WaveGlow

Ως γνωστόν το πρόβλημα για τη σύνθεση φωνής από κείμενο (text-to-speech) μπορεί να διασπαστεί σε δύο επιμέρους προβλήματα. Το πρώτο αφορά τη μετατροπή του κειμένου σε ορισμένα χαρακτηριστικά, όπως είναι το φασματογράφημα στην κλίμακα mel και μπορεί να προκύψει χρησιμοποιώντας αρχιτεκτονικές όπως το Tacotron2 [She+18]. Στη συνέχεια αναλαμβάνει ένα μοντέλο vocoder προκειμένου να μετατρέψει τα χαρακτηριστικά αυτά στα αντίστοιχα δείγματα μιας κυματομορφής ήχου. Στην Ενότητα αυτή εστιάζουμε σε ένα μοντέλο vocoder που ονομάζεται WaveGlow [PVC19]. Το WaveGlow είναι ένα γενετικό μοντέλο βασισμένο σε ροή (flow-based generative network), που αποτελείται από μία σειρά αντίστροφων μετασχηματισμών για την παραγωγή δειγμάτων μιας κυματομορφής. Χρησιμοποιώντας πληροφορία από ένα φασματογράφημα στην κλίμακα mel μπορεί να παράξει ηχητικά δείγματα πολύ καλής ποιότητας, όπως θα δούμε παρακάτω και από τα σχετικά αποτελέσματα. Η δομή του συνδυάζει στοιχεία από τα μοντέλα Glow [KD18] και WaveNet [Oor+16], αλλά σε αυτή την περίπτωση πρόκειται για ένα μη-αυτοπαλινδρομικό (non-autoregressive) μοντέλο, με την έννοια ότι για την παραγωγή ενός δείγματος σε μια δεδομένη χρονική στιγμή, δε λαμβάνονται υπόψιν μόνο τα προηγούμενα δείγματα[1]. Το βασικό θετικό στοιχείο του μοντέλου είναι ότι για την εκπαίδευσή του χρειάζεται μόνο η ελαχιστοποίηση μιας απλής συνάρτησης κόστους, δηλαδή του αρνητικού λογαρίθμου της πιθανοφάνειας των δεδομένων, που όπως θα δούμε υπολογίζεται εύκολα λόγω των αντίστροφων μετασχηματισμών που συνθέτουν το μοντέλο και με χρήση του θεωρήματος αλλαγής μεταβλητών. Αν και πρόκειται για ένα απλό δίκτυο ως προς την υλοποίηση με πολύ καλά αποτελέσματα, στα αρνητικά του συγκαταλέγεται ο μεγάλος χρόνος που απαιτείται κατά την εκπαίδευση και την συμπερασματολογία (inference). Προτού περιγράψουμε την αρχιτεκτονική του, παραθέτουμε ορισμένες βασικές έννοιες στις οποίες στηρίζεται το μοντέλο WaveGlow.

Normalizing Flow

Ένα μοντέλο ομαλοποιημένης ροής (normalizing flow) [RM15] μετασχηματίζει μία απλή κατανομή σε μία σύνθετη μέσω μιας σειράς αντίστροφων μετασχηματισμών, με στόχο τον ευκολότερο υπολογισμό της σύνθετης κατανομής των τελικών δεδομένων που προκύπτουν. Αυτό ακριβώς συμβαίνει και στο μοντέλο WaveGlow αφού για τη σύνθεση μιας κυματομορφής ήχου, χρειάζεται πρώτα να πάρουμε δείγματα από την κανονική κατανομή και στη συνέχεια μέσω ορισμένων μετασχηματισμών (steps of flow) να παράγουμε τα τελικά δείγματα ήχου. Με το μοντέλο normalizing flow και με τη χρήση του θεωρήματος αλλαγής μεταβλητών, ο υπολογισμός της πιθανοφάνειας των ηχητικών δειγμάτων ανάγεται εύκολα στον υπολογισμό της πιθανοφάνειας των δειγμάτων που προέρχονται από την κανονική κατανομή.

Αρχικά ας υποθέσουμε ότι έχουμε δύο συνεχείς τυχαίες μεταβλητές (τ.μ.) X, Y με τιμές στο $\mathbb{R}$, οι οποίες συνδέονται μέσω ενός αντίστροφου μετασχηματισμού $f : \mathbb{R} \to \mathbb{R}$ σύμφωνα με τη σχέση

$$y = f(x) \iff x = f^{-1}(y), \text{ όπου } X \sim p_X(x) \text{ και } Y \sim p_Y(y). \tag{0.1}$$

Έστω ότι είναι γνωστή η συνάρτηση πυκνότητας πιθανότητας (σ.π.π.) $p_X(x)$ της τ.μ. X και μπορούμε εύκολα να πάρουμε δείγματα από αυτήν (π.χ. μια κανονική κατανομή). Για να υπολογίσουμε τη σ.π.π. της τ.μ. Y έχουμε:

[1] Η autoregressive φύση στο μοντέλο WaveNet οφείλεται στη χρήση casual convolutions στην αρχιτεκτονική του, ενώ το WaveGlow δε χρησιμοποιεί casual convolutions, γεγονός που το καθιστά ένα non-autoregressive μοντέλο.

$$F_Y(y) = \mathbb{P}[Y \leq y] = \mathbb{P}[f(X) \leq y] = \begin{cases} \mathbb{P}[X \leq f^{-1}(y)] = F_X(f^{-1}(y)), & \text{αν } f^{-1} \text{ αύξουσα} \\ \mathbb{P}[X \geq f^{-1}(y)] = 1 - F_X(f^{-1}(y)), & \text{αν } f^{-1} \text{ φθίνουσα,} \end{cases}$$

επομένως λαμβάνοντας την παράγωγο ως προς y παίρνουμε

$$p_Y(y) = \frac{dF_Y(y)}{dy} = \begin{cases} p_X(f^{-1}(y))\dfrac{\mathrm{d}f^{-1}}{\mathrm{d}y}, & \text{αν } f^{-1} \text{ αύξουσα} \\ -p_X(f^{-1}(y))\dfrac{\mathrm{d}f^{-1}}{\mathrm{d}y}, & \text{αν } f^{-1} \text{ φθίνουσα.} \end{cases}$$

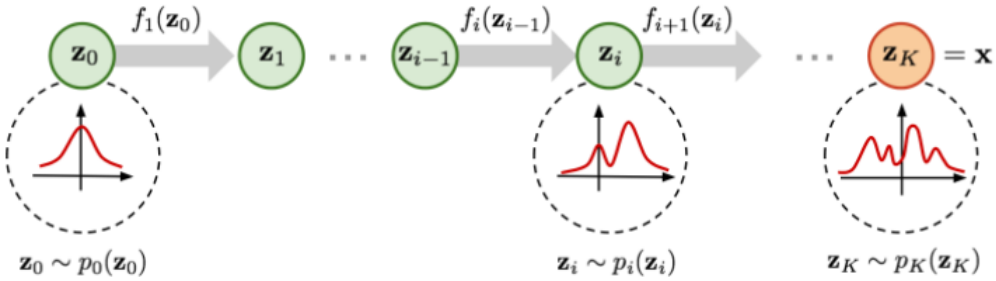
Τελικά η σ.π.π. της τ.μ. Y θα είναι $p_Y(y) = p_X(f^{-1}(y))\left|\dfrac{\mathrm{d}f^{-1}}{\mathrm{d}y}\right|$. Για την περίπτωση των πολυδιάστατων κατανομών προκύπτει η εξής ανάλογη σχέση

$$p_Y(\mathbf{y}) = p_X(f^{-1}(\mathbf{y}))\left|\det\left(\mathbf{J}(f^{-1}(\mathbf{y}))\right)\right|, \tag{0.2}$$

όπου $\det(\mathbf{J}(f^{-1}(\mathbf{y})))$ είναι η ορίζουσα του Ιακωβιανού (Jacobian) πίνακα των παραγώγων πρώτης τάξης του αντίστροφου μετασχηματισμού f^{-1}. Έστω τώρα ότι έχουμε μια σειρά από αντίστροφους μετασχηματισμούς, όπως φαίνεται στο Σχήμα 1, όπου η αρχική τ.μ. Z_0 συνδέεται με την τ.μ. $Z_k = X$ μέσω της σχέσης

$$\mathbf{x} = \mathbf{z}_k = f_k \circ f_{k-1} \cdots \circ f_1(\mathbf{z}_0) \Longleftrightarrow \mathbf{z}_0 = f_1^{-1} \circ f_2^{-1} \cdots \circ f_k^{-1}(\mathbf{x}). \tag{0.3}$$

Όμως από τη σχέση 0.2 έχουμε $p_i(\mathbf{z}_i) = p_{i-1}(f_i^{-1}(\mathbf{z}_i))\left|\det\left(\mathbf{J}(f_i^{-1}(\mathbf{z}_i))\right)\right|$, για κάθε $i = 1, \ldots, k$.



Σχήμα 1: Μοντέλο normalizing-flow. Τα τελικά δεδομένα $\mathbf{x}$ υπολογίζονται μέσω μιας σειράς αντίστροφων μετασχηματισμών f_i από τα αρχικά δεδομένα $\mathbf{z}_0$. Ο υπολογισμός της κατανομής των δεδομένων $\mathbf{x}$ ανάγεται στην κατανομή των δεδομένων $\mathbf{z}_0$ [Wen18].

Επιπλέον ισχύει:

$$\mathbf{J}(f_i^{-1}(\mathbf{z}_i)) = \frac{\partial f_i^{-1}(\mathbf{z}_i)}{\partial \mathbf{z}_i} = \frac{\partial \mathbf{z}_{i-1}}{\partial \mathbf{z}_i} = \left(\frac{\partial \mathbf{z}_i}{\partial \mathbf{z}_{i-1}}\right)^{-1} = \left(\frac{\partial f_i(\mathbf{z}_{i-1})}{\partial \mathbf{z}_{i-1}}\right)^{-1} = (\mathbf{J}(f_i(\mathbf{z}_{i-1})))^{-1} \tag{0.4}$$

και για έναν αντιστρέψιμο πίνακα A είναι

$$\det\left(AA^{-1}\right) = \det(\mathbb{I}) = 1 \Rightarrow \det\left(A^{-1}\right) = (\det(A))^{-1}. \tag{0.5}$$

Από τις σχέσεις 0.4 και 0.5 συμπαιρένουμε ότι οι σ.π.π. δύο διαδοχικών τ.μ. σε ένα μοντέλο normalizing-flow δίνονται από τη σχέση

$$p_i(\mathbf{z}_i) = p_{i-1}(f_i^{-1}(\mathbf{z}_i))\left|\det\left((J(f_i(\mathbf{z}_{i-1})))^{-1}\right)\right| = p_{i-1}(\mathbf{z}_{i-1})|\det(J(f_i(\mathbf{z}_{i-1})))|^{-1},$$

με λογάριθμο

$$\log p_i(\mathbf{z}_i) = \log p_{i-1}(\mathbf{z}_{i-1}) - \log|\det(\mathbf{J}(f_i(\mathbf{z}_{i-1})))|.$$

Υπολογίζοντας αναδρομικά την κατανομή των τελικών δεδομένων $\mathbf{x}$ λαμβάνουμε

$$\log p_X(\mathbf{x}) = \log p_k(\mathbf{z}_k) = \log p_{k-1}(\mathbf{z}_{\kappa-1}) - \log|\det(\mathbf{J}(f_k(\mathbf{z}_{k-1})))|$$

$$\Rightarrow \log p_X(\mathbf{x}) = \log p_0(\mathbf{z}_0) - \sum_{i=1}^{k} \log|\det(\mathbf{J}(f_i(\mathbf{z}_{i-1})))|.$$

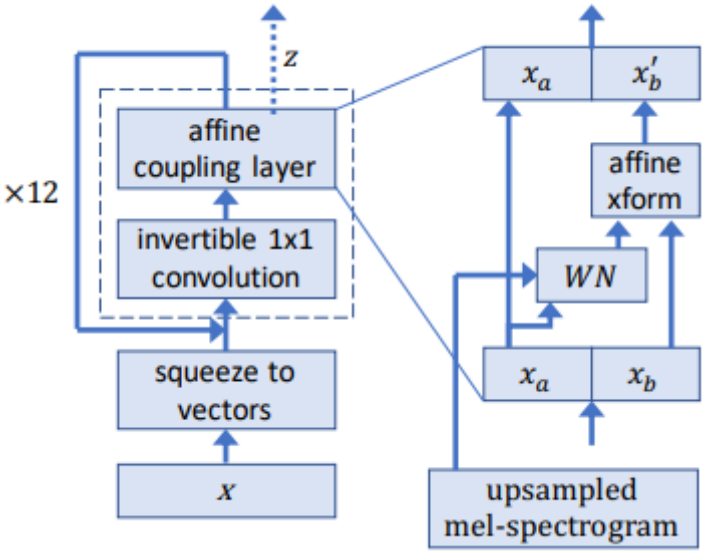
Για να είναι εύκολος ο υπολογισμός του λογαρίθμου της πιθανοφάνειας των τελικών δεδομένων, θα πρέπει κάθε μετασχηματισμός f_i στην ομαλοποιημένη ροή να είναι εύκολα αντιστρέψιμος και επιπλέον να μπορεί να υπολογιστεί εύκολα η ορίζουσα όλων των Ιακωβιανών πινάκων. Στην περίπτωση του μοντέλου WaveGlow οι μετασχηματισμοί που χρησιμοποιούνται ικανοποιούν και τις δύο παραπάνω ιδιότητες. Στη συνέχεια περιγράφουμε την αρχιτεκτονική του μοντέλου.

Αρχιτεκτονική του μοντέλου

Σε πρώτη φάση για να παράγουμε μια κυματομορφή ήχου με το μοντέλο WaveGlow, λαμβάνουμε δείγματα $\mathbf{z} \sim N(\mathbf{0}, \mathbb{I})$ από την πολυδιάστατη κανονική κατανομή με μέση τιμή $\mathbf{0}$ και μοναδιαίο πίνακα συνδιασποράς. Τα δείγματα αυτά έχουν ίδια διάσταση με τη διάσταση της εξόδου, δηλαδή με το πλήθος των σημείων που αποτελούν την κυματομορφή. Εν συνεχεία ακολουθεί μια σειρά αντιστρέψιμων μετασχηματισμών που επεξεργάζονται τα δείγματα $\mathbf{z}$ καθώς και το φασματογράφημα στην κλίμακα mel και παράγουν τα τελικά δείγματα ήχου. Κατά τη φάση της εκπαίδευσης το δίκτυο δέχεται ως είσοδο μια κυματομορφή, η οποία ομαδοποιείται ανά 8 δείγματα. Για παράδειγμα αν η κυματομορφή αποτελείται από 16000 σημεία, τότε η είσοδος στο μοντέλο θα έχει διάσταση $(8, 2000)$. Η ομαδοποίηση σε groups των 8 δειγμάτων αναφέρεται ως «squeeze» από τους συγγραφείς. Στη συνέχεια ακολουθεί μια σειρά από 12 μετασχηματισμούς (steps of flow) μέχρι και την τελική έξοδο $\mathbf{z}$ του δικτύου. Η αρχιτεκτονική του μοντέλου παρουσιάζεται στο Σχήμα 2.

Τα δύο βασικά modules που δομούν το μοντέλο WaveGlow είναι το αντιστρέψιμο 1×1 συνελικτικό επίπεδο (invertible 1×1 convolution) και το επίπεδο affine coupling [DSB16]. Το invertible 1×1 convolution είναι ένα απλό συνελικτικό επίπεδο με μέγεθος πυρήνα (kernel size) 1, όπου ο πίνακας βαρών $\mathbf{W}$ της συνέλιξης είναι αντιστρέψιμος. Ο πίνακας αυτός αρχικοποιείται κατά τέτοιο τρόπο ώστε να είναι ορθομοναδιαίος και κατά συνέπεια αντιστρέψιμος. Ένας πίνακας $\mathbf{W}$ ονομάζεται ορθομοναδιαίος όταν οι γραμμές και οι στήλες του είναι ορθομοναδιαία διανύσματα, δηλαδή έχουν μέτρο 1 και είναι ανά δύο κάθετα. Για έναν ορθομοναδιαίο πίνακα ισχύει ότι $\mathbf{W}\mathbf{W}^T = \mathbb{I} \Rightarrow \det\mathbf{W} \cdot \det\mathbf{W}^T = 1 \Rightarrow \det\mathbf{W} \neq 0$, επομένως ο πίνακας αντιστρέφεται. Το αποτέλεσμα του invertible 1×1 convolution για μια είσοδο $\mathbf{x}$ δίνεται από τη σχέση $\mathbf{y} = f_{conv}(\mathbf{x}) = \mathbf{W}\mathbf{x} \Leftrightarrow \mathbf{x} = \mathbf{W}^{-1}\mathbf{y}$. Ο Ιακωβιανός πίνακας αυτής της συνέλιξης είναι απλά ο $\mathbf{W}$, επομένως θα ισχύει $\log|\det(\mathbf{J}(f_{conv}(\mathbf{x})))| = \log|\det\mathbf{W}|$.

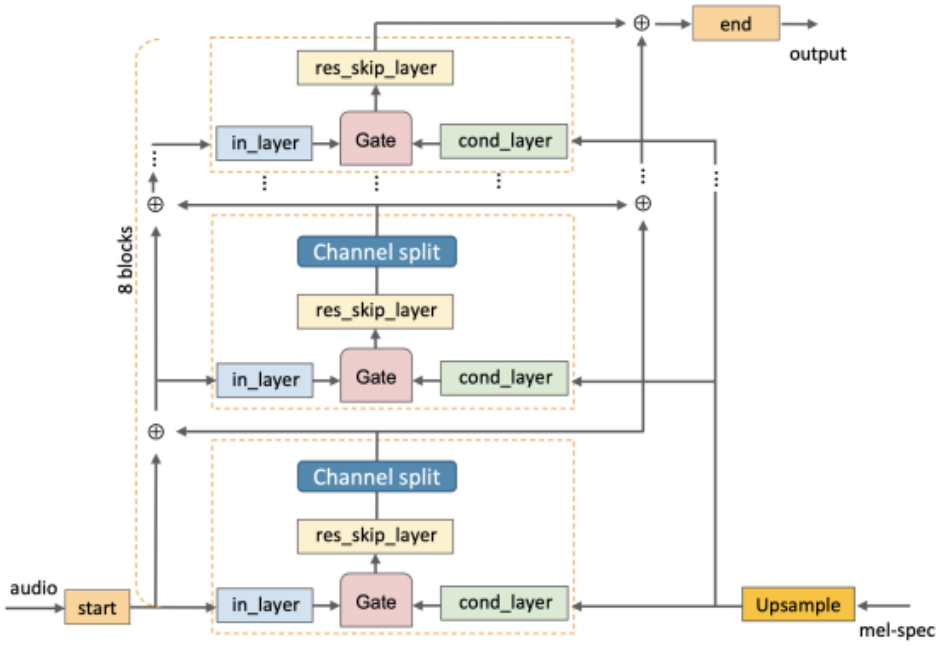
Στη συνέχεια εφαρμόζεται ένα επίπεδο affine coupling, όπως φαίνεται στο δεξί μέρος του Σχήματος 2. Η έξοδος $\mathbf{x}$ από το invertible 1×1 convolution διασπάται αρχικά σε δύο μέρη $\mathbf{x}_a$ και $\mathbf{x}_b$ με ίδιο πλήθος καναλιών το καθένα. Επιπλέον γίνεται ένα upsampling του φασματογραφήματος με χρήση ενός ανάστραφου (transpose) συνελικτικού επιπέδου, προκειμένου να έρθει στην ίδια κλίμακα με το διάνυσμα $\mathbf{x}$. Αν δηλαδή η διάσταση του $\mathbf{x}$ είναι $(8, 2000)$ τότε το upsampled



Σχήμα 2: Αρχιτεκτονική του μοντέλου WaveGlow.

φασματογράφημα θα έχει διάσταση (mel channels, 2000), όπου mel channels = 80. Ύστερα το μέρος $\mathbf{x}_a$ μαζί με το upsampled φασματογράφημα περνούν ως είσοδος σε ένα δίκτυο WN τύπου Wavenet, του οποίου η αρχιτεκτονική φαίνεται στο Σχήμα 3. Το δίκτυο WN αποτελείται από 8 blocks με συνελίξεις που χρησιμοποιούν διαστολή (dilation), η οποία διπλασιάζεται σε κάθε block. Αυξάνοντας το dilation, αυξάνεται το πεδίο λήψης (receptive field), επομένως για τον υπολογισμό ενός στοιχείου στην έξοδο, συνεισφέρουν πολλά περισσότερα δείγματα από την ακολουθία εισόδου, αντί να χρησιμοποιούσαμε απλές συνελίξεις. Πιο συγκεκριμένα, το διάνυσμα $\mathbf{x}_a$ (στο Σχήμα 3 ονομάζεται audio) περνά από ένα απλό convolution layer (start) ώστε να προκύψει ένα στοιχείο με 256 κανάλια. Ύστερα κάθε block περιλαμβάνει ένα συνελικτικό επίπεδο με όνομα in_layer και dilation $= 2^b$ (ανάλογα το block $b = 0, \ldots, 7$), του οποίου η έξοδος έχει 512 κανάλια. Επίσης το upsampled φασματογράφημα διέρχεται από το συνελικτικό επίπεδο με όνομα cond_layer, ώστε η έξοδος να έχει επίσης 512 κανάλια. Εν συνεχεία ο μετασχηματισμός Gate εφαρμόζεται στο άθροισμα των εξόδων του in_layer και του cond_layer. Στα πρώτα μισά (256) κανάλια εφαρμόζεται η υπερβολική εφαπτομένη (tanh) και στα υπόλοιπα 256 η σιγμοειδής συνάρτηση ενεργοποίησης (sigmoid) και τα αποτελέσματα πολλαπλασιάζονται μεταξύ τους. Στη συνέχεια ακολουθεί ένα συνελικτικό επίπεδο με όνομα res_skip_layer ώστε η έξοδός του να έχει 512 κανάλια (εκτός από το τελευταίο block όπου η έξοδος έχει 256 κανάλια). Έπειτα τα κανάλια διαχωρίζονται εκ νέου στα δύο (Channel split), όπου τα πρώτα 256 αθροίζονται στην είσοδο πριν το in_layer του επόμενου block και τα υπόλοιπα 256 αθροίζονται μέχρι και την έξοδο του τελευταίου res_skip_layer. Μετά το τελευταίο block ακολουθεί ένα απλό συνελικτικό επίπεδο (end) που «επαναφέρει» τα κανάλια στην αρχική διάσταση της εισόδου στο affine coupling layer. Το affine coupling layer περιγράφεται από τις παρακάτω σχέσεις:

$$
\begin{aligned}
\mathbf{x}_a, \mathbf{x}_b &= \text{split}(\mathbf{x}) \\
(\log \mathbf{s}, \mathbf{t}) &= \text{WN}(\mathbf{x}_a, \text{mel-spectrogram}) \\
\mathbf{x}'_b &= \mathbf{s} \odot \mathbf{x}_b + \mathbf{t} \\
\mathbf{x}'_a &= \mathbf{x}_a \\
f_{coupling}(\mathbf{x}) &= \text{concat}(\mathbf{x}'_a, \mathbf{x}'_b).
\end{aligned}
$$



Σχήμα 3: Αρχιτεκτονική WN, τύπου Wavenet που χρησιμοποιείται στο μοντέλο WaveGlow σε ένα επίπεδο affine coupling.

Από την έξοδο του δικτύου WN προκύπτουν τα $(\log \mathbf{s}, \mathbf{t})$, όπου το $\log \mathbf{s}$ αντιστοιχεί στα πρώτα μισά κανάλια της εξόδου και το $\mathbf{t}$ στα υπόλοιπα. Η τελική έξοδος του affine coupling layer προκύπτει από τη συνένωση των $\mathbf{x}'_a$ και $\mathbf{x}'_b$. Το $\mathbf{x}'_a$ ισούται απλά με το $\mathbf{x}_a$, δηλαδή το πρώτο μισό ως προς τα κανάλια της εισόδου $\mathbf{x}$, ενώ το $\mathbf{x}'_b$ είναι ένας γραμμικός μετασχηματισμός του $\mathbf{x}_b$ με βάση τα $\mathbf{s}$ και $\mathbf{t}$. Εύκολα μπορούμε να δούμε ότι το επίπεδο affine coupling αποτελεί έναν αντίστροφο μετασχηματισμό. Συγκεκριμένα αν έχουμε τα $\mathbf{x}'_a$ και $\mathbf{x}'_b$ τότε η αρχική είσοδος $\mathbf{x} = \text{split}(\mathbf{x}_a, \mathbf{x}_b)$ μπορεί να ανακτηθεί από τις παρακάτω σχέσεις

$$\begin{aligned} \mathbf{x}_a &= \mathbf{x}'_a \\ (\log \mathbf{s}, \mathbf{t}) &= \text{WN}(\mathbf{x}'_a, \text{mel-spectrogram}) \\ \mathbf{x}_b &= \frac{\mathbf{x}'_b - \mathbf{t}}{\mathbf{s}}. \end{aligned}$$

Παρατηρούμε ότι για τον υπολογισμό των $(\log \mathbf{s}, \mathbf{t})$ αρκεί ένα forward pass από το δίκτυο WN αφού $\mathbf{x}_a = \mathbf{x}'_a$. Επομένως το δίκτυο WN δε χρειάζεται να είναι αντιστρέψιμο και μπορεί να έχει οποιαδήποτε σύνθετη μορφή. Επιπλέον το επίπεδο affine coupling συνεισφέρει στην συνάρτηση κόστους μέσω της σχέσης $\log|\det(\mathbf{J}(f_{coupling}(\mathbf{x})))| = \log|\mathbf{s}|$. Τελικά, ύστερα από την εφαρμογή όλων των επιπέδων invertible 1×1 convolution και affine coupling ο λογάριθμος της πιθανοφάνειας των δεδομένων θα είναι

$$\log p_\theta(\mathbf{x}) = -\frac{\mathbf{z}^T\mathbf{z}}{2\sigma^2} + \sum_{i=1}^{\#\text{conv}} \log\det|\mathbf{W}_i| + \sum_{i=1}^{\#\text{coupling}} \log|s_i|, \tag{0.6}$$

όπου θ είναι οι παράμετροι του μοντέλου και ο όρος $-\frac{\mathbf{z}^T\mathbf{z}}{2\sigma^2}$ είναι ο λογάριθμος της σ.π.π. της κανονικής κατανομής $N(\mathbf{0}, \sigma^2\mathbb{I})$.

Οι συγγραφείς αναφέρουν ότι κατά τη διάρκεια της εκπαίδευσης και του inference το μοντέλο έδινε ως έξοδο 2 κανάλια έπειτα από κάθε 4 affine coupling layers. Συγκεκριμένα, εφόσον το μοντέλο αποτελείται από 12 coupling layers και η είσοδος μετά το «squeeze» έχει 8 κανάλια, έπειτα από τα πρώτα 4 layers η έξοδος θα έχει 2 κανάλια, έπειτα από τα 8 layers θα έχει επίσης 2 κανάλια και έπειτα από το 12ο layer η έξοδος θα έχει 4 κανάλια (στο τελευταίο επίπεδο τα κανάλια παραμένουν τα μισά από τα αρχικά κανάλια που ήταν 8). Τέλος γίνεται μια συνένωση των δύο ενδιάμεσων εξόδων και της τελικής και προκύπτει το τελικό αποτέλεσμα $\mathbf{z}$ με 8 κανάλια. Η διαδικασία αυτή χρησιμεύει προκειμένου το μοντέλο να διατηρεί πληροφορία από την αρχική είσοδο καθώς αυτή μετασχηματίζεται από τα layers του δίκτυο.

Εκπαίδευση, Αξιολόγηση και Αποτελέσματα

Το μοντέλο WaveGlow εκπαιδεύτηκε πάνω στο σύνολο δεδομένων LJ Speech [IJ17], το οποίο περιλαμβάνει 13100 ηχητικά δείγματα συνολικής διάρκειας περίπου 24 ωρών μαζί με τα αντίστοιχα κείμενα (transcriptions). Από τα ηχητικά δείγματα έγινε εξαγωγή των φασματογραφημάτων στη κλίμακα mel χρησιμοποιώντας τις εξής παραμέτρους: FFT length = 1024, hop length = 256 και window length = 1024. Οι συγγραφείς αναφέρουν ότι για την εκπαίδευση του δικτύου χρειάστηκαν 8 GPUs. Το δίκτυο εκπαιδεύτηκε για 580.000 εποχές με χρήση του βελτιστοποιητή Adam [KB14], μέγεθος batch ίσο με 24 και αρχικό βήμα εκμάθησης (learning rate) 10^{-4}, το οποίο στη συνέχεια μειώθηκε στο $5 \cdot 10^{-5}$. Επίσης κατά την εκπαίδευση για τη διασπορά των μεταβλητών $\mathbf{z}$ της κανονικής κατανομής χρησιμοποιήθηκε η τιμή $\sigma^2 = 0.5$, ενώ κατά το inference η τιμή της διασποράς επιλέχθηκε ως $\sigma^2 = 0.36$.

Το μοντέλο αξιολογήθηκε βάσει της μετρικής του MOS (Mean Opinion Score). Κάθε βαθμολογητής έδινε ένα score στην κλίμακα 1 έως 5 για ορισμένες εκφωνήσεις οι οποίες δεν περιέχονταν στο σύνολο εκπαίδευσης. Το μοντέλο συγκρίθηκε και με άλλες μεθόδους όπως ο αλγόριθμος Griffin-Lim [GL84] και το μοντέλο WaveNet και παρουσίασε το υψηλότερο score με τιμή MOS ≈ 3.96. Η τιμή του MOS για τις πραγματικές εκφωνήσεις ήταν περίπου 4.27, για το Griffin-Lim 3.82 και για το WaveNet 3.88. Όπως φαίνεται από τα αποτελέσματα το μοντέλο WaveGlow παράγει ηχητικά δείγματα υψηλής ποιότητας. Στα θετικά του επίσης συγκαταλέγεται η ευκολία κατά την εκπαίδευση αφού χρησιμοποιεί μια απλή συνάρτηση κόστους. Τέλος, αν και η ταχύτητά του κατά το inference είναι περίπου 25 φορές μεγαλύτερη από real-time με τιμή 520kHz σε μία GPU, εντούτοις παρουσιάζει μεγαλύτερη ταχύτητα και από τον αλγόριθμο Griffin-Lim (507kHz) και από το WaveNet (0.11kHz). Τα παραπάνω καθιστούν το WaveGlow ένα state of the art μοντέλο vocoder που μπορεί να χρησιμοποιηθεί σε ένα end-to-end σύστημα για την παραγωγή φωνής.

Αναφορές

[DSB16] Dinh, L., Sohl-Dickstein, J., and Bengio, S. "Density estimation using real nvp". In: *arXiv preprint arXiv:1605.08803* (2016).

[GL84] Griffin, D. and Lim, J. "Signal estimation from modified short-time Fourier transform". In: *IEEE Transactions on acoustics, speech, and signal processing* 32.2 (1984), pp. 236–243.

[IJ17] Ito, K. and Johnson, L. *The LJ Speech Dataset*. https://keithito.com/LJ-Speech-Dataset/. 2017.

[KB14] Kingma, D. P. and Ba, J. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014).

[KD18] Kingma, D. P. and Dhariwal, P. "Glow: Generative flow with invertible 1x1 convolutions". In: *arXiv preprint arXiv:1807.03039* (2018).

[Oor+16] Oord, A. v. d. et al. "Wavenet: A generative model for raw audio". In: *arXiv preprint arXiv:1609.03499* (2016).

[PVC19] Prenger, R., Valle, R., and Catanzaro, B. "Waveglow: A flow-based generative network for speech synthesis". In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2019, pp. 3617–3621.

[RM15] Rezende, D. and Mohamed, S. "Variational inference with normalizing flows". In: *International conference on machine learning*. PMLR. 2015, pp. 1530–1538.

[She+18] Shen, J. et al. "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions". In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2018, pp. 4779–4783.

[Wen18] Weng, L. "Flow-based Deep Generative Models". In: *lilianweng.github.io/lil-log* (2018). URL: http://lilianweng.github.io/lil-log/2018/10/13/flow-based-deep-generative-models.html.