Τον αριθμό των συνεχόμενων εκτελέσεων

Πρόβλεψη βέλτιστης δομής με βάση την είσοδο

Σκοπός

- Βέλτιστη εκτέλεση για κάθε είσοδο.
- Δυνατότητα αξιοποίησης όλων των state-of-the-art υλοποιήσεων
- Εύκολη αφομοίωση νέων μεθόδων/δομών

Πρόβλεψη βέλτιστης δομής με βάση την είσοδο

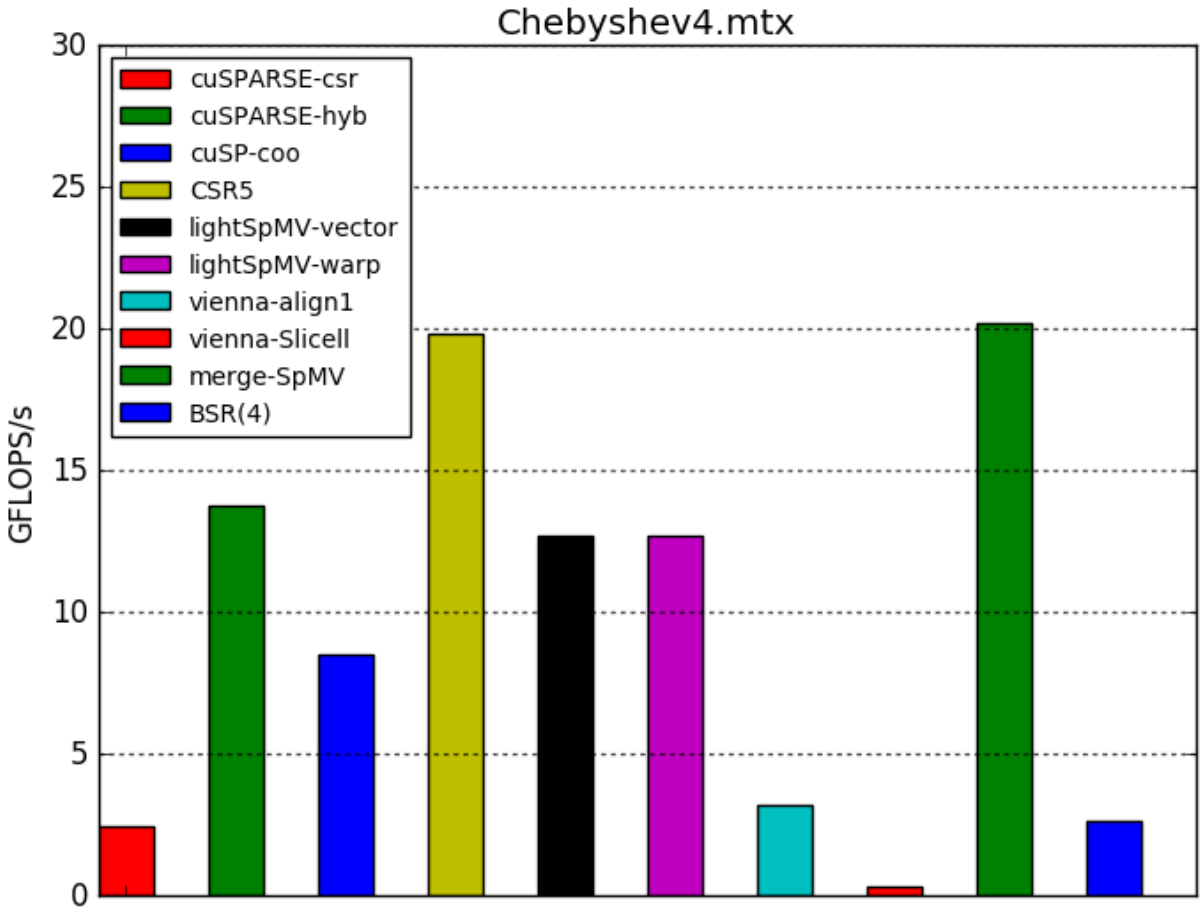
Σκοπός

- Βέλτιστη εκτέλεση για κάθε είσοδο.
- Δυνατότητα αξιοποίησης όλων των state-of-the-art υλοποιήσεων
- Εύκολη αφομοίωση νέων μεθόδων/δομών

Προβλήματα

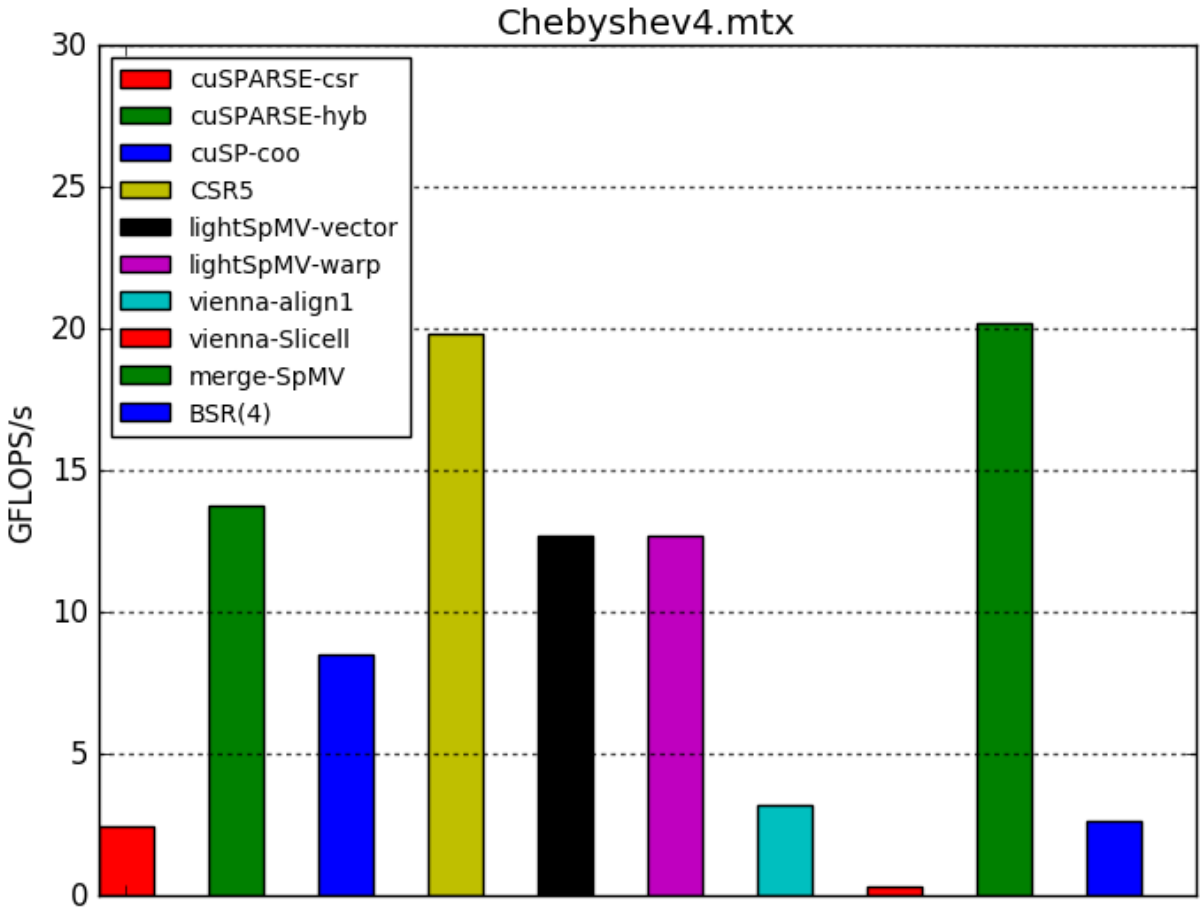
- Κόστος προεπεξεργασίας
- Χαμηλή ακρίβεια πρόβλεψης
- Κόστος χειρότερης περίπτωσης

Παράδειγμα



- Μεγάλη η ανάγκη πρόβλεψης... ΑΛΛΑ

Παράδειγμα



- Μεγάλη η ανάγκη πρόβλεψης... ΑΛΛΑ
- Επίσης τεράστιο πιθανό κόστος λάθος πρόβλεψης

Συνελικτικά Νευρωνικά Δίκτυα (CNNs)

Τι μπορούν να προσφέρουν;

- Υψηλή ακρίβεια πρόβλεψης

Συνελικτικά Νευρωνικά Δίκτυα (CNNs)

Τι μπορούν να προσφέρουν;

- Υψηλή ακρίβεια πρόβλεψης
- Σχετικά μικρό χρόνο προεπεξεργασίας
  - Η εκπαίδευση γίνεται σε ξεχωριστό χρόνο.
  - Η διάρκεια πρόβλεψης είναι σχετικά μικρή.

Συνελικτικά Νευρωνικά Δίκτυα (CNNs)

Τι μπορούν να προσφέρουν;

- Υψηλή ακρίβεια πρόβλεψης
- Σχετικά μικρό χρόνο προεπεξεργασίας
  - Η εκπαίδευση γίνεται σε ξεχωριστό χρόνο.
  - Η διάρκεια πρόβλεψης είναι σχετικά μικρή.
- Ευκολία υλοποίησης
  - Δεν απαιτείται ιδιαίτερη μελέτη των χαρακτηριστικών εισόδου
  - Το νευρωνικό 'ανακαλύπτει' περίπλοκες σχέσεις μεταξύ της εισόδου και της βέλτιστης δομής
- Μεταφερσιμότητα
  - Transfer Learning

Προϋπάρχουσες Υλοποιήσεις με CNNs

"Bridging the gap between deep learning and sparse matrix format selection" Zhao et al, PPoPP'18

- Πρώτη προσέγγιση με CNN

Προϋπάρχουσες Υλοποιήσεις με CNNs

"Bridging the gap between deep learning and sparse matrix format selection" Zhao et al, PPoPP'18

- Πρώτη προσέγγιση με CNN
- Πρόταση 3 διαφορετικών απεικονίσεων ως είσοδο
  - Δυαδική εικόνα
  - Εικόνα πυκνότητας
  - Ιστογράμματα απόστασης από τη διαγώνιο

Προϋπάρχουσες Υλοποιήσεις με CNNs

"Bridging the gap between deep learning and sparse matrix format selection" Zhao et al, PPoPP'18

- Πρώτη προσέγγιση με CNN
- Πρόταση 3 διαφορετικών απεικονίσεων ως είσοδο
  - Δυαδική εικόνα
  - Εικόνα πυκνότητας
  - Ιστογράμματα απόστασης από τη διαγώνιο
- Εφαρμογή σε GPUs και CPUs
- Επικέντρωση στο κομμάτι της δομής του νευρωνικού

Αποτελέσματα Zhao et al, 2018

Θετικά

- Μεγάλη ακρίβεια πρόβλεψης
  - 93% με ιστογράμματα, 90% με συνδυασμό δυαδικών + πυκνωτικών εικόνων σε CPUs,
  - 90% με ιστογράμματα σε GPUs
- Πρόταση μοντέλων με δυνατότητα *Transfer learning*

Αποτελέσματα Zhao et al, 2018

Θετικά

- Μεγάλη ακρίβεια πρόβλεψης
  - 93% με ιστογράμματα, 90% με συνδυασμό δυαδικών + πυκνωτικών εικόνων σε CPUs,
  - 90% με ιστογράμματα σε GPUs
- Πρόταση μοντέλων με δυνατότητα *Transfer learning*

Αρνητικά

- Μη αντιπροσωπευτικό dataset
  - Μεγάλο ποσοστό πολύ μικρών πινάκων
  - Μη-ισορροπημένες κλάσεις
- Αποτελέσματα πρόβλεψης σε συνθετικό dataset
- Χρήση περιορισμένων δομών
  - Πολύ πιο εύκολη η κατηγοριοποίηση

Η προσέγγισή μας

- Πειραματισμός σε GPUs
- Πρόβλεψη μεταξύ state-of-the-art υλοποιήσεων δομών
  - Πολύ μεγαλύτερη δυσκολία κατηγοριοποίησης

Η προσέγγισή μας

- Πειραματισμός σε GPUs
- Πρόβλεψη μεταξύ state-of-the-art υλοποιήσεων δομών
  - Πολύ μεγαλύτερη δυσκολία κατηγοριοποίησης
- Τα βήματα της υλοποίησής μας
  1. Δημιουργία συνθετικού dataset
  2. Εκτέλεση SpMV για κάθε δομή
  3. Υλοποίηση απεικονίσεων για την εκπαίδευση
  4. Επιλογή καλύτερων δομών
  5. Επιλογή 3 συνελικτικών δικτύων
  6. Εκπαίδευση και αξιολόγηση των αποτελεσμάτων

Training set

Προβλήματα

- Μικρός αριθμός πινάκων $> 100MB$
- Αδυναμία χρήσης πινάκων $> 1GB$
  - Λόγο χώρου (ανάγκη $30TB$ δεδομένων)
  - Λόγο επεξεργαστικής δύναμης (μήνες επεξεργασίας σε 1 GPU)
- Μη-ισορροπημένο dataset
- Ανάγκη πολύ περισσότερων πινάκων (10000+) για τα βαθέα νευρωνικά

Training set

Προβλήματα

- Μικρός αριθμός πινάκων $> 100MB$
- Αδυναμία χρήσης πινάκων $> 1GB$
  - Λόγο χώρου (ανάγκη $30TB$ δεδομένων)
  - Λόγο επεξεργαστικής δύναμης (μήνες επεξεργασίας σε 1 GPU)
- Μη-ισορροπημένο dataset
- Ανάγκη πολύ περισσότερων πινάκων (10000+) για τα βαθέα νευρωνικά

Λύση

Συνθετικοί πίνακες

Δημιουργία συνθετικών πινάκων

- Υλοποιήσαμε 3 αλγορίθμους μετασχηματισμού
  1. Μεγέθυνσης μέσω δημιουργίας μπλοκ
  2. Αυξομείωσης απόστασης από διαγώνιο(DDVT)
  3. Αντικατοπτρισμού και αντιγραφής (Mirroring)

Δημιουργία συνθετικών πινάκων

- Υλοποιήσαμε 3 αλγορίθμους μετασχηματισμού
  1. Μεγέθυνσης μέσω δημιουργίας μπλοκ
  2. Αυξομείωσης απόστασης από διαγώνιο(DDVT)
  3. Αντικατοπτρισμού και αντιγραφής (Mirroring)
- 17962 συνθετικοί πίνακες ($1.4TB$)
  - $PWL_{cluster}$: 495 ταξινομημένοι power-law graph
  - $PWL_{seq.}$: 1350 τυχαίοι power-law graph
  - $DDVT_{resized}$:7830 από τους αλγορίθμους 1 και 2
  - $Mirror_{aug.}$: 6655 από τον αλγόριθμο 3
  - $Block_{aug.}$: 1632 με μπλοκ από τον αλγόριθμο 3

Επιλογή υλοποιήσεων δομών

Στοιχεία εκτέλεσης

- Nvidia Tesla K40 GPU
- 19 υλοποιήσεις δομών
- Κοινός χρονομετρητής, μέσος όρος 100 εκτελέσεων με προθέρμανση
- 4 εβδομάδες συνολική εκτέλεση

Επιλογή υλοποιήσεων δομών

Στοιχεία εκτέλεσης

- Nvidia Tesla K40 GPU
- 19 υλοποιήσεις δομών
- Κοινός χρονομετρητής, μέσος όρος 100 εκτελέσεων με προθέρμανση
- 4 εβδομάδες συνολική εκτέλεση

Τελικές Δομές

- CSR5
- merge-SpMV
- lightSpMV$_{warp}$
- cuSPARSE-CSR
- cuSPARSE-HYB
- cuSPARSE-BSR (Block size 4)

Είσοδος Συνελικτικών Νευρωνικών

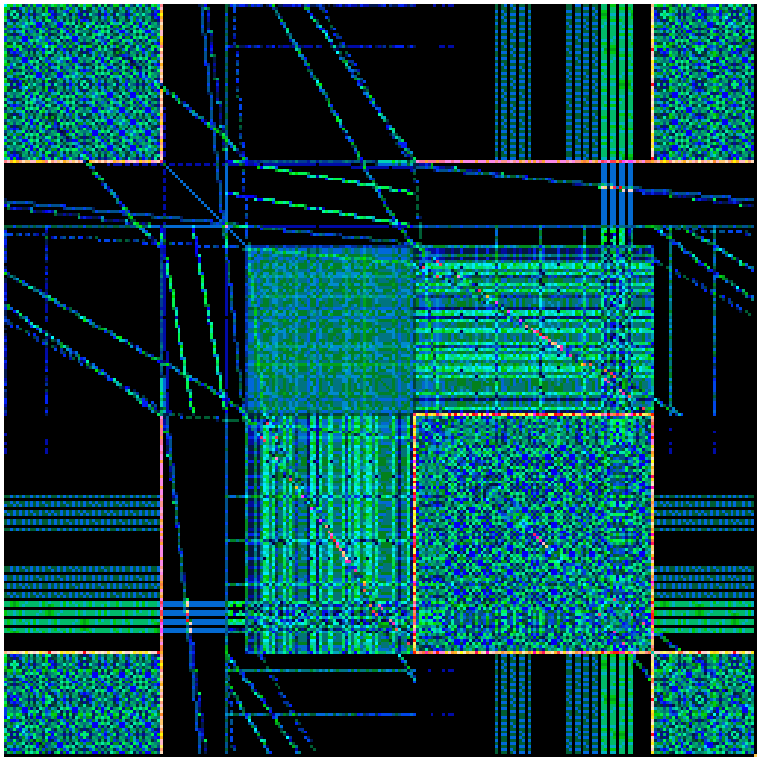
- Τα Συνελικτικά Νευρωνικά Δίκτυα κατηγοριοποιούν εικόνες
- Στην περίπτωσή μας απεικονίσαμε τη δομή των μη-μηδενικών στοιχείων του κάθε πίνακα με 2 τρόπους

Είσοδος Συνελικτικών Νευρωνικών

- Τα Συνελικτικά Νευρωνικά Δίκτυα κατηγοριοποιούν εικόνες
- Στην περίπτωσή μας απεικονίσαμε τη δομή των μη-μηδενικών στοιχείων του κάθε πίνακα με 2 τρόπους

Δυαδική απεικόνιση

Απεικόνιση RGB πυκνότητας



Επιλογή Συνελικτικών Δικτύων

Επιλέξαμε 3 διαφορετικά δίκτυα

Lenet

- 5 κρυφά επίπεδα, 10000 επαναλήψεις
- Γρήγορη εκπαίδευση, σχεδιασμένο για μικρές εικόνες

Επιλογή Συνελικτικών Δικτύων

Επιλέξαμε 3 διαφορετικά δίκτυα

Lenet

- 5 κρυφά επίπεδα, 10000 επαναλήψεις
- Γρήγορη εκπαίδευση, σχεδιασμένο για μικρές εικόνες

CaffeNet (AlexNet)

- 8 κρυφά επίπεδα, 100000 επαναλήψεις
- Αργή εκπαίδευση, μεγαλύτερο βάθος

Επιλογή Συνελικτικών Δικτύων

Επιλέξαμε 3 διαφορετικά δίκτυα

Lenet

- 5 κρυφά επίπεδα, 10000 επαναλήψεις
- Γρήγορη εκπαίδευση, σχεδιασμένο για μικρές εικόνες

CaffeNet (AlexNet)

- 8 κρυφά επίπεδα, 100000 επαναλήψεις
- Αργή εκπαίδευση, μεγαλύτερο βάθος

GoogleNet

- 22 κρυφά επίπεδα, 200000 επαναλήψεις
- Γρήγορη εκπαίδευση λόγω μείωσης παραμέτρων με Inception Layers
- Μεγάλο βάθος, τρία επίπεδα εξόδου

Εκπαίδευση με δυαδικές εικόνες

Binary Lenet

- Μέγεθος εικόνων 372×372 απαγορευτικό για το δίκτυο
- Σμίκρυνση σε $256 \times 256 -> $ απώλεια πληροφορίας-τοπικό ελάχιστο

Εκπαίδευση με δυαδικές εικόνες

Binary Lenet

- Μέγεθος εικόνων 372×372 απαγορευτικό για το δίκτυο
- Σμίκρυνση σε $256 \times 256 -> $ απώλεια πληροφορίας-τοπικό ελάχιστο

Binary CaffeNet (AlexNet)

- Μη-ισορροπημένο train set πρόβλημα για το δίκτυο
- Σε όλες τι περιπτώσεις οδηγούνταν σε τοπικό ελάχιστο (CSR5)

Εκπαίδευση με δυαδικές εικόνες

Binary Lenet

- Μέγεθος εικόνων 372×372 απαγορευτικό για το δίκτυο
- Σμίκρυνση σε $256 \times 256 -> $ απώλεια πληροφορίας-τοπικό ελάχιστο

Binary CaffeNet (AlexNet)

- Μη-ισορροπημένο train set πρόβλημα για το δίκτυο
- Σε όλες τι περιπτώσεις οδηγούνταν σε τοπικό ελάχιστο (CSR5)

Binary GoogleNet

- Υψηλή ακρίβεια πρόβλεψης σε συνθετικό test set
- Σχεδόν βέλτιστο speedup σε αυτό
- Πρόβλημα ακρίβειας σε πραγματικούς αραιούς πίνακες

Binary GoogleNet results

Binary Googlenet + Mirror train set						
		Accuracy		Speedup over cuSPARSE-CSR		
Sub-set	Size	Top 1	Top 2	Predicted	CSR5	Max
Block_{aug}	1632	0.81	0.95	1.11	0.97	1.14
$\text{PWL}_{cluster}$	495	0.74	0.88	1.76	1.66	1.77
$\text{PWL}_{seq.}$	1350	0.82	0.96	29.77	29.15	29.91
Mirror_{aug}	6655	0.94	0.97	1.46	1.36	1.50
$\text{Test}_{synthetic}$	2500	0.90	0.97	**5.28**	**4.03**	5.35
Test_{real}	416	0.44	0.71	**1.23**	**1.30**	1.57
Test_{large}	208	0.33	0.50	**1.19**	**1.27**	1.42

- Binary Googlenet **30%** καλύτερο από CSR5 σε συνθετικούς πίνακες
- CSR5 **6%** καλύτερο σε πραγματικούς πίνακες

Εκπαίδευση με εικόνες πυκνότητας

RGB Density Lenet

- Υψηλή ακρίβεια πρόβλεψης σε συνθετικό test set
- Μικρότερο speedup σε συνθετικούς πίνακες από Googlenet
- Καλύτερο speedup απο Binary Googlenet σε πραγματικούς πίνακες

Εκπαίδευση με εικόνες πυκνότητας

RGB Density Lenet

- Υψηλή ακρίβεια πρόβλεψης σε συνθετικό test set
- Μικρότερο speedup σε συνθετικούς πίνακες από Googlenet
- Καλύτερο speedup απο Binary Googlenet σε πραγματικούς πίνακες

RGB Density CaffeNet (AlexNet)

- Όμοιο πρόβλημα μη-ισορροπημένου train set ανεξάρτητο εικόνων
- Και πάλι οδηγούνταν σε τοπικό ελάχιστο (CSR5)

Εκπαίδευση με εικόνες πυκνότητας

RGB Density Lenet

- Υψηλή ακρίβεια πρόβλεψης σε συνθετικό test set
- Μικρότερο speedup σε συνθετικούς πίνακες από Googlenet
- Καλύτερο speedup απο Binary Googlenet σε πραγματικούς πίνακες

RGB Density CaffeNet (AlexNet)

- Όμοιο πρόβλημα μη-ισορροπημένου train set ανεξάρτητο εικόνων
- Και πάλι οδηγούνταν σε τοπικό ελάχιστο (CSR5)

RGB Density GoogleNet

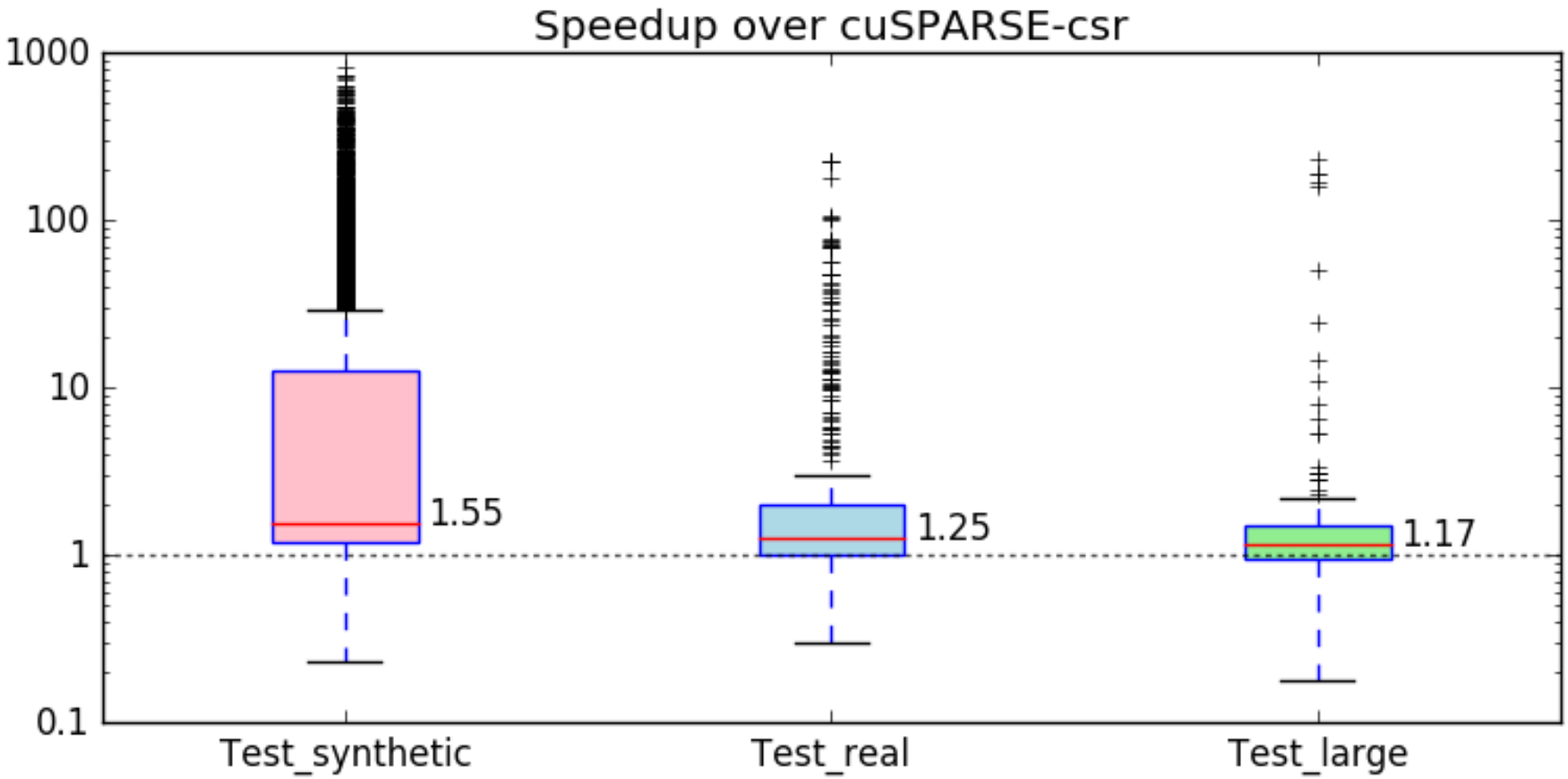
- Μεγαλύτερη ακρίβεια πρόβλεψης σε **όλα** τα test set
- Καλύτερη απόδοση από οποιαδήποτε μεμονωμένη υλοποίηση δομής
- Αξιοποίηση μεγαλύτερου train set (DDVT + Mirror)

RGB Density GoogleNet results

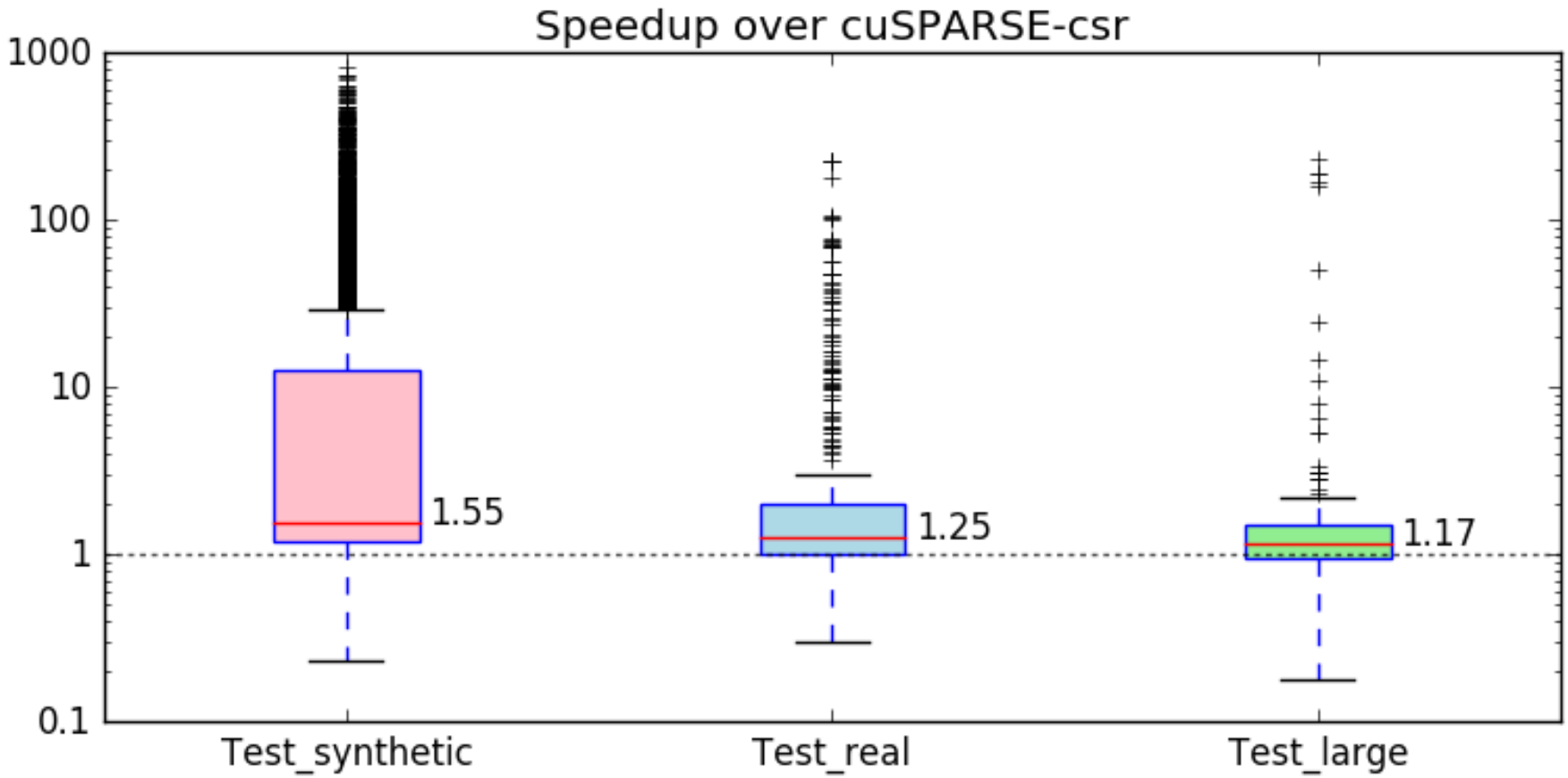
RGB Density GoogleNet						
		Accuracy		Speedup over cuSPARSE-CSR		
Sub-set	Size	Top 1	Top 2	Predicted	CSR5	Max
Block_{aug}	1632	0.91	0.97	1.11	0.97	1.14
$\text{DDVT}_{resized}$	495	0.86	0.95	2.16	1.99	2.21
$\text{PWL}_{cluster}$	495	0.80	0.92	1.77	1.66	1.77
$\text{PWL}_{seq.}$	1350	0.81	1.00	29.81	29.15	29.91
Mirror_{aug}	6655	0.96	0.99	1.50	1.36	1.50
$\text{Test}_{synthetic}$	2500	0.91	0.97	**5.32**	**4.03**	5.35
Test_{real}	416	0.57	0.73	**1.35**	**1.30**	1.57
Test_{large}	208	0.43	0.66	**1.23**	**1.27**	1.42

- Density Googlenet **32%** καλύτερο απο CSR5 σε συνθετικούς **και 4%** σε πραγματικούς πίνακες
- 10% αύξηση σε σχέση με δυαδική απεικόνιση, ακόμα πρόβλημα σε μεγάλους πραγματικούς πίνακες

RGB Density GoogleNet results



RGB Density GoogleNet results



- Πολλοί πίνακες είχαν τεράστια speedups σε σχέση με το baseline cuSPARSE-csr
- Εδώ δεν συμπεριλαμβάνονται στον υπολογισμό του μέσου

Σύγκριση

Binary *vs* RGB Density

- Η υλοποίηση πυκνότητας ξεπέρασε σημαντικά την δυαδική
  - Περισσότερη πληροφορία
  - Μεγαλύτερη ακρίβεια πρόβλεψης
  - Αποτελεσματική εκπαίδευση και σε Lenet
  - Μικρότερο μέγεθος (256×256 vs 372×372)

Σύγκριση

Binary *vs* RGB Density

- Η υλοποίηση πυκνότητας ξεπέρασε σημαντικά την δυαδική
  - Περισσότερη πληροφορία
  - Μεγαλύτερη ακρίβεια πρόβλεψης
  - Αποτελεσματική εκπαίδευση και σε Lenet
  - Μικρότερο μέγεθος (256×256 vs 372×372)

Lenet *vs* GoogleNet

- Ταχύτητα εκπαίδευσης ή ακρίβεια πρόβλεψης;
  - Μέγιστη διαφορά τάξης **5%** στην ακρίβεια πρόβλεψης
  - GoogleNet δέκα φορές περισσότερο χρόνο εκπαίδευσης
  - Περίπου διπλάσιο κόστος πρόβλεψης με GoogleNet
- Η απάντηση δεν είναι καθόλου ξεκάθαρη αλλά...

Σύγκριση

Binary *vs* RGB Density

- Η υλοποίηση πυκνότητας ξεπέρασε σημαντικά την δυαδική
  - Περισσότερη πληροφορία
  - Μεγαλύτερη ακρίβεια πρόβλεψης
  - Αποτελεσματική εκπαίδευση και σε Lenet
  - Μικρότερο μέγεθος (256×256 vs 372×372)

Lenet *vs* GoogleNet

- Ταχύτητα εκπαίδευσης ή ακρίβεια πρόβλεψης;
  - Μέγιστη διαφορά τάξης **5%** στην ακρίβεια πρόβλεψης
  - GoogleNet δέκα φορές περισσότερο χρόνο εκπαίδευσης
  - Περίπου διπλάσιο κόστος πρόβλεψης με GoogleNet
- Η απάντηση δεν είναι καθόλου ξεκάθαρη αλλά...
- Το GoogleNet αναδεικνύει περισσότερες δυνατότητες βελτίωσης

Σύνοψη

Βελτιστοποίηση του SpMV σε GPUs με CNNs αρκετά υποσχόμενη

- Η έρευνα και η εφαρμογή της απαιτεί πολλούς πόρους
- Μπορεί όμως να προσφέρει πολύ υψηλά επίπεδα πρόβλεψης

Future work

- Δημιουργία μεγαλύτερων σετ εκπαίδευσης
- Υλοποίηση πιο αντιπροσωπευτικών απεικονίσεων
- Δοκιμή πιο σύγχρονων δικτύων (VGG, Inception)
- Προσθήκη σε υπάρχοντα εργαλεία/βιβλιοθήκες

Ερωτήσεις